2019

Spatial models for infants HIV/AIDS incidence using an integrated nested Laplace approximation approach

Mutua, Susan Nzula
Strathmore Institute of Mathematical Sciences (SIMs)
Strathmore University

Follow this and additional works at https://su-plus.strathmore.edu/handle/11071/6790

Spatial Models for Infants HIV/AIDS Incidence Using an Integrated Nested Laplace Approximation Approach

Susan Nzula Mutua

Submitted in partial fulfillment of the requirements for the Masters of Science in Statistical Science at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2019

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by 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 except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Susan Nzula Mutua

....

June 6, 2019

Approval

The dissertation of Susan Nzula Mutua was reviewed and approved by the following:

Professor Thomas Achia

Lecturer, Strathmore Institute of Mathematics (SIMS)

Strathmore University

Ferdinand Othieno,

Dean, Strathmore Institute of Mathematical Sciences,

Strathmore University

Professor Ruth Kiraka,

Dean, School of Graduate Studies,

Strathmore University

Abstract

Background

Kenya has made significant progress in the elimination of mother to child transmission of HIV through increasing access to HIV treatment and improving the health and well-being of women and children living with HIV. Despite this progress, broad geographical inequalities in infant HIV outcomes still exist. This study aimed at assessing the spatial distribution of HIV amongst infants, areas of abnormally high risk and associated risk factors for mother to child transmission of HIV.

Methods

Data were obtained from the Early infant diagnosis (EID) database that is routinely collected for infants under one year for the year 2017. We performed both areal and point-reference analysis. Bayesian hierarchical Poisson models with spatially structured random effects were fitted to the data to examine the effects of the covariates on infant HIV risk. Spatial random effects were modelled using Conditional autoregressive model (CAR) and stochastic partial differential equations (SPDEs). Inference was done using Integrated Nested Laplace Approximation. Posterior probabilities for exceedance were produced to assess areas where the risk exceeds 1. The Deviance Information Criteria (DIC) selection was used for model comparison and selection.

Results

Among the models considered, CAR model (DIC = 306.36) performed better in terms of modelling and mapping HIV relative risk in Kenya. SPDE model outperformed the spatial GLM model based on the DIC statistic. The map of the spatial field revealed that the spatial random effects cause an increase or a decrease in the expected disease count in specific regions.

Highly active antiretroviral therapy (HAART) and breastfeeding were found to be negatively and positively associated with infant HIV positivity respectively [-0.125, 95% Credible Interval = -0.348, -0.102], [0.178, 95% Credible Interval -0.051, 0.412].

Conclusion The study provides relevant strategic information required to make invest-

ment decisions for targeted high impact interventions to reduce HIV infections among infants in Kenya.

Keywords

Markov Chain Monte Carlo, Integrated Nested Laplace Approximation, Stochastic Partial Differential Equation, Early Infant Diagnosis

Table Of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Name
AIDS	Acquired Immune Deficiency Syndrome
CAR	Conditional Autoregressive Model
EID	Early Infant Diagnosis
eMTCT	Elimination of Mother-to-Child Transmission
GMRF	Gaussian Markov Random Field
GRF	Gaussian Random Field
HAART	Highly Active Antiretroviral Therapy
HEI	HIV Exposed Infants
HIV	Human Immunodeficiency Virus
INLA	Integrated Nested Laplace Approximation
LGM	Latent Gaussian Models
MCMC	Markov Chain Monte Carlo
MOH	Ministry of Health
MTCT	Mother-to-Child Transmission
NASCOP	National AIDS and STI Control Programme
PCR	Polymerase Chain Reaction
PEPFAR	President's Emergency Plan for AIDS Relief
PMTCT	Prevention of Mother-to-Child Transmission
SPDE	Stochastic Partial Differential Equation
UNAIDS	The Joint United Nations Programme on HIV AIDS

Acknowledgements

Chapter 1

Introduction

1.1 Background

The global super-fast-track framework launched by The Joint United Nations Programme (UNAIDS) and President's Emergency Plan for AIDS Relief (PEPFAR) in 2016 aims at eliminating new HIV infections among children to less than 20,000 (start free), preventing new infections among adolescents and young women (Stay free) end paediatric and adolescents AIDS (AIDS free) by 2020 (PEPFAR et al., 2016). This framework builds on the previous global framework HIV "Global Plan towards the elimination of new HIV infections among children by 2015 and keeping their mothers alive".

Globally, there has been a remarkable progress in elimination of mother to child HIV transmission. There was a decline in the number of new infections among children from 290,000 (250,000-350,000) in 2010 to 150,000 (110,000-190,000) in 2015 reflecting the scale-up of coverage of prevention of mother-to-child transmission services (PMTCT) (UNAIDS, 2016).

In Sub-Saharan countries, new HIV paediatric infections declined marginally between 2010 to 2015. A decline of 66% and was reported in Southern and Eastern Africa while 31% was reported in Western and central Africa (UNAIDS, 2016).

Despite these achievements, concerted efforts are required to reach the UNAIDS targets

by 2020 (PEPFAR et al., 2016). This can be achieved through ensuring all pregnant women living with HIV receive lifelong ART, scale up the provision of early diagnosis, treatment optimization and care for HIV exposed infants to prevent mother to child transmission of HIV (WHO, 2016b).

Elimination of mother to child transmission (eMTCT) of HIV would directly contribute to the attainment of Sustainable Development Goal targets of reduction of maternal mortality ratio, ending preventable deaths of newborns and children under 5 years and ending the AIDS epidemic (Griggs et al., 2013).

In Kenya, an estimated 6,613 infants were infected with HIV through mother to child transmission in 2015 (NASCOP, 2015). This is a decline in mother to child transmission rates from 16% in 2013 to 8.3% in 2015 (NASCOP, 2015). These gains are attributed to the increase in HIV treatment coverage for women and infants. Kenya is one of the 22 priority countries focused for reduction of mother-to-child Transmission (MTCT) of HIV and has been chosen to be validated for the pre-elimination of MTCT of HIV and Syphilis by 2021.

The government of Kenya through the Ministry of Health (MOH) established the PMTCT program as part of continuing strategies for dealing with the epidemic and interventions to reduce MTCT of HIV. The goal of the program is to reduce the rate of MTCT of HIV to less than 5% and reduce maternal mortality by 50% in line with the "Global Plan towards the elimination of new HIV infections among children by 2015 and keeping their mothers alive".

The program implements the four-pronged strategy to prevent mother to child transmission of HIV :

1. Prevention of HIV transmission to HIV-negative women of reproductive age (primary prevention),

2. Prevention of unintended pregnancies among HIV-positive women by using family planning,

3. Prevention of HIV transmission from HIV-positive mothers to their babies during pregnancy, labour and delivery and infant feeding,

4. Provision of treatment, care and support to women infected with HIV, their children, and their families.

In Kenya, the early infant diagnosis (EID) program, under the umbrella of the PMTCT program, is responsible for making HIV diagnosis in HIV exposed infants (HEI) and young children under 18 months of age. Polymerase chain reaction (PCR) is the most common virological test used in PMTCT settings for HIV exposed babies and there are seven laboratories nationally with capacity to conduct PCR testing of HIV. According to the current HEI testing guidelines, DNA PCR should be conducted 6 weeks or at first contact thereafter, 6 months and 12 months for those breastfeeding. HIV antibody test 6 weeks after cessation of breastfeeding and HIV antibody test at 18 months and every 6 months thereafter until complete cessation of breastfeeding.

Despite the marked progress in elimination of mother to child transmission (eMTCT) of HIV, few counties in Kenya are still registering high number of infections. Some of the counties include; Laikipia, Taita taveta, Kwale, Mombasa, Mandera, Wajir, Lamu, Mandera. New HIV infections amongst infants exhibits marked geographical disparities with counties contributing disproportionately high number of new infections annually. To accelerate and achieve a drastic reduction in the new HIV infections, focussed efforts needs to be directed to the non performing counties.

Disease mapping has developed immensely in the recent years due to availability of geo-referenced data with advances in computing, geographical information systems (GIS) and statistical methodology. The interest lies in providing estimates of the relative risks of a disease across a geographical study area, assessing clustering and clusters of disease and assessing geographical distribution of disease in relation to potential risk factors (spatial regression). Disease mapping is often used as an exploratory tool to calculate and visualize disease risk across space allowing for an ad-hoc way of examining areas of higher risk and their risk factors through the use of statistical techniques.

Disease clustering is a technique used to assess whether a disease is clustered and where the clusters are located to identify potential environmental hazards. There is extensive use of scanning techniques for disease cluster detection after (Kulldorff & Nagarwalla, 1995) developed these methods. One of the first representations of disease on a map was by John Snow (Snow, 1854) who studied the geographical distribution of cholera victims through creating a dot map related to the location of the broad street pump.

Bayesian methods which offer a flexible and robust approach are increasingly being utilized in disease mapping. In disease mapping, a popular spatial model is the conditional autoregressive model (CAR) (Besag et al., 1991). In the conditional autoregressive model (CAR), the conditional distribution is modelled via the neighbourhood structure where the random effects in a region given all the others is the weighted average of all the other random effects. The weights are based on the neighbouring structure on adjacent areas. Extensions of the CAR include the Besag-York-Mollie (BYM) model, the proper CAR and the Leroux model.

1.2 Problem Statement

Bayesian approach requires integration over high dimensional probability distributions to make inference about a parameter of interest and predictions. A limitation of the Bayesian approach has been the Intractabilities involved in the calculation of the posterior distribution (Ntzoufras, 2011). The advent of the MCMC algorithms in the early 1990's for obtaining posterior distribution in combination with the rapid evolution of computers sparked an increase in application of Bayesian statistics in statistical research. MCMC is essentially Monte Carlo Integration using Markov chains (Gilks et al., 1995). The technique enables simulation from a probability distribution of the unknown quantities through construction of a Markov chain that eventually converges to the target (equilibrium/stationary) distribution which should be the posterior distribution.

The primary ways of constructing these chains are metropolis-Hastings algorithm and

Gibbs sampler.

In the last few years, MCMC methods have boosted the implementation of fixed effects and hierarchical models particularly in spatial and spatial temporal field (Ntzoufras, 2011). Despite the progress made in bayesian computing, MCMC methods are not without potential problems. MCMC samplers involve computationally and time intensive simulations especially for high dimensional models such as hierarchical models. The computational issue is due to the infeasibility of the linear algebra operations involving big dense covariance matrices when large spatial datasets are present (Brooks et al., 2011). Furthermore, Parameter estimation might be impossible and the algorithms may induce large Monte Carlo standard errors if they're not run for many iterations (Gilks et al., 1995). Additionally, MCMC methods present issues with convergence of the algorithm to the posterior distribution as well as choice of prior distributions.

Analysis of large datasets with vast level of spatial disaggregation could lead to long computation time to perform Bayesian inference via MCMC. Cross validation tests and sensitivity analyses might be impractical because of the computational demands that come with MCMC. This can translate to poor interpretation of the results.

The Integrated Nested Laplace Approximation (INLA) proposed by (Rue et al., 2009) provides a alternative approach to handling these complexities and providing precise and consistent estimates within a short computational time. Whereas MCMC algorithms take hours or days to run, INLA will take seconds or minutes to run. This is due to the fact that INLA is parallelized thus making it possible to exploit the new trend of having multicore processors. In addition, this approach permits automation of the inference process and can be used to analyse latent gaussian models (Rue et al., 2009). INLA is designed for latent Gaussian models ranging from generalised linear (mixed) models, generalised additive (mixed) models, geoadditive models and time series models (Martino & Rue, 2009). The development of R package named R-INLA has proved valuable for the implementation of INLA. In addition, INLA can be integrated with the Stochastic Partial Differential Equation (SPDE) approach proposed by (Lindgren et al., 2011) to execute

spatial and spatio-temporal models for geostatistical data.

There has been a gradual increase in the utilization of these methods in analysis of epidemiological and public health data particularly in spatial and spatial temporal models. This project describes how spatial models can be fitted to areal and point-referenced data to assess the association between maternal/infants covariates and HIV sero-conversion among HEI using INLA.

1.3 Objectives

1.3.1 Main Objective

To develop a Bayesian model for HEI born to HIV positive mothers in Kenya using INLA.

1.3.2 Specific Objectives

i. To review and investigate the relative performance of conditional autoregressive models for analysing areal data within the integrated nested laplace approximation.

ii. To review and investigate the statistical properties of the spatial partial differential equations (SPDE) approach with INLA for the analysis of point reference spatial data.

iii. To develop a conditional autoregressive model that assesses the relationship between HIV positivity in infants born to women in Kenya and various mother specific, facility specific and county specific covariates using INLA.

iv. To develop a smooth map for HIV positivity for Kenyan EID data within the SPDE framework in INLA.

Chapter 2

Epidemiological Literature Review

2.1 Established risk factors for mother to child HIV transmission

Mother-to-child transmission of HIV can occur in utero (intrauterine), during labour, delivery (intrapartum) and through breastfeeding (postpartum). In the absence of PMTCT interventions, transmission rates ranges from 15% to 45% (WHO, 2016b). Various factors are known to increase the risk of HIV transmission from mother to infant. Understanding these factors is crucial in identifying potential interventions against disease progression and transmission. Risk factors for vertical transmission can be grouped into three categories; maternal, infant and obstetric factors. A study by (Landesman et al., 1996) showed that duration of ruptured amniotic membranes is significantly associated with the risk of HIV transmission. The results indicated prolonged (more than four hours) rupture of membranes doubled the risk of transmission of HIV. According to the WHO guidelines (WHO, 2016a), HIV positive mothers on ART should exclusively breastfeed their infants for the first six months of life, introducing complementary foods thereafter and continue breastfeeding until 24 months or beyond. Breastfeeding should stop once a nutritionally adequate and safe diet without breast milk can be provided. Longer durations (24 months or beyond) of breastfeeding has been shown to improve HIV free survival among HIV

exposed infants particularly in settings where pneumonia and diarrhoea are significant causes of child mortality. However, the risk of transmission is high if a woman is not receiving ART or when maternal ART adherence is inconsistent (WHO, 2016a).

Postnatal transmission through breastfeeding was observed in a infant who was breast fed by a seropositive wet nurse (Colebunders et al., 1988). A randomized clinical trial to assess infant HIV-infection and mortality rates between breastfeeding and formula feeding intervention groups conducted in Nairobi Kenya indicated that breast milk accounted for 44% of all infant HIV infections among those exposed to breast milk (Nduati et al., 2000). In addition, according to the study, 75% of breast milk transmission occurred within 6 months of breastfeeding.

A couple of studies have reported that maternal viral load is a strong risk factor for both utero and intapartum transmission. A high maternal viral load increases the likelihood of perinatal transmission of HIV (Thea et al., 1997; Mock et al., 1999; Bryson, 1996; Ioannidis et al., 2001). Low CD4 (<1000 cells/ml) count has been shown to be significantly associated with increased risk of MTCT of HIV (Semba et al., 1999; Ngwende et al., 2013; Louis et al., 1993; Fawzi et al., 2001).

The mode of delivery has been shown to be associated with the rate of MTCT of HIV. Evidence suggests that caesarian section in comparison with vaginal delivery increases an infants exposure or susceptibility to acquiring HIV (DeHovitz et al., 2000).

Consistent use of highly active antiretroviral therapy (HAART) has been shown to suppress viral replication and increase CD4 counts substantially (DeHovitz et al., 2000; Ledergerber et al., 1999; Lucas et al., 1999). In addition, the use of infant prophylaxis is recommended to decrease the risk of HIV transmission (Bernstein et al., 2001).

In 2015, Kenya adopted lifelong ART option B+ as the preferred regimen for pregnant and breastfeeding women regardless of WHO clinical age and at any CD4 cell count as per the World health organization recommendations (World Health Organization(WHO), 2015).

2.2   INLA for Areal data

Aggregated data collected over space on irregular/regular polygons with well defined administrative boundaries are referred to as lattice/areal data. Areal data is used when the main interest lies in smoothing or mapping an outcome over space to assess the spatial distribution of a disease and identify areas characterized by unusually high or low relative risk (Lawson, 2013). INLA which is a novel numerical inference approach has been used widely in areal data analysis to implement Bayesian hierarchical spatial models.

(Beguin et al., 2012) investigated the performance of INLA and MCMC methods in analysis of spatially autocorrelated ecological data on the distribution of woodland caribou in Eastern Canada. Findings of the study indicated that INLA performs better in terms of accuracy of results and rapidity as compared to MCMC. The CAR model fitted in GIBBS took many hours to converge whereas INLA took 5 seconds. The fast algorithms allowed for both sensitivity analyses on priors and cross-validation tests to be performed. Furthermore, similar results for parameter estimates for the Bayesian CAR models were obtained from the two algorithms. (Moraga et al., 2017) proposed a joint Bayesian model for fusion of concentration of fine particulate matter $PM_{2.5}$ data obtained at point and areal resolutions. The models were fitted using INLA and SPDE approaches. Although the models were limited to Gaussian data and could not be extended to non-Gaussian data, they reported quicker results by these methods. Unlike MCMC based methods that require convergence of chains, INLA uses numerical approximations to obtain the variables of interest. A recent study revealed that there is some substantial differences in the performance of OpenBUGS and INLA (Carroll et al., 2015). Comparisons of the two algorithms was accomplished by calculating the number of effective parameters, mean Squared error (MSE), mean squared predictive error (MSPE) and DIC for the models fitted. INLA outperformed OpenBUGS in computation time though as shown by the MSE, OpenBUGS outperformed INLA with respect to the precision parameter estimates for the spatial random effects. In a simulation study done by (Konstantinoudis et al., 2018) to investigate the performance of discrete and continuous domain models to recover risk

surfaces and identify high-risk areas, inference was conducted using INLA. They used root mean integrated squared error evaluated on a fine grid to gauge the potential of the models to capture the true risk and found out that log-Gaussian Cox processes (LGCPs) models outperformed BYM models in quantifying disease risk over space and identifying areas of high risk. (Okuto, n.d.) investigated the spatial distribution of tuberculosis treatment outcomes in Kenya given risk factors associated with the disease from January 2014 to March 2014. Test for spatial autocorrelation was done using Moran's I and Inference was done using INLA.

2.3   INLA for Point-reference data

In a recent paper by (Simpson et al., 2012) that compared two approximations to gaussian random fields (GRFs) with Matérn covariance functions : the kernel convolution approximation and the stochastic partial differential equation approach, the SPDE approach seemed to outperform the kernel approximations in terms of accuracy and computational efficiency for approximating the class of Matérn random fields. Kernel methods proved to be unstable and the approximations depend on the scale and smoothness parameters unlike finite element basis functions which do not have this problem. SPDE approach was employed to estimate and predict a spatial-temporal hierarchical model that involved a Gaussian field and a state process characterized by first order autoregressive dynamics for particulate matter concentration (Cameletti et al., 2013). Computational strength of the SPDE clearly stood out and there were no issues of convergence and mixing. (Musenge et al., 2013) investigated the effects of different covariates on zero inflated spatiotemporal HIV/TB child mortality data using INLA and SPDE approach. They fitted zero inflated Poisson and Binomial spatial temporal models to the data and assessed the accuracy of the posterior marginals using the effective number of parameters and DIC. The analysis demonstrated that the big "n" problem can be resolved with the aid of SPDE and INLA.

Chapter 3

Critical Review of Statistical Methods

3.1 Bayesian Inference

Implementation of the MCMC algorithms in combination with the rapid evolution of computers has made Bayesian approach to gain popularity over the years amongst researchers. One of the advantages of Bayesian approach is that it takes into account the uncertainty in the estimates and predictions and its potential to deal with issues like missing data (Blangiardo & Cameletti, 2015).

Bayesian statistics differ from the classical statistics since all unknown parameters are considered to be random variables. Prior distribution which expresses the information available to the researcher must be defined initially. The goal of Bayesian inference is the calculation of the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ of the parameters $\boldsymbol{\theta}$ given the observed data y where $\boldsymbol{\theta}$ denotes the vector of the model parameters. This distribution which is based on the Bayes theorem can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \propto p(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}), \tag{3.1}$$

where $p(\boldsymbol{\theta})$ is the prior distribution and $\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ which is the constant of proportionality ensures the posterior distribution integrates to 1.

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}),$$

is the likelihood of the model. The combination of the likelihood and the prior results to the posterior distribution.

Prior specification is crucial in Bayesian inference since it influences the posterior inference. To assess the robustness of the posterior distribution to the selection of the prior distribution, we conduct a sensitivity analysis in which we use different priors and assess changes in the posterior distribution. Bayesians often use non-informative priors to ensure that all information about the posterior comes from the data and not the prior. When prior distributions for the parameters have parameters controlling its form and since they are regarded as stochastic, then these parameters must also be characterized by a distribution. The parameters are known as hyperparameters and the distributions are known as hyperprior distributions. This idea of parameters arising from distributions leads to the fundamental feature of hierarchies which is common in disease mapping models.

Consider data vector **y** and the latent field **x** with a vector of hyperparameters $\boldsymbol{\theta}$. the three stages of the hierarchical model are as follows:

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \prod_i p(y_i|\boldsymbol{x},\boldsymbol{\theta}),$$

$$x|\boldsymbol{\theta} \sim p(\boldsymbol{x}|\boldsymbol{\theta}),$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}),$$

with $\prod_i p(y_i|\boldsymbol{x},\boldsymbol{\theta})$ denoting the data model, $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a GMRF and $p(\boldsymbol{\theta})$ hyperprior (random effects covariance parameters). The corresponding posterior distribution becomes:

$$p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_i p(y_i|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.2}$$

3.2   Inference with Monte Carlo Methods

Bayesian inference requires evaluation of complex integrals of the type

$$E[g(\boldsymbol{\theta}|\boldsymbol{y})] = \int_{\Omega} g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{3.3}$$

to obtain quantities of interest such as the mean $E[g(\boldsymbol{\theta}|\boldsymbol{y})]$ where g(.) is some function of the parameter of interest $\boldsymbol{\theta}$.

One of the proposed solutions in the literature is based on random samples generation and computing the statistical unbiased estimate the sample mean (Ntzoufras, 2011).

MCMC methods works by generating samples from the posterior distribution $p(\boldsymbol{\theta}|y)$ to estimate quantities such as mean and variance. These techniques are based on construction of a Markov chain that eventually converges to the target distribution (stationary or equilibrium) which is the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$.

A Markov chain is a stochastic process of a set of random variables $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ..., \boldsymbol{\theta}^T$ such that

$$p(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t, ..., \boldsymbol{\theta}^1) = p(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t) \tag{3.4}$$

the distribution of $\boldsymbol{\theta}$ at sequence t+1 given all the preceding values of $\boldsymbol{\theta}$ depends only on the value of $\boldsymbol{\theta}$ of the previous sequence t.

As t$\rightarrow \infty$, the distribution of $\boldsymbol{\theta}^t$ converges to its equilibrium distribution if the Markov chain is aperiodic, irreducible and positive-recurrent (Egordic) (Ntzoufras, 2011).

The combination of the Monte Carlo methods on samples from a Markov chain led to the development of MCMC methods. MCMC methods are not without limitations including lack/slow convergence of the algorithm, length of burn-in periods and choice of starting values (Brooks et al., 2011). Some of the ways to assess convergence include : Monitoring the MCMC error and autocorrelation plot, monitoring trace plots and evolution of the ergodic mean, running multiple chains with different starting points. Two most common MCMC algorithm are the Gibbs sampler and the Metropolis Hastings algorithm. In the

Metropolis Hastings algorithm, one samples from the proposal distribution and based on the acceptance probability, creates a Markov chain (Metropolis et al., 1953). Gibbs sampler (Geman & Geman, 1984) is a special case of the Metropolis Hastings that requires that one samples from full conditional distributions. Convergence using Gibbs sampler is expected to be faster than the Metropolis Hastings algorithm (Carlin et al., 2014).

3.3 Latent Gaussian Models

INLA algorithm used to carry out approximate Bayesian inference is used with latent Gaussian models which are a sub class of structured additive regression models.

In structured additive regression models, the response variable y_i belongs to the exponential family of distributions such as Binomial, Gaussian, Poisson and many others. The structured additive predictor η_i defines the mean (μ_i) through a link function g(.) such that g(μ_i) = η_i. The linear predictor η_i is defined as:

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m t_{mi} + \sum_{l=1}^{L} f_l z_{li} + \varepsilon_i, \tag{3.5}$$

where β_0 is the model intercept, β_m's quantify the linear effects of covariates $\mathbf{t} = (t_1, t_2, \ldots, t_M)'$, $\mathbf{f} = (f_1(.), f_2(.), \ldots, f_L(.))'$ is a function of the covariates, $\mathbf{z} = (z_1, z_2, \ldots, z_L)'$ and the ε_i's are the error terms. $f_l(.)$ can assume different forms such as spatial random effects, spline, time trends, non linear effects of covariates (Blangiardo & Cameletti, 2015).

Latent Gaussian models are under these class of models with a structured additive predictor η_i (1). The latent (unobserved) variables, denoted by $\mathbf{x} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\varepsilon})$ are assigned a Gaussian prior whereas $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ denotes the vector of K hyperparameters. When the latent field $\boldsymbol{x}$ has Markov properties. That is, it satisfies the conditional independence assumptions, then it is referred to as a Gaussian Markov random field (GMRF). GMRFs are exceedingly useful in practice because of the lower computational cost owing to the sparseness of the precision matrix and the structure of its non-zero terms see (Rue & Held, 2005) for a detailed account.

Assuming conditional independence, the likelihood of the data is given by

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i}^{n} p(y_i/x_i, \boldsymbol{\theta}).$$

Assume a multivariate Normal prior on **x** with mean **0** and precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ (inverse covariance). That is,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \sim Normal(\boldsymbol{0}, \boldsymbol{Q}^{-1}),$$

with density function given by $p(\boldsymbol{x}|\boldsymbol{\theta}) = (2\pi)^{-n/2} \mid \boldsymbol{Q}(\boldsymbol{\theta}) \mid^{1/2} exp(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x})$. Assuming that $y_i, i = 1, 2, \ldots, n$ are conditionally independent, the joint posterior distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}) \times p(\boldsymbol{x}|\boldsymbol{\theta}) * p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}).$$

3.4   Integrated Nested Laplace Approximations

INLA is a deterministic Bayesian inference approach which works by approximating the marginal posterior distributions for the parameter vectors. Marginal distributions are obtained by factoring out a term from the joint distribution. This means we need to solve the following integrals:

$$p(x_i|\boldsymbol{y}) = \int p(x_i, \boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}. \tag{3.6}$$

$$p(\boldsymbol{\theta}_k|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k}. \tag{3.7}$$

.

Using the conditional probability rules (3.2) becomes

$$p(x_i|\boldsymbol{y}) = \int p(x_i, \boldsymbol{\theta}|\boldsymbol{y}) * p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}. \tag{3.8}$$

We need to compute

   (i) $p(\boldsymbol{\theta}|\boldsymbol{y})$ from which all the $p(\boldsymbol{\theta}_k|\boldsymbol{y})$ can be obtained;

   (ii) $p(x_i, \boldsymbol{\theta}|\boldsymbol{y})$. This is required to compute the posterior marginals of the latent field.

INLA performs numerical approximation to the posteriors of interest based on the Laplace approximations (Tierney & Kadane, 1986). The INLA scheme proceeds in three successive steps:

   1. Compute the posterior marginals of the hyperparameters p($\boldsymbol{\theta}$|y)

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&\propto \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&\approx \left. \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}*(\boldsymbol{\theta})} \\
&=: \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}).
\end{aligned} \tag{3.9}
$$

Here $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{x}$ and $\boldsymbol{x}*(\boldsymbol{\theta})$ is the mode of $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ given $\boldsymbol{\theta}$. Laplace approximation proposed by (Tierney & Kadane, 1986) is used for $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. For a univariate x with a density function that can be written as $\exp f(x)$, we can evaluate the integral $\int \exp$ f(x) by representing f(x) as a Taylor series expansion evaluated at $x = x_0$. Setting $x_0 = x*$ to be the

mode of the function f(x), the integrand is the kernel of a Normal distribution that is, $\sim$ N(x*,$[d^2 log f(x)/dx^2]^{-1}$).

$\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is used to select good evaluation points to integrate out the uncertainty with respect to $\boldsymbol{\theta}$ when approximating the posterior marginals of each parameter x_i. (Rue et al., 2009) highlighted the importance of representing $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ in a non parametric way to ensure accuracy of the approximations.

To select the values of the hyperparameters $\boldsymbol{\theta}$, (Rue et al., 2009) suggested to perform a grid exploration of $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ by locating its mode through some quasi-Newton algorithm. This is followed by computation of the negative Hessian matrix $\mathbf{H} > 0$ at the modal configuration to construct principal components. Finally performing a grid search to locate the mass of the probability distribution. Due to the computational cost associated with this exploration scheme, (Blangiardo & Cameletti, 2015) suggested using this approach when the dimension of $\boldsymbol{\theta}$ is < 4 otherwise central composite design (CCD) which is the default option in R-INLA) should be used. In the above procedure we have evaluated $\tilde{p}(\boldsymbol{\theta}_q|\boldsymbol{y})$ on q points. Posterior marginals for $\boldsymbol{\theta}_j$ can be obtained from $\tilde{p}(\boldsymbol{\theta}_q|\boldsymbol{y})$ through numerical integration.

2. Build a Laplace approximation (or or its simplified version) to p($x_i|\boldsymbol{\theta}$,**y**).

   To approximate the posterior marginals of the latent field for selected values of the hyperparameters, (Rue et al., 2009) proposed three approximations that is Gaussian, Laplace and simplified Laplace approximation.

   $p(x_i|\boldsymbol{\theta}$,**y**) can often be well approximated with a Gaussian distribution derived from $\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ by matching the mode and the curvature at the mode.

   While this is a simple and fast approximation to $p(x_i|\boldsymbol{\theta}, \mathbf{y})$ it present errors in shape (lack of skewness) and location (Rue & Martino, 2007).

Laplace approximation given by

$$
\begin{aligned}
p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) = \frac{p(x_i, x_{-i})|\boldsymbol{\theta}, \boldsymbol{y})}{p(x_{-i}|(x_i, \boldsymbol{\theta}, \boldsymbol{y})} &= \frac{p(x, \boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})p(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} = \frac{p(x, \boldsymbol{\theta}|\boldsymbol{y})}{p(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\
&\approx \frac{p(x, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{p}(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\Bigg|_{x_{-i}=x*_{-i}(x_i,\boldsymbol{\theta})} \\
&= \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y}).
\end{aligned}
$$

$$(3.10)$$

$\tilde{p}(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ is the Laplace Gaussian approximation to $p(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ and $x*_{-i}$ $(x_i, \boldsymbol{\theta})$ is its modal configuration. $\tilde{p}(x_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ must be computed for each value of $\boldsymbol{\theta}$ and x rendering the approximation computationally intensive.

A computationally efficient alternative is the simplified Laplace approximation $\tilde{p}_{SLA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ which corrects for skewness and location in Gaussian approximation. $\tilde{p}_{SLA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is derived by performing a Taylor series expansion of $\tilde{p}_{LA}(x_i|, \boldsymbol{y})$ around $x_i = \mu(\boldsymbol{\theta})$ and corrects by including a mixing term (For example, cubic splines) to increase the fit to the required distribution (Blangiardo & Cameletti, 2015).

3. Numerical integration (finite sum) to obtain approximations of the marginal of interest x_i

The marginal posterior distributions $p(x_i|\boldsymbol{\theta})$ are approximated by

$$\tilde{p}(x_i|\boldsymbol{y}) \approx \int \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{3.11}$$

which is integrated numerically with respect to $\boldsymbol{\theta}$ through a finite weighted sum:

$$\tilde{p}(x_i|\boldsymbol{y}) \approx \Sigma_q \tilde{p}(x_i|\boldsymbol{\theta}^q, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}^q|\boldsymbol{y})\triangle_q, \tag{3.12}$$

where $\boldsymbol{\theta}^q$ are the integration points with corresponding set of weights $\triangle_q$.

3.5   Stochastic Partial Differential Equation

Geostatistical data are realizations of a stochastic/random process $Z(s) : s \in D$ where D is a fixed subset of R^d with positive d-dimensional volume and the spatial index s varies continuously in the domain D (Cressie, 1993).

Let s be a continuously indexed GRF, X is a covariate that varies spatially, β is the effect of covariate X, s is the data location and ϵ_n is Gaussian noise that is i.i.d for the observations.

$$Z(s_n) = \boldsymbol{X}(s_n)\boldsymbol{\beta} + s(s_n) + \epsilon_n. \tag{3.13}$$

The random process is a Gaussian field if the collection of the random variables $Z(s_1), Z(s_2), ..., Z(s_n)$ have a multivariate Normal distribution for integer n and set of locations s_n with mean $\boldsymbol{\mu} = (\mu(s_1, ..., \mu(s_n))$ and covariance matrix $\boldsymbol{\Sigma}$ that is Cov(z(s), z(s')).

The process $Z(s) : s \in D$ is weakly stationary if E [Z(s)] = μ for some finite constant μ and Cov (Z(s), Z(s') = c($\| s - s' \|$), where $\| . \|$ denotes the Euclidean distance. Isotropy means that the covariance function is directionally invariant so that only the distance between two points determines their covariance and not the direction you have to travel to get from the first to the second.

For a finite set of locations, the covariance function must induce a positive definite matrix. Some of the isotropic covariance functions used in geostatistics include; Exponential, Gaussian, Powered exponential and Matérn.

Hierarchical models used in large spatial and spatial temporal datasets require repeated evaluations of the likelihood and the conditional densities arising from the process. Such computations involve computation of the inverse and determinant of the covariance matrix. When n is large, the computation of these matrices becomes very slow or even infeasible. Spatial covariance matrices are in general dense and the Cholesky factorization require about $O(\frac{n^3}{3})$ flops. This problem is known as "the big n problem" (Carlin et al., 2014). Beyond the computational issues, Gaussian random field can present other challenges such as introduction of non-stationarity in the covariance function, creation of

a covariance function for other geometries such as spheres (Bakka et al., 2018). To circumvent this problem, (Carlin et al., 2014) proposed two approaches that is approximate likelihood approaches and developing models that can can be fitted with large values of n. One of the approximate likelihood approaches is approximating a Gaussian process with a Gaussian Markov random field (GMRF). GMRFs are commonly in spatial discrete models and a practical question would be how do we extend this to continuously indexed spatial models owing to their Markov property. Gaussian Markov random fields (GMRFs) is a finite-dimensional random vector following a multivariate Gaussian distribution with conditional independence properties, hence termed as Markov. We use undirected graphs to represent the conditional independence structure in a GMRF.

More specifically, a GMRF is a random vector $\boldsymbol{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n$ wrt to an undirected graph $\mathcal{G} = (\mathbb{V}, \varepsilon)$ with mean μ and precision matrix $\mathbf{Q} > 0$ with a density

$$p(\boldsymbol{x}) = (2\pi)^{\frac{-n}{2}}|\mathbf{Q}|^{\frac{1}{2}}exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\mathbf{Q}(\boldsymbol{x} - \boldsymbol{\mu})). \tag{3.14}$$

x_i and x_j are conditionally independent given variables x_{-ij} that is $x_1 \perp x_2 | x_{-ij}$ for $i \neq j$ thus $Q_{ij} = 0$ (Rue & Held, 2005).

The density of the matrix is given by the number of non-zero elements divided by the number of elements. Most precision matrices for GMRF are sparse where only $\mathbb{O}(n)$ of the terms of the n^2 in $\mathbf{Q}$ are not zeros. This means sparse matrices are mostly characterized by zeros and majority of them are not stored in the computer. This translates to fast numerical factorization of $\mathbf{Q}$ as SS' where S is the lower triangular , matrix and S' is the conjugate transpose of S. It is easy to produce random samples and compute log density of equation above and marginal variances from a GMRF using the Cholesky triangle (Rue, 2001).

The computational cost involved for two dimensional GMRFs is $\mathbb{O}(n^{\frac{3}{2}})$ which is a speed up compared to $\mathbb{O}(n^3)$ of the GF where n is the dimension of the GMRF. Furthermore,

beyond the fast computations, GMRF are quite stable with respect to conditioning. An example of GMRF is the autoregressive process of order 1 in time series.

(Lindgren et al., 2011) provided an explicit link between Gaussian fields with the Matérn covariance function for certain choices of the smoothing parameter (ν) and GMRFs.

The SPDE approach consists in representing a continuous indexed GF spatial process using a discretely indexed GMRF (Blangiardo & Cameletti, 2015) which in turn produces substantial computational advantages. This approach extends the work of (Besag, 1981) who approximated a Gaussian field when $\nu \to 0$ in the Matérn correlation function. Gaussian field with Matérn covariance is a stationary solution to the linear fractional Stochastic partial differential equation (SPDE) (Whittle, 1963).

$$(\kappa^2 - \triangle)^{\frac{\alpha}{2}} (\tau x(s)) = \omega(s) \qquad u \in \mathbb{R}^d, \qquad \alpha = \nu + \frac{d}{2}, \qquad \kappa > 0, \nu > 0, \quad (3.15)$$

where $\triangle$ is the Laplacian given by:

$$\triangle = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}. \tag{3.16}$$

κ is the scale parameter, α controls the smoothness, τ controls the variance and ω is spatial Gaussian white noise process.

By solving a certain SPDE, (Lindgren et al., 2011) showed that a GRF with a Matérn correlation function and $\nu = 1$ or $\nu = 2$ has a GMRF representation. Matérn covariance function between locations $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ is defined as

$$Cov(X(s), X(t)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa \mid\mid t - s \mid\mid)^\nu K_\nu(\kappa \mid\mid t - s \mid\mid), \tag{3.17}$$

where $\mid\mid . \mid\mid$ denotes the Euclidean distance in $\mathbb{R}^d$, K_ν is the Bessel function of the second kind and order $\nu > 0$. $\kappa > 0$ is a scale parameter with the dimensions of distance and σ^2 is the marginal variance. The smoothness parameter $\nu > 0$ is a shape parameter that is usually fixed due to poor identifiability and it determines the differentiability of the

underlying process. The scaling parameter can be represented by a range parameter ρ. Where $\rho = \sqrt{(8\nu)}/\kappa$ corresponding to correlations near 0.1 at distance ρ, the Matérn covariance function becomes (Blangiardo & Cameletti, 2015):

$$Cov(s,t) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\sqrt{8\nu}\,||\,t - s\,||)/\rho)^{\nu}K_{\nu}(\sqrt{8\nu}\,||\,t - s\,||)/\rho). \qquad (3.18)$$

It is worth noting that when $\nu = 0.5$, the Matérn correlation function reduces to the exponential whilst as $\nu \to \infty$ to the Gaussian correlation function.

The link between the Matérn covariance function and the SPDE is given by $\nu = \alpha - \frac{d}{2}$ and the marginal variance is

$$\Sigma^2 = \frac{\Gamma(\nu)}{\Gamma\alpha(4\phi)^{\frac{d}{2}}\kappa^{2\nu}\tau^2}. \qquad (3.19)$$

For spatial sites are on irregular grid, (Lindgren et al., 2011) proposed overlaying a irregular grid (mesh) and constructing a finite element representation of the solution to the SPDE through a basis function that is defined on the triangulation of the domain D.

$$g(u) = \sum_{k=1}^{n} \phi_k(u)w_k, \qquad (3.20)$$

where ϕ_k is the set of basis functions (deterministic) and w_k are zero mean Gaussian weights and n is the number of vertices in the triangulation. w_k determine the values of the field at the vertices. The joint distribution of w = $w_1, ..., w_n$ is chosen so that the distribution of the functions g_u approximates the distribution of solutions to the SPDE on the domain. This distribution for the weights in the basis function is a GMRF.

To obtain a Markov structure, we use ϕ_k that are piecewise linear with compact support in each triangle defined such that $\phi_k = 1$ (sampling location on vertex k) and 0 (sampling location is not on vertex k) (Blangiardo & Cameletti, 2015).

As noted earlier, the construction is done by projecting the SPDE onto the basis representation in what is essentially a finite element method. The default value of α in R-INLA is

2 and the precision matrix $\mathbf{Q}$ for the Gaussian weight vector is:

$$\mathbf{Q} = \tau^2(\kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}), \qquad (3.21)$$

where C is a diagonal matrix and G is a sparse matrix.

R-INLA implements the Matérn field by creating a mesh that comprises of triangles. The mesh covers the entire domain and extends a bit more to account for boundary effects. The corners/edges of the triangles are known as vertices/nodes. It is important to note that when creating a mesh the triangles shouldn't be big or irregularly shaped as this will give poor interpolation results. Define the weighting factors w_k also called the projector matrix. For each of the vertices, R-INLA will estimate a w_k, Define the SPDE and the spatial field. Define a stack where we inform R-INLA at which sampling locations we have the data for the response and covariate. Finally we specify the model formulae and run the spatial model in R-INLA.

SPDE approach is flexible and can be extended to models on manifolds, non-stationary and non-separable models with small changes to the differential operator in the SPDE. The main objectives of geostatistical analysis are estimation (inference about the parameters of a stochastic model for the data) and prediction (inference about the unknown stochastic process at unknown locations.

Chapter 4

Methods

4.1 Study area and Data

The study was conducted in Kenya, a country situated in East Africa. The country is divided into 47 administrative units (counties). We used early infant diagnosis (EID) program data collected routinely by the Ministry of Health to investigate the spatial distribution patterns of HIV amongst infants in Kenya. Additionally, we aim to model infants HIV positivity as a function of maternal prophylaxis and breastfeeding. HEI are tested through EID which gives an opportunity for early identification of HIV and linkage to care and treatment services. The facility-level data comprised of 68,600 PCR tests types (Initial, 2nd, 3rd and confirmatory), PCR test results, sex and age of infant, infant and maternal prophylaxis, mother HIV status, infant breastfed, Entry point, testing laboratory, date samples received, tested and dispatched collected in the 47 counties in Kenya. The analyses were restricted to infants under one year born to HIV positive women. The data were aggregated to provide county level summaries that were used for areal analyses. The facility level EID data was used for geostatistical analysis. Sampling took place in 2,547 locations (see Figure 5.5) in 2017.

4.2   Statistical analysis

Variables

The outcome variable we considered in these analyses was the count of infant HIV positive test result. The covariates included in the model were proportion of infants breastfed and HAART. We excluded infant prophylaxis since it is highly correlated with HAART.

4.2.1   Statistical models

We developed Bayesian hierarchical Poisson regression models to investigate the spatial heterogeneity of HIV across the counties and assess the effects of the selected covariates. For the i-th district, the count of HIV cases y_i is modelled as $y_i \sim \text{Poisson}\,(E_i\theta_i)$, where the mean θ_i is the risk of infection and E_i is the expected number of infants infected with HIV in district i, i = 1,2,...,47.

The linear predictor is defined on the logarithmic scale:

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m t_{mi} + u_i,$$

where β_0 denotes HIV outcome rate for all the 47 counties and u_i are the spatially structured random effects that capture spatial variation modelled using conditional intrinsic autoregressive model (ICAR) prior distribution. The ICAR model is given as

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim \text{Normal}\left(\frac{1}{n_i}\sum_{j\sim i}^{n}\phi_j, \frac{\sigma^2}{n_i}\right) \tag{4.1}$$

(Besag & Kooperberg, 1995).

The conditional expectation of random effect ϕ_i is the average of the effects of its neighbours. The conditional variance depends on its number of neighbours n_i. An area with many neighbours will have a smaller variance.

Turning to the continuous spatial field, the linear predictor is defined on the logarithmic scale:

$$\eta_i = \log(\pi) = \beta_0 + \sum_{m=1}^M \beta_m t_{mi} + S(x_i),$$

where β_0 is the intercept, β are the covariates effects, S = S(x_1),..., S(x_n) is a gaussian process,

$$\mu(x) = E[S(x)] \tag{4.2}$$

$$\rho(h) = Cov S(x], S(x+h] = \sigma^2 Matern(|h|/\phi, \nu). \tag{4.3}$$

$\rho(h)$ is a symmetric definite function depending on the marginal variance σ^2 and a range parameter (ϕ), beyond which correlations fall beyond a certain threshold of 0.1. ν is the smoothness parameter and |h| is the distance between two points on the plane (Diggle et al., 1998).

To investigate the association between the outcome variable and the covariates, we fitted four Poisson models : **Model 1** the ordinary Poisson regression model, **Model 2** the generalized linear mixed model spatially unstructured random effects, **Model 3** the generalized linear mixed model spatially structured random effects, **Model 4** the generalized linear mixed model with both spatially unstructured and structured random effects.

For the geostatistical data, we investigated spatial dependency in the Pearson residuals by making a sample variogram. We proceeded with fitting two models: **Model 1** the spatial component model (GLM Poisson model), **Model 2** the model without spatial component (SPDE model).

Bayesian inference was done in R software (Team et al., 2013) using Integrated Nested Laplace approximation (R-INLA) package. A stochastic partial differential equation (SPDE) with INLA was employed to estimate the posterior marginals in the geostatistical analysis.

Model comparison and selection was done using the deviance information criterion (DIC) that takes into account the trade off between model fit and complexity (Spiegelhalter et al., 2002). The best model was given by the model with the smallest value of DIC.

Chapter 5

Results

5.1   Summary Results

Table 5.1 presents the summary statistics of the variables in the data. In total, of the 68,600 HIV exposed infants who provided their blood samples for PCR HIV testing, 2,363 (3.4%) turned positive for HIV. Of the 2,363 HIV positive women, 86.33% breast-fed their infants while the remaining 13.24% did not breastfeed their infants. Majority (77.06%) of these women accessed the health facility through the MCH/PMTCT. 61.32% of the infants who turned out positive were on prophylaxis while 63.98% of their mothers were on maternal prophylaxis. 81.17% of the infants who turned HIV positive had their first diagnosis after 2 months.

Table 5.1: Summary of infant and maternal risk factors associated with mother to child transmission of HIV in Kenya 2017

Variable	Total (N = 68,600)	HIV Positive (N = 2,363)
Entry Point		
CCC/PSC	6,544 (9.53)	233 (9.86)
IPD	293 (0.42)	87 (3.68)
Maternity	409 (0.59)	18 (0.76)
MCH/PMTCT	59,816 (87.19)	1,821 (77.06)
OPD	774 (1.12)	145 (6.13)
Other	350 (0.51)	35 (1.48)
Infant breastfed		
BF	57,889 (84.39)	2,040 (86.33)
NBF	10,711 (15.61)	313 (13.24)
Sex		
Female	34,651 (50.51)	1,197 (50.65)
Male	33,188 (48.38)	1,132 (47.91)
Infant prophylaxis		
Yes	51,681 (75.33)	1,449 (61.32)
No	16,919 (24.66)	914 (38.67)
Maternal prophylaxis		
Yes	62,769 (91.57)	1,512 (63.98)
No	5,831 (8.50)	851 (36.02)
Age at diagnosis		
Under 8 weeks	14255 (20.78)	445 (18.83)
Over 8 weeks	54345 (79.22)	1918 (81.17)

5.2 Areal data analysis

The results of fitting the hierarchical models are shown in table 5.2. Based on these results, we deduce that the model with spatially structured random effects (Model 3) offered a better fit (DIC 306.36).

Increased use of HAART is associated with mother to child transmission of HIV. -0.80 (95% credible interval:-2.19, 0.60). Breast feeding is positively associated with mother to child transmission of HIV 0.5576 (95% credible interval:-1.62,-2.77). The results however suggest that the covariates are not significant.

Table 5.2: Posterior statistics of the four models

| Variable	Model 1	Model 2	Model 3	Model 4
β_0	3.3285 (1.8445, 4.7958)	3.9853 (1.2337, 6.6126)	3.0724 (0.3483, - 5.6721)	3.7619 (1.0254, 6.3732)
HAART	-0.3422 (-1.1703, 0.4926)	-0.5565 (-2.0945, 0.9918)	-0.8020 (-2.1955, 0.6030)	-0.6588 (-2.1714, 0.8674)
BF	-0.4925 (-1.7174, -0.7381)	-0.1736 (-2.4892, -2.1779)	0.5576 (-1.6219, -2.7776)	0.0399 (-2.2514, -2.3716)
medianage	-0.3028 (-0.4115, -0.1922)	-0.3777 (-0.5661, -0.1791)	-0.3173 (-0.5125 , -0.1114)	-0.3632 (-0.5504, -0.1660)
τ_v		20.47 (9.400, 39.498)		22.754 (10.084, 45.226)
τ_u			4.703 (2.134, 9.079)	8622.668 (142.832, 39713.666)
DIC	393.11	307.39	306.36	307.67
pD	4.031	27.14	27.55	26.93
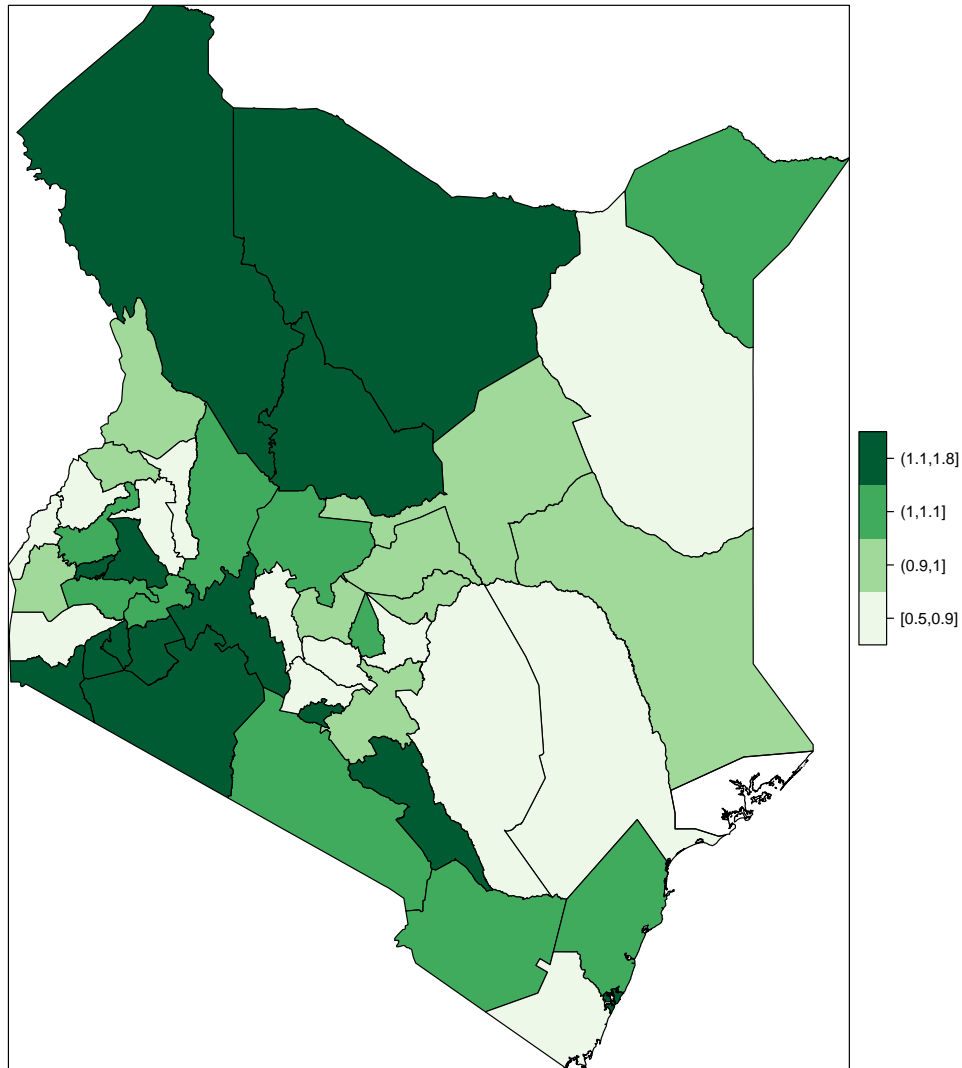
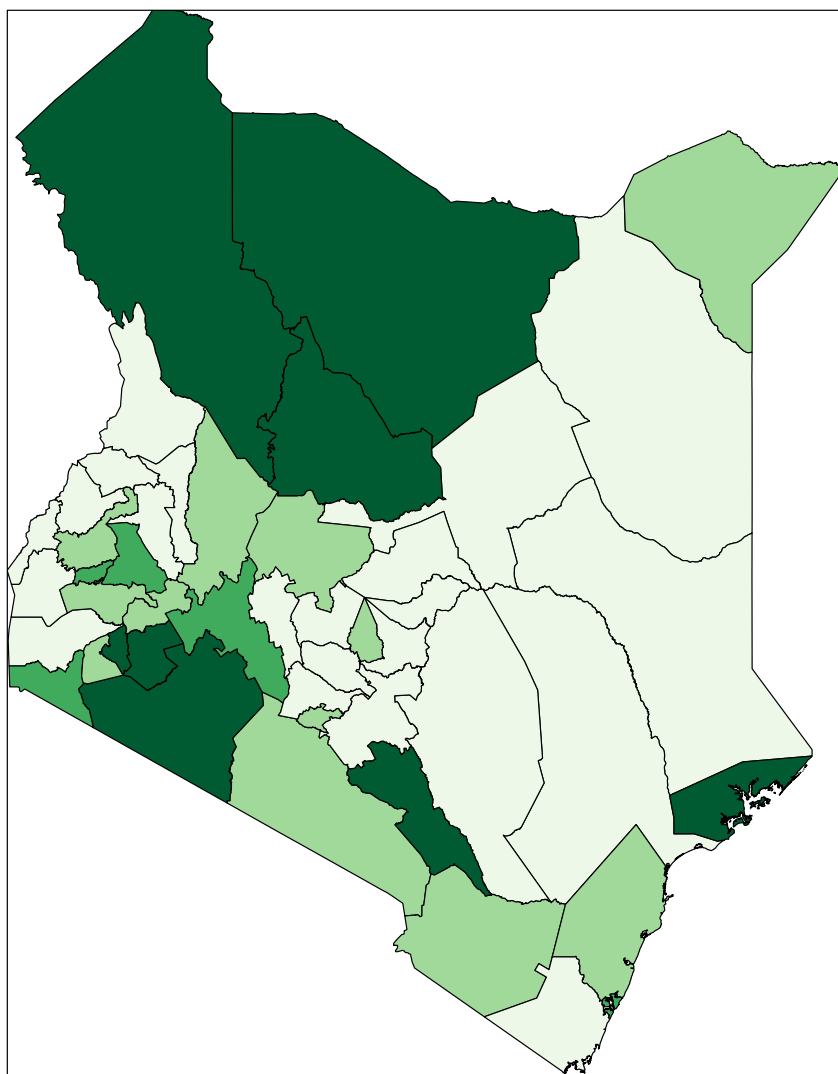Figure 5.1: Infants HIV Relative risk in Kenya in 2017

Figure 5.2: Posterior relative risk exceedance probability map p (ξ > 1|y)

Figure 5.1 shows the relative risk map of the counties based on the best fitting model. The relative risk ranges from 0.71 to 1.71. Counties that were predicted to have high risk of Infant HIV infection were Makueni, Lamu, Turkana, Marsabit, Samburu, Mombasa, Nairobi, Narok, Taita Taveta, Kilifi. Bungoma, Embu, Muranga, Garissa, Tana river, El-geiyo Marakwet, Homabay and Nyandarua were associated with lower risk of infant HIV infection (RR < 0.8). The exceedence probability map in figure 5.2 displays counties with relative risk above the national risk (RR > 1). This map confirms counties associated with high risk of HIV infection (darker color). The distribution of spatial random effects (Fig 5.3) revealed strong spatial patterns at multiple scales. According to these patterns,

the risk of HIV infection is associated with living in the following counties; Makueni, Turkana, Marsabit, Lamu. This map shows a similar pattern to the relative risk map signifying counties highly affected by spatially structured random effects.

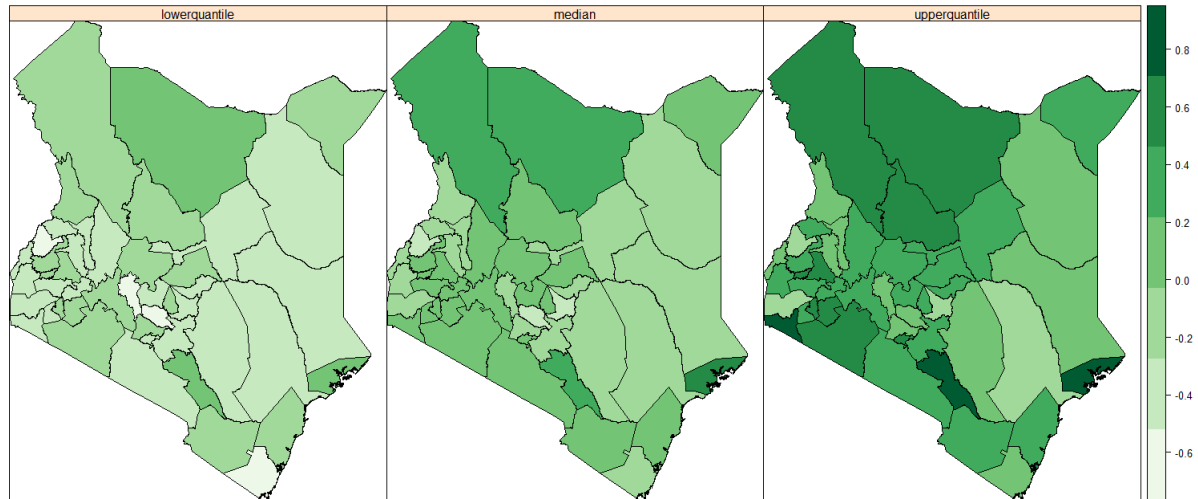Clustering of risk and elevated risk can be observed in the North-west, south west counties of Kenya.

Figure 5.3: Estimates of the random effects from the spatially structured model

5.3 Geostatistical analyses

The variogram in figure 5.4 obtained from the Poisson GLM indicates presence of spatial correlation in the data. This means that we should fit spatial models to the data. Table 5.3 displays the posterior estimates of the Poisson GLM and the SPDE models.

Table 5.3: Posterior statistics of the two models

Variable	Poisson GLM	Poisson GLM + SPDE
β_0	-0.059 (-0.27727402, 0.15261972)	-0.609 (-0.904, -0.322))
HAART	-0.129 (-0.313, 0.058)	-0.125 (-0.348, 0.102)
BF	0.135178 (-0.062, 0.338)	0.178 (-0.051, 0.412)
DIC	8463.834	5591.732
σ		0.019 (0.000, 0.003)
κ		0.495 (0.008, 1.388)
Range		14.621 (0.623, 48.579)

The best fitting model was the SPDE model (DIC 5591.732). Consistent with the results of the areal analysis, HAART was found to be negatively associated with infant positivity whereas breastfeeding was positively associated with infant positivity. The associations were however not significant.

Figure 5.4: Sample variogram of the pearsons residuals of the Poisson GLM

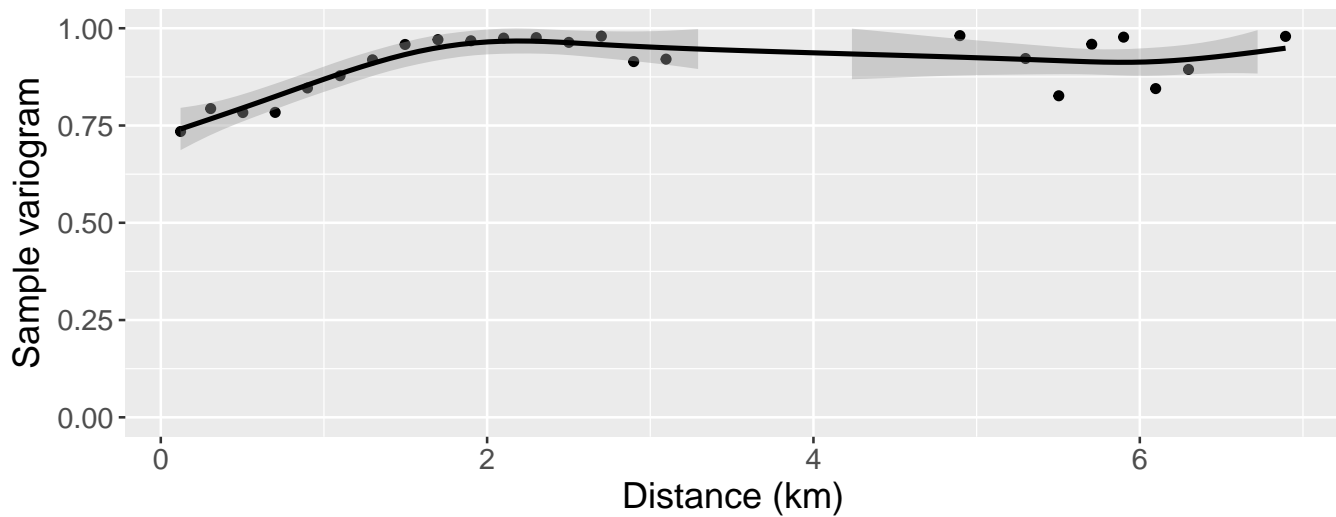

Figure 5.5 shows a map of the sampled locations and the selected mesh is shown in figure 5.6. The triangulation produced a mesh with 1249 vertices. To avoid the boundary effects, we used a mesh that extends the study region. The choice of mesh is a trade-off between the GMRF representation accuracy and the computational costs (Blangiardo & Cameletti, 2015).
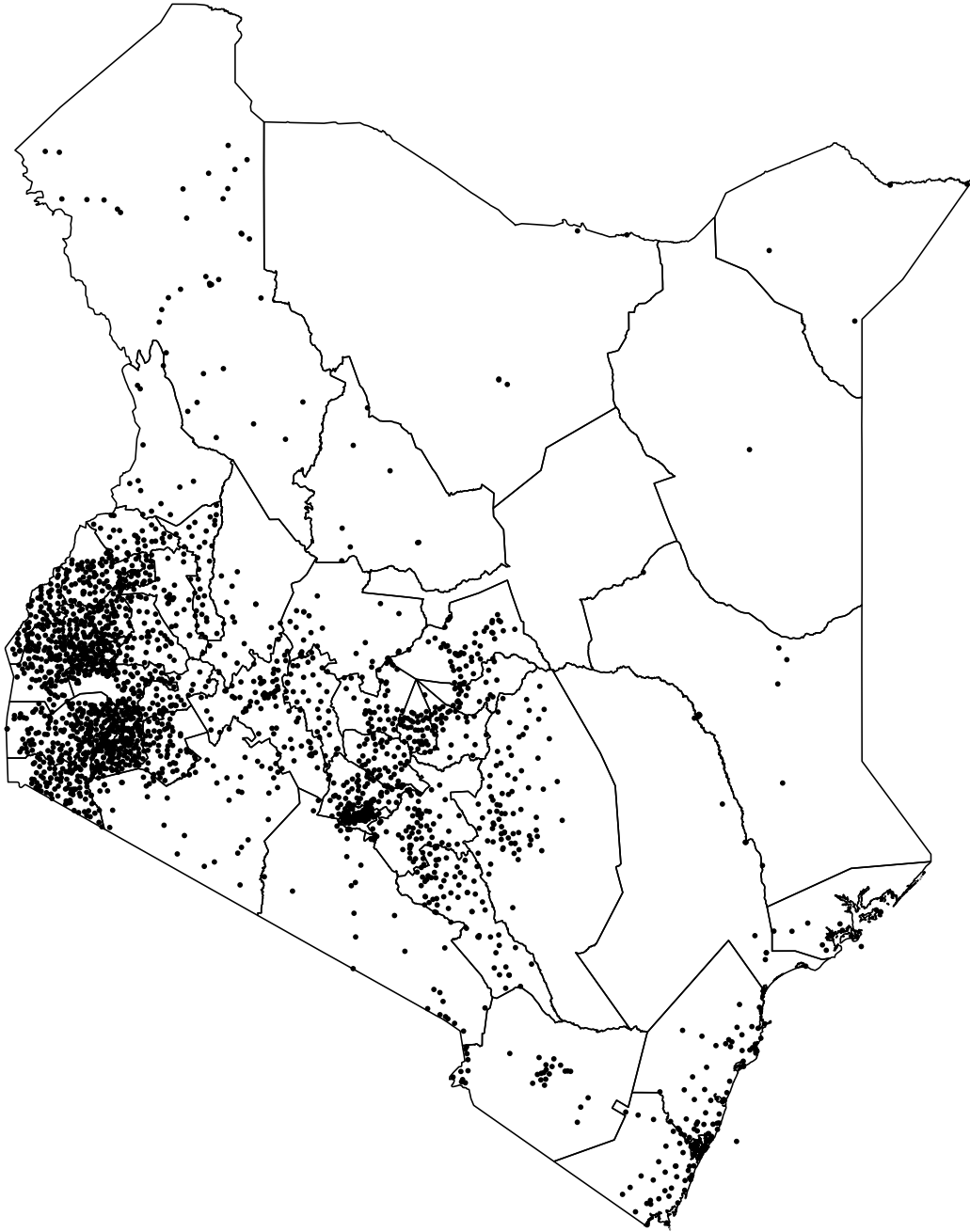
Figure 5.5: Location of facilities

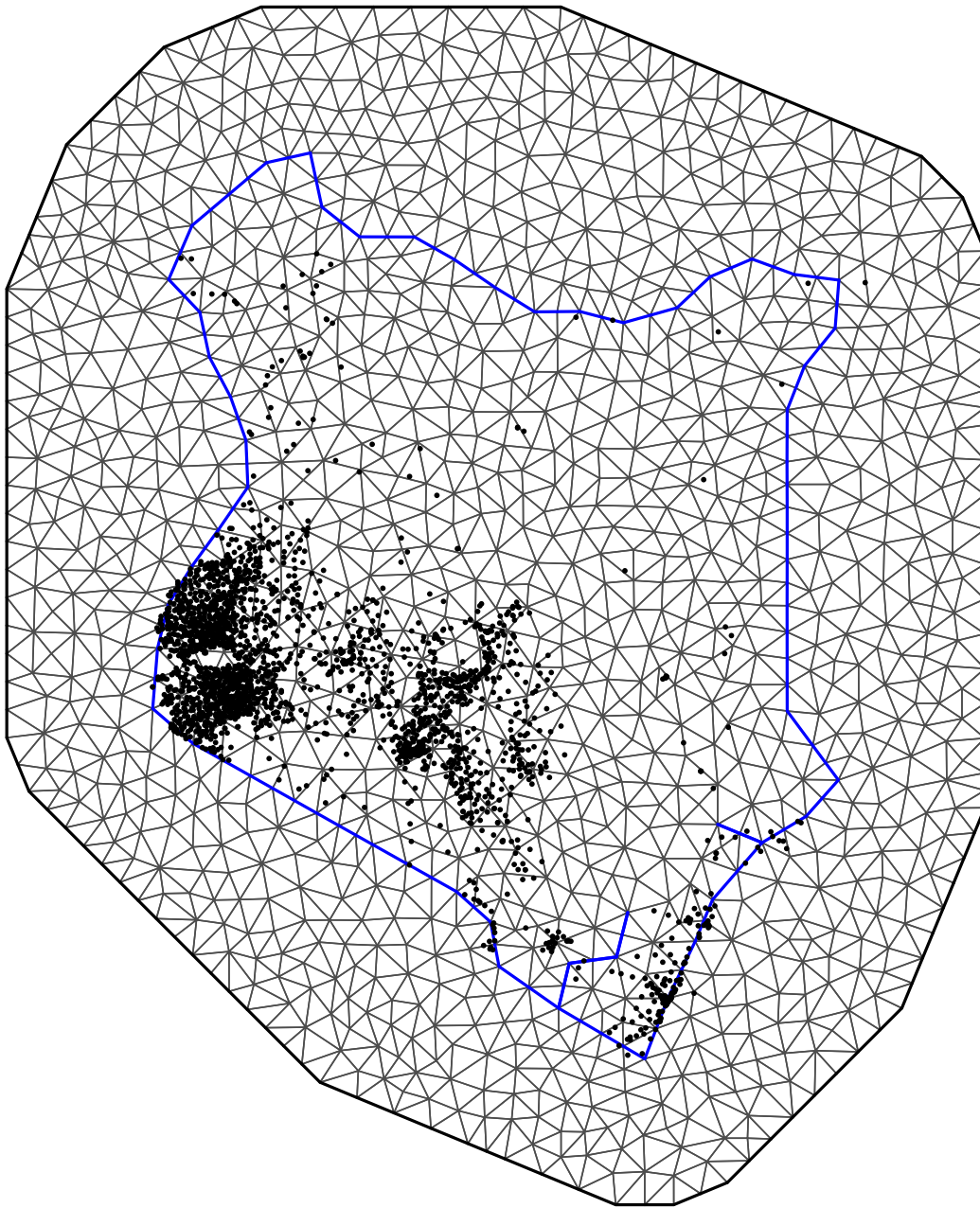**Constrained refined Delaunay triangulation**

Figure 5.6: Mesh with sampled locations
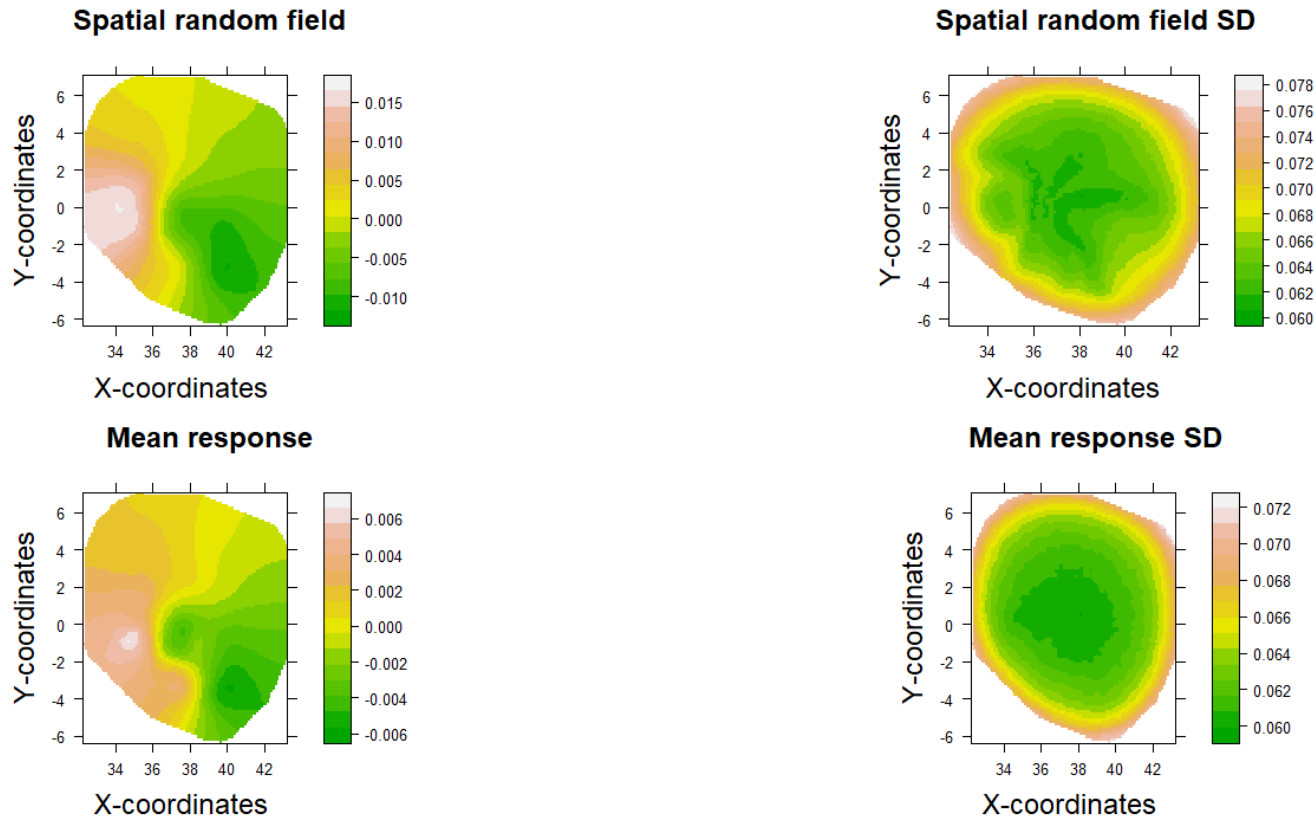
Figure 5.7: Spatial random field (top) and the predicted mean response (bottom)

The map of the spatial field (Figure 5.7 top left) reveals that the spatial random effects causes an increase or decrease in the expected disease counts in specific regions.

The spatial pattern of the posterior mean of the latent field and of the mean response are similar as seen in figure 4.5 whereas the latent field has a higher variability than the response. The spatial effect ranges from 0.010 to 0.015 whereas standard deviation ranges from 0.060 to 0.078.

Matérn correlation function is displayed in figure 5.8 which shows strong correlation is upto about 0.495 km. The distance at which correlation is close to 0.1 is 14.621 km.

Figure 5.8: Matérn correlation function

Chapter 6

Discussion and conclusions

6.1 Discussion

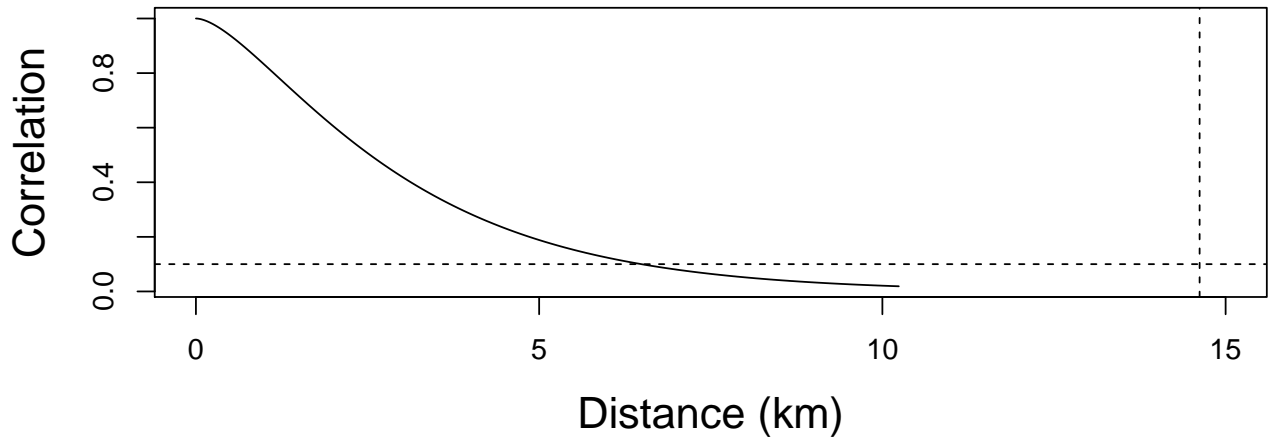In this work, deterministic Bayesian approaches (INLA and INLA-SPDE) were used for analysis of areal and point reference data to assess the spatial distribution of HIV amongst infants and associated risk factors for mother to child transmission. We employed INLA to fit Bayesian hierarchical spatial models within the R library INLA. CAR model was identified to be suitable for modeling and mapping relative risk of HIV amongst infants in Kenya.

The findings revealed that breastfeeding increases mother to child transmission of HIV. Previous studies (Nduati et al., 2000); (Colebunders et al., 1988); (Miotti et al., 1999), noted a substantially higher risk of infection among breast fed infants within the first months of breastfeeding compared to later months. A high early transmission rate might be explained by the milk which is rich in HIV infected cells and the immaturity of the infant's immune system (Southern, 1998). In resource limited settings complementary feeding increases the risk of morbidity and mortality from infectious disease. Until recently, WHO recommended that HIV positive mothers breastfeed exclusively for the first six months and continue breastfeeding with appropriate complementary foods for at least 12 months while taking their antiretroviral to reduce risk of post-natal transmission. (Slater

et al., 2010) showed that perinatal transmission can be greatly reduced in breastfeeding populations on antiretroviral therapy.

The results of the study indicated that the use of HAART by the mother reduces the transmission rate of HIV. The use of HAART clearly stands out as a key determinant of MTCT risk as has been consistently reported in many studies (Townsend et al., 2008),(Ledergerber et al., 1999), (Becquet et al., 2009). The risk of HIV infection among infants was found to be high in Makueni, Lamu, Turkana, Marsabit, Samburu, Mombasa, Nairobi, Narok, Taita Taveta, Kilifi counties suggesting the need of geographical prioritization of infant HIV prevention interventions to optimise reduction of new HIV infections. The random effects map revealed residual variation suggesting unaccounted variation after including the covariates in the model. Additionally, the spatial patterns in the random effects suggested possible covariates omitted from the model. INLA which is a promising alternative to the MCMC is a recent methodology for Bayesian inference used in hierarchical models. The numerical inference approach has gained popularity due to the fast and accurate estimates to the posterior distributions produced by the methodology. This alleviates one of the most important bottlenecks associated with MCMC which is computationally intensive especially for models that are complex in nature. INLA methodology can be used in the free R software with the package R-INLA. The methodology provides several quantities for Bayesian model choice and selection such as effective number of parameters(pD) and Deviance information Criterion (DIC).

We employed INLA-SPDE approach to perform Bayesian inference on spatial hierarchical Gaussian fields. The SPDE model was found to be suitable for modelling and predicting point reference data as compared to the Poisson GLM. The posterior covariates effects were similar to those of the areal analysis. It is apparent from the maps of the random effects and the spatial random field that inclusion of spatial random effects yields strong latent spatial patterns that were not explained by the explanatory variables. As shown in Figure 5.7, the predicted risk of infant HIV infection is high in the western part of Kenya and lowest in the eastern regions. The possible reason for this could be

that the data exhibits clustering in the western regions with little or no observations in the other parts. The standard errors are high on the peripherals which should be expected since they are extensions of the triangulation. One advantage of INLA-SPDE is its capability of estimation and prediction of the posterior marginal distributions of the model parameters and the model responses without carrying out extensive simulations. A major limitation with INLA-SPDE is that it becomes computationally intensive when dealing with non-Gaussian likelihood (Lindgren et al., 2011).

Limitations of this study include the use of routinely collected data which may not be of quality due to missing information. Additionally, the data has limited information on potential explanatory variables that could be important in explaining the positivity rates. Lack of data on important covariates such as virological status and maternal immunologic which could influence transmission rates.

6.2 Conclusion and Recommendations for further research

The aim of this study was to investigate the geographical variation of HIV and identify risk factors among infants in Kenya in 2017. INLA which is a computationally effective alternative to MCMC was used for Bayesian inference. This study has shown the determinants of infant HIV infection and the spatial distribution through mapping of HIV risk. It is apparent that there are still geographical disparities in infant HIV infection and targeted PMTCT interventions and resources should be directed to counties that have high infection rates. It is hoped that the presented outcome variations will stimulate further research efforts to investigate the reasons underlying the disparities and inform policy. Many aspects of this study could be extended. Further research is required to determine suitable spatio-temporal models for modeling and mapping relative risk of HIV among infants in Kenya. This will allow to explain evolution of the relative risk of HIV in space and time. Incorporating other covariates for mother to child transmission of HIV that were not captured in the analysis could yield important insights on the spatial distribution

of the disease.

References

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Krainski, E., . . . Lindgren, F. (2018). Spatial modelling with r-inla: A review. *arXiv preprint arXiv:1802.06350*.

Becquet, R., Ekouevi, D. K., Arrive, E., Stringer, J. S., Meda, N., Chaix, M.-L., . . . others (2009). Universal antiretroviral therapy for pregnant and breast-feeding hiv-1-infected women: towards the elimination of mother-to-child transmission of hiv-1 in resource-limited settings. *Clinical infectious diseases*, *49*(12), 1936–1945.

Beguin, J., Martino, S., Rue, H., & Cumming, S. G. (2012). Hierarchical analysis of spatially autocorrelated ecological data using integrated nested laplace approximation. *Methods in Ecology and Evolution*, *3*(5), 921–929.

Bernstein, L., Brooke, K., Collado, D., Crain, M., Birmingham, A., DiPaolo, C., . . . others (2001). *Guidelines for the use of antiretroviral agents in pediatric hiv infection.*

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 302–309.

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

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r-inla*. John Wiley & Sons.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Bryson, Y. J. (1996). Perinatal hiv-1 transmission: recent advances and therapeutic interventions. *AIDS (London, England)*, *10*, S33–42.

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, *97*(2), 109–131.

Carlin, B. P., Gelfand, A. E., & Banerjee, S. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Carroll, R., Lawson, A., Faes, C., Kirby, R., Aregay, M., & Watjou, K. (2015). Comparing inla and openbugs for hierarchical poisson modeling in disease mapping. *Spatial and spatio-temporal epidemiology*, *14*, 45–54.

Colebunders, R., Kapita, B., Nekwei, W., Bahwe, Y., Lebughe, I., Oxtoby, M., & Ryder, R. (1988). Breastfeeding and transmission of hiv. *Lancet*, *2*(8626-8627), 1487.

Cressie, N. A. (1993). *Statistics for spatial data: Wiley series in probability and mathematical statistics*. John Wiley & Sons.

DeHovitz, J. A., Kovacs, A., Feldman, J. G., Anastos, K., Young, M., Cohen, M., . . . Greenblatt, R. M. (2000). The relationship between virus load response to highly active antiretroviral therapy and change in cd4 cell counts: A report from the women's interagency hiv study. *The Journal of infectious diseases*, *182*(5), 1527–1530.

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

Fawzi, W., Msamanga, G., Renjifo, B., Spiegelman, D., Urassa, E., Hashemi, L., ... Hunter, D. (2001). Predictors of intrauterine and intrapartum transmission of hiv-1 among tanzanian women. *Aids*, *15*(9), 1157–1165.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. Chapman and Hall/CRC.

Griggs, D., Stafford-Smith, M., Gaffney, O., Rockström, J., Öhman, M. C., Shyamsundar, P., ... Noble, I. (2013). Policy: Sustainable development goals for people and planet. *Nature*, *495*(7441), 305.

Ioannidis, J. P., Abrams, E. J., Ammann, A., Bulterys, M., Goedert, J. J., Gray, L., ... others (2001). Perinatal transmission of human immunodeficiency virus type 1 by pregnant women with rna virus loads< 1000 copies/ml. *The Journal of infectious diseases*, *183*(4), 539–545.

Konstantinoudis, G., Schuhmacher, D., Rue, H., & Spycher, B. (2018). Discrete versus continuous domain models for disease mapping. *arXiv preprint arXiv:1808.04765*.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, *14*(8), 799–810.

Landesman, S. H., Kalish, L. A., Burns, D. N., Minkoff, H., Fox, H. E., Zorrilla, C., ... Tuomala, R. (1996). Obstetrical factors and the transmission of human immunodeficiency virus type 1 from mother to child. *New England Journal of Medicine*, *334*(25), 1617–1623.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC.

Ledergerber, B., Egger, M., Opravil, M., Telenti, A., Hirschel, B., Battegay, M., . . . others (1999). Clinical progression and virological failure on highly active antiretroviral therapy in hiv-1 patients: a prospective cohort study. *The Lancet*, *353*(9156), 863–868.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Louis, M. E. S., Kamenga, M., Brown, C., Nelson, A. M., Manzila, T., Batter, V., . . . others (1993). Risk for perinatal hiv-1 transmission according to maternal immunologic, virologic, and placental factors. *Jama*, *269*(22), 2853–2859.

Lucas, G. M., Chaisson, R. E., & Moore, R. D. (1999). Highly active antiretroviral therapy in a large urban clinic: risk factors for virologic failure and adverse drug reactions. *Annals of internal medicine*, *131*(2), 81–87.

Martino, S., & Rue, H. (2009). Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Miotti, P. G., Taha, T. E., Kumwenda, N. I., Broadhead, R., Mtimavalye, L. A., Van der Hoeven, L., . . . Biggar, R. J. (1999). Hiv transmission through breastfeeding: a study in malawi. *Jama*, *282*(8), 744–749.

Mock, P. A., Shaffer, N., Bhadrakom, C., Siriwasin, W., Chotpitayasunondh, T., Chearskul, S., . . . others (1999). Maternal viral load and timing of mother-to-child hiv transmission, bangkok, thailand. *Aids*, *13*(3), 407–414.

Moraga, P., Cramb, S. M., Mengersen, K. L., & Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, *21*, 27–41.

Musenge, E., Chirwa, T. F., Kahn, K., & Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal hiv/tb child mortality data through the inla and spde approaches: Applied to data observed between 1992 and 2010 in rural north east south africa. *International journal of applied earth observation and geoinformation*, *22*, 86–98.

NASCOP. (2015). *Kenya framework for elimination of mother to child transmission of hiv 2016-2021*. n.p.

Nduati, R., John, G., Mbori-Ngacha, D., Richardson, B., Overbaugh, J., Mwatha, A., . . . others (2000). Effect of breastfeeding and formula feeding on transmission of hiv-1: a randomized clinical trial. *Jama*, *283*(9), 1167–1174.

Ngwende, S., Gombe, N. T., Midzi, S., Tshimanga, M., Shambira, G., & Chadambuka, A. (2013). Factors associated with hiv infection among children born to mothers on the prevention of mother to child transmission programme at chitungwiza hospital, zimbabwe, 2008. *BMC public health*, *13*(1), 1181.

Ntzoufras, I. (2011). *Bayesian modeling using winbugs* (Vol. 698). John Wiley & Sons.

Okuto, E. (n.d.). Spatial areal data analysis with application to various tuberculosis outcomes in kenya.

PEPFAR, U., UNICEF, et al. (2016). Who. start free. stay free. aids free. *A Super-Fast-Track Framework for Ending AIDS Among Children, Adolescents and Young Women by*, *2020*.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 325–338.

Rue, H., & Held, L. (2005). *Gaussian markov random fields: theory and applications*. CRC press.

Rue, H., & Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, *137*(10), 3177–3192.

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

Semba, R. D., Kumwenda, N., Hoover, D. R., Taha, T. E., Quinn, T. C., Mtimavalye, L., . . . others (1999). Human immunodeficiency virus load in breast milk, mastitis, and mother-to-child transmission of human immunodeficiency virus type 1. *The Journal of infectious diseases*, *180*(1), 93–98.

Simpson, D., Lindgren, F., & Rue, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, *23*(1), 65–74.

Slater, M., Stringer, E. M., & Stringer, J. S. (2010). Breastfeeding in hiv-positive women. *Pediatric Drugs*, *12*(1), 1–9.

Snow, J. (1854). The cholera near golden-square, and at deptford. *Medical Times and Gazette*, *9*, 321–322.

Southern, S. (1998). Milk-borne transmission of hiv. characterization of productively infected cells in breast milk and interactions between milk and saliva. *Journal of human virology*, *1*(5), 328–337.

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.

Team, R. C., et al. (2013). R: A language and environment for statistical computing.

Thea, D. M., Steketee, R. W., Pliner, V., Bornschlegel, K., Brown, T., Orloff, S., . . . others (1997). The effect of maternal viral load on the risk of perinatal transmission of hiv-1. *Aids*, *11*(4), 437–444.

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

Townsend, C. L., Cortina-Borja, M., Peckham, C. S., de Ruiter, A., Lyall, H., & Tookey, P. A. (2008). Low rates of mother-to-child transmission of hiv following effective pregnancy interventions in the united kingdom and ireland, 2000–2006. *Aids*, *22*(8), 973–981.

UNAIDS, U. (2016). *Prevention gap report*. UNAIDS Geneva.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, *40*(2), 974–994.

WHO. (2016a). Guideline: Updates on hiv and infant feeding: the duration of breast-feeding,and support from health services to improve feeding practices among mothers living with hiv.

WHO. (2016b). *Mother-to-child transmission of hiv.* `www.who.int/hiv/topics/mtct/about/en/`.

World Health Organization(WHO). (2015). Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for hiv.

APPENDIX

AREAL DATA ANALYSIS

```
library(ape)

library(mapview)

library(spdep)

library(INLA)

library(maptools)

library(rgdal)

library(maptools)

library(RColorBrewer)

library(classInt)

library(classInt)

library(scanstatistics)

setwd("E:/MSC Project/EID_DATA")

data<-read.csv("Data.csv")

kenyacount<-readOGR("E:/MSC Project/EID_Data/County.shp")

plot(kenyacount,axes=T)

nbkenya<-poly2nb(kenyacount)

nbkenya

nb2INLA(file="kenya.graph",nbkenya)

nb2INLA("LDN.graph", temp)
```

```
LDN.adj <- paste(getwd(),"/kenya.graph",sep="")

H <- inla.read.graph(filename="kenya.graph")

image(inla.graph2matrix(H),xlab="",ylab="")

data<-cbind(data,region=as.numeric(data$id),

            region.struct=as.numeric(data$id)))

View(data)

str(data)

#MODEL1-A generalized linear model (Full model)

fit1<- inla(Cases~HAART+BF,family="poisson",

            data=data, E=Expectedcases,

            control.predictor=list(compute=TRUE),

            control.compute=list(dic=TRUE,cpo=TRUE))

summary(fit1) #418.88 #Pd 3.032


##################################################

#MODEL 2-Model with spatially unstructured random effects

fit2a<- Cases ~ HAART+BF+f(region)

fit2<- inla(fit2a,family="poisson",

            data=data,E=Expectedcases,

            control.predictor=list(compute=TRUE),

            control.compute=list(dic=TRUE,cpo=TRUE))


summary(fit2) #309.99  29.37


##################################################

#MODEL 3-Model with spatially structured random effects

names(data)

View(data)
```

```
fit3a<- Cases ~ HAART+BF+f(region.struct,

                           adjust.for.con.comp=TRUE,

                           model="besag",constr=T,graph.file= H)

                              #hyper = list(theta=list(prior="loggamma"


fit3<- inla(fit3a,family="poisson",

            data=data,E=Expectedcases,

            control.predictor=list(compute=TRUE),

            control.compute=list(dic=TRUE,cpo=TRUE))
summary(fit3) # 308.56  28.17


#MAPPING RELATIVE RISK
csi <- fit3$marginals.random$region.struct[1:47]
zeta <- lapply(csi,function(x) inla.emarginal(exp,x))
zeta
ecxd2<-zeta<0.8
ecxd2
 #Define the cutoff for zeta
 zeta.cutoff <- c(0.5, 0.9, 1.0, 1.1, 1.8)
#Transform zeta in categorical variable
cat.zeta <- cut(unlist(zeta),breaks=zeta.cutoff,
                    include.lowest=TRUE)
cat.zeta
#Create a dataframe with all the information needed for the map
maps.cat.zeta <- data.frame(id=data$region.struct, cat.zeta=cat.zeta)
maps.cat.zeta


data.boroughs <- attr(kenyacount, "data")
```

```
data.boroughs


#sp@data = data.frame(sp@data, df[match(sp@data[,by], df[,by]),])

attr(kenyacount, "data") <- merge(data.boroughs, maps.cat.zeta,

                                      by="id", sort=FALSE)


attr(kenyacount, "data")

#names(kenyacount)

#View(kenyacount)

#Map zeta

#kenyacount

col.r <- rev(rainbow(20, alpha = 0.5))

my.palette <- brewer.pal(n = 7, name = "Greens")

spplot(obj=kenyacount, zcol= "cat.zeta",

            col.regions=my.palette,cuts = 6)\\ #main="Posterior mean rel


#Exceedance probability >1

a<-0

exc <- lapply(csi, function(x) {1 - inla.pmarginal(a, x)})

ecxd<-exc>1.

ecxd

cutoff <- c(0,0.5,0.8,0.9,1)

#Transform zeta in categorical variable

exc.zeta <- cut(unlist(exc),breaks=cutoff,

                include.lowest=TRUE)

#Create a dataframe with all the information needed for the map

maps.exc.zeta <- data.frame(id=data$region.struct, exc.zeta=exc.zeta)

maps.exc.zeta
```

51

```
#Add the categorized zeta to the spatial polygon
#View(kenyacount)
data.boroughs <- attr(kenyacount, "data")
attr(kenyacount, "data") <- merge(data.boroughs, maps.exc.zeta,
                                  by="id", sort=FALSE)
spplot(obj=kenyacount, zcol= "exc.zeta",
       col.regions=my.palette, cuts = 6)#main="Probability exceedance",
kenyacount$median=as.numeric(fit3$summary.fitted.values[,4])
kenyacount$median
kenyacount$q025=as.numeric(fit3$summary.fitted.values[,3])
kenyacount$q975=as.numeric(fit3$summary.fitted.values[,5])
kenyacount$mean=as.numeric(fit3$summary.fitted.values[,1])
at=c(0, 0.5,1.0, 1.5,2.0, 2.5,5.0,11)
#my.palette <- brewer.pal(n = 8, name = "Blues")
spplot(kenyacount,c("q025","median","q975"),
       names.attr = c( "2.5% quantile","Median","97.5% quantile"),
       as.table = TRUE,col.regions=my.palette, cuts=6)


kenyacount$mean=fit3$summary.random$region[,2]
kenyacount$median=fit3$summary.random$region[,5]
kenyacount$lowerquantile=fit3$summary.random$region[,4]
kenyacount$upperquantile=fit3$summary.random$region[,6]


plot(density(kenyacount$median))


#my.palette <- brewer.pal(n = 5, name = "YlOrRd")
spplot(kenyacount,c("lowerquantile","median",
"upperquantile"),
```

```
            names.attr = c("lowerquantile","median","upperquantile"),

            as.table = TRUE,col.regions = my.palette,

             cuts = 6)
#spplot(kenyacount,c("lowerquantile","upperquantile",

"mean","median"))
#spplot(kenyacount,c("lowerquantile","upperquantile",

"mean","median"))


###################################################


#MODEL 4- Convolution model
fit4a<- Cases ~ HAART+BF+f(region)+f(region.struct,model="bym",graph.fi


fit4<- inla(fit4a,family="poisson",

            data=data,E=Expectedcases,

            control.predictor=list(compute=TRUE),

            control.compute=list(dic=TRUE,cpo=TRUE))


summary(fit4) # 310.65 pD  28.96


names(fit4)
# Disease risk estimates/fitted values
kenyacount$median=as.numeric(fit4$summary.fitted.values[,4])

kenyacount$median

kenyacount$q025=as.numeric(fit4$summary.fitted.values[,3])

kenyacount$q975=as.numeric(fit4$summary.fitted.values[,5])
```

```r
my.palette <- brewer.pal(n = 7, name = "YlOrRd")
spplot(kenyacount,c("median","q025","q975"),
       names.attr = c("Median", "2.5% quantile","97.5% quantile"),
       as.table = TRUE,col.regions = my.palette, cuts = 6)
######################################################
######################################################
SPDE

 library(spdep)
library(INLA)
library(maptools)
library(rgeos)
library(maptools)
library(RColorBrewer)
library(classInt)
library(lattice)
library(classInt)
library(scanstatistics)
library(ggplot2)
library(gstat)
library(INLA)
library(sp)
library(inlabru)
library(fields)
library(rgdal)

setwd("E:/MSC Project/EID_DATA/SPDE_APPROACH_ANALYSIS")
kenyacount<-readOGR("E:/MSC Project/EID_Data/SPDE_APPROACH_ANALYSIS/Cou
```

```r
data<-read.csv("spde_data.csv")

str(data)

data$PropBF<-as.numeric(levels(data$PropBF))[data$PropBF] # convert pro

plot(kenyacount,asp=1)# main = 'Map of Kenya and Sampled locations

(dots)')

data$Xkm = data$Longitude

data$Ykm = data$Latitude

points(x= data$Xkm,data$Ykm, col=1,pch=16, cex=0.5 ) # Map of sampled l

#View(data)

I1 <- inla(Positive ~ PropBF + PropHAART,

            family = "poisson",

            control.predictor = list(compute = TRUE),

            data = data)

             #Spatial dependency?

 #Let's make a variogram of the Pearson residuals.

# Sample-variogram with distances up to 100 km

MyData <- data.frame(E1 = E1,  Xkm = data$Xkm,  Ykm = data$Ykm)

coordinates(MyData)    <- c("Xkm", "Ykm")

Vario <- variogram(object = E1 ~ Xkm + Ykm,

                 data = MyData,

                 cressie = TRUE,

                  cutoff = 7,

                  width = 0.2)


p <- ggplot(data = Vario, aes(x = dist, y = gamma))

p <- p + geom_point()

p <- p + geom_smooth(method = "gam",

                  formula = y ~ s(x, bs = "cs"),
```

```
                        colour = "black")

p <- p + ylim(0,1)

p <- p + theme(text = element_text(size = 15))

p <- p + xlab("Distance (km)") + ylab("Sample variogram")

p




#INLA ANALYSIS
# Creating a mesh-Triangulation with a SpatialPolygonsDataFrame
border<-unionSpatialPolygons(kenyacount, rep(1, nrow(kenyacount)))
bordery<-inla.sp2segment(border)


#mesh<- inla.mesh.2d(boundary = bordery,cutoff = 0.02,
                    #max.edge = c(1,1))
#mesh<- inla.mesh.2d(boundary = bordery, cutoff = 0.5,
                     #max.edge = c(0.3,0.3))
#mesh<- inla.mesh.2d(boundary = bordery,cutoff = 0.1,
                    #max.edge = c(0.1, 0.1))
mesh<- inla.mesh.2d(boundary = bordery,cutoff = 0.5,
                    max.edge = c(0.05, 0.5))


plot(mesh, asp=1)


# max.edge: maximum allowed triangle edge lengths in
#           the inner domain and in the outer extension
# cutoff: minimum allowed distance between points. Points
# at a closer distance than the supplied value are
# replaced by a single vertex
```

```
mesh$n #1249

points(x= data$Longitude,data$Latitude, col=1,pch=16, cex=0.5 ) # sampl

Loc <- cbind(data$Xkm, data$Ykm)

#what are the distances between the points?
D <- dist(Loc)
par(mfrow = c(1,2), mar = c(5,5,2,2), cex.lab = 1.5)
hist(D,
     freq = TRUE,
     main = "",
     xlab = "Distance between sites (km)",
     ylab = "Frequency")
text(2, 35000, "A", cex = 1.5)


plot(x = sort(D),
     y = (1:length(D))/length(D),
     type = "l",
     xlab = "Distance between sites (km)",
     ylab = "Cumulative proportion")
text(2, 1, "B", cex = 1.5)


# sampling locations match the points on the mesh
A5<-inla.spde.make.A(mesh,loc = Loc)
dim(A5) #2457 1249
#2457 observations on a 4089 grid
#       A is a weight matrix
head(A5)
```

```
??inla.spde2.pcmatern

#View(data)

#SPDE MODEL

spde<-inla.spde2.pcmatern(mesh, alpha=2,prior.range=c(8,0.5),

                    prior.sigma=c(log(1.1),0.01))

w.index<-inla.spde.make.index(name = 'w',

                        n.spde = spde$n.spde,

                        n.group = 1,

                        n.rep1 = 1)



N<-nrow(data)

X<-data.frame(Intercept = rep(1,N),

            Positive= data$Positive,

            PropBF= data$ PropBF,

            PropHAART = data$ PropHAART,

            iidx=1:nrow(data))

N # sampled locations

X<-as.data.frame(X)

str(X)

# Tell INLA at which mesh points the covariates are sampled.

stk<-inla.stack(tag = 'est',

            data= list(y = data$Positive),

            A= list(A5, 1),

            effects =list (w= w.index, #spatial field

                        X=X)) #covariates



dim(inla.stack.A(stk)) #2457 2199#Define the Matern correlation
on the mesh
```

```
#defining the stack- dataframe that contains intercept
 and all covariates
inla.stack.data(stk)


# Specfying models to fit
#Model without spatial componenent
model1<-y~ -1 +Intercept + PropBF + PropHAART
#MODEL with spatial component
model2<- y ~ -1 + Intercept +PropBF + PropHAART + f(w,model= spde)+
 f(iidx, model="iid")


# Run R-INLA
fit1<-inla(model1,
           family = "poisson",
           data = inla.stack.data(stk),
           control.compute = list(dic= TRUE),
           control.predictor = list( A= inla.stack.A(stk)))


fit2<-inla(model2,
           family = "poisson",
           data = inla.stack.data(stk),
           control.compute = list(dic= TRUE),
           control.predictor = list( A= inla.stack.A(stk)))


fit2$summary.fixed #Summary of beta0
fit1$summary.fixed
```

```
fit2$marginals.hyperpar




# Visualise the differences of parameter estimates in the two models:
Combined <- rbind(fit1$summary.fixed[, c("mean", "0.025quant",
"0.975quant")],
fit2$summary.fixed[, c("mean", "0.025quant", "0.975quant")]
)
Combined$WhichModel <- rep(c("GLM", "spatial GLM"), each = 3)
Combined$WhichVariable <- rep(rownames(fit2$summary.fixed), 2)
colnames(Combined) <- c("Mean", "Lo", "Up", "WhichModel",
"WhichVariable")
Combined
p <- ggplot()
p <- p + geom_point(data = Combined,
                    aes(x = WhichModel,
                        y = Mean)
)

p <- p + geom_errorbar(data = Combined,
                       aes(x = WhichModel,
                           ymax = Up,
                           ymin = Lo),
                       width=0.2)



p <- p + xlab("Parameters") + ylab("Values")
```

```r
p <- p + theme(text = element_text(size=15))

p <- p + facet_wrap( ~ WhichVariable, scales = "free_y")

p <- p + theme(legend.position="none")

p <- p + geom_hline(yintercept = 0, linetype = 2)

p



# Comparing the two models using DIC

fit1$dic$dic #  8463.834

fit2$dic$dic # 5591.732 # Model with spatial correlation performs well


fit2$summary.fixed

#posterior mean values of the spatial random field

w.pm<-fit2$summary.random$w$mean

w.pm

w.sd<-fit2$summary.random$w$sd # posterior standard deviation of the sp

length(w.pm)

# This is the spatial field calculated at all the mesh points


# Now we calculate the spatial field on a grid.

# Prediction of the mean values of random field  on the grid

wproj <- inla.mesh.projector(mesh)

# This is a 100 by 100 field...and we can plot it,

# using standard 3-D software tools.


# Plot the spatial random field

# The function inla.mesh.project can then

# be used to project the 1829 posterior mean
```

```
# values on this grid. By default a lattice

# of 100 by 100 is used.

w.pm100_100 <- inla.mesh.project(wproj, w.pm)

w.sd100_100 <- inla.mesh.project(wproj, w.sd)


# This w.pm100_100 is of dimension 100 by 100

# and is a projection (interpolation and extrapolation)

# of the random field w. We can use the levelplot

# function from the lattice package to plot w.pm100_100


Grid <- expand.grid(Xkm = wproj$x,

                    Ykm = wproj$y)

Grid$w.pm <- as.vector(w.pm100_100)

Grid$w.sd <- as.vector(w.sd100_100)


col.l <- colorRampPalette(c('red', 'green'))(30)

col.r <- rev(rainbow(20, alpha = 0.5))


plot.wpm <- levelplot(w.pm ~ Xkm * Ykm,

                    data = Grid,

                    aspect = "iso",

                    col.regions= terrain.colors(20),

                    scales = list(draw = TRUE),

                    xlab = list("X-coordinates ", cex = 1.5),

                    ylab = list("Y-coordinates", cex = 1.5),

                    main = list("Spatial random field", cex = 1.5))

plot.wpm
```

```
# And do the same for the posterior standard deviation

plot.wsd<-levelplot(w.sd ~ Xkm * Ykm,

        col.regions= terrain.colors(20),

        data = Grid,

        scales = list(draw = TRUE),

        xlab = list("X-coordinates", cex = 1.5),

        ylab = list("Y-coordinates", cex = 1.5),

        main = list("Spatial random field SD", cex = 1.5), asp=1)


plot.wsd
##########################################
# And this is the correlation function that we are imposing
# on the residuals:


SpFi.w <- inla.spde2.result(inla = fit2,

                    name = "w",

                    spde = spde,

                    do.transfer = TRUE)


Kappa <- inla.emarginal(function(x) x,

                SpFi.w$marginals.kappa[[1]] )
inla.hpdmarginal(0.95, SpFi.w$marginals.kappa[[1]])
Kappa 0.4953149
#                low     high
#level:level:0.95 0.007974941 1.388208


sigmau <- inla.emarginal(function(x) sqrt(x),

                SpFi.w$marginals.variance.nominal[[1]] )
```

63

```
inla.hpdmarginal(0.95, SpFi.w$marginals.variance.nominal[[1]])
sigmau 0.01910129
#                    low      high
#level:0.95 5.207278e-08 0.002615157


range <- inla.emarginal(function(x) x,
                   SpFi.w$marginals.range.nominal[[1]] )
inla.hpdmarginal(0.95, SpFi.w$marginals.range.nominal[[1]])
range
#                   low       high
#level:0.95 0.6234094 48.57928


c(Kappa, sigmau, range)
#[1]0.49531489   0.01910129 14.62118831


#Show correlation structure
D <- as.matrix(dist(Loc[,1:2]))
d.vec <- seq(0, max(D), length = 1000)
d.vec
Cor.M <- (Kappa * d.vec) * besselK(Kappa * d.vec, 1)
Cor.M[1] <- 1


par(mfrow=c(1,1), mar = c(5,5,2,2), cex.lab = 1.5)
plot(x = d.vec,
     y = Cor.M,
     pch = 16,
     type = "l",
     cex.lab = 1.5,
```

```r
      xlab = "Distance (km)",

      ylab = "Correlation",

      xlim = c(0, 15))

abline(h = 0.1, lty = 2)

abline(v = range, lty = 2)



dim(A5)

str(data)

#### Predicting the response.

stk.pred <- inla.stack(tag='pred', A=list(A5, 1), data=list(y=NA), ## r

effects=list(s=1:spde$n.spde, data.frame(X=X)))



stk.full <- inla.stack(stk, stk.pred)

p.res <- inla(model2, data=inla.stack.data(stk.full), ## full stack

              control.predictor=list(compute=TRUE, ## compute the predi

                                  A=inla.stack.A(stk.full))) ## usin

p<-p.res$summary.random$w$mean

p

s<-p.res$summary.random$w$sd

wproj <- inla.mesh.projector(mesh)



w.pm100_100 <- inla.mesh.project(wproj,p)

w.sd100_100 <- inla.mesh.project(wproj,s)

Grid <- expand.grid(Xkm = wproj$x,

                    Ykm = wproj$y)

Grid$p <- as.vector(w.pm100_100)

Grid$s <- as.vector(w.sd100_100)
```

```r
col.l <- colorRampPalette(c('red', 'green'))(30)

col.r <- rev(rainbow(20, alpha = 0.5))


pm <- levelplot(p ~ Xkm * Ykm,

                data = Grid,

                aspect = "iso",

                col.regions= terrain.colors(20),

                scales = list(draw = TRUE),

                xlab = list("X-coordinates ", cex = 1.5),

                ylab = list("Y-coordinates", cex = 1.5),

                main = list("Mean response", cex = 1.5))

pm


wpm <- levelplot(s ~ Xkm * Ykm,

                data = Grid,

                aspect = "iso",

                col.regions= terrain.colors(20),

                scales = list(draw = TRUE),

                xlab = list("X-coordinates ", cex = 1.5),

                ylab = list("Y-coordinates", cex = 1.5),

                main = list("Mean response SD", cex = 1.5))

wpm



require(gridExtra)

grid.arrange(plot.wpm,

             plot.wsd,

          pm,
```

```
wpm,

nrow=2)
```