GEOSTATISTICAL MODELLING AND ANALYSIS OF UNDER FIVE MALARIA RISK IN MALAWI

MSc (BIOSTATISTICS) THESIS

JAMES J. CHIROMBO

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

August, 2012

GEOSTATISTICAL MODELLING AND ANALYSIS OF UNDER FIVE MALARIA RISK IN MALAWI

MSc (Biostatistics)

By

JAMES J. CHIROMBO

BSc (Statistics and Mathematics)

Thesis submitted to the Mathematical Sciences Department, Faculty of Science,
in partial fulfillment of the requirements for the degree of
Master of Science (Biostatistics)

UNIVERSITY OF MALAWI
CHANCELLOR COLLEGE

August, 2012

Declaration

I the undersigned hereby declare that this thesis/dissertation is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

_____

Full Legal Name

_____

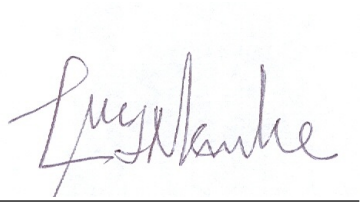
Signature

_____

Date

Certificate of approval

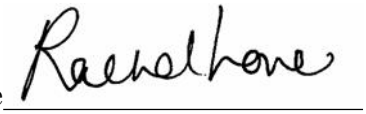
The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.



Signature_____ Date August 27, 2012

Lawrence Kazembe, PhD (Associate Professor)

Main supervisor



Signature_____ Date August 27, 2012

Rachel Lowe, PhD (Postdoctoral Scientist)

Co-supervisor

Signature_____ Date_____

Jupiter Simbeye, MSc (Lecturer)

Course coordinator

Signature_____ Date_____

Levis Eneya, PhD (Senior Lecturer)

Head of Department

Dedication

To my family, for everything. I owe it to you

Acknowledgment

I am very grateful to my supervisors for all their support and ideas. Firstly, to Dr Lawrence Kazembe for without him, I would not have come this far. He was very patient with me and was always available whenever I had questions that needed his attention. I have learnt a lot of things over the last couple of months from him. Special thanks to Dr Rachel Lowe of the Catalan Insitute for Climatic Sciences (IC3) for imparting in me such a wealth of knowledge and helping me with data analysis especially with the R software. She has been an inspiring figure in my quest to be excellent in statistics. Deric Zanera made it all possible by kindly providing me with the data.

My gratitude also goes to the HRCSI programme at NCST for giving me a scholarship to attend this course. I was very secure in terms of materials needed to pursue the Biostatistics course. To all the lecturers who taught me over the past two yeas, I say thank you very much for the knowledge gained.

Secondly to Chris Moyo, head of M&E in the Ministry of Health, Willie Kachaka, George Chapotera, Macleod Mwale, Patrick Naphini, Ferdinand Khunga and the entire HMIS family for their patience during the last two years. I acted as though work was not important at times but these people still put up with me. In particular, Chris gave me the chance to go for the further studies even though I was very new at the department.

My deepest gratitude goes to the Mwadzaangati family at Mulunguzi in Zomba for their support during my time in Zomba.

Abstract

Malaria is one of the most important diseases in tropical and subtropical areas, with sub-Saharan Africa including Malawi being the region most burdened. The region has the right combination of biotic and abiotic components, including socioeconomic, climatic and environmental factors that sustain transmission of the disease. Heterogeneity in these conditions across the country consequently leads to spatial variation in risk of the disease. Analysis of nationwide survey data that takes into account this spatial variation is crucial in a resource constrained country like Malawi for targeted allocation of scare resources in the fight against malaria.

We used the 2010 Malaria Indicator Survey, which provides point referenced data for the analysis. Structured additive logistic regression models with spatial correlation were utilized to model the presence of parasitaemia in children while adjusting for child, household level and climatic covariates, environmental factors and personal interventions. The resultant model was then used to produce a malaria risk map for Malawi.

Children from poor households were over twice at risk of malaria than those from the richest households (OR=2.07, CI: 1.72-2.78). However, the results indicated a possible nonlinear relationship. On the other hand, the youngest children aged between 0 and 1 year are about 76% less likely to contract malaria than children aged between 4 and 5 (OR=0.244,CI:0.196,0.281). Those aged between 3 and 4 are only 28% less likely to have malaria (OR=0.717, CI:0.667-0.818). There is a general increase in risk as the child approaches the age of five which could be

explained by a decline in maternal immunity. Average total rainfall in the three months preceding the survey did not show a strong association with the disease risk while minimum temperatures shows an association with disease risk. The predicted malaria risk map produced by the model was in conformity with the expected disease pattern whereby central plain areas have higher risk than the high altitude districts in the north.

Our risk maps show an improved estimation at local level than previous efforts which were based on limited data collected from small surveys. It is hoped that this study can help reveal areas that require more attention from the authorities in the continued fight against childhood malaria.

Contents

List of Figures

List of Tables

Appendices

List of abbreviations

AIC	Akaike Information Criterion
BCI	Bayesian Confidence Interval
BIC	Bayesian Information Criteria
CAR	Conditional Autoregressive
DIC	Deviance Information Criterion
GAM	Generalized Additive Models
GAMM	Generalized Additive Mixed Models
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GoM	Government of Malawi
GPS	Global Positioning System
IRS	Indoor Residual Spraying
ITN	Insecticide Treated Net
MBG	Model Based Geostatistics
MCMC	Markov Chain Monte Carlo
MRF	Markov Random Fields
MIS	Malaria Indicator Survey
NGO	Non Governmental Organization
NMCP	National Malaria Control Programme
NSO	National Statistical Office
PMI	President's Malaria Initiative
RBM	Roll Back Malaria
RDT	Rapid Diagnostic Test
REML	Restricted (Residual) Maximum Likelihood
SEA	Standard Enumeration Area
STAR	Structured Additive Regression models

Chapter 1

Introduction

1.1 Global burden of malaria

Malaria is one of the most important diseases in the world today and it is common in tropical and subtropical areas, with sub-Saharan Africa being the region most burdened. It is a vector borne disease caused by the protozoan *Plasmodium*. About 90% of all malaria cases are reported in this part of Africa. The region has the right combination of biotic and abiotic factors including socioeconomic, climatic and environmental variables that sustain transmission of the disease. The parasite *Plasmodium falciparum* is the cause of the fatal type of malaria. The other types malaria parasites (*Plasmodium vivax, Plasmodium ovale and Plasmodium malariae*) are less likely to cause fatal episodes of malaria.

The disease has both serious social and economic implications in the countries where it is endemic. On the economic front, the disease is estimated to cost about $12 billion every year in lost GDP in Africa. In the year 2007, 2.73 billion people lived in areas at any risk of *Plasmodium falciparum* which is the malaria causing parasite and almost all *P.falciparum* prevalence rates above 50% were reported in Africa (Guerra et al., 2008). Every year, malaria accounts for 16% of all under-five deaths in Africa (Yoko and Rifat, 2011).

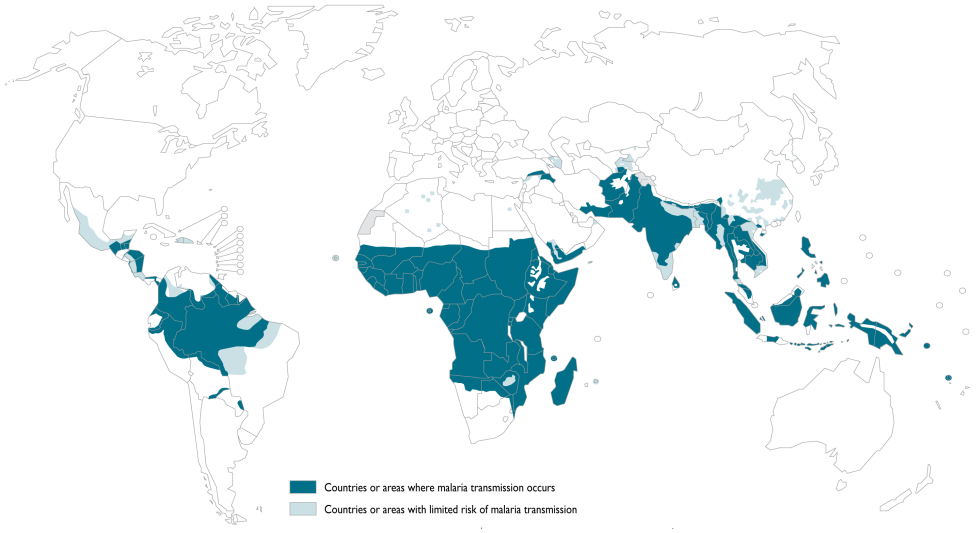


Figure 1.1: World map showing malaria endemic regions in 2010. *Source*: WHO

1.2 Local malaria burden

Malawi has not been spared the burden of the disease either with an estimated 6 million cases annually. The disease is the single biggest killer of children under the age of five in Malawi. Children in this age group are at risk of other co-infections as a result of malaria such as anaemia. This age group is also at a higher risk of developing cerebral malaria which is a severe form of the disease.

The government of Malawi (GoM) together with its development partners as well as the Global Fund have pooled together a lot of resources in the battle against malaria. The total expenditure on malaria as a percentage of total expenditure increased from 10.28% in 2002/03 to 12.18% in 2005/06. This increase in malaria financing is in line with the National Malaria Strategic Plan that was concerned with scaling up malaria interventions in order to achieve a malaria free Malawi.

1.3 Malaria epidemiology and transmission

In Malawi, the disease is mainly caused by *Plasmodium falciparum* accounting for 98% of all malaria cases (Ministry of Health (MOH), 2010). Mosquitoes act as the vectors especially those of the *Anopheles funestus*, *A. gambiae* and *A. arabiae*. Transmission is through bites of mosquitoes carrying the parasite.

Malaria has been known to be climate driven as the vector activities are determined by the prevailing conditions. Higher temperature, humidity and rainfall provide the optimum conditions for the breeding and development of these vectors. A 2006 study found that malaria risk in Malawi was significantly associated with climatic factors such as rainfall, maximum temperature (Kazembe et al., 2006). Consequently, malaria incidence peaks during the period October to April which coincides with the rainy season and thus the right environmental and climatic conditions. The lower Shire districts of Nsanje and Chikhwawa have also been identified as having the best combination of climatic and geographical features that increases malaria transmission (Djinjalamala, 2006).Temperature further dictates the latitudinal and altitudinal ranges of the vector (Westbrook et al., 2010). On the other hand, extreme climatic conditions are not suitable for the life cycle development of mosquitoes (Gemperli, 2003).

The transmission and range of the disease are being projected to change due to climate change. This hypothesis has already been tested in West Africa and Europe using the scenarios of the International Panel on Climate Change Annual Report 4 (IPCC AR4)(Usher, 2010).

1.4 Malaria interventions

In a bid to stop malaria, the Roll Back Malaria (RBM) partnership was founded in order to halve malaria burden by the year 2010. Some of the strategies advanced

by RBM to combat malaria include country strategic plans, country partnerships and health systems delivery. The NMCP strategic plan for 2011-15 aims to achieve universal coverage in the prevention and treatment of malaria.

There are many preventive measures for malaria that are currently in use in Malawi. The most commonly used and perhaps the cheapest are insecticide treated nets (ITNs). The President's Malaria Initiative (PMI) has been supporting ITN distribution and indoor residual spraying (IRS) in the country. This initiative distributed 2, 370, 831 ITNs between 2007 and 2011 countrywide. It has also sprayed 97, 329 homes countrywide protecting 364, 349 people in the process (*President's Malaria Initiative (PMI) Country Profile: Malawi*, 2011). These methods are targeted at killing the vectors thereby stopping transmission.

1.5 Distribution of malaria

In Malawi, all people are at risk of malaria (WHO, 2010). Malaria prevalence and incidence greatly varies in the country as a result of a wide range of factors including socio-economic and climatic. Socioeconomic factors play a role in this disparity as revealed by differences in rural and urban disease burden. Children from rural and less privileged families are more vulnerable to malaria attacks and have a higher risk of developing severe malaria than children from urban areas. According to the 2010 DHS report, 30.7% of urban children had fever in the preceding two weeks before the survey as compared to 35.1% of rural children (National Statistical Office (NSO) and ICF Macro, 2011). The 2010 MIS survey reported malaria parasitaemia prevalence of 14.7% in urban areas and 46.9% prevalence in rural areas.

On the other hand, rainfall, temperature and humidity are known to have an impact on malaria. These conditions directly affect the vectors both negatively and positively. Very high and low temperatures for example slow down life cycle de-

velopment of the mosquitoes thereby reducing or stopping transmission at certain temperatures. Generally, malaria prevalence is lower in highland areas where temperatures are much lower. Such areas include Nyika Plateau, Dedza among others (Kazembe, 2007). On the other hand, low lying areas with higher temperatures are associated with higher risks of malaria. The lakeshore and Lower Shire areas are examples of such places.

1.6 Problem statement

From the foregoing discussion, it is obvious that there should be a spatial variation in the risk of malaria across the country. Such variation could be between places such as rural and urban, in terms of elevation among other factors. However, there has not been much research on the spatial statistical analysis of malaria data in the country. Kazembe is one of the researchers that have worked in this area (Kazembe et al., 2006; Kazembe, 2007). These papers however did not use comprehensive malaria data from nationwide survey like the MIS.

The lack of geo-referenced data that is required for this kind of analysis has also contributed to less utilization of spatial statistical methods. As a result, the issue of disease mapping that is now crucial among epidemiologists has not been adequately addressed thus leading to a lack of an empirical malaria risk map for the country. Moreover, the predictive nature of these spatial statistical models has also not been fully exploited thereby denying policy makers a head start in the fight against the disease. Much of the knowledge about malaria distribution has most times been based only on expert opinion. Such information, though very useful in targeted malaria interventions needs also to be supported by data from representative surveys like the MIS.

1.7 Research questions

It is hoped that this analysis will answer the following critical questions regarding childhood malaria:

1. How does malaria risk in Malawi vary spatially?.
2. To what extent can spatial variation in malaria risk be accounted for by household and geographical variables as well climate variations?.
3. To what extent do interventions for malaria, such as ITN distribution, affect the spatial distribution of malaria among children.
4. Can the developed models predict where increases in malaria risk are more likely to occur?.

1.8 Research objectives

The main objective of the study was to predict malaria incidence in children in areas where no survey observations were made. The specific objectives were as follows:

1. To analyse, predict and map malaria prevalence in Malawi.
2. To develop models for predicting malaria risk.
3. To investigate risk factors for malaria.
4. To assess the impact of different malaria interventions on disease risk.

1.9 Justification and significance of the study

Through the Mapping Malaria Risk in Africa (MARA) project (MARA, 2004), the first coordinated efforts to map malaria risk in Africa were made. Based on

this project, many countries and researchers have produced risk maps for different countries in Africa that clearly show the malaria endemic areas on the continent as well as the different levels in the risk of the disease. These maps show malaria to be endemic to Malawi with the highlands having the lowest risk. The Malaria Atlas project (MAP) (MAP, 2006) is another attempt to map malaria prevalence with the purpose of effective allocation of scarce resources.

However, both MARA and MAP products, though informative are not very useful in predicting malaria incidence at the local level because they have a coarse resolution. The maps developed are climate based and as a result, they do not take into consideration the other factors that are connected to malaria risk and distribution. Both MARA and MAP did not use empirical data for Malawi.

Furthermore, different surveys that have been conducted in the past did not cover the whole country so as to be useful for predictive purposes. The MIS is the first nationally representative malaria survey to be conducted that provides rich information on geographical as well as socioeconomic variables. With data from MIS, a comprehensive risk map for malaria based on a wide range of variables is possible.

Chapter 2

Theory

In this chapter, the development of statistical models for the analysis of spatial data sets is discussed. The opening section looks at the linear model and its development and later its evolution into more general models that are capable of handling data where spatial correlation is present.

2.1 The linear model

Linear statistical modeling is one of the central ideas of statistics. In linear regression, a mathematical relationship between a response and explanatory variable is defined as a linear function. In the presence of k explanatory variables and n observations, a multiple linear regression model can be specified in this form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \qquad i = 1, \ldots n \tag{2.1}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients and x_{ij} are explanatory variables. In general, the linear model is often represented in matrix notation as follows:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.2}$$

In the above model, $\mathbf{y}$ is a vector of observations, $\mathbf{X}$ a matrix of explanatory variables known as the design matrix, $\boldsymbol{\beta}$ a vector of regression coefficients and lastly, $\boldsymbol{\varepsilon}$ is a vector of random errors. The following assumptions are made about the linear regression model.

1. Response variable y is continuous.
2. Errors are independent and identically distributed as normal with mean zero and constant variance, i.e. $\varepsilon \sim N(0, \sigma^2)$.

The method of least squares is widely used to estimate the model parameters $\boldsymbol{\beta}$.

2.2 Generalized linear model

The Generalized linear model (GLM) (McCullagh and Nelder, 1989) is an extension to the linear regression models. The classic regression models specified in the preceding section are not sufficient as they are restricted by the major assumptions above. In medical research for example, it is not uncommon to come across a response variable that is not continuous in nature. For example, whether a child has received a vaccination or not could be used as a response. Clearly, these models are not suitable in these situations. The GLM is a unifying model framework in which non-normal responses can be modelled and allows for a more complicated relationship between the response and the explanatory variables other than a simple linear relationship (Dobson, 2002).

The GLM has got three parts: random component, link function and the systematic component (Agresti, 2006). The random component identifies the response variable and its underlying distribution. The systematic component on the other hand specifies the explanatory variables. The explanatory variables are usually written as a linear combination in the form $\eta = \boldsymbol{X}\boldsymbol{\beta}$ to form the linear predictor. This term is related to the expected value of the response variable through the link function, $g(.)$. The choice of a link function depends on the distribution of

the response variable. For a binomial distribution such as in this study where the response is binary, the logit link $\log(\frac{\mu}{1-\mu})$ is used. In other words,

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

where $\mu = E(Y)$. In a GLM the probability distribution of the response variable is a member of the exponential family of distributions which can be written in the form:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]. \tag{2.3}$$

The θ in equation 2.3 is the canonical parameter representing the location and the ϕ is the dispersion parameter (Faraway, 2006).

2.3 Generalized linear mixed model

It happens in some instances that there is correlation between the observations that are being modelled. In longitudinal studies for example, where repeated measurements are made on the same individual over time, the data is also usually correlated (Jiming, 2007). In spatial statistics, observations are also usually correlated due to spatial autocorrelation which means that observations close in space are more similar than those further apart (Lawson, 2008). The GLM looked at earlier assumes that the observations are independent which is not the case in this situation (Hedeker and Gibbons, 2006). As a solution, linear mixed models can be utilised to take into account the random effects.

A generalized linear mixed model (GLMM) (Breslow and Clayton, 1993) is an extension of the GLM in the sense that random effects are accommodated in addition to the fixed effects. In general, a mixed model has the form,

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}.$$

where $\mathbf{y}$ is the response vector, $\boldsymbol{u}$ is a vector of random effects which are usually assumed to follow a normal distribution with mean $\mathbf{0}$ and some variance-covariance matrix $\boldsymbol{\Sigma}$, that is $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. A logistic regression with random effects may be written as,

$$\text{logit}(p_i) = \boldsymbol{X_i\beta} + \boldsymbol{Z_i u}. \tag{2.4}$$

2.4 Types of spatial data

Spatial data is finding increasing usage in medical research and other disciplines. In this section, the different classes of spatial data that are encountered in practice are briefly discussed.

2.4.1 Areal data

In areal (lattice) data, it is thought that there exists a regular or irregular subset D divided into a finite number of areal units. These units possess well-defined boundaries that enclose the spatial regions.

2.4.2 Point-pattern data

Point-pattern data is obtained when the subset D is random, its index set gives the locations of random events that are in the spatial point pattern. $Y(s)$ equals 1 for all $s \in D$.

2.4.3 Point-referenced data

This kind of data is obtained when the sampled points x_i are georeferenced by either latitude-longitude or northing-easting systems. Point referenced (also known as geostatistical) data has its own dedicated branch of spatial statistics known as

geostatistics which is concerned with this kind of data. In this stufy, the focus is to use malaria risk observed at spatial locations x_i to predict the risk throughout the study region which is the whole of Malawi. The use of these points for prediction is necessary since it may not be feasible to measure the risk at each and every location. The malaria risk is then represented by a spatially continuous stochastic process $S(x)$ which is a function of location x_i.

To model geostatistical data such as the MIS dataset, Gaussian stochastic processes are widely used (Diggle and Ribeiro, 2007). A Gaussian spatial process, $S(x) : x \in \mathbb{R}^2$ is a stochastic process with the property that for all the spatial locations $x_1, \ldots, x_n$ with each $x_i \in \mathbb{R}^2$, the joint distribution of $S = S(x_1), \ldots, S(x_n)$ is multivariate Gaussian (Diggle and Ribeiro, 2007). Obviously there is a discrepancy between the true risk, $S(x_i)$ and Y_i which is the measured risk during the survey and this has to be taken into consideration in the model. In its simplest form, geostatistical data is represented by $(x_i, y_i) : i = 1, \ldots, n$ where x_i is the spatial location and y_i is the measured value at x_i. The stationary Gaussian model has these two assumptions (Diggle and Ribeiro, 2007):

1. $\{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean μ, variance $\sigma^2 = \text{Var}\{S(x)\}$ and a correlation function $\rho(u) = \text{Corr}\{S(x), S(x')\}$ where $||x - x'||$ is the distance between spatial locations x and x';

2. Conditional on $S(x) : x \in \mathbb{R}^2$, the y_i are the realizations of mutually independent random variables Y_i, normally distributed with conditional means $E[Y_i|S(.)] = S(x_i)$ and conditional variances τ^2

The model can then be represented as the equation

$$Y_i = S(x_i) + Z_i : i = 1, \ldots, n \tag{2.5}$$

where $S(x) : x \in \mathbb{R}^2$ is defined as above and the Z_i are mutually independent $N(0, \tau^2)$ random variables.

The MIS dataset is an example of geostatistical data that fits this description. In this survey, malaria status of a child measured by Rapid Diagnostic Test (RDT) was measured at sampled locations across the country. The country can be thought of as the entire spatial region and the sampled locations as realizations of an unobservable spatial process. Furthermore, all households with children who were eligible for testing were referenced by latitudes and longitudes. In theory, malaria can be detected everywhere in Malawi. However, in practice, the observed data are just a partial realization of the spatial process observed at $\{x_1, x_2, \ldots, x_n\}$, which represents the sampled households.

Point referenced data is increasingly been collected in Malawi mainly due to the need of linking health outcomes to specific locations for targeted interventions and increased access to GPS technology. The fact that data is typically collected at a subset of all the locations in the area of interest makes inference about the process $S(.)$ at new unsampled locations the primary objective (Banergee and Finley, 2009). The survey data is presented in the figure below.

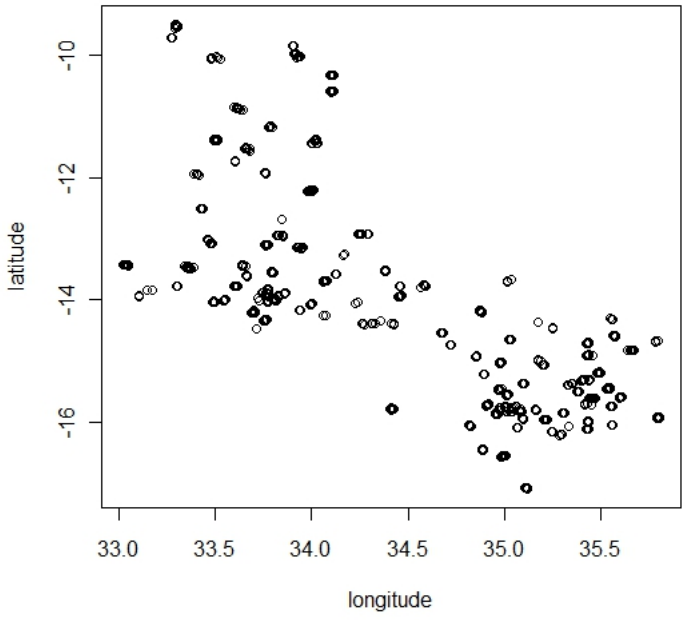


Figure 2.1: Point referenced data from MIS survey

2.5 Analysis of spatial data

The commonly used statistical models falling in the family of GLMs assume that the observations are independent. This obviously is violated when there is a correlation between the observations. For spatial data, a phenomenon known as spatial autocorrelation is usually seen. This refers to the idea that values close to each other in space are more similar than those separated by a greater distance. In this project, the malaria status of a child at a location is a binary covariate hence necessitating a logistic regression. As a result, the simple GLM has to be modified to take into account this autocorrelation. This modification leads to the use of GLMM and Bayesian techniques are often employed for inference.

In the next two sections, we look at different approaches to analyzing the two commonly encountered types of spatial data in epidemiology. We look at how the GLM is adapted to take into consideration the inherent spatial correlation in the data.

2.5.1 Lattice spatial modeling

The analysis of areal data is quite common in the literature and such problems occur quite often in public health. If for example, the presence or absence of malaria at an enumeration area (EA) level in the country can be designated as $Y(s)$ where s is the EA and the EAs are the grid cells, then a logistic regression model that includes a spatial autocorrelation can be defined as:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i'\beta + u_i. \tag{2.6}$$

For this equation, the spatial random effect u_i is associated with each of the grid cells and adjusts the probability of presence of the variable of interest depending on the value of p in the cell's spatial neighbourhood (Latimer et al., 2006). The

Gaussian conditional autoregressive model is used to capture this process (Besag and Kooperberg, 1995). It is assumed that the conditional distribution of the spatial random effect in a given cell i given values of the spatial random effect in the other grids $j \neq i$ depends only on the neighbouring cells of i (Latimer et al., 2006). Cells i and j are defined as neighours if their boundaries intersect. The spatial effect for any given cell depends on the value of u for the cells in its neighbourhood, i.e.

$$u_i|u_j \approx N\left(\frac{\sum_{j\in\delta_i} a_{ij}u_j}{a_{i+}}, \frac{\sigma_u^2}{a_{i+}}\right); \qquad j \neq i. \tag{2.7}$$

In this CAR model, a_{i+} denotes the total number of neighbouring cells of i and $a_{i+} = 1$ if two sites i and j share the same boundaries and 0 otherwise. The variance term σ_u^2 is assigned an inverse gamma prior which has mean b_u with infinite variance, i.e. $\sigma_u^2 \sim IG(2, b_u)$. The β are assigned non-informative normal priors with mean 0 and a large variance.

2.5.2 Point-level statistical modelling

In the preceding section, an example of a common approach to analysing spatial binary data at grid level was discussed. In this section, we go a step further to look at data that is point referenced. The data is referenced to exact locations where the observations were made and not to grid cells. Unlike the approach used in lattice model where the spatial structure is modelled between neighbouring cells, the spatial autocorrelation in point level models is directly modelled between the points in the dataset. The distances between the points are typically used to model this spatial dependence.

Point referenced data have been extensively analysed using model based geostatistics (Diggle et al., 1998). Model based geostatistics (MBG) is a framework that utilises explicit parametric stochastic models to geostatistical data. Since their

development, MBG models, which are typically implemented under a Bayesian framework, have enjoyed increasing usage among researchers. For instance, a Bayesian logistic geostatistical analysis of Human African Tryoanosomiasis was done in Uganda (Wardrop et al., 2010). In West Africa, Schistosomiasis prevalence was analysed using the Bayesian geostatistical models (Schur et al., 2011). A similar study was conducted to predict intensity of infection with Schistosoma misoni in East Africa (Clements et al., 2006). Bayesian geostatistical models were also used to analyse malaria indicator survey data in Angola and Zambia (Gosoniu et al., 2010; Riedel et al., 2010). Finally, a Bayesian binary logistic geostatistical model was used to study risk factors for childhood malaria in the Gambia (Rowlingson et al., 2002).

For this study, let the outcome of a test for malaria at household i be y_i. If $y_i = 1$ for a diseased child and $y_i = 0$ for an under five child without the disease and p_i is the probability of testing positive, then

$$y_i \sim Bernoulli(p_i).$$

An ordinary GLM with a logit link can be fitted to the data to model the relationship between malaria status and different covariates (household level, environmental and climatic etc)

$$\text{logit}(p_i) = \alpha + \beta_i' x_i. \tag{2.8}$$

In the model α is the intercept, β_i is the vector of regression coefficients and x_i is a vector of explanatory variables. This base model is fitted in order to find which variables are associated with malaria prevalence in the area.

In order to capture the spatial autocorrelation that may be overlooked by the ordinary GLM, a GLMM with spatial random effects has to be fitted. In the absence of this spatial random effect term, there is danger of attributing unobserved spatial variation to the random error. The following equation is a GLMM with a

spatial random effect.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_i' x_i + S_i. \tag{2.9}$$

A fully Bayesian approach is used to estimate model parameters. Each parameter in the model 2.9 above is assigned a prior distribution that will be updated to obtain the posterior. The spatial component $\mathbf{S} = (s_1, \ldots, s_n)^T$ is assumed to be distributed as a multivariate normal with mean 0 and covariance matrix between any two locations s_i and s_j is,

$$\Sigma_{ij} = \sigma^2 \exp\left(\frac{-d_{ij}}{\rho}\right).$$

In this formulation, σ^2 stands for the spatial variation while ρ controls the rate of decay of spatial autocorrelation (Gosoniu et al., 2006). The term $d_{ij} = x_i - x_j$ measures the shortest distance between two locations x_i and x_j. The coefficients are typically given non informative uniform priors while the σ^2 and ρ are assigned vague inverse gamma priors .i.e. $p(\beta) \propto 1$, $p(\sigma^2) = IG(a_1, b_1)$ and $p(\rho) = IG(a_2, b_2)$ respectively (Gemperli, 2003).

2.6 Structured additive regression models

Until now, we have been looking at both GLMs and GLMMs as they are applied in geostatistics. In both these models the assumption of linearity of the covariate effects is usually made. GLMMs for instance are used to model covariate effects using a parametric mean function while accommodating correlation and overdispersion by adding random effects to the linear predictor (Lin and Zhang, 1999).

Generalized additive mixed models (GAMM) extend the GLMM by adding unknown smooth functions of continuous and spatial covariates among others (Lang

and Fahrmeir, 2001). The parametric mean assumption may not be appropriate since the functional form of these covariates may not be known (Lin and Zhang, 1999). In general, the GAMM and the generalized additive model (GAM) (Hastie and Tibshirani, 1990) are special cases of a broader group of additive models known as structured additive regression models (STAR) which will be our focus for the remainder of this thesis. Just like GAMMs, GAMs extend the GLM by adding smooth functions in addition to the linear terms in the predictor (Kelsall and Diggle, 1998). Another class of models belonging to the STAR are geoadditive models (Kammann and Wand, 2003) which were derived by merging additive and geostatistical models. These STAR models are increasingly being used mainly due to the accessibility of powerful computers which have made the analyses somewhat easy.

The presence or absence of the malaria parasite in a child can potentially be modelled as a STAR model in any of its forms. The models considered earlier possess the linear predictor component

$$\sum x_{ij}\beta_j,$$

which can be replaced by the additive component

$$\sum f_j(x_{ij}).$$

In our study, we are concerned with the problem of fitting a logistic regression to model the outcome of a malaria test in a child but taking into account the spatial correlation observed. The ordinary GLM without any spatial effects is given by,

$$\log\left(\frac{p(y_i|x_{i1},\ldots,x_{ip})}{1-p(y_i|x_{i1},\ldots,x_{ip})}\right) = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip} \tag{2.10}$$

where $p(y_i|x_{i1},\ldots,x_{ip})$ is the probability of testing positive given a set of different covariates $x_{i1},\ldots,x_{ip}$. The GAM is obtained by simply adding the additive

component to model 2.10 and consequently becomes

$$\log\left(\frac{p(y_i|x_{i1},\ldots,x_{ip})}{1-p(y_i|x_{i1},\ldots,x_{ip})}\right) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}). \tag{2.11}$$

In equation 2.11, the functions $f_1,\ldots,f_p$ are the smooth functions. The GAMM is just an extension to the GAM and for our logistic model can be written as,

$$\text{logit}(p_i) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + z_i^T b. \tag{2.12}$$

The geoadditive model, also belonging to STAR models was defined by (Kammann and Wand, 2003) by combing the simple universal kriging model 2.39 with the an additive model $y = X\beta + Zb + \varepsilon$ to yield

$$y_i = \beta_0 + f(s_i) + g(t_i) + \beta_1'(x_i) + s(x_i) + \varepsilon.$$

Like GAMs and GAMMs these models are also being used in practice. For instance geoadditive modeling of malaria was done in Burundi (Nkurunziza et al., 2010). This study used cubic splines to model effects of continuous covariates. These models have also been used in South Africa to assess nonlinear geographical variation in HIV prevalence while controlling for demographic and sexual risk factors (Wand et al., 2011). Kandala examined the spatial variation in under five malnutrition and risk factors for child morbidity by using these models (Kandala et al., 2007, 2011).

In general, the equation below represents the general form of a STAR model which unifies all these special cases.

$$\eta = f_1(x_1) + f_2(x_2) + \cdots + f_j(x_j) + \cdots + f_p(x_p) + \mathbf{u}'\gamma. \tag{2.13}$$

In equation 2.13, x_j are the covariates of different types and f_j are functions of the covariates and contain the non-linear effects of the continuous covariates and γ is

the vector of regression coefficients of linear effects. However, it is necessary to represent the smooth effects of the metric covariates in some way. These functions are approximated by a linear combination of basis functions which are defined as follows (Belitz et al., 2009*b*);

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x). \tag{2.14}$$

The B_k are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a vector of regression parameters to be estimated. The incorporation of the smoothing functions in the general STAR model 2.13 leads to a linear model framework (Belitz et al., 2009*b*). Model 2.13 then becomes:

$$\eta = \mathbf{X_1}\beta_1 + \cdots + \mathbf{X_p}\beta_p + \mathbf{U}\boldsymbol{\gamma} + \varepsilon, \tag{2.15}$$

where $\mathbf{U}$ is the design matrix for linear effects, $\boldsymbol{\gamma}$ is the vector of regression coefficients for linear effects and $\boldsymbol{\varepsilon}$ is the vector of errors (Belitz et al., 2009*b*). A roughness penalty is imposed on the regression coefficients to avoid over fitting since a large number of basis functions is usually specified. We employ the same quadratic penalty of the form $\beta'\mathbf{P}(\lambda)\beta$ as in (Belitz et al., 2009*b*) where $\mathbf{P}(\lambda) = \lambda\mathbf{K}$ is the penalty matrix and λ is a scalar smoothing parameter that determines the smoothness of fit.

2.7 Parameter estimation

In this section, we take a closer look at the major approaches to inference of STAR models. We discuss the expanded approach to estimation and Bayesian methods. However, challenges are faced when it comes to parameter estimation. In the expanded approach, we briefly discuss marginal and joint distribution methods.

2.7.1 Expanded approach to parameter estimation

The marginal estimation methods are concerned with the fixed effects. The focus is to come up with the marginal distribution from which the parameters can be estimated. However, obtaining the marginal distribution of the form $f(y;\beta) = \int f(y,\beta,u)du$ from the mixed model 2.4 poses a great challenge. Approximate methods utilized include the quasi likelihood, Laplace approximation, Gauss-Hermite quadrature and Markov Chain Monte Carlo. In general, the likelihood of a GLMM involves the following integral (Dalgaard, 2006)

$$\int \prod_{j=1}^{n_i} f(y_{ij}|b_i,\beta,\phi)f(b_i|D)db_i. \tag{2.16}$$

Integral 2.16 cannot be exactly evaluated hence the need for better methods of parameter estimation.

2.7.2 Bayesian statistical modeling

In Bayesian approach to statistical analysis, prior knowledge is incorporated to come up with new estimates. The parameters are thought to be random with distributions. The whole idea is to update observed data with prior knowledge to come up with posterior beliefs that can then be used for inference. All the inferences such as estimating means are carried out on the posterior distributions. The posterior mean for example, is a weighted mean of the prior and the observed data. At the cornerstone of this method is the Bayes Theorem which for a continuous probability density function can be written in the form below:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}. \tag{2.17}$$

In this equation, $f(\theta|y)$ is the conditional distribution of the parameters given the data known as the posterior distribution. The prior distribution is given by $f(\theta)$

while $f(y|\theta) = \int f(y|\theta)f(\theta)d\theta$ is the likelihood function. In most cases however, the expression is written in the form $f(\theta|y) \propto f(y|\theta)f(\theta)$.

As an example, suppose one has drawn a single observation y where $y \sim N(\theta, \sigma^2)$. If an assumption is made that σ^2 is known, then

$$f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-1}{2\sigma^2}(y-\theta)^2\right). \tag{2.18}$$

Further, suppose that the prior of θ is normal distribution, i.e. $p(\theta) = N(\theta|\mu, \tau)$ where μ and τ^2 are known parameters. Then the posterior is computed as

$$p(\theta|y) \propto N(\theta|\mu, \tau^2) \times N(y|\theta, \sigma^2) \tag{2.19}$$

After some algebraic manipulations, the posterior comes up to:

$$p(\theta|y) = N\left(\theta|\frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right). \tag{2.20}$$

The maximum likelihood estimation (MLE) can be used to estimate the posterior parameters for inference. However, this is not feasible for high dimensional integrals that often result from combining the likelihood with the prior. The posterior becomes complicated making evaluation under the MLE approach difficult. Several numerical techniques have been developed to aid in the analysis of data in a Bayesian approach.

2.7.3 MCMC simulation

The evaluation of the posterior distribution $p(\theta|y)$ to obtain statistics such as the mean is computationally demanding and requires indirect methods of evaluating the integral. Other methods such as Monte Carlo and Laplace are limited in their scope (Smith and Roberts, 1993). To get around this problem, Markov Chain Monte Carlo (MCMC) simulation techniques are utilized and are responsible for

the rapid rise in use of Bayesian analysis since the early 1990's. These methods combine Monte Carlo Simulation and Markov chain ideas hence the name.

In MCMC, the idea of a Markov Chain comes in which is defined as:

$$\mathbb{P}(X_n \in A|X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n \in A|X_{n-1})$$

for $\{X_n, n \in \mathbb{N}\}$ defined on $(m, h(m))$, i.e. future states are independent of past states given the present state. The MCMC techniques simulate draws from the complex distribution of interest which is usually the posterior distribution. The idea is to learn from the posterior by repeatedly sampling from it and then summarizing the draws. To compute the posterior mean for example, the following integral has to be evaluated.

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta$$

In order to make inference about the posterior, a sequence of G random draws

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$$

from the posterior $p(\theta|y)$ is drawn. Then the posterior is computed as a mean of the G draws. In short, this integral is evaluated via Monte Carlo integration and the simulation is through Markov chains i.e,

$$E(\theta|y) = \int_{\Theta} \theta p(\theta|y) d\theta \approx \frac{1}{G} \sum_{g=1}^{G} \theta^{(g)}.$$

Suppose the draw $\theta^{(t)}$ is the present state at iteration t. The next draw $\theta^{(t+1)}$ depends only on the current draw $\theta^{(t)}$ and not on any other past draws thus giving rise to the Markov chain;

$$p(\theta^{t+1}|\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}) = p(\theta^{(t+1)}|\theta^{(t)})$$

The generated chain is made up of draws and each is slightly dependent on the previous one and it converges to the target distribution $p(\theta|y)$ under any sampling scheme regardless of the starting point.

There are two sampling schemes used in MCMC, Gibbs and Metropolis-Hastings algorithms. In Metropolis-Hastings algorithm (M-H), a function proportional to the density function is required and has to be calculated. On the other hand, the Gibbs sampler does not need to calculate this function and it is generally faster than the M-H algorithm since it works on weaker assumptions. The software package WinBUGS[1] uses the Gibbs sampler for analysis. In our study, we use the M-H algorithm as implemented by Bayes X (Belitz et al., 2009*a*). In general, the estimator $\frac{1}{G}\sum_{g=1}^{G}\theta^{(g)}$ converges to $E(\theta|y)$. More details on the MCMC algorithms are included in the appendix.

2.7.4 Fully Bayesian approach

The unknown functions f_j in the STAR model are assumed to be random and hence have their own distributions. The Bayesian approach to statistical inference demands that prior knowledge be incorporated in the inference process. Without any prior knowledge, the most appropriate priors for the fixed effects parameters are the diffuse priors, i.e.

$$p(\theta_i) \propto const \tag{2.21}$$

Priors for unknown functions f_j depend on the type of covariates and on prior beliefs about the smoothness of f_j. By expressing the vector of function evaluations $\mathbf{f_j} = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ of a function f_j as the product of a design matrix and a vector of unknown parameters $\boldsymbol{\beta_j}$, the general STAR model can be written in this manner;

$$\boldsymbol{\eta} = \mathbf{X_1}\boldsymbol{\beta_1} + \cdots + \mathbf{X_p}\boldsymbol{\beta_p} + \mathbf{U}\boldsymbol{\gamma}, \tag{2.22}$$

[1] `http://www.mrc-bsu.cam.ac.uk/bugs/`

where $\mathbf{U}$ is the design matrix for fixed effects and $\mathbf{f_j} = \mathbf{X_j}\boldsymbol{\beta_j}$. The prior for the function f_j is defined by specifying a suitable design matrix $\mathbf{X_j}$ and a prior distribution for the vector $\boldsymbol{\beta_j}$ of unknown parameters which has the general form

$$p(\boldsymbol{\beta_j}|\tau_j^2) \propto \frac{1}{(\tau_j^2)^{rank(\mathbf{K_j}/2)}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta_j}'\mathbf{K_j}\boldsymbol{\beta_j}\right), \tag{2.23}$$

where $\mathbf{K_j}$ is a penalty matrix. The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a penalized likelihood approach and controls the trade off between flexibility and smoothness. The unknown variance parameters τ_j^2 are assigned hyperpriors. For this study, we assign the usual non informative dispersed inverse Gamma priors $p(\tau_j^2) \sim IG(a_j, b_j)$ where,

$$\tau_j^2 \propto (\tau_j^2)^{-a_j-1}\exp(-b_j/\tau_j^2) \tag{2.24}$$

Bayesian inference is based on the posterior.

$$p(\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2, \gamma|\mathbf{y}) \propto L(\mathbf{y}, \beta_1, \ldots, \beta_p, \gamma) \prod_{j=1}^{p} (p(\beta_j|\tau_j^2)p(\tau_j^2)), \tag{2.25}$$

where $L(.)$ is the likelihood which is simply the product of individual likelihood contributions.

2.7.5 Emperical Bayesian inference

The empirical Bayesian approach to inference is based on mixed model methodology. The STAR model is reparameterized as a GLMM (Fahrmeir et al., 2004) and then restricted maximum likelihood estimation approach (REML) is used. Parameter estimation in REML approach involves decomposing the vector of regression coefficients $\boldsymbol{\beta_j}$ in the STAR model 2.13 into penalized and unpenalized parts. For a parameter vector $\boldsymbol{\beta_j}$ with dimension $K_j \times 1$ and the penalty matrix $\mathbf{K}_j$ with

rank j, the decomposition process yields

$$\boldsymbol{\beta}_j = \mathbf{X}_j^{unp}\boldsymbol{\beta_j}^{unp} + \mathbf{X}_j^{pen}\boldsymbol{\beta_j}^{pen}. \tag{2.26}$$

From the decomposition, the following is obtained

$$\frac{1}{\tau_j^2}\beta_j'\mathbf{K}_j\beta_j = \frac{1}{\tau_j^2}(\beta_j^{pen})'\beta_j^{pen}$$

The general prior for β_j in expression 2.23 leads to the following;

$$p(\beta_{jm}^{unp}) \propto const, \qquad m = 1, \ldots, K_j - k_j$$

and

$$\beta_j^{pen} \sim N(\mathbf{0}, \tau_j^2\mathbf{I}).$$

The last step the reparameterization process involves defining matrices $\tilde{\mathbf{U}}_j = \mathbf{X}_j\mathbf{X}_j^{unp}$ and $\tilde{\mathbf{X}}_\mathbf{j} = \mathbf{X_j}\mathbf{X}_j^{pen}$ thus leading to the predictor $\eta = \sum_{j=1}^{p}\mathbf{X}_j\beta_j + \mathbf{U}\gamma$ defined in equation 2.22 changing to;

$$\eta = \sum_{j=1}^{p}(\tilde{\mathbf{U}}_j\beta_j^{unp} + \tilde{\mathbf{X}}_j\beta_j^{pen} + \mathbf{U}\gamma) \tag{2.27}$$

$$= \tilde{\mathbf{U}}\beta^{unp} + \tilde{\mathbf{X}}\beta^{pen}. \tag{2.28}$$

The final GLMM has fixed effects β^{unp} and random effects $\beta^{pen} \sim N(\mathbf{0}, \mathbf{\Lambda})$ where $\mathbf{\Lambda} = diag(\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_p^2, \ldots, \tau_p^2)$. The variances τ_j^2 are assumed to be unknown constants that have to be estimated from their marginal likelihood. The posterior is given as follows.

$$p(\beta^{unp}, \beta^{pen}|\mathbf{y}) \propto L(\mathbf{y}, \beta^{unp}, \beta^{p}en)\prod_{j=1}^{p}(p(\beta_j^{pen}|\tau_j^2)), \tag{2.29}$$

2.8 Modelling spatial effects in STAR models

In the preceding sections, we looked at approaches widely used to model spatial effects in the parametric geostatistical models for both point referenced and gridded data. In STAR models, methods for point and grid data also exist. More details are discussed below.

2.8.1 Markov random fields

The Markov random fields (MRF) are the CAR models discussed under the parametric models. The similarity lies in the fact that we are modelling spatial effects on a lattice. In both these methods, conditional distribution on a cell depends on the neighbouring cells. The sites in $\mathbb{S}$ are related to one another via the neighbourhood system which is defined as $\mathbb{N} = \{N_i, i \in \mathbb{S}\}$, where N_i is a set of sites neighbouring i and $i \notin N_i$. Moreover, this relation holds true, i.e. $i \in N_j$ implies that $j \in N_i$.

Suppose $s \in \{1, \ldots, S\}$ is the location or site in connected spatial regions. In our context, these spatial regions can either be enumeration areas or districts. These regions are labeled consecutively for simplicity reasons and we assume that neighbouring sites are more similar than the distant ones. Any two sites s_i and s_j are considered neighbours if they share a common boundary. The simplest and commonest spatial prior for the function evaluations $f(s) = \beta_s$ is,

$$\beta_{s_i} | \beta_{s_j}, s_i \neq s_j, \tau^2 \sim N\left(\frac{1}{N_{s_i}} \sum_{s_j \in \partial_{s_i}} \beta_{s_j}, \frac{\tau^2}{N_{s_i}}\right) \tag{2.30}$$

where N_{s_i} is the number of adjacent sites and $s_j \in \partial_{s_i}$ denotes that site s_j is a neighbour of site s_i. The conditional mean of β_{s_i} is therefore an unweighted average of function evaluations at neighbouring sites. The MRF is a direct generalization of a first order random walk prior.

2.8.2 Polynomial splines

A polynomial spline of degree p is defined in this manner; $f : [a, b] \to \mathbb{R}$ where $p \in \mathbb{N}$ with equally spaced knots:

$$a = \kappa_0 < \kappa_1 < \cdots < \kappa_{m-1} < \kappa_m = b. \tag{2.31}$$

A simple linear model of the form $y_i = m(x_i) + \varepsilon$ can be written in terms of a polynomial spline in order to approximate $m(.)$ as:

$$m(x; \beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)_+^p, \tag{2.32}$$

where p is the degree of the polynomial, $\kappa_1 < \cdots < \kappa_K$ is a set of K knots and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p+K})$ (Ruppert et al., 2003). Types of polynomial splines include the P-splines and the B-splines among others.

Penalised splines based on the linear combination of basis functions in equation 2.14 are utilized to solve the problem of overfitting and underfitting which sometimes arises. For each regression coefficient, β_k, there is an associated smooth term of the form $\mathbf{P(\lambda)} = \lambda K$ where λ is a scalar parameter that controls smoothing. P-splines are another approach used to model the effects of continuous covariates (Eilers and Marx, 1996). An assumption that there is an underlying unknown smooth function f of a covariate x is made. It is further assumed that this smooth function can be approximated by a polynomial spline of degree l defined by equally spaced knots, i.e.

$$x_{min} = \kappa_0 < \kappa_1 < \cdots < \kappa_{m-1} < \kappa_m = x_{max} \tag{2.33}$$

within the domain of x. Written in terms of $K = m + 1$ B-spline basis functions,

B_k, we have;

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x) \tag{2.34}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown regression coefficients and the $n \times K$ design matrix $\mathbf{X}$ consists of the basis functions evaluated at the observations x_i, i.e. $\mathbf{X}[i,k] = B_k(x_i)$. If a few knots are used, the spline may not capture the variability of the data. For large number of knots, overfitting is a concern in the model fitting process. To get around this problem, Eilers and Marx suggested using moderate number of equally spaced knots, usually between 20 and 40 to ensure flexibility and to define a roughness penalty based on first and second order differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curve (Eilers and Marx, 1996). This leads to penalized likelihood estimation with penalty terms. First and second order random walks are used as priors for the regression coefficients.

$$\lambda \sum_{k=r+1}^{K} (\Delta^r \beta_k)^2, r = 1, 2 \tag{2.35}$$

where λ is the smoothing parameter. First order differences penalize abrupt jumps $\beta_k - \beta_{k-1}$ between successive parameters while second order differences penalize deviations from the linear trend $2\beta_{k-1} - \beta_{k-2}$.

2.8.3 Random walks

One approach of modelling the effects of continuous variables is through random walks. Suppose x is a time scale or a continuous covariate with equally spaced ordered observations such that

$$x^{(1)} < x^{(2)} < \cdots < x^{(K)}$$

where $K \leq n$ denotes the number of different observed values of x. An estimate for one β_k can be made for each $x^{(k)}$, i.e. $f(x^{(x)}) = \beta_k$ and a penalty is imposed for abrupt jumps between successive parameters using random walk priors. The

first and second order random walks are given as follows

$$\beta_k = \beta_{k-1} + \varepsilon_k, \qquad \beta_k = 2\beta_{k-1} - \beta_{k-2} + \varepsilon_k \tag{2.36}$$

with normally distributed errors, i.e. $\varepsilon_k \sim N(0, \tau^2)$, diffuse priors $p(\beta) \propto const$, and $p(\beta_1)$ and $P(\beta_2) \propto const$ for initial values respectively.

These specifications act as smoothness priors that penalize too rough functions f_j. The first order random walk penalizes abrupt jumps $\beta_k - \beta_{k-1}$ between successive states. The second order random walk on the other hand penalizes deviations from the linear trend $2\beta_{k-1} - \beta_{k-2}$. The joint distribution of the regression parameters $\boldsymbol{\beta}$ is computed as the product of conditional densities defined in equations 2.36 above and can also be brought into the more general form as defined by (2.23). The penalty matrix has the form $\mathbf{K} = \mathbf{D^T D}$ where $\mathbf{D}$ is the first or second order difference matrix. The penalty matrix is given by,

$$\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{2.37}$$

2.9 Kriging

Kriging is a geostatistical technique that is used for spatial prediction at unobserved locations (Waller and Gotway, 2004) utilizing the observation values at nearby locations in the process (Shyu et al., 2011). Unlike the CAR and MRF which deal with grided data in parametric and nonparametric setting respectively, kriging is most useful in point level data (Latimer et al., 2006). In the 2010 MIS, data was collected at specific points with the coordinates of sampled households

being collected by GPS. Consequently, point models are most useful in this context.In this section, we look at kriging for geostatistical models and then extend to STAR models. In kriging, the variogram is a widely used statistic to represent the spatial continuity or the roughness of the data. The empirical variogram as defined by (Banergee and Finley, 2009) is,

$$\gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2, \tag{2.38}$$

where $\parallel s_i - s_j \parallel$=t and $|N(t)|$ is the number of points in $N(t)$. Let (x_i, y_i), $1 \leq i \leq n$, where the y_i is a scalar and $x_i \in \mathbb{R}^2$ is a geographical space, then the universal kriging formula is given as below (Kammann and Wand, 2003).

$$y_i = \beta_0 + \beta_1' + S(x_i) + \varepsilon_i, \tag{2.39}$$

where $\{S(x) : x \in \mathbb{R}^2\}$ is a stationary spatial process with mean 0. The error terms ε are assumed to have mean 0 and a common variance σ_ε^2. To predict at an unobserved location $x_0 \in \mathbb{R}^2$, the formula becomes:

$$\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1' + \hat{S}(x_0), \tag{2.40}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ being estimates of β_0 and β_1 respectively in equation (2.39) above. The $\hat{S}(x_0)$ is the empirical best linear predictor of $S(x_0)$. Finally, the equation becomes,

$$\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1' x_0 + \hat{c_0'}(\mathbf{C} + \sigma_\varepsilon^2 \mathbf{I})(\mathbf{y} - \hat{\beta}_0')(\mathbf{y} - \hat{\beta}_0 - \hat{\beta}_1' x_0) \tag{2.41}$$

where

$$\mathbf{C} = (\text{cov}\{\mathbf{S}(\mathbf{x_i}), \mathbf{S}(\mathbf{x_j})\}) \qquad 1 \leq i, j \leq n, \tag{2.42}$$

$$\mathbf{c_0}' = (\text{cov}\{\mathbf{S}(\mathbf{x_0}), \mathbf{S}(\mathbf{x_i})\}) \qquad 1 \leq i \leq n. \tag{2.43}$$

The spatial component is assumed to follow a zero mean Gaussian random field with variance τ^2 and an isotropic correlation function,

$$\text{cov}\{\mathbf{S}(\mathbf{x}), \mathbf{S}(\mathbf{x}')\} = \mathbf{C}_\theta(\parallel x - x' \parallel). \tag{2.44}$$

In equation 2.44, the $\mathbf{C}_\theta$ belongs to the Matern family of covariance functions which are defined as follows:

$$\mathbf{C}_\theta(\mathbf{r}) = \sigma_x^2(1+ \mid r \mid /\rho)exp(- \mid r \mid /\rho). \tag{2.45}$$

In this formulation, ρ controls how fast correlations die out with increasing distance. As a rule, the choice of ρ is such that the scale invariance of the estimates is ensured (Belitz et al., 2009*b*)

$$\hat{\rho} = \max_{i,j} ||x_i - x_j||/c \tag{2.46}$$

The constant $c > 0$ has to be chosen so that $C(c)$ is small. The formulation of the kriging methodology is still faced with the problem of larger datesets (Cressie and Johannesson, 2008). In this study, the sample size of $n = 2094$ made the evaluation of equation 2.41 a complex task. The difficulty comes in due to the computation burden rendered by the $n \times n$ covariance structure. We are faced with the dilemma where we would like to use all the points in the dataset for inference, but at the same time, computation efficiency is required. To get around this problem known as the big N problem, low rank kriging (Royle and Nychka, 1998) is used.

2.9.1 Low rank kriging and space filling algorithm

Space filling designs are sampling plans that optimize a distance based criterion (Royle and Nychka, 1998). These designs do not depend on the covariance struc-

ture of the process to be sampled hence their computation efficiency. Let the points $\{y_1, \ldots, y_k\}$ obtained through the space filling algorithm be a representative sample of points $\{x_1, \ldots, x_n\}$. These points, y_i's, are also known as knots. It means that for our inference purposes, we are going to use only a subset of all the sampled locations.

2.10 Model selection

The need to select a model is of great importance in statistics. The observed data is usually from an unknown probability distribution. As a result, several models are fitted in order to the find the best. Those that are not very close to the actual distribution have to be discarded then. We now take a look at the different statistics that will be used in the model selection procedure.

2.10.1 The Likelihood function

Given a sample realization of $x_1, \ldots, x_n$ from a distribution with the density function $p_i(x)$, the likelihood function is

$$L_n(p_i; x) = \prod_i^n p_i(x_i).$$

It is the joint probability density function of observable random variables and it is viewed as the function of the parameters given the realized random variables. The likelihood function is of great importance as it is widely used by several model selection statistics.

2.10.2 Akaike Information criterion

The Akaike information criterion (AIC) is one statistic used to select the best model.

$$\text{AIC} = -2\text{logL}(\hat{\theta}|y) + 2k$$

where $L(\hat{\theta})$ is the likelihood function and k is the number of estimated parameters. The AIC is calculated for each model under consideration using the same data and the model with the lowest AIC is chosen. The term $2k$ is a penalty to be paid for over fitting and this discourages adding too many variables in the models which always leads to a smaller likelihood. This provides the trade off between over fitting and optimum model fit.

2.10.3 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is another model selection statistic that is based on the empirical log-likelihood and is independent of priors. Due to this fact, the BIC is favoured in situations where the priors are difficult to set. It is related to the AIC and both statistics penalize model complexity. Mathematically, the BIC is,

$$\text{BIC} = -2\ln f(y|\hat{\theta}_k) + k\ln n.$$

The penalty term in the BIC is more strigent than the penalty term of AIC and this leads to BIC favouring smaller models than the AIC.

2.10.4 Deviance Information Criterion

The deviance information criterion (DIC) (Spiegelhalter et al., 2002) is a generalization of the AIC and the BIC and is widely used in model selection where MCMC simulation is used. The DIC only works when the posterior is approxi-

mately distributed as multivariate normal. The deviance is given as;

$$D(y,\theta) = -2\log P(y|\theta) + C \tag{2.47}$$

where $p(y|\theta)$ is the likelihood, θ is the unknown parameter and C is a constant that cancels out in model comparison. The DIC has two components with one measuring the goodness of fit and the other component is a penalty for increasing model complexity. The average deviance,

$$\bar{D} = E[D(\theta)] \tag{2.48}$$

over the true sampling distribution measures how well the model fits the data. The effective number of parameters is the component that measures model complexity and is given by,

$$p_D = \bar{D} - D(\bar{\theta}). \tag{2.49}$$

The DIC is then calculated as,

$$\text{DIC} = p_D + \bar{D}. \tag{2.50}$$

In model selection, the general rule is that models with smaller DIC be preferred over models with a larger DIC. The DIC decreases as the number of parameters in the model increases

Chapter 3

Methodology

3.1 Study area characteristics

Malawi is a small country in Southern Africa bordered by Mozambique, Tanzania and Zambia. As of 2011, the total population was 14 million and over 2 million are aged less than five years (National Statistical Office (NSO), 2008). It is administratively divided into three regions and further into 28 districts. Malawi lies within the tropical regions with two distinct dry and wet seasons. Malaria is endemic to most parts of the country and peaks during the rainy season that falls between November and April. Altitude plays a significant role in the observed differences in risk across the country as highland areas with cooler weather conditions have lower disease risk. Areas such as Nyika and Zomba Plateaus fall in this category. On the other hand, low lying areas along the lake and the lower Shire Valley have higher prevalence of the disease. The country also has areas with potential environmental characteristics that favour malaria transmission.

A large proportion of the Malawi population lives in the rural areas and children from these families are at a higher risk of the disease than their urban counterparts (National Statistical Office (NSO) and ICF Macro, 2011). In general, children and pregnant women are at the highest risk of the disease than the other groups.

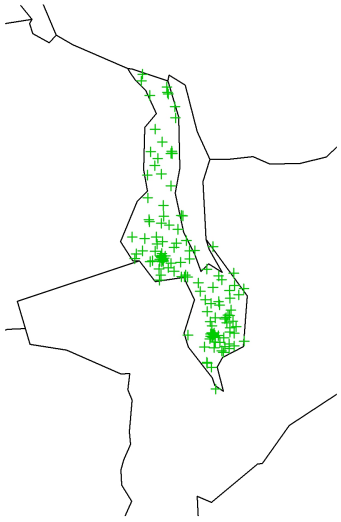


Figure 3.1: Location of enumeration areas

Intervention efforts like ITNs continue to be applied in the country with special emphasis to protect the vulnerable groups. Figure 3.1 shows the locations of sampled households across Malawi. It shows that the survey data was collected at points scattered all over the country hence the need to predict the results at unobserved locations

3.2 Data sources and characteristics

3.2.1 Data collection

The MIS took place between March and April 2010. The survey was nationally representative and a two stage cluster sampling approach was used. 3,500 households were surveyed from 140 standard enumeration areas (SEAs) randomly selected from all the districts in the country. The malaria status of the child between 0 and 5 years was determined by rapid diagnostic tests (RDT). Among other pieces of information collected were knowledge of malaria by the parents,

information on ITN use and wealth status of the household. Coordinates of each visited household were also collected by the data collection team by means of hand held GPS receivers thus yielding point referenced data. Due to the nature of the survey, both NSO and Ministry of Health were involved.

3.2.2 Data management

Child and women data files were first merged by matching their household numbers which reduced the number of observations from 3,500 to 2741. Children with no matching test result were removed from analysis causing the sample to drop to 2094. In most of the households, only a single child who met the requirements was tested. However, some households had more than one child tested for malaria. All children were included in the analysis, with a common spatial random effect at EA level. Data cleaning was carried out mainly in MS Excel to remove duplicates and missing data. The data was then exported to Bayes X and R for further analysis.

One aspect of the data cleaning process was to reduce the number of variables in the merged dataset. Variables such as age, wealth index, gender among others were used in the analysis. Some of the key variables used are presented in table 3.1.

3.2.3 Climatic data

The climatic data used in the analysis was obtained from the Department of Meteorological Services and Climate Change. Specifically, the data was obtained from the network of weather stations that the department has across the country. However, the challenge was that the data was from weather stations that did not match the exact location of the data points. In other words, there was a mismatch of observations. This was solved by using the interpolation method known as the nearest neighbour method. Interpolation is a method of approximating the value

of a non given point in space. The approximating algorithm considers the value of the nearest point in order to calculate the missing value. The network of weather stations provided the known points for the interpolation and the distance was given by the coordinates. It was thus possible to calculate the approximate rainfall, temperature and humidity of each household.

The advantage of using this method is that we have relatively reliable estimates as it matches a data point to the nearest weather station thereby minimizing errors. The method assumes linear interpolation between locations but non-linear variations between places exist. Despite the presence of some errors, the calculated estimates are quite reliable. Moreover, this methods uses actual readings recorded by the MET department unlike other methods of obtaining point data through simulation. Three month average (January-March) in order to assess how climatic factors during the peak of the rainy season in Malawi affect malaria risk in children.

3.2.4 Low rank kriging

Figure 3.1 illustrates how the data points are scattered across the country. The parametric geostatististical models use all the points for kriging leading to a computation burden and slow convergence. Thus low rank kriging is used in this analysis through the REML procedure to predict malaria risk at unsampled locations. Only a representative subset of these points is used for the kriging procedure.

3.3 Data analysis

The data is analyzed in two major software packages, Bayes X (Belitz et al., 2009*a*) and R (R Development Core Team, 2011). Bayes X was used to formulate and estimate the STAR models and for the prediction. However, further handling of the results was done in the R package Bayes X (Kneib et al., 2011). This is due to the fact that the stand alone Bayes X has limited graphical capabilities.

The epicalc package in R (Chongsuvivatwong, 2011) is also used primarily for the production of aggregate plots in exploratory analysis.

3.3.1 Description of key variables

Table 3.1 belows shows some of the variables used in the modeling.

Table 3.1: Description of key variables

Covariate	Description
Age	Age of the child in years
Age category	Age category of child (categorical)
Altitude	Height above sea level measured in m
Wealth Index	Index showing the well being of the household (1=poorest, 2=poorer, 3=medium, 4=richer, 5=richest)
ITN	Variable showing if child slept in treated net (0=did not, 1: slept in net)
Latitude	The location of the sampled household in degrees
Longitude	Location of sampled household given in degrees
District	District where child was tested for malaria
EA	Enumeration area where households are located
Rainfall	Three month average rainfall in mm (i.e. Jan to March)
Min. temp	Three month average minimum temperature in °C
Humidity	Three month average humidity

3.3.2 Model specification

Let Y_{ij} be the malaria status of a child j at household i. Then Y_{ij} follows a Bernoulli distribution. That is

$$Y_{ij} \sim Bernoulli(p_{ij}).$$

The first step was to find the variables that were significantly associated with malaria risk. Bivariate tests were also carried out to identify these variables which were later put into the spatial models for further analysis. The following additive

logistic regression models are fitted.

$$A1 : \eta_i = \beta_i'\gamma + \sum_{k=1}^{q} x_{ik}$$
$$A2 : \eta_i = \beta_i'\gamma + \sum_{k=1}^{q} f_k(x_{ik})$$
$$A3 : \eta_i = \beta_i'\gamma + \sum_{k=1}^{q} x_{ik} + \Phi(S_i)$$
$$A4 : \eta_i = \beta_i'\gamma + \sum_{k=1}^{q} f_k(x_{ik}) + \Phi(S_i)$$

where η_i is the predictor, β' is a vector of regression coefficients, γ is a vector of categorical variables such as wealth index and x_{ik} are the continuous covariates. The first model, A1, fitted all variables including the climatic ones as fixed linear effects without any random effects. Secondly, in model A2, all the q continuous covariates including the climatic variables were fitted as non linear terms in order to assess the need of having these non linear terms in the model.

The second step in the model building process involved the inclusion of random effects of district and enumeration area. A3 was fitted as a linear effects model with random effects. Finally, model A4 fitted both the climatic and non climatic metric covariates as non-linear terms in addition to the random effects. All the four models were fitted in a full Bayesian framework.

For the spatial effect, a two dimensional P-spline prior was assumed. In order to fit the non-linear effects, random walk priors of order two were employed. Trace plots and autocorrelation plots were used for monitoring convergence. For all the models, 10,000 iterations were run with a burn in sample of 1000 where the first 1000 iterations were discarded and a thinning parameter of 50. The thinning parameter gives the sampling interval so that every 50th draw is stored and used for calculating parameters. Model selection using the DIC was then carried out at the end to choose the best model which could be used for inference. Model with the lowest DIC was chosen as the best model.

3.3.3 Priors for the fully Bayesian models

Two types of models are fitted, fully Bayesian and empirical Bayesian. Under the full Bayesian approach, prior distributions were assigned to all model parameters. The priors used in the modelling have the general form of equation 2.23. Specifically this approach uses inverse gamma priors with parameters $a = 0.001$ and $b = 0.001$.

Chapter 4

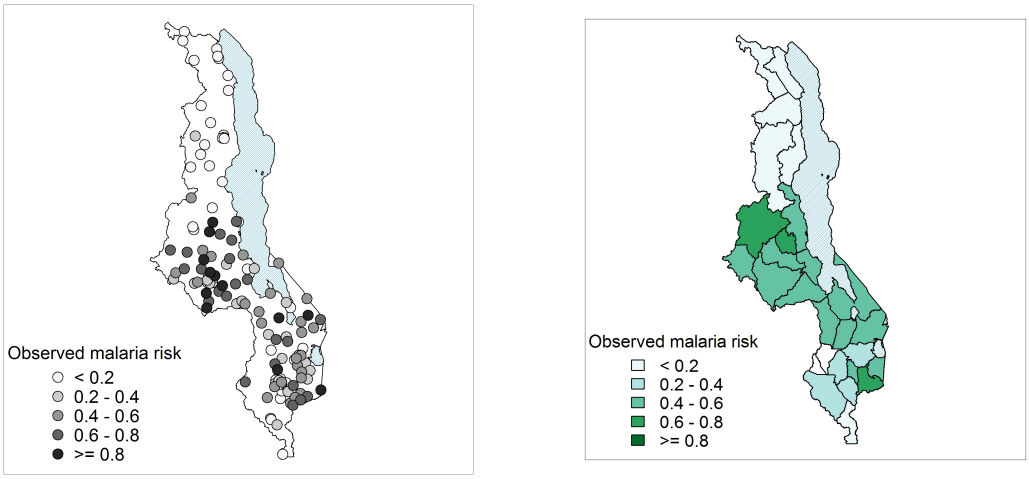
Results and discussions

In this chapter, we present the results of the analysis of the MIS dataset. We first present the exploratory data analysis results before moving on to make further inferences based on the statistical models built.

4.1 Exploratory data analysis

The MIS was done across the country at EA and down to the household level. The maps below show the differences in observed risks across the country.

Figure 4.1a shows that the northern part of Malawi has the lowest risk based on the selected sample. The central and southern regions have higher malaria risk than the northern part. In particular, central region had children at greater risk of malaria with risk mainly over 0.6. These are just observed risks and there are likely to be underlying reasons for these observed disparities across the country. Figure 4.1b generalizes the observed risk to the district level which again shows lower risks mainly in the northern region.



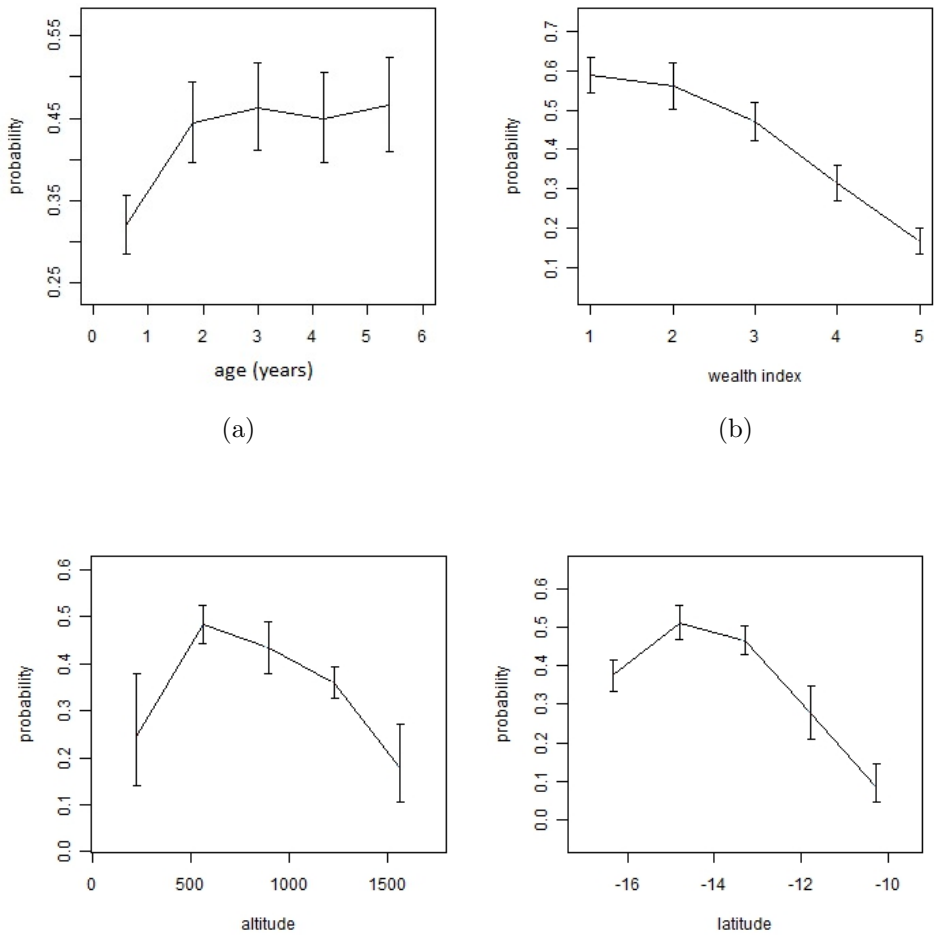
(a) (b)

Figure 4.1: Observed parasitaemia risk (a) EA level (b) district level

.

4.2 Differences in malaria risk

Figure 4.2 below shows how malaria risk varies with changes in different covariates at household level. It is evident that there is a general decline in the probability of a child developing malaria as the wealth status of the house improves. Children from these households are more resistant to the disease than their counterparts from poorer families. It is also observed that the risk of malaria generally increases as the age of the child increases. The youngest tested children in the survey, those less than 12 months show the lowest probability of contracting malaria. These probabilities were calculated by epicalc package by diving number of children with malaria parasite with the number tested. The risk however steadily increases as the age is increasing. Altitude also seems to play a key role in the possible distribution of the disease. It can be seen from figure 4.2c that the risk drops as the altitude increases from about 500m. The highest malaria risk was observed in medium lying areas at around 500m and the lowest at around 1500m. These highland areas are associated with lower temperatures that may not be very conducive for malaria development. On the other hand, the low lying areas with observed low



(a) (b)

(c) (d)

Figure 4.2: Aggregate plots: (a) malaria against age, (b) malaria by wealth, (c) malaria by altitude and (d) malaria by latitude

.

probabilities of the disease could be because of extremely high temperatures typical of these areas. It is known that the biology and ecology of malaria vectors is heavily dependent on prevailing factors such as precipitation, humidity and temperature (Githeko et al., 2000). For instance, mosquitoes are very active at the temperature range of between 22°C and 30°C (Gemperli, 2003). Lower temperatures slow down the life cycle development and temperatures above 34°C generally have a negative impact on the survival of parasites.

Table 4.1: Association between parasitaemia risk and selected variables

		Malaria		
Selected variables	# of children	Yes (%)	No (%)	p-value
Age (years)				<0.001
0	268	68(25.4)	200(74.6)	
1	414	150(36.2)	264(63.8)	
2	418	186(44.5)	232(55.5)	
3	353	164(46.5)	189(53.5)	
4	331	149(45.0)	182(55.0)	
5	306	144(47.1)	162(52.9)	
Location				<0.001
Urban	555	91(16.4)	464(83.6)	
Rural	1538	770(50.1)	768(49.9)	
ITN				0.002
Yes	1502	586(39.0)	916(61)	
No	591	275(46.5)	316(53.5)	
Sex				0.649
Male	1045	435(41.6)	610(58.4)	
Female	1048	426(40.6)	622(59.4)	
Wealth index				<0.001
Poorest	503	297(59.0)	206(41.0)	
Poorer	286	161(56.3)	125(43.7)	
Medium	423	125(43.7)	224(53.0)	
Richer	395	124(31.4)	271(68.6)	
Richest	486	80(16.5)	406(83.5)	

4.2.1 Association between malaria and covariates

A Pearson Chi Square test showed that the age of a child is strongly associated with the risk of the disease ($p<0.001$) which is in agreement with the aggregate plot above. Similarly, a strong association exists between malaria risk and the wealth status of the household ($p<0.001$). The area of residence, whether rural or urban has a significant association with malaria with rural children more at risk than their urban counterparts ($p<0.001$). Gender is however not linked to the disease. This shows that all children are equally likely to have malaria bouts regardless of the sex. Lastly, there is an observed association between ITN use and malaria ($p = 0.002$) which needs to further be assessed to quantify its effect as a malaria prevention strategy. These observed associations will be investigated in relation with other variables later on under statistical models.

4.3 Full Bayesian analysis results

In this section, we look at the different models built under a full Bayesian approach and choose the best to be used in subsequent analysis. The chosen model will be used to answer the first three objectives which are concerned with using the models for answering questions about malaria.

4.3.1 Model choice

In table 4.2 below, the four models fitted under the full Bayesian approach are compared in terms of their DIC.

Table 4.2: DIC of fully Bayesian models

	A1	A2	A3	A4
Deviance	2162.31	1948.90	1726.42	1738.77
P_D	15.46	38.48	100.45	77.79
DIC	2193.23	2025.85	1927.36	1894.35

The four different models were fitted in order to find the most parsimonious. The model which fitted climatic covariates as non linear effects as well as incorporating random effects of district and enumeration area (i.e. A4) was the best fitting model (DIC=1894.35). The model with these components is chosen as the best model since it has the smallest DIC among the four models. This model however competes very well with linear model but with random effects (A3)(DIC=1927.33). On the other hand, A2 which fitted the continuous covariates as non linear functions without random effects has DIC=2025.85.

These results show that fitting the continuous covariates as non linear functions is plausible and leads to better model. Consequently, model A4 is used for parameter estimation. Furthermore, this model fitting reveals there is an effect of location district and the enumeration area that are captured by the random effects.

Table 4.3: Posterior estimates for model with non linear climatic effects and random effects

Explanatory variable	Non linear effects model (A4)		
	OR	2.5% Quantile	97.5% Quantile
Intercept	0.628	0.207	2.86
Location			
Urban	0.218	0.130	0.397
Rural	1	1	1
Region			
South	1	1	1
Centre	1.37	0.564	5.59
North	1.01	0.0752	5.58
Age category (yrs)			
0-1	0.244	0.196	0.281
1-2	0.452	0.357	0.575
2-3	0.717	0.667	0.818
3-4	1.110	0.955	1.420
4-5	1	1	1
Wealth index			
Poorest	2.07	1.72	2.78
Poorer	4.25	0.093	3118
Medium	0.476	0.00088	25.5
Richer	1.12	0.473	0.783
Richest	1	1	1
Interventions			
Bed nets	0.628	0.473	0.783

4.3.2 Risk factors for malaria

Table 4.3 shows that age of a child is very much associated with malaria and the risk increases with age. For instance, the odds of infection for the age group 0-1 years are lower than the 4-5 year group (adjusted odds ratio[OR]=0.244, BCI: (0.196,0.281)). In other words, younger children aged between 0 and 1 year are less likely to be infected with the disease. Those children aged between 1 and 2 years have slightly increased odds than the children aged less than 1 year but still much less than those aged 4-5 (adjusted OR=0.452, BCI:(0.357,0.575)). Following the same trend, the age group 2-3 have lower odds than children aged between 4 and 5 (adjusted OR=0.717: BCI(0.667,0.818)). It is very clear from these results than there is indeed an increased risk of malaria as the child's age increases. This could

be explained by a drop in the inherited immunity from the mother as the child grows up. During this period, other supplementary foods are typically introduced to the child. These results corroborates the observation seen in the exploratory analysis where aggregate plots showed increasing risk with age. (See Figure 4.2a).

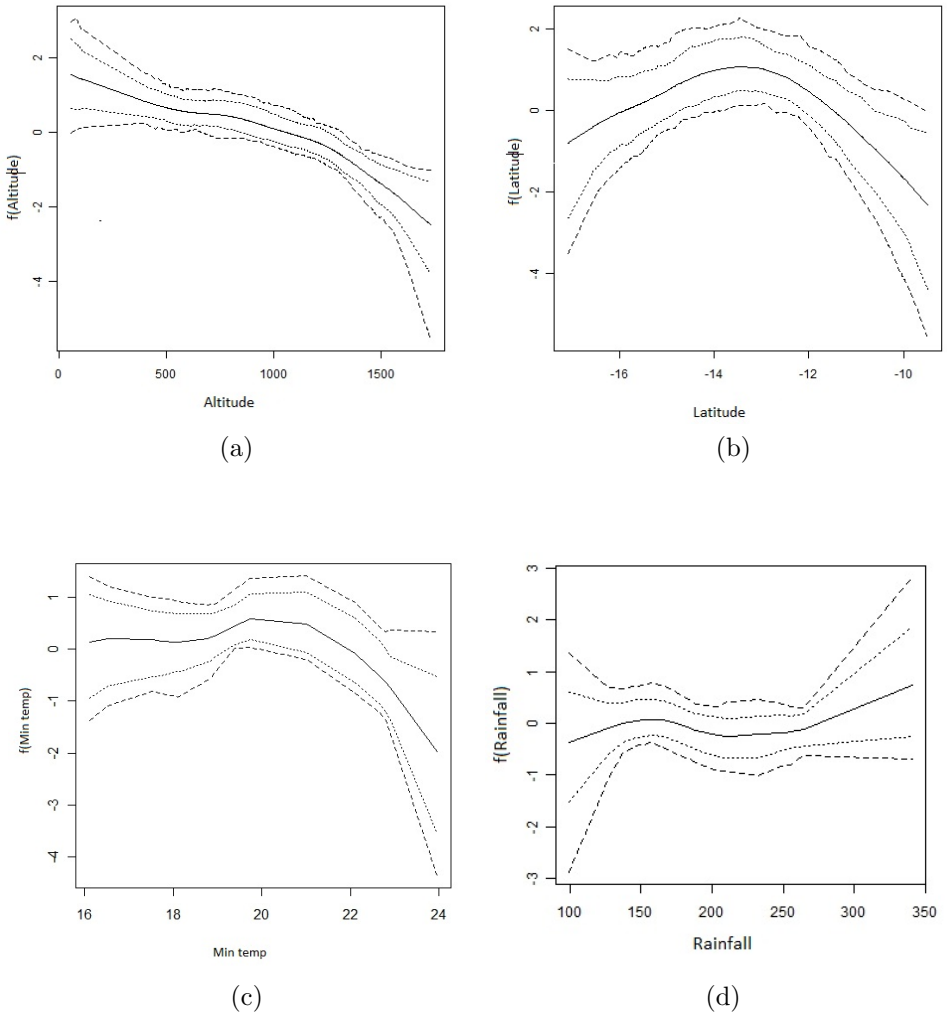
On a different note, area of residence also has an effect on the household malaria risk with the odds of contracting malaria for urban children much lower than for rural children (adjusted OR=0.218, BCI:(0.130,0.397)). There is also an observed non-linear relationship between malaria risk and the wealth status of a household. The odds of malaria infection drops as the wealth status goes up. From the fitted climate and random effects model, children from poorest households have twice the risk of malaria infection than the children from the richest household (adjusted OR=2.07: BCI(1.72,2.78)). For the fourth quintile, there is lack of association with the risk (adjusted OR=1.12, BCI:(0.473,0.783)). The strong association dies down to almost non existence as shown in Figure 4.2b.

The region also has some effect particularly the central region which shows slightly higher odds of malaria infection than the south (adjusted OR=1.37,BCI:(0.56,5.59)). The use of bed nets as an intervention yielded positive results as children sleeping under an ITN are less likely to have malaria attack (adjusted OR=0.628, BCI:(0.473,0.783)). This is in agreement with a recent study which found a significant reduction in asexual parasitaemia in under five children who slept under an ITN (Skarbinski et al., 2011).

4.4 Non linear effects of continuous covariates

This section looks at the non-linear effects of continuous variables encountered in the model.The figures plots the non-linear function of the continuous covariates on the log odds scale against the covariate values of altitude, latitude, minimum temperature and rainfall. In addition, the figures show the 80% and 95% confi-

dence intervals of the posterior estimates. From figure 4.3a below, it can be seen



(a) (b)

(c) (d)

Figure 4.3: Non linear effects of continuous covariates: (a) altitude (b) latitude (c) minimum temperature (d) rainfall

that malaria risk again steadily drops as the altitude increases. Households in low lying areas especially those between 0-500m are seen to be highly at risk than their counterparts in upland areas. These high risk areas in Malawi are likely to be the low altitude areas along the lake shore and the lower shire which are generally below 800m above sea level. These areas are about 8 times likely to have their young ones contract malaria as households in upland areas (OR:exp [2]=7.84). The figure shows very low risk for areas above 1500 m to be as low as 0.0183. Areas in the country with such observed low risks mainly include the highland areas in the north such as Nyika Plateau, Dedza in the Central areas

among others (Kazembe, 2007).

4.4.1 Effect of latitude

The latitude also shows an effect on malaria with higher risk being observed in the mid lying latitudes. Malawi is long and narrow in shape and generally falls between $-10°$S in the North and around $-17°$S in the Southern Region. The observed high risk mid latitudes corresponds to areas in the Central Region. The high risk here could be due to several reasons one of which is the elevation. The two biggest plains in Malawi, Lilongwe and Kasungu are found in this region thus there are fewer mountainous areas compared with the other two regions. The central areas have about twice the risk as northern most regions (exp[0.5]=1.65). Although the south itself has low areas in the Lower Shire, temperatures there are usually very high which may hinder larval development and consequently the transmission. This perceived high risk area in Nsanje and Chikwawa is also home to a number of NGOs that are involved in the distribution of ITN especially targeting children and pregnant mothers. This may have an overall reduction in the risk for the south by neutralizing the Lower Shire effect to an extent. The lower populations in the north may also be to the advantage of malaria intervention initiatives. A risk map produced by (Kazembe et al., 2006) shows large portions of the Central Region with high prevalences ranging from 0.739 to 0.944.

4.4.2 Effects of rainfall and temperature

The peak climatic variables during the period January to March show varying season effects on the risk of the disease. For minimum temperature, there is an observed drop as the minimum temperatures increase towards mid twenties. There is an observed lack of association between lower seasonal temperatures (i.e. from 16°C to 21°C) and malaria risk. It has been observed that malaria parasites are very inactive around these temperatures which could help explain this lack

of a association (Sachs and Malaney, 2002). The highest risk is observed 19°C and 21°C (OR=exp[0.1]=1.11) and drops to around (OR=exp[-1.8]=0.165). As temperatures start rising above 21°C, there is an observed association between minimum temperature and malaria risk (OR=0.135). Minimum temperatures during the peak rainfall period have a generally weak association with malaria risk. The odds ratios are approximated from the nonlinear effects in figure 4.3.

On the other hand, the non linear effect of rainfall is also not very significant with the odds of infection staying relatively constant ($OR \approx 1.00$) as the average rainfall in the three months preceding the survey increases. This clearly shows lack of association between the two variables. However, there is an increase in risk as average rainfall exceeds 260mm. (OR increases from 1.00 at 230mm to about 2.2 at 350mm). This doubling in risk may have been brought about by widespread availability of conducive breeding sites for the malaria vectors. Even though this is the case, the relatively wide Bayesian credible interval (BCI) as rainfall amount increases shows that there is a lot of uncertainty in the calculated estimate. Comparatively, there is a narrow BCI associated with the less rainfall amounts which suggests that the observed lack of association between rainfall and malaria risk shown by the odds ratio close to the value1 ($OR \approx 1.00$) is quite valid. This lack of association has been found before by Kazembe et al among others (Kazembe et al., 2006; Gosoniu et al., 2006). However, the climatic variables are found to be not significant after accounting for altitude and latitude which are associated with rainfall and temperature.

4.5 Sensitivity analysis in fully Bayesian models

In Bayesian analysis, the estimates may be influenced by several settings whilst running the MCMC algorithm. This necessitates that model validation be carried out on the chosen model by changing a few parameters and then comparing the observed changes in the estimates. In this thesis, we are going to achieve this by

changing the priors and other parameters in our models and then compare the fit. In the fitted models, we used the default inverse gamma priors with both parameters equal to 0.001, i.e.

$$p(\theta) \sim \text{IG}(a = 0.001, b = 0.001)$$

For the chosen model, we fit three models by changing the prior distributions where model M1 has priors (a=0.00001,b=0.00001), M2 has priors (a=0.0005,b=0.0005), M3 has (a=1,b=0.005) and lastly M4 has the default priors (a=0.001,b=0.001). The predicted means of the three models are plotted in the boxplot below; The

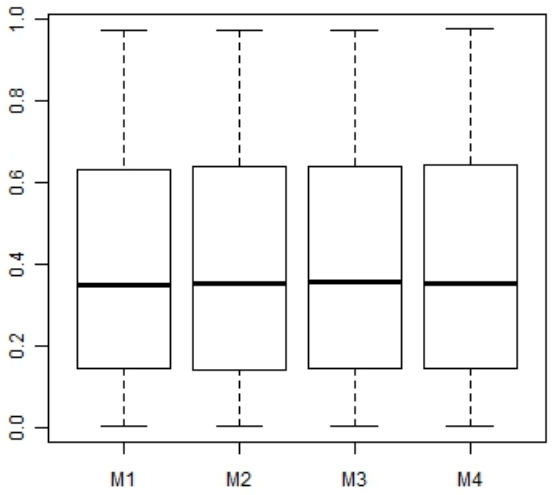


Figure 4.4: Box plot showing distribution of predicted means using the four models

predicted means are almost identical which shows that the model is not affected by changes to the priors. This observation gives confidence in the model performance. This observation is supported by the almost similar DIC which shows that the model's performance is not affected by changes in the prior distributions. Changing the prior distribution does not seem to affective the model.

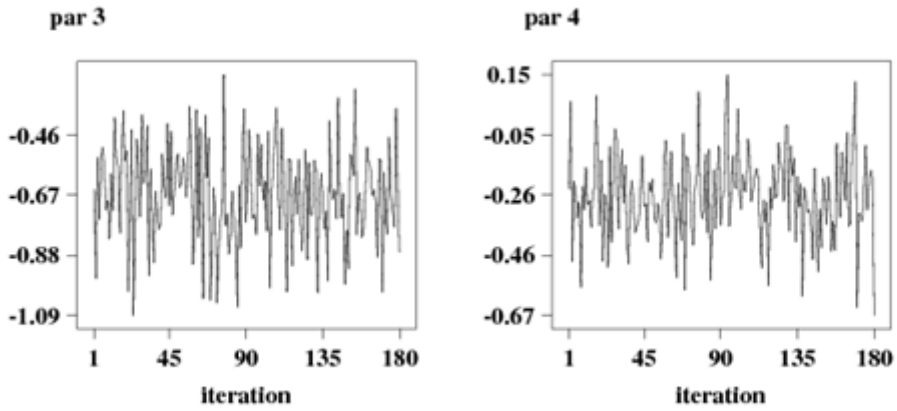
Table 4.4: Table showing comparative predictive power given different priors

	M1	M2	M3	M4
Deviance	1727.73	1727.60	1732.53	1727.98
p_D	95.25	93.65	92.47	93.06
DIC	1918.23	1914.89	1917.45	1914.15

4.6 Model diagnostics

In this section, we take a further look at the model performance by looking at how it converges. The trace below is of two parameters in the model. The results shows a converging Markov Chain as the number of iterations increases. It mixes quite well and jumps almost between two bounds. More trace plots are presented in the appendix.

Figure 4.5: Trace plots of two parameters in the model



4.7 Empirical Bayesian analysis

This section is primarily concerned with presenting kriging results. Due to the large number of observation points, the space filling algorithm is used where only a representative sample of the points (knots) is sampled for inference. In the results that follow, we present results of the same model but using different numbers of knots. Later on, the kriging results will be presented and discussed.

The prediction surface below presents the results of the empirical Bayesian analysis. Figure 4.6 is a surface interaction plot of the various covariates. This plot

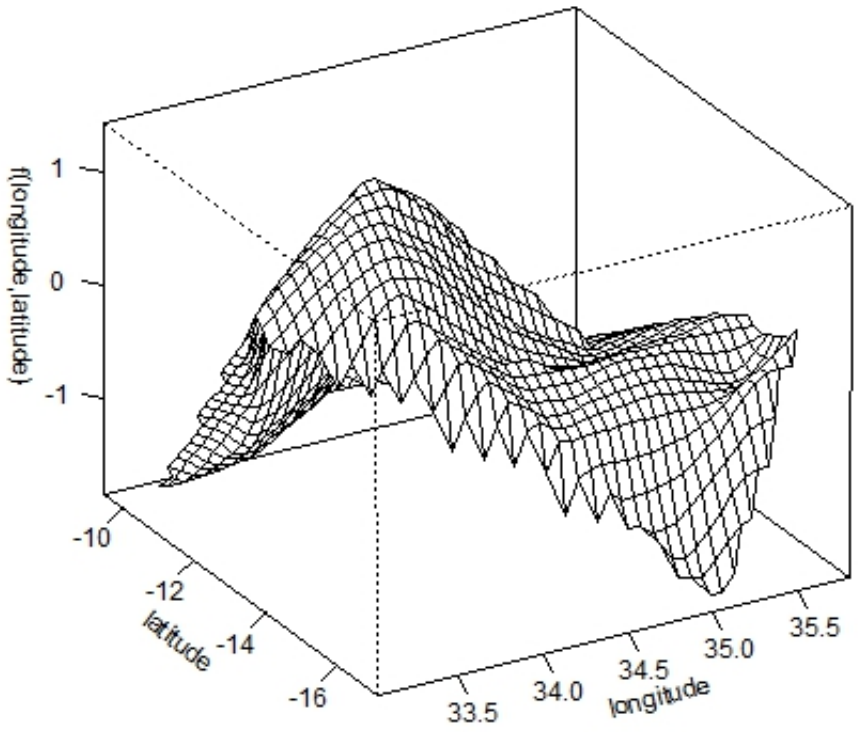


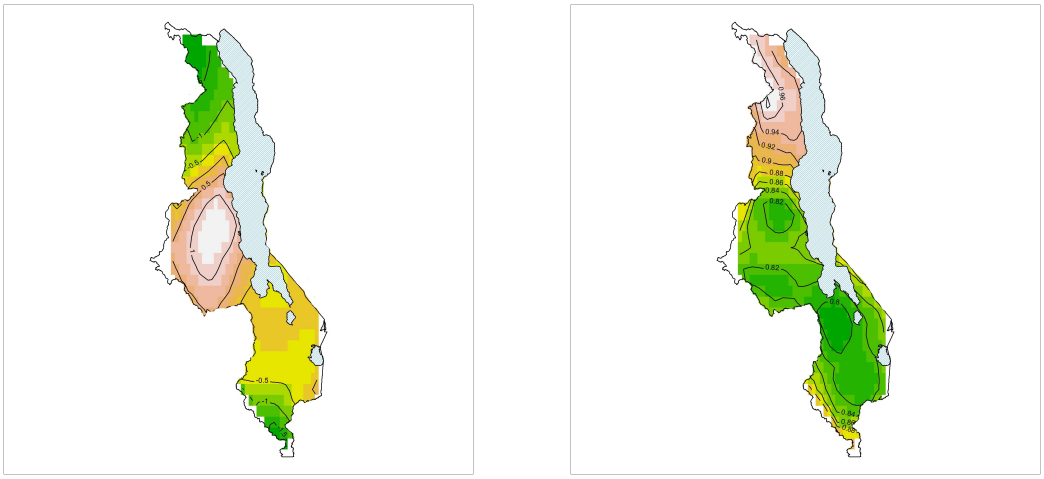
Figure 4.6: Predictive surface of under five malaria risk in Malawi

represents risk factors not directly observed but having an impact on the risk of malaria in children aged less than five. The mid latitude areas which corresponds to Central region areas have highest risk as shown by the peak. The north registers the lowest risk at the low latitudes between $-10°$ and $-12°$. The perspective plot also reveals that malaria risk stays relatively constant across the width of the country which is represented by the longitudes. This can be explained by the limited range of longitude for a narrow country like Malawi (Kazembe, 2007). There is less spatial variation in disease risk across the breadth of the country.

Figure 4.7a shows the kriging surface with the map of Malawi overlaid. The central

areas in particular are more likely to have the highest risk. This is in agreement with the observed risk map which showed the central areas to be at the highest risk. Similarly, the lowest malaria risks are predicted in the northern region. This is in agreement with a recent study using point referenced malaria prevalence data that found high and low risks in the centre and north respectively (Kazembe, 2007). In particular, these northern areas with low risk are higher altitude areas such as Mzimba and Rumphi. In the centre, Lilongwe, Kasungu, Salima have particularly higher risks as shown on the map. The south also has high malaria risk in areas such as Phalombe plain and Mangochi and surrounding areas. One interesting finding is that lower risks are predicted in the the Lower Shire. Higher temperatures which are known to reduce transmission are very prevalent in this area and that could potentially reduce the malaria prevalence. Mass distribution of ITNs as a preventive measure with special emphasis on usage by children and pregnant mothers is another reason that could explain the apparent lower risks. The under sampling in this region coupled by total lack of sampling in Neno and Mwanza may also help explain this observation. In this analysis, we used smoothing approach and as a result, kriging parameters of sill, range and nugget parameters are not explicitly estimated.

In order to show the precision, a map of standard errors is plotted. The errors in this map are relatively large ranging from 0.8 to around 0.96. Even though this is the case, this narrow variation points to a quite precise prediction map. In figure 4.7b, there are higher errors in the northern part of Malawi shown by warmer colours as compared to the central and southern areas. This observation also coincides with the sampling density in these areas. The north had the lowest sampling density in the 2010 MIS. The general observation is that areas closest to the sampled points have lower standard errors than those areas that are far as shown by the contours. Areas around Lilongwe and Blantyre fall within the same standard error owing to the high sampling density in these two areas.



(a) (b)

Figure 4.7: (a) Map showing predicted risk based on the posterior median of the prediction model (b) Map showing the prediction standard errors

4.7.1 Sensitivity analysis in empirical Bayesian models

We fitted the same models three times by increasing the number of knots used in the space filling algorithm. We started with a model with 300 knots then moved to 400 before finally settling at 500 knots. The AIC and BIC for the three models used for model selection are shown in the table below. From Table 4.5, it can be

Number of knots	AIC	BIC
300	2107.95	2661.41
400	2108.75	2664.43
500	2108.15	2664.43

Table 4.5: AIC and BIC of three different models

seen that all three models are very similar in their predictive power. However, the model with 300 knots which represents 20.5% of the observation points is marginally better than the other two and was consequently used to produce the prediction surface. The parameter estimates realised from the models are also almost identical.

Chapter 5

Conclusions and recommendation

5.1 Conclusions

The study has shown that geostatistical data such as the MIS are usually spatially correlated and requires adaptation of the usual GLMs to take into consideration the inherent correlation in the data. Environmental, topographical and climatic variables are usually associated with malaria in the malaria endemic zones including Malawi. Accounting for these variables in the model leads to more accurate estimates which may not be the case when spatial autocorrelation is omitted. In such a scenario, parameter estimates tend to be overstated.

In this thesis, we used structured additive regression models that extend the GLM by modeling the nonlinear effect of the continuous covariates. This approach has helped reveal some complex relationships between the response variable and the continuous covariates that may be missed in the GLM which usually assumes a linear association between the two.

In particular, peak seasonal climate variables of rainfall and minimum temperatures were shown to be not very significantly associated with malaria risk.

Over and above these observations, the empirical risk map in figure 4.7a can be

used in intervention activities by identifying areas that are likely to have higher risks. Since the map is based on country representative survey, the maps produced in this thesis are more credible than what has been done in the past and can be trusted for use in control initiatives. These results, coupled with expert opinion which is widely utilized in the absence of empirically produced maps can lead to the best understanding of the spatial distribution of malaria and hence better approaches in the fight against the disease in young children.

The 2010 MIS will also act as a baseline upon which subsequent surveys will be built. This is crucial in that it is possible to monitor trends in malaria risk among children as well as exploring newer and complex relationships between parasitaemia risk and environmental, climatic and socio-economic factors among others. At the same time, this kind of analysis makes possible the evaluation of different interventions so that they can be improved upon in the subsequent surveys.

5.2 Recommendations

The continued efforts to fight malaria in children should place focus on the Central Regions plains that show high malaria risk. Traditionally, emphasis has been placed on the Lower Shire and areas along the Lake Shore with similar climatic and environmental attributes. These areas, though having high risk should not take focus away from the central plains. Empirically obtained findings from representative surveys together with the predicted risk maps should be used in the planning of malaria interventions. In this way, the allocation of scarce resources can be effectively planned well in advance. There is need to make full use of the risk maps and cultivate the spirit of attaching more importance to survey findings which currently is not widely done.

There was also under sampling in many of the enumeration areas and this can

be improved in the next MIS. Being a nationally representative survey, the MIS should be very comprehensive resulting in more data that can yield more reliable results thereby contributing to effective malaria control interventions. The observed conflict with expert opinion can be partly attributed to this under sampling thus rendering the prediction not very effective in some parts of the country.

For a long time, the analysis of epidemiological data in Malawi has not utilized spatial statistical methods mainly due to the unavailability of geographically referenced datasets. We recommend that in the national surveys geographical coordinates of the sampled locations be collected. This will promote the area of spatial epidemiology that is gaining widespread use elsewhere but has not been fully exploited in Malawi. The fight against malaria in children needs targeted interventions that can be achieved by having this kind of data.

References

Agresti, A. (2006), *Categorical Data Analysis*, second edn, John Wiley.

Banergee, S. and Finley, A. O. (2009), 'Introduction to spatial data and models', `http://blue.for.msu.edu/JBC_10/SC/slides/CourseSlides.pdf`.

Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009*a*), *BayesX - Software for Bayesian inference in structured additive regression models, Version 2.0.1.*
URL: *http://www.stat.uni-muenchen.de/ bayes2.0.1*

Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009*b*), *BayesX - Software for Bayesian inference in structured additive regression models, Version 2.0.1: Methodology manual.*

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

Breslow, N. E. and Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**(421), 9–25.
URL: *http://www.jstor.org/stable/2290687*

Chongsuvivatwong, V. (2011), *Epicalc: Epidemiological Calculator.* R package version 2.14.0.0.
URL: *http://CRAN.R-project.org/package=epicalc*

Clements, A. C. A., Moyeed, R. and Brooker, S. (2006), 'Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa',

Parasitology **133**(Pt 6), 711–719.

URL: *http://journals.cambridge.org/abstract_S0031182006001181*

Cressie, N. and Johannesson, G. (2008), 'Fixed rank kriging for very large spatial data sets', *Journal of the Royal Statistical Society Series B Statistical Methodology* **70**(1), 209–226.

URL: *http://www3.interscience.wiley.com/journal/119418562/abstract*

Dalgaard, P. (2006), 'Generalized Linear Mixed Models: Mixed Models in R', `http://staff.pubhealth.ku.dk/~pd/mixed-jan.2006/glmm.pdf`.

Diggle, P. J. and Ribeiro, P. J. (2007), *Model Based geostatistics*, Springer, New York.

Diggle, P., Moyeed, R. A. and Tawn, J. A. (1998), 'Model-based geostatistics', *Applied Statistics* **47**, 299–350.

Djinjalamala, F. (2006), Epidemiology of malaria in Malawi, *in* E. Guebbels and C. Bowie, eds, 'The Epidemiology of Malawi', chapter 3.

Dobson, A. (2002), *An Introduction to Generalized Linear Models*, second edn, Chapman & Hall/CRC.

Eilers, P. H. C. and Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical Science* **11**(2), 89–121.

Fahrmeir, L., Kneib, T. and Lang, S. (2004), 'Penalized additive regression for space-time data: a Bayesian perspective', *Statistica Sinica* **14**, 731–761.

Faraway, J. J. (2006), *Extending the Linear Model with R:Generalized Linear,Mixed Effects and Nonparametric Regression Models*, Taylor and Francis.

Gemperli, A. (2003), Development of Spatial Statistical Methods for Modelling Point-Referenced Spatial Data in Malaria Epidemiology, PhD thesis, University of Basel.

Githeko, A. K., Lindsay, S. W., Confalonieri, U. E. and Patz, J. A. (2000), 'Climate

change and vector borne diseases: A regional analysis', *Bulletin of the World Health Organization* **78**(9), 1136–1147.

Gosoniu, L., Veta, A. M. and Vounatsou, P. (2010), 'Bayesian geostatistical modeling of malaria indicator survey data in Angola', *PLoS ONE* **5**(3), 9.

Gosoniu, L., Vounatsou, P., Sogoba, N. and Smith, T. (2006), 'Bayesian modelling of geostatistical malaria risk data', *Geospatial Health* **1**(1), 127–139.

Guerra, C. A., Gikandi, P. W., Tatem, A. J., Noor, A. M., Smith, D. L., Hay, S. I. and Snow, R. W. (2008), 'The limits and intensity of Plasmodium falciparum transmission: Implications for malaria control and elimination worldwide', *PLoS Medicine* **5**(2).

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

Hedeker, D. and Gibbons, R. D. (2006), *Longitudinal Data Analysis*, John Wiley & Sons.

Jiming, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

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

Kandala, N.-B., Ji, C., Stallard, N., Stranges, S. and Cappuccio, F. P. (2007), 'Spatial analysis of risk factors for childhood morbidity in Nigeria', *The American Journal of Tropical Medicine and Hygiene* **77**(4), 770–779.

Kandala, N.-B., Madungu, T. P., Emina, J. B., Nzita, K. P. and Cappuccio, F. P. (2011), 'Malnutrition among children under the age of five in the Democratic Republic of Congo (DRC): does geographic location matter?', *BMC Public Health* **11**(1), 261.

Kazembe, L. N. (2007), 'Spatial modelling and risk factors of malaria incidence in Northern Malawi', *Acta Tropica* **102**(2), 126–137.

Kazembe, L. N., Kleinschmidt, I., Holtz, T. H. and Sharp, B. L. (2006), 'Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data', *International Journal of Health Geographics* **5**(1), 41.

Kelsall, J. E. and Diggle, P. J. (1998), 'Spatial variation in risk of disease: A non-parametric binary regression approach', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **47**(4), 559–573.
URL: *http://www.jstor.org/stable/2986082*

Kim, J., Lawson, A. B., Suzanne and Aelion, M. C. (2010), 'Bayesian spatial modeling of disease risk in relation to multivariate environmental risk fields', *Statistics in Medicine* **29**(1), 142–157.

Kneib, T., Heinzl, F., Brezger, A. and Bove, D. S. (2011), *BayesX: R Utilities Accompanying the Software Package BayesX*. R package version 0.2-5.
URL: *http://CRAN.R-project.org/package=BayesX*

Lang, S. and Fahrmeir, L. (2001), 'Bayesian generalized additive mixed models. a simulation study', *Sonderforschungsbereich* **386**(230).
URL: *http://epub.ub.uni-muenchen.de/*

Latimer, A. M., Wu, S., Gelfand, A. E. and Silander, J. A. (2006), 'Building statistical models to analyze species distributions', *Ecological Applications* **16**(1), 33–50.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/16705959*

Lawson, A. B. (2008), *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, Chapman & Hall/CRC.

Lin, X. and Zhang, D. (1999), 'Inference in generalized additive mixed models by using smoothing splines', *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **61**(2), 381–400.

MAP (2006), 'Malaria Atlas Project', http://www.map.ox.ac.uk/.

MARA (2004), 'Mapping Malaria Risk in Africa', http://www.mara.org.za/.

McCullagh, P. and Nelder, J. A. (1989), *Genarized Linear Models*, second edn, Chapman & Hall.

Ministry of Health (MOH) (2010), 'Malawi National Malaria Indicator Survey'.

National Statistical Office (NSO) (2008), *Malawi Population Projections*, NSO, Zomba.

National Statistical Office (NSO) and ICF Macro (2011), 'Malawi demographic and health survey 2010'.

Nkurunziza, H., Gebhardt, A. and Pilz, J. (2010), 'Bayesian modelling of the effect of climate on malaria in Burundi', *Malaria Journal* **9**(1), 114.

President's Malaria Initiative (PMI) Country Profile: Malawi (2011).
URL: *http://pmi.gov/countries/profiles/malawi_profile.pdf*

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Riedel, N., Vounatsou, P., Miller, J. M., Gosoniu, L., Chizema-Kawesha, E., Mukonka, V. and Steketee, R. W. (2010), 'Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modelling of the 2006 Zambia national malaria indicator survey (ZMIS)', *Malaria Journal* **9**(37).

Rowlingson, B., Diggle, P., Moyeed, R. and Thomson, M. (2002), 'Childhood malaria in the gambia: A case-study in model-based geostatistics', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **51**(4), 493–506.

Royle, J. A. and Nychka, D. (1998), 'An algorithm for the construction of spatial coverage designs with implementation in splus', *Computers & Geosciences* **24**(5), 479–488.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Sachs, J. and Malaney, P. (2002), 'The economic and social burden of malaria', *nature* **415**(6872), 680–685.

Schur, N., Hrlimann, E., Garba, A., Traor, M. S., Ndir, O., Ratard, R. C., Tchuent, L.-A. T., Kristensen, T. K., Utzinger, J. and Vounatsou, P. (2011), 'Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged $\leq$ 20 years in West Africa', *PLoS neglected tropical diseases* **5**(6), 17.

Shyu, G.-S., Cheng, B.-Y., Chiang, C.-T., Yao, P.-H. and Chang, T.-K. (2011), 'Applying factor analysis combined with kriging and information entropy theory for mapping and evaluating the stability of groundwater quality variation in taiwan', *International Journal of Environmental Research and Public Health* **8**(4), 1084–1109.
URL: *http://www.mdpi.com/1660-4601/8/4/1084/*

Skarbinski, J., Mwandama, D., Luka, M., Jafali, J., Wolkon, A., Townes, D., Campbell, C., Zoya, J., Ali, D. and Mathanga, D. P. (2011), 'Impact of health facility-based insecticide treated bednet distribution in Malawi: Progress and challenges towards achieving universal coverage', *PLoS ONE* **6**(7), 11.

Smith, A. F. M. and Roberts, G. O. (1993), 'Bayesian computation via the gibbs sampler and related markov chain monte carlo methods', *Journal of the Royal Statistical Society. Series B (Methodological)* **55**(1), 3–23.
URL: *http://www.jstor.org/stable/2346063*

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

Usher, P. K. (2010), 'Modelling Malaria Transmission Potential for Climate Scenarios in West Africa and Europe', *Earth & E-nvironments* **5**, 40–65.

Waller, L. A. and Gotway, C. A. (2004), *Applied Spatial Statistics for Public Health Data*, John Wiley & sons, New Jersey.

Wand, H., Whitaker, C. and Ramjee, G. (2011), 'Geoadditive models to assess spatial variation of HIV infections among women in local communities of Durban,South africa', *International Journal of Health Geographics* **10**(1), 28.

Wardrop, N. A., Atkinson, P. M., Gething, P. W., Fèvre, E. M., Picozzi, K., Kakembo, A. S. L. and Welburn, S. C. (2010), 'Bayesian geostatistical analysis and prediction of Rhodesian Human African Trypanosomiasis', *PLoS neglected tropical diseases* **4**(12), e914.

URL: *http://eprints.soton.ac.uk/178437/*

Westbrook, C. J., Reiskind, M. H., Pesko, K. N., Greene, K. E. and Lounibos, L. P. (2010), 'Larval environmental temperature and the susceptibility of Aedes albopictus skuse (Diptera: Culicidae) to Chikungunya virus', *Vector borne and zoonotic diseases Larchmont NY* **10**(3), 241–247.

WHO (2010), 'World health organization: World malaria report', `http://www.who.int/malaria/world_malaria_report_2010/en/index.html`.

Yoko, A. and Rifat, A. (2011), 'Effect of investment in malaria control on child mortality in sub-Saharan Africa in 2002 − 2008', *PLoS ONE* **6**(6), 12.

Appendix A

Software

A.1 Bayes X

This object oriented software for structured additive regression models was developed at the Ludwig-Maximilians-Universität München in Germany and is the primary software used in the analysis of the data for this thesis. `Bayes X` can be freely downloaded from this site (`http://www.stat.uni-muenchen.de/~bayesx/`). The current version 2.0.1 was developed in October 2009. It supports both fully Bayesian and Empirical Bayesian estimation approaches

A.1.1 Bayes X syntax

For the full Bayesian, the syntax is as follows: The data has to be converted into an ASCII format before it is read into `Bayes X`. Being an object oriented, an object that holds the dataset must be defined at the beginning. Here we create an object holding the MIS data which will be called `mis`

```
> dataset mis
> mis.infile using C:/ ...          # This line of code reads in the data.
> mis.describe        # Allows visualiation of the data in a spread sheet
```

```
> bayesreg model1        #Make a bayesreg object
> model1.outfile = C:/...        # Specifies output location
```

Now the actual model is specified in this manner

```
> model1.regress rdt=agecat1+agecat2+agecat3+agecat4+urban+
north+centre+w1+w2+w3+w4+sleptnet+alt(psplinerw2)+Rainfall(psplinerw2)+
Min.temp(psplinerw2)+latitude(psplinerw2)+distr(random)+
ea(random),family=binomial iterations=10000 burnin=1000 step=50
predict using mis
```

The empirical Bayesian syntax The bayesreg object is simply replaced with the remlreg, i.e.

```
> remlreg model2
> model2.regress rdt=agecat1+agecat2+agecat3+agecat4+urban+
north+centre+w1+w2+w3+w4+sleptnet+alt+Rainfall+
Min.temp+latitude*longitude(kriging,nrknots=500),
family=binomial lowerlim=0.001 eps=0.001 using mis
```

A.2 R

R is a freely available statistical package widely used in statistical analysis. People are free to download `R` and its contributed packages at `http://www.r-project.org/`. The software is very versatile and has all the capabilities of commercial software and much more. The `R` package `BayesX` provides additional graphing capabilities. Results from the stand alone Bayes X software can be visualized in `R` thereby making use its excellent graphing facilities to plot the nonlinear effect of continuous variables.

```
> library(BayesX)        #Load the package first
```

The `plotnonp` function plots nonparametric effects from the `Bayes X` results directory

```
> plotnonp("C:/.../results altitude pspline.res",ylab="s(altitude)")
```

Other functions for visualizing `Bayes X` results include `plotsurf`, `plotautocorr`. For instance, the `plotsurf` can be used to visualize surfaces.

```
> plotsurf("C:/.../results latitude*longitude.res",
ylab="s(latitude*longitude)")
```

A.3 MCMC convergence

A.3.1 Trace plots

The check for convergence of the Markov chain in Bayesian analysis is a necessary step in model diagnostics of fully Bayesian models. This is done by looking at the trace plots of the model parameters. In this section, we present trace plots of some of the model parameters.

Fig A.1 show that the chains converge quite well for the model parameters.

A.3.2 Autocorrelation functions

Fig A.2 below shows the autocorrelation functions of sampled parameters in the Markov Chain. The plots show clearly that there is no autocorrelation in the sampled parameters as the line moves around zero.

A.4 Metropolis-Hastings algorithm

Let $\theta = (\theta_1, \ldots, \theta)$ are the parameters to be estimated, then a value for θ at each iteration is sampled. The algorithm discussed by (Kim et al., 2010) is as follows;

1. Assign initial values to $\theta^{(0)}$. The starting point can be arbitrary provided $f(\theta^0|y) > 0$

2. Set θ_i^* from the density $J_i(\theta_i^*;\theta_i^{(t)}) for 1 \leq i \leq n$

3. Compute the ratio of densities

$$r = \frac{p(\theta_i^*;y)J_i(\theta_i^{(t)};\theta_i^*)}{p(\theta_i^{(t)};y)J_i(\theta_i^*;\theta_i^{(t)})}$$

where p is the full conditional distribution of θ_i

4. Set

$$\theta_i^{(t+1)} = \begin{cases} \theta_i^* & \text{with prob. min}(r,1) \text{ and,} \\ \theta_i^{(t)} & \text{otherwise} \end{cases}$$

A.5 R code for prediction surfaces

This section gives the R code used to produce the predictive surface for malaria risk. The data files are extracted from Bayes X after running a REML model #Plot predictive posterior means

```
>data=read.table("kriging.txt",header=T)
>x=data$longitude
>y=data$latitude
>z=data$pmode
>library(akima)
>data.interp <- interp(x,y,z,duplicate="mean")
>jpeg("mappmode.jpg",width=(20*0.39),height=(20*0.39),units="in",
res=300,quality=100)
>par(cex=1.2,mfrow=c(1,1),cex.lab=1.2,cex.axis=1,tcl=NA,bty="n",
mar=c(3,3,3,3),mgp=c(1,0.1,0))
>plot(map,col="transparent",bg="white")
>image(data.interp,axes=F,col=terrain.colors(12),xlim=c(31,37),
ylim=c(-19,-8),add=T)
>contour(data.interp,add=T,axes=F)
```

```
>plot(map,add=T)
>plot(lake,col="white",add=T)
>plot(lake,col="lightblue",density =50,add=T)
>plot(map,col="transparent",bg="white",add=T)
>box(lwd=0.3,bty="o")
>dev.off()
```

#Plot of standard errors

```
>x=data$longitude
>y=data$latitude
>z=data$std
>library(akima)
>data.interp <- interp(x,y,z,duplicate="mean")
>jpeg("mapstd.jpg",width=(20*0.39),height=(20*0.39),units="in",
res=300,quality=100)
>par(cex=1.2,mfrow=c(1,1),cex.lab=1.2,cex.axis=1,tcl=NA,bty="n",
mar=c(3,3,3,3),mgp=c(1,0.1,0))
>plot(map,col="transparent",bg="white")
>image(data.interp,axes=F,col=terrain.colors(12),xlim=c(31,37),
ylim=c(-19,-8),add=T)
>contour(data.interp,add=T,axes=F)
>plot(map,add=T)
>plot(lake,col="white",add=T)
>plot(lake,col="lightblue",density =50,add=T)
>plot(map,col="transparent",bg="white",add=T)
>box(lwd=0.3,bty="o")
>dev.off()
```

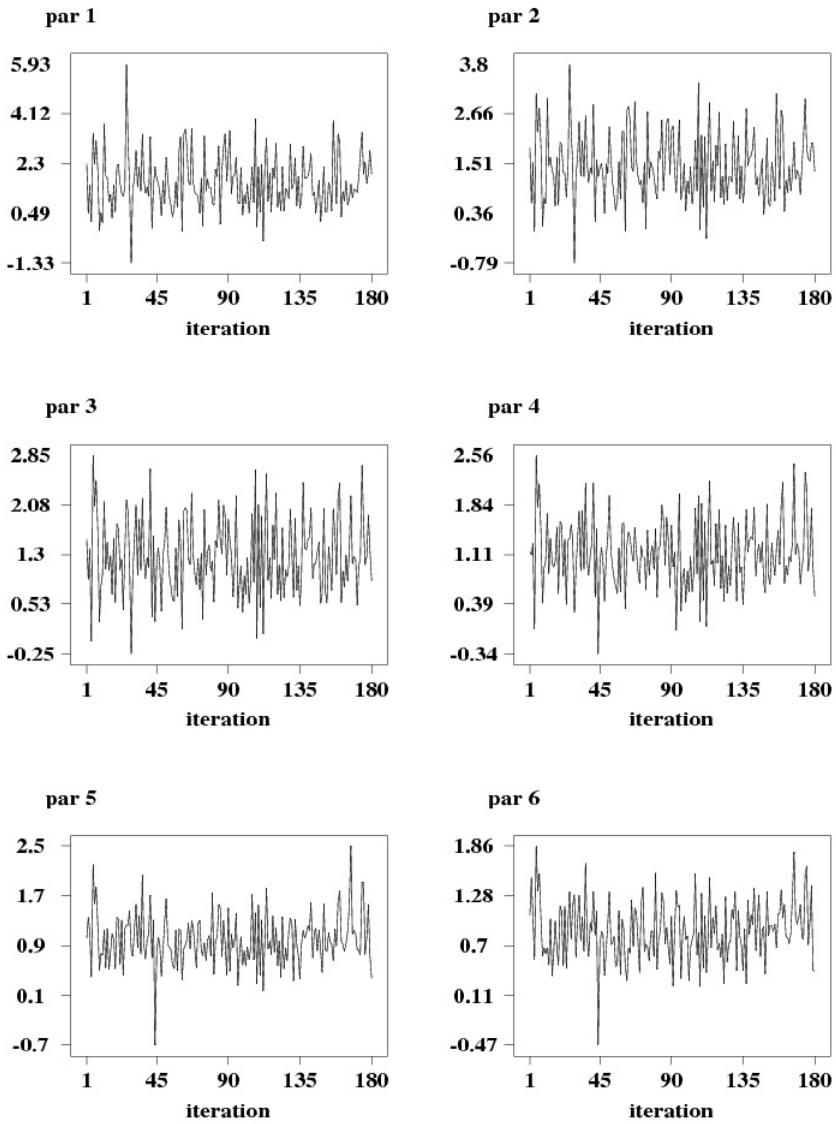


Figure A.1: Trace plots for some of the model parameters

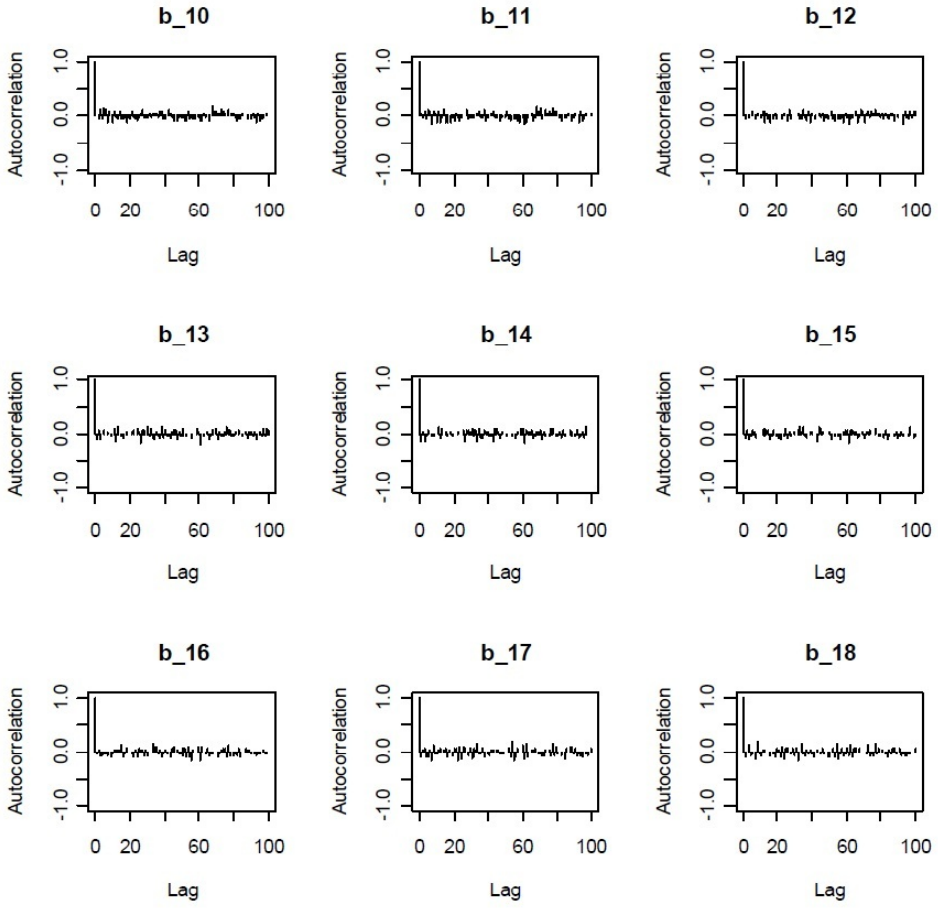


Figure A.2: Autocorrelation functions for sampled parameters

Appendix B

Ethical statement

The coordinates of the households have been used only for purposes of data analysis and will NOT be used in any other way. No attempt to identify the locations of the households using the coordinates will be made. Furthermore, consent was sought from the respondents before interviewing them during the data collection exercise and the confidentiality was guaranteed. The detailed consent form is presented below.

Appendix C

Consent form

C.1 Introduction

The Ministry of Health wants to learn how well the malaria prevention programme is working in Malawi. We would like to ask you some questions about bed net use in your home, and also some general questions about your children's health.

We are also doing a survey of malaria in children. To do this, we will test children for malaria parasites in the blood. One way to test for malaria parasites in the blood includes taking a small sample of blood by finger prick and examining under a microscope and in a laboratory. Another way is to look at anaemia (low levels of blood), by taking a small sample of blood by finger prick and examining with a HemoCue machine.

C.2 Purpose of the survey

We want to see if our country's malaria programme works. We will ask you some questions about bednet use in your home and also about your children's health. We will also see how common malaria is among young children in the community

by testing for parasites in the blood and also by testing for low levels of blood. We will visit people in their homes and look at people that come to health facilities. This will help us learn how best to measure the effects of malaria control in the community.

C.3 Procedures

If you agree to take part, we will ask you a few questions, and a nurse will take a small amount of blood from your child's finger.

We will ask you questions about bed net use in your home, and about other things that are linked to malaria. We will also ask some questions about your health and about your children's health. This should only take about 30 minutes.

We will take only up to eight drops of blood from your child. One drop of blood will be wiped off. The second drop of blood will be used to test for malaria in the lab using a microscope. The third drop of blood will be used to test for low levels of blood (anaemia) here in the house. The fourth drop will be used for a rapid malaria diagnostic test here in the house. The remaining four drops of blood may be put on paper for additional laboratory analysis of malaria.

The results for low levels of blood and for the rapid malaria diagnostic test will be given to you today. If your child has low levels of blood, malaria, or history of fever, we will give you treatment. This will be the same treatment your child would get if you went to your health centre. This will cost you and your family nothing. If the nurse thinks that your child is very ill, we will assure transport to the nearest health clinic to provide your child with the necessary health care.

Lab test results will be ready after one week. If your child has malaria, a survey staff member will return to your house to give treatment for malaria to your child. This will only happen if your child has not already been treated today. Even if you do not wish to take part, you can still ask to see the nurse and get the correct

treatment. Even if you do not agree to take part, if your child is ill, you should visit the nearest health clinic if your child is not better in three days or is worse over time.

C.4 Risks and benefits

Your child will feel a pinch that lasts a few seconds when we take the blood tests. For any malaria health problem that we find, the nurse will give the treatments that the Ministry of Health suggests. These drugs are proven safe and effective, but any drugs can cause side effects in a small number of patients. The nurse will discuss these with you.

C.5 Voluntariness

It is your choice to be in this survey. It will not affect the care that the nurse will give you or your children should you wish to receive it. If you do agree to take part, your answers to all questions and your child's test results will be kept private to the extent the law allows. If you agree to take part, you can also decide not to answer any of the questions that you do not want to, and you can refuse the blood tests.

If you have any questions or clarification pertaining to this survey please feel free to contact Mrs Doreen Ali, 0889374043 or Dr D. Kathyola, 088834443.

Thank you very much for your time. Would you like to take part in this survey?

Statement of Parental Permission for malaria surveillance (signature or thumbprint required) The above has been read to me, and I agree to let my child take part.

Signature_____ Date_____

Thumb print:

Participant's name_____

For persons who cannot sign

The above consent was read and the person agreed to take part

Signature_____ Date_____