Spatial Analysis and Modeling of All Fever and Self-reported Malaria Fever in Kenya and their correlation with *Plasmodium falciparum* prevalence

By:

Damaris Kinyoki – W62/77749/2009

Project Submitted to the Institute of Tropical & Infectious Diseases (UNITID), in fulfilment of degree of Masters in Medical Statistics

University of Nairobi

2011

DECLARATION

I hereby declare that this research is my original work and has not been presented for any degree award in any institution or university.

Signature.. Date. 24/11/2011

Name: Damaris Kinyoki

Reg. No. W62/77749/2009

RECOMMENDATION

This work has been submitted with our approval as supervisors:

Supervisors

Signature.................................. Date 24/11/2011

Dr. Thomas Achia,

School of Mathematics,

University of Nairobi

Signature.................................. Date 24/11/2011

Dr. Abdisalan Noor,

Department of Geospatial and Space Technology,

University of Nairobi

Director, Institute of Tropical & Infectious Diseases (UNITID)

Signature.................................. Date 25/11/2011

Prof. Benson Estambale,

Institute of Tropical & Infectious Diseases (UNITID),

University of Nairobi

DEDICATION

I dedicate this study to my parents, for establishing an academic base and moral support and more importantly their love that gave me confidence and strength during my studies.

ACKNOWLEDGEMENT

My sincere thanks to my supervisors, Dr. Thomas Achia and Dr. Abdisalan Noor whose initial encouragement and subsequent support and supervision shaped my thinking and gave me more insight into the completion of this project. Their thinking and suggestions greatly influenced my view of the research problem and led to successful completion of this study.

I also acknowledge Kenya Medical Research Institute/Welcome Trust Research Programme in collaboration with Kenya National Bureau of Statistics for availing the data for this study and more so providing all the equipment and softwares needed for statistical analysis.

Lots of thanks go to all the academic staff at the UNITID, College of Health Science, and University of Nairobi for their dedicated training in all aspects of medical statistics.

Finally, my most profound thanks to God, my faithful shepherd.

TABLE OF CONTENT

ABSTRACT

Introduction: Fever has been used as the presumptive marker for malaria in Kenya for a long time. The Kenya National Malaria Strategy 2001-2010 states that all fevers should be treated as early and as close to a patient's home as possible, with acceptable quality and correct dosages of the first line anti-malarial and supportive treatment. The aetiology of fevers in malaria endemic areas has been the subject of considerable basic and applied public health research for many years. Plasmodium parasite is not the only cause of fever and only one of many pathogens that cause identical pyrogenic responses in Kenya. Although fever usually has a high sensitivity for the diagnosis of malaria it suffers from poor specificity and critically depends on the prevalence of both asymptomatic infection and the overall prevalence of fever

Objectives: The overall objective of this study was to model the risk of self-reported fever in Kenya and self-reported malaria and examine its relationship with the modeled estimates of P. falciparum parasite prevalence.

Methodology: This was a cross-sectional study that sought to model the risk of all self-reported reported fevers and self-reported malaria fevers in Kenya at district level against selected covariates for the study and their relationship with actual risk of malaria infection. Household level data assembled during the Kenya Integrated Household Budget Survey of 2005/06 was used. Data were aggregated at the districts level. Semi-parametric regression models which allowed joint analysis of nonlinear effects of some covariates, spatially structured variation, unstructured heterogeneity, and other fixed covariates were developed. Modeling and inference used fully Bayesian approach via Markov Chain Monte Carlo (MCMC) simulation techniques.

Results: The risk of all fevers and self-reported fever increases with increase in distance to the health facility, parasite prevalence, proportion of people with no toilets and under-fives. The results also indicate that the risk of fever decreases with increase in the proportion of Male, the proportion of the people using protected sources of water. The results also indicate significant differences in both structured and unstructured spatial effects.

Conclusion: This study emphasizes that the methodological framework used provides a useful tool for analyzing the data at hand and of similar structure.

LIST OF ABBREVIATIONS/ACRONYMS

AIC - Akaike Information Criterion

ASAL - Arid and semi-arid lands

CBS - Chromosome Banding Patterns

CI – Confidence Interval

DHS - Demographic and Health Surveys

DIC - Deviance Information Criterion

EA - Enumeration Area (Population Census)

GIS – Geographical Information Systems

HIV – Human Immune Deficiency Virus

HSRC - Human Sciences Research Council

IG - Inverse Gamma (IG) Distribution

KIHBS - Kenya Integrated Household Budget Survey

KNMS - Kenyan National Malaria Strategy

MARA - Mapping Malaria Risk in Africa

MBG - Model-based Geostatistical

MCMC - Markov chain Monte Carlo

MOH – Ministry of Health

NASSEP - National Sample Survey and Evaluation Programme

PCR - Polymerase Chain Reaction

PET - Potential Evapotranspiration

***Pf*PR**$_{2-10}$ - P*lasmodium falciparum* parasite rate data

PPS - Probability Proportional to Size

PSUs - Primary Sampling Units

TB - Tuberculosis

UNICEF - United Nations Children's Fund

WHO – World Health Organization

LIST OF TABLES

LIST OF FIGURES

DEFINITION OF OPERATIONAL TERMS

Burn-in period – This is the adaptive phase of the Bayesian model. All iterations before a model convergence is achieved are eliminated from the sample in order to avoid the influence of the initial values. If the generated sample is large enough, the effect of this period on the calculation of posterior summaries is minimal.

Convergence of the algorithm - With the term convergence of an MCMC algorithm, refers to situations where the algorithm has reached its equilibrium and generates values from the desired target distribution. Generally it is unclear how much to run an algorithm to obtain samples from the correct target distributions. Several diagnostic tests have been developed to monitor the convergence of the algorithm.

Equilibrium distribution - This is called the stationary or target distribution of the MCMC algorithm. The notion of the equilibrium distribution is related to the Markov chain used to construct the MCMC algorithm. Such chains stabilize to the equilibriud stationary distribution after a number of time sequences t > B. Therefore, in a Markov chain, the distribution of 8 (t) and 8(''') will be identical and equal to the equilibrium stationary distribution. Equivalently, once it reaches its equilibrium (distribution), an MCMC scheme generates dependent random values from the corresponding stationary distribution (Robert and Casella, 2004, pp. 206-207).

Fever - An increase in body temperature above the normal temperature i.e. above an oral temperature of 37.5°C. This is according to the national guidelines for the diagnosis, treatment and prevention of Malaria in Kenya 2010.

Fixed Effects model: A statistical model that represents the observed quantities in terms of explanatory variables that are treated as if the quantities were non-random so that model is of the form: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$; $j = 1,2,..,n_i, \sum_i n_i = n_T, \varepsilon_{ij} \sim N(0,\sigma^2)$ the errors are i.i.d

Iteration - refers to a cycle of the algorithm that generates a full set of parameter values from the posterior distribution. It is frequently used to denote an observation of simulated values.

Iterations kept, T'. These are the number of the iterations retained after discarding the initial burn-in iterations (that is, $T' = T - B$). If we also consider a sampling lag $L > 1$, then the total number of iterations kept refers to the final independent sample used for posterior analysis. MCMC output. This refers to the MCMC generated sample. We often refer to the MCMC output as the sample after removing the initial iterations (produced during the burn-in period) and considering the appropriate lag. Output analysis. This refers to analysis of the MCMC output sample. It includes both the monitoring procedure of the algorithm's convergence and analysis of the sample used for the description of the posterior distribution and inference about the parameters of interest;

Initial values of the chain - Starting values used to initialize the chain are simply called initial values. These initial values may influence the posterior summaries if they are far away from the highest posterior probability areas and the sample size of the simulated sample T is sufficient to eliminate its effect. To mitigate or avoid the influence of the initial values is done by removing the first iterations of the algorithm or letting the algorithm run for a large number of iterations or obtain different samples with different starting points.

Malaria - An infectious disease characterized by cycles of chills, fever, and sweating, caused by a protozoan of the genus *Plasmodium* in red blood cells, which is transmitted to humans by the bite of an infected female anopheles mosquito.

Random Effects model: A statistical model that assumes that the dataset being analyzed consists of a hierarchy of different populations whose differences relate to that hierarchy. In random effects models the hierarchies are selected randomly so that inference is made about the population of factor levels.

Spatial Autocorrelation - 'Spatial autocorrelation' is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional (2-D) surface, introducing a deviation from the independent observations assumption of classical statistics. In this study the topographical covariate is split into correlated (structured) and uncorrelated (unstructured) parts.

Thinning interval or sampling lag - the final MCMC generated sample is not independent. For this reason, there is need to monitor the autocorrelations of the generated values and select a sampling lag $L > 1$ after which the corresponding autocorrelation are low. Then, we can produce an independent sample by keeping the first generated values in every batch of L iterations. Hence, if we consider a lag (or thin interval) of three iterations then we keep the first every three iterations (that is, we keep observations 1, 4, 7,...). This tactic is also followed to save storage space or computational speed in high-dimensional problems.

Total number of iterations T. This refers to the total number of the iterations of the MCMC algorithm.

CHAPTER ONE

1 INTRODUCTION

Fever, defined by Schaffner A. 2006, is a phylogenetically ancient host reaction to invading microorganisms and other noxious stimuli. Poikylothermic organisms can reach febrile temperatures by seeking a hot environment in response to a higher set point in their thermoregulatory center. Endothermic organisms produce febrile temperatures through endogenous heat production at the expenditure of a higher metabolic rate. Fever is a complex physiological response that is aimed at facilitating survival of the host (Schaffner A. 2006). Other terms used synonymously with fever are pyrexia or controlled hyperthermia.

Fever is induced by endogenous inflammatory mediators, such as prostaglandins and pyrogenic cytokines that are released by immune cells activated by exogenous pyrogens. Although the pathways (humoral and/or neuronal) responsible for transfer of the pyretic signals from the blood to the brain are still under discussion, it is generally accepted that they act on the level of the anterior hypothalamus to raise the thermoregulatory set-point (Soszynski 2003).

Fever has traditionally served as the entry point for presumptive treatment of malaria in African community (Okiro and Snow 2010). However, recent downward transition in the epidemiology of malaria across many places in Africa would suggest that the predictive accuracy of a fever history as a marker of disease has changed prompting calls for the change to diagnosis-based treatment strategies (Okiro and Snow 2010).

According to Kenya National Malaria Strategy 2001-2010, all fevers should treated promptly and therefore as close to a patient's home as possible, with acceptable quality and correct dosages of the first line anti-malarial and supportive treatment. The aetiology of fevers in malaria endemic areas has been the subject of considerable basic and applied public health research for many years.

Plasmodium falciparum, the main malaria parasite in sub-Saharan Africa, is not the only cause of fever (Kallander, *et al.*, 2004). Although fever usually has a high sensitivity for the diagnosis of malaria it suffers from poor specificity and critically depends on the prevalence of both asymptomatic malaria infection and the overall prevalence of other fever conditions. Presumptive treatment of all fevers has, therefore, been the most risk-adverse approach to managing "malaria" across Africa and is enshrined in the recommendations proposed by the Integrated Management of Childhood Illnesses (IMCI) (Gove S 1997). There is, however, increasing evidence that the intensity of *P. falciparum* transmission is declining across many parts of Africa (Hay, *et al.*, 2007). World Health Organization (WHO) Guidelines for the treatment of malaria (2010) has now moved away from presumptive treatment in Africa to one that recommends parasitological diagnosis (WHO 2010).

Interest in mapping the global distribution of malaria is motivated by a need to define populations at risk for appropriate resource allocation and to provide a robust framework for evaluating its global economic impact. A study done by Snow et al 2005, estimated that there were 515 (range 300–660) million episodes of clinical *P. falciparum* malaria in 2002. These global estimates are up to 50% higher than those reported by the World Health Organization (WHO) and 200% higher for areas outside Africa, reflecting the WHO's reliance upon passive national reporting for these countries. Without an informed understanding of the cartography of malaria risk, the global extent of clinical disease caused by *P. falciparum* will continue to be underestimated. The WHO estimated that in

2008 there were 250 million cases malaria leading to approximately 850,000 malaria deaths. While malaria is endemic within most tropical and subtropical regions of the world, 90 per cent of all malaria deaths currently occur in sub-Saharan Africa and most of these deaths are among children under five years of age. Approximately 1 in every 6 child deaths (16%) in Africa is due to malaria.

In Kenya, malaria is one of the leading causes of morbidity and mortality, particularly in children under five years of age in Kenya. *Plasmodium falciparum* is the commonest cause of malaria (National Guideline of Diagnosis, Treatment and Prevention of Malaria in Kenya 2010).The malaria disease is debilitating, affecting millions of Kenyans each year and fatal to many thousands. The toll it exacts must be viewed not only in terms of the physical, financial and emotional burden) but also by its macroeconomic impact. Malaria accounts for 30% of all outpatient attendance and 19% of all admissions to our health facilities. An estimated 170 million working days are lost to the disease each year (Kenya National Malaria Strategy 2001-2010). Approximately 25 million out of a population of 39 million people in Kenya in 2009 are at risk of malaria ((Noor, *et al.,* 2009). An estimated 170 million working days are lost to the disease each year (MOH 2001).

As efforts to control malaria are expanded across the world, understanding the role of transmission intensity in determining the burden of clinical malaria is crucial to the prediction and measurement of the effectiveness of interventions to reduce transmission. Furthermore, studies comparing several endemic sites led to speculation that as transmission decreases morbidity and mortality caused by severe malaria might increase. A study done in Kilifi, Kenya, aimed at assessing the epidemiological characteristics of malaria in Kilifi during a period of decreasing transmission intensity with 18 years (1990–2007) of surveillance data from a paediatric ward in a malaria-endemic region of Kenya, found out that Hospital admissions for malaria decreased from 18·43 per 1000 children in 2003 to 3·42 in 2007. Over the 18 year surveillance period, the incidence of cerebral malaria initially increased. However, malaria mortality decreased overall because of a decrease in incidence of severe malarial anaemia since 1997 (4·75 to 0·37 per 1000 children) and improved survival among children admitted with non-severe malaria. Parasite prevalence, the mean age of children admitted with malaria, and the proportion of children with cerebral malaria began to change 10 years before hospitalization for malaria started to fall (Okiro *et al.*, 2008).

To align its strategy with the new international agenda the Government of Kenya has developed the 10-year Kenyan National Malaria Strategy (KNMS) 2009-2017 which was launched 4th November 2009. The National Malaria Strategy is based on and carries forward an inclusive partnership between the two ministries responsible for health, other line ministries of the Government of Kenya, and our development and implementing partners in malaria control. It is a product of extensive consultation and collaboration with all stakeholders and establishes a strategic framework for the delivery of malaria control interventions, along with monitoring and evaluating performance. National scale-up of parasitological diagnosis of febrile cases before treatment and the universal coverage of all vulnerable populations with malaria prevention interventions are seen as key factors to achieve the strategic goals. Not only will diagnosis reduce the cost of treating malaria by cutting down on drug wastage but will also increase the chances of diagnosing other causes of infections among non-malarious patients and thereby offering appropriate treatment and decreasing the risk of severe and/or fatal outcomes.

In this study the burden of general morbidity at the district level in Kenya is estimated by modeling self-reported fevers among all age groups using Bayesian geostatistical approaches. The burden of self-reported malaria fevers is then modeled similarly. District level estimates of self-reported fevers and self-reported malaria fevers are compared with estimated prevalence of confirmed malaria infections at the district level. Estimate of malaria infections were extracted for each district from a map developed by Noor et al (2009).

CHAPTER TWO

2 LITERATURE REVIEW

2.1 Background Information on Fever and Malaria

Fever has been recognized as an accompaniment of infection since the time of the early Sumerians *circa* 4,000 B.C. The earliest surviving description of febrile illnesses was transcribed by Hippocrates *circa* 425 B.C. Fever is an elevation of temperature above the normal daily variation. It is commonly caused by infection, but noninfectious causes such as neoplastic and immunologically-mediated disease may also have fever as a primary clinical manifestation (Briedis 2008).

Individuals maintain their body temperature within a narrow range around 37°C despite wide variations in environmental temperatures. During a 24-hour period, body temperature varies (up to +/- 0.6°C) in a diurnal or *circadian* rhythm from a low point in the early morning to the highest levels in late afternoon or early evening. Most fevers are induced by polypeptide molecules called *endogenous pyrogens*. These are produced by the host in response to infection, injury, inflammation, or antigenic challenge. These polypeptides cause fever by triggering biochemical changes in the hypothalamus, particularly to stimulate hypothalamic prostaglandin synthesis (Briedis 2008).

At the beginning, gradual increase in body temperature is observed together with muscle shivering, vasoconstriction in the skin, and piloerection. This situation is called chills. Increased body temperature is achieved by lowered loss of heat. Vasoconstriction in the skin and subcutaneous tissue is the cause of pale color and dryness, the affected person has a feeling of coldness. At the same time the production of heat in the organism increases. The muscle tonus increases, the spasms occur. Spasms may occur mainly in children. When the vasodilatation starts in the skin, the feeling of warmth and sweating occurs (Bornstein, 1963).

Malaria is a disease caused by parasites of the genus Plasmodium. Nationally, *Plasmodium falciparum* is the predominant species (98.2 per cent) while *P. malariae*, *P.ovale* is 1.8 per cent often occurring as mixed infections. *P.vivax* may account for up to 40-50 per cent of infections (often mixed with P.falciparum) in the Northern and North Eastern parts of Kenya (Hamel *et al.*, 2001)

Kenya has four malaria epidemiological zones, with diversity in risk determined largely by altitude, rainfall patterns and temperature. The zones are:

Endemic: Areas of stable malaria have altitudes ranging from 0 to 1,300 metres around Lake Victoria in western Kenya and in the coastal regions. Rainfall, temperature and humidity are the determinants of the perennial transmission of malaria. The vector life cycle is usually short and survival rates are high because of the suitable climatic conditions. Transmission is intense throughout the year, with annual entomological inoculation rates between 30 and100 (Kenya National Guidelines for the Diagnosis, Treatment and Prevention of Malaria 2010).

Seasonal transmission: Arid and semi-arid areas of northern and south-eastern parts of the country experience short periods of intense malaria transmission during the rainfall seasons. Temperatures are usually high and water pools created during the rainy season provide breeding sites for the malaria vectors. Extreme climatic conditions like the El Niño southern oscillation lead to flooding in these areas, resulting in epidemic outbreaks with high morbidity rates owing to the low immune status of the population (Kenya National Guidelines for the Diagnosis, Treatment and Prevention of Malaria 2010).

Epidemic prone areas of western highlands of Kenya: Malaria transmission in the western highlands of Kenya is seasonal, with considerable year-to-year variation. Epidemics are experienced when climatic conditions favor sustainability of minimum temperatures around 18°C. This increase in minimum temperatures during the long rains favours and sustains vector breeding, resulting in increased intensity of malaria transmission. The whole population is vulnerable and case fatality rates during an epidemic can be up to ten times greater than those experienced in regions where malaria occurs regularly.

Low risk malaria areas: This zone covers the central highlands of Kenya including Nairobi. The temperatures are usually too low to allow completion of the sporogonic cycle of the malaria parasite in the vector (Kenya National Guidelines for the Diagnosis, Treatment and Prevention of Malaria 2010).

Malaria is a climate sensitive disease and climate information can be used to monitor and predict aspects of its spatial distribution seasonality year-to-year variability and longer term trends. Furthermore, climate information is increasingly recognized as necessary to enable accurate impact evaluations of malaria interventions. The biology of malaria transmission is markedly complex, involving interactions between multiple, constantly changing, extrinsic and intrinsic factors, many of which cannot be easily measured and are therefore challenging to model. Mathematical models of malaria transmission are highly sensitive to the non-linear response of both the vector and parasite to variations in temperature. Thus, the issue of temperature variability and change is often considered central to the discussion of whether malaria transmission is likely to increase if global temperatures rise (Hamel *et al.*, 2001).

2.2 Modeling of Infectious Diseases in Africa

Bayesian statistical approaches have gained widespread use in infectious disease mapping, especially malaria. Bayesian inference was implemented via a Markov chain Monte Carlo algorithm using the model-based geostatistics framework of Diggle, *et al.*, (1998). Prediction of risk based on point-referenced data presents some challenges when the data are sparsely distributed. Such data often exhibit autocorrelation, such that locations close to each other have similar risk. Models should allow for spatial correlation, failing which, the significance of risk factors is overstated (Thomsom, *et al.*, 1999 and Boyd, *et al.*, 005). Analyses of point-referenced data have been carried out using geostatistical models (Cressie, *et al.*, 1993), for optimal prediction. Recently, a model-based geostatistical (MBG) approach has been applied (Diggle, *et al.*, 1998). The approach permits simultaneous modelling of related issues such as risk assessment, spatial dependence, prediction and quantification of uncertainty (Diggle, *et al.*, 2002).

In the last decade, maps have been produced at different geographical scales in sub-Saharan Africa (Omumbo, *et al.*, 2005), following the Mapping Malaria Risk in Africa (MARA) project (MARA 1998), with the aim of identifying areas where greatest control effort should be focused. It is important to characterize malaria risk based on empirical evidence using a malaria-specific indicator, in this case, malaria prevalence of infection in children, and assess its relationship with environmental risk factors. A benchmark indicator by which malaria risk is modeled and mapped in Africa is the parasite rate (PfPR), which is the proportion of a random sample of population with malaria parasites in their peripheral blood, used frequently to define transmission intensity since the 1950's and has a predictable mathematical relationship to the rarely sampled measures of entomological inoculation rate (EIR) and the basic reproductive number (R_o). The PfPR has therefore become the benchmark indicator by which malaria risk is modeled and mapped in Africa (Noor *et al.*, 2009)

Mzolo (2008) used Bayesian approach in estimating risk determinants of infectious diseases. In his study the data was clustered at different level. By controlling for both fixed and random risk factors any excess association between HIV & TB was quantified. Bayesian methods require prior information to estimate the posterior distribution. These methods involved integrating high-dimensional functions. The focus was on the MCMC methods of simulating data. The roots of the MCMC methods come from the Metropolis Algorithm (Metropolis & Ulam 1949; Metroplis 1953).

Mikael, *et al.*, 2010 carried out a study on mapping malaria incidence distribution that accounted for environmental factors in Maputo Province – Mozambique. This study formulated a Bayesian hierarchical model to malaria count data aggregated at district level over a two years period. This model made it possible to account for spatial area variations. The model was extended to include environmental covariates temperature and rainfall. Study period was then divided into two climate conditions: rainy and dry seasons. The incidences of malaria between the two seasons were compared. Parameter estimation and inference were carried out using MCMC simulation techniques based on Poisson variation. Model comparisons are made using DIC.

Moyeed, *et al.,* 2005 used Bayesian geostatistical prediction to provide an explanation of the over dispersion in the data and in particular to assess whether the over-dispersion is spatially structured in his study on the intensity of infection with *Schistosoma mansoni* in East Africa. This study followed the approach of Alexander, *et al.,* (2000). In this, the total egg count of each individual was modelled as a negative binomial variate with over-dispersion parameter $k > 0$ which incorporates extra-Poisson variation. Larger values of k indicate less variability, with the limiting case $k = \infty$ corresponding to the Poisson distribution. The logarithm of the mean of the distribution as an additive function of the individual-level covariate sex, the two school-level covariates elevation and distance to nearest inland perennial water body and a spatially-structured school-level random-effect was modelled. The spatial random-effect was modelled as a stationary Gaussian process with mean 0, variance σ^2 and correlation function $\exp^{(-d_{ij}/\alpha)}$, where d_{ij} is the distance between villages i and j and the parameter α measures the rate at which the spatial correlation decays over distance, with α $\log_2$ being a characteristic length, which we call the 'half-distance', over which the correlation reduces by half, and 3α being the distance at which the correlation reduces to 0.05.

Several studies have shown that malaria infection is influenced by environmental factors such as temperature, rainfall, humidity and elevation. Specifically, temperature and rainfall act as limiting factors on the development of *Anopheles* mosquitoes which are the intermediate hosts in the transmission of malaria parasites (Cox, *et al.,* 1999). In tropical settings, temperature and rainfall conditions are nearly always favourable for transmission. Humidity is also suitable for transmission because it affects the survival rate of mosquitoes. Similarly, elevation above sea level (asl) is known to define the ecology of malaria transmission through temperature (Bødker, 2003). At certain altitudes malaria transmission does not occur because of extreme temperatures that inhibit the mosquito and parasite life-cycle. For small countries like Kenya, topography remains a single most important factor that defines large-scale differences in malaria risk because climatic variables change little over the limited range of latitude.

Zacarias, *et al.,* 2010, analyzed the relationship between environmental factors and malaria cases, a Poisson model in Statistical Package R was fitted. A Bayesian hierarchical model to malaria count data aggregated at district level over a two year period wass formulated. This model made it possible to account for spatial area variations. The model was extended to include environmental covariates temperature and rainfall. Study period was then divided into two climate conditions: rainy and dry seasons. The incidences of malaria between the two seasons were compared. Parameter estimation and inference were carried out using MCMC simulation techniques based on Poisson variation. Model comparisons are made using DIC.

In the study on Spatial patterns of infant mortality in Mali: the effects of malaria endemicity by Gemperli, *et al.*, 2000, Logistic regression models were fitted to infant mortality, using SAS version 8.2 software to identify significant socioeconomic, demographic, and birth-related covariates. Variables showing a significant bivariate association with infant mortality were selected for subsequent spatial multivariate analysis: type of region, mother's education, sex, birth order, and preceding birth interval. Bayesian hierarchical models were fitted to estimate the amount of spatial heterogeneity in infant mortality as well as associations between risk factors and infant mortality in the presence of spatial correlation. Three spatial Bayesian models were fitted. A baseline model (model 0) included no covariates but overall constant and site-specific random effects. Model 1 was an extension of the baseline model with the inclusion of year of birth and socioeconomic and demographic variables as potential risk factors. Model 2 included the same parameters as did model 1 but, in addition, adjusted for levels of malaria endemicity. In addition, a Bayesian non-spatial analog of model 2 was fitted for comparative purposes. The model-based geostatistical methods were applied to analyze and predict malaria risk in areas where data were not observed. Topographical and climatic covariates were added in the model for risk assessment and improved prediction. A Bayesian approach was used for model fitting and prediction. It confirmed that mother's education, birth order and interval, infant's sex, residence, and mother's age at infant's birth had a strong impact on infant mortality risk in Mali (Gemperli, *et al.*, 2000). The residual spatial pattern of infant mortality showed a clear relation to well-known foci of malaria transmission, especially the inland delta of the Niger River. No effect of estimated parasite prevalence could be demonstrated. Possible explanations include confounding by unmeasured covariates and sparsity of the source malaria data. Spatial statistical models of malaria prevalence are useful for indicating approximate levels of endemicity over wide areas and, hence, for guiding intervention strategies. However, at points very remote from those sampled, it is important to consider prediction error.

Spatial prediction is a procedure for prediction at an unobserved location, using data at observed locations, optimized with reference to a specific error criterion. The criterion is the squared prediction error at the unobserved location – averaged over a conceptual class of spatial prediction problems that have the same configuration of observed and unobserved locations. The specification of this averaging class is the model under which the optimization is carried out and the estimation error is reported. The usual model under which kriging calculations are done is that of a spatial stochastic process that generates spatial fields over the geographical region of interest. A stochastic process model is selected with characteristics that reflect characteristics of the available data. With this averaging model, the stated kriging properties are purely conceptual – they refer to *average* prediction errors that would be seen if the same kriging procedure were applied to the same prediction problem on spatial fields generated repeatedly by the selected stochastic process. Locations of the observed sites within the geographical domain are fixed under this averaging model, but not the values of the observations themselves (Cressie, N. 1988). The fact that the kriging averaging model does not fix the values of the observations can be seen as a limitation. An alternative to this stochastic process averaging model treats the whole spatial field as fixed and considers the configuration of observed and unobserved sites as one configuration from a specified class of possible configurations. The error associated with spatial estimation is then the average error associated with the entire class of specified configurations. However, for estimating (predicting) field values at specified sites, as in interpolation and mapping, an averaging model that uses only randomization of the observation sites would not be meaningful for the computation of estimation error (Neath & Cavanaugh 2010).

In spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data, Kazembe, *et al.*, 2006, used Point-referenced prevalence ratio data of children aged 1–10 years, obtained at 73 survey sites across the country. Data were abstracted from grey or published literature based on collection methods outlined in MARA technical report. Response Variable: Malaria Risk Predictor Variable: elevation, mean annual maximum temperature, PET and rainfall. The model-based geostatistical methods were applied to analyze and predict malaria risk in areas where data were not observed. Topographical and climatic covariates were added in the model for risk assessment and improved prediction. A Bayesian approach was used for model fitting and prediction. Bivariate models showed a significant association of malaria risk with elevation, annual maximum temperature, rainfall and potential evapotranspiration (PET). However in the prediction model, the spatial distribution of malaria risk was associated with elevation, and marginally with maximum temperature and PET. The resulting map broadly agreed with expert opinion about the variation of risk in the country, and further showed marked variation even at local level. High risk areas were in the low-lying lake shore regions, while low risk was along the highlands in the country.

In the study of estimating risk determinants of Infectious diseases using Bayesian approach in, University of KwaZulu (Mzolo 2008), used data from a household based second-generation surveillance survey of HIV conducted by HSRC in 2005. The survey design applied a multi-stage disproportionate, stratified sampling approach based on a master sample of 1000 EAs. The sample was stratified by province and locality type of the EAs whereas in urban areas race was used as a third stratification variable. The master sample allowed for reporting of results at the level of province, type of locality, age and race group. The data is clustered at an EA level. By controlling for both fixed and random risk factors we will be able to quantify any excess association between HIV & TB. Bayesian methods require prior information to estimate the posterior distribution. These methods involve integrating high-dimensional functions. The focus was on the MCMC methods of simulating data. The roots of the MCMC methods came from the Metropolis Algorithm (Metropolis & Ulam 1949; Metroplis 1953). The Gibbs sampler (Geman & Geman 1984) is a MCMC method that was widely applicable. Priors for fixed effects were assumed multivariate normal centered at zero. Priors for random effects (EA) were assumed to follow a normal distribution. The burn-in period of 2000 iterations were allowed for both models. An estimated intraclass correlations for HIV and TB are _HIV = 0.169 and _TB = 0.249, respectively.

Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa by Moyeed, *et al.*, 2005 used Individual-level data on intensity of *Schistosoma mansoni* infection which were obtained from cross-sectional random samples of school children from dedicated school surveys conducted between 1999 and 2004 at 459 locations by national research teams under the auspices of the Schistosomiasis Control Initiative (SCI) in Tanzania (Clements *et al.*, 2006) and in Uganda (Kabatereine, *et al.*, 2004) and by research projects in western Kenya (Brooker, *et al.*, 2001; Clarke, *et al.*, 2005). A Bayesian geostatistical model was developed to predict the intensity of infection with *Schistosoma mansoni* in East Africa. Epidemiological data from purposively-designed and standardized surveys were available for 31,458 school children (90% aged between 6-16 years) from 459 locations across the region and used in combination with remote sensing environmental data to identify factors associated with spatial variation in infection patterns. The geostatistical model explicitly took into account the highly aggregated distribution of parasite distribution by fitting a negative binomial distribution to the data and accounted for spatial correlation. Results identified the role of environmental risk factors in explaining geographical heterogeneity in infection intensity and show how these factors can be used to develop a predictive map.

Gosoniu, *et al.*, 2006 carried out a bayesian modeling of geostatistical malaria risk data in Angola. The model was based on the logistic regression method. The assumption was that the number of those found with malaria parasite in a blood sample aroused from a Binomial distribution, that is

$Y_i \sim Bin(N_i, p_i)$ with parameter pi measuring malaria risk at location s_i and modeled the relation between the malaria risk and environmental covariates X_i via the logistic regression

$$\log it(P_i) = X_i^{'}\beta$$

where $\beta = (\beta_1, \beta_2, \dots \beta_p)\mathrm{T}$ are the regression coefficients. This model assumed independence between the surveys. However, the geographical location introduced correlation since the malaria risk at nearby locations was influenced by similar environmental factors and therefore it was expected that the closer the locations the similar the way malaria risk varies. To account for spatial variation in the data, an error term (random effect) φ_i was introduced at each location s_i. That is

$$\log it(p_i) = X_i T\beta + \varphi_i$$

and modeled the spatial correlation on the φ_i parameters $Q_i \sim MVN(0, \varepsilon)$. The φ_i's are not independent but are derived from a distribution which models the correlation or equivalently the covariance between every pair of random effects. They adopted a Multivariate Normal distribution for the φ_i's since they represent error terms and therefore, are defined on a continuous scale. That is, $\varphi_i = (\varphi_1, \varphi_2, \dots \varphi_n)' \sim N(0, \Sigma)$, where Σ is a matrix with elements $\sum_{ij}$ quantifying the covariance $Cov(\varphi_i, \varphi_i)$ between every pair (φ_i, φ_j)at locations s_i and s_j respectively. The distribution of random effect ϕdefined Gaussian spatial process. Results indicate that the stationarity assumption is important because it influenced the significance of environmental factors and the corresponding malaria risk maps.

Bayesian Geostatistical Modeling of Malaria Indicator Survey Data in Angola showed that the categorical model was able to better capture the relationship between parasitaemia prevalence and the environmental factors. Model fit and prediction were handled within a Bayesian framework using Markov chain Monte Carlo (MCMC) simulations. Combining estimates of parasitaemia prevalence with the number of children under 5 were obtained estimates of the number of infected children in the country. The population-adjusted prevalence ranges from 3:76% in Namibe province to 32:65% in Malanje province. The odds of parasitaemia in children living in a household with at least 0:2 ITNs per person was by 41% lower (CI: 14%, 60%) than in those with fewer ITNs. The estimates of the number of parasitaemic children produced in this paper are important for planning and implementing malaria control interventions and for monitoring the impact of prevention and control activities.

Spatial modeling and risk factors of malaria incidence in northern Malawi by Kazembe, *et al.*, 2007 used ecological spatial regression models to profile spatial variation of malaria risk and analyzed possible association of disease risk with environmental factors at sub-district level in northern Malawi. Using malaria incidence data collected between January 2002 and December 2003, applied and compared Bayesian Poisson regression models assuming different spatial structures. For each model environmental covariates were adjusted initially identified through bivariate non-spatial models. The model with both spatially structured and unstructured heterogeneity provided a better fit, guided by the model comparison criteria. Malaria incidence was associated with altitude, precipitation and soil water holding capacity. The risk increased with and precipitation. Smoothed map showed less spatial variation in risk, with slightly higher estimates of malaria risk (RR > 1) in low-lying areas mostly situated along the lakeshore regions, in particular in Karonga and Nkhatabay districts, and low risk in high-lying areas along Nyika plateau and Vwaza highlands. The results suggested that the spatial variation in malaria risk in the region is a combination of various environmental factors, both observed and unobserved, and the map only highlighted the overall effect of these factors. The results also identified areas of increased risk, where further epidemiological investigations can be carried out.

2.3 Modeling of Infectious Diseases in Kenya

Kenya is one of very few countries that have a surplus of malaria risk data, spanning over 30 years. The earliest attempts to describe the spatial distribution of malaria risk in Kenya were based on expert opinion of malaria seasons and climate (Holtz, *et al.*, 2002).

Hay, *et al.*, 2010 states that a total of 174 sites in Kenya reported the presence of the *An. gambiae* complex without specification of the sibling species. One hundred and fifty three survey locations reported the presence of *An. gambiae* and these were largely located in areas of Western and Nyanza Provinces closest to Lake Victoria and in the Coast Province with few presences reported in the more central regions of the country. Out of these reports 17 *An. gambiae* were identified using morphology only and the remainder identified using species-specific chromosomal PCR and cytogenetic techniques involving analysis of polytene chromosome banding patterns (CBS). The majority (120, 78%) of reported *An. gambiae* presences were based on adult catches. *Anopheles arabiensis* was more ubiquitous in its reported distribution with observations along the coast, across Western Kenya and central Kenya including the arid areas of the northwest in Turkana district with 244 unique spatial incidences of this sibling species reported since 1990. *Anopheles arabiensis* larvae were sampled at 124 (51%) sites, adult catches were conducted at 110 (45%) sites and a combination of larval and adult sampling methods were used at ten (4%) sites.

Omumbo, *et al.*, 1998, reports that risks of infection with *Plasmodium falciparum* among Kenyan children, estimated from combinations of parasitological, geographical, demographic and climatic data in a GIS platform, appear to be low for 2.9 million, stable but low for another 1.3 million, moderate for 3.0 million and high for 0.8 million. (Estimates were not available for 1.4 million children.) Whilst the parasitological data were obtained from a variety of sources across different age-groups and times, these markers of endemicity remained relatively stable within the broad dentitions of high, moderate and low transmission intensity. Models relating ecological and climatic features to malaria intensity and improvements in our understanding of the relationships between parasite exposure and disease outcome will hopefully provide a more rational basis for malaria control in the near future.

A study done by Snow *et al.* (1998), states that climate operates to affect the vectorial capacity of P. falciparum transmission and this is particularly important in the Horn of Africa and parts of East Africa. A logic climate suitability model has been used to define areas of Kenya unsuitable for stable transmission. Kenya's unstable transmission areas can be divided into areas where transmission potential is limited by low rainfall or low temperature and, combined, encompass over 8 million people. Among areas of stable transmission empirical data on P. falciparum infection rates among 124 childhood populations in Kenya has been used to develop a climate-based statistical model of transmission intensity. This model correctly identified 75% (95% confidence interval CI 70-85) of 3 endemicity classes (low, < 20%; high, > or = 70%; and intermediate parasite prevalences). The model was applied to meteorological and remote sensed data using a geographical information system to provide estimates of endemicity for all of the 1080 populated fourth level administrative regions in Kenya. National census data for 1989 on the childhood populations within each administrative region were projected to provide 1997 estimates. Endemicity-specific estimates of morbidity and mortality were derived from published and unpublished sources and applied to their corresponding exposed-to-risk childhood populations. This combined transmission, population and disease-risk model suggested that every day in Kenya approximately between 72 and 400 children below the age of 5 years either die or develop clinical malaria warranting in-patient care, respectively. Despite several limitations, such an approach goes beyond 'best guesses' to provide informed estimates of the geographical burden of malaria and its fatal consequences in Kenya.

Snow *et al.* (1998) used an electronic and national search that was undertaken to identify community-based parasite prevalence surveys in Kenya. Data from these surveys were matched using ArcView 3.2 to extract spatially congruent estimates of the FCS values generated by the MARA model. Levels of agreement between three classes used during recent continental burden estimations of parasite prevalence (0%, >0 – <25% and ≥25%) and three classes of FCS (0, >0 – <0.75 and ≥0.75) were tested using the kappa (κ) statistic and examined as continuous variables to define better levels of agreement. Two hundred and seventeen independent parasite prevalence surveys undertaken since 1980 were identified during the search. Overall agreement between the three classes of parasite prevalence and FCS was weak although significant ($\kappa = 0.367$, $p < 0.0001$). The overall correlation between the FCS and the parasite ratio when considered as continuous variables was also positive (0.364, $p < 0.001$). The margins of error were in the stable, endemic (parasite ratio ≥25%) class with 42% of surveys represented by an FCS <0.75. Reducing the FCS value criterion to ≥0.6 improved the classification of stable, endemic parasite ratio surveys. Zero values of FCS were not adequate discriminators of zero parasite prevalence.

Okara 2010 carried out a study on distribution of the main malaria vectors in Kenya. Survey locations were geo-positioned using national digital place name archives and on-line geo-referencing resources. The geo-located species-presence data were displayed and described administratively, using first-level administrative units (province), and biologically, based on the predicted spatial margins of Plasmodium falciparum transmission intensity in Kenya for the year 2009. Each geo-located survey site was assigned an urban or rural classification and attributed an altitude value. A total of 498 spatially unique descriptions of Anopheles vector species across Kenya sampled between 1990 and 2009 were identified, 53% were obtained from published sources and further communications with authors. More than half (54%) of the sites surveyed were investigated since 2005. A total of 174 sites reported the presence of An. gambiae complex without identification of sibling species. Anopheles arabiensis and An. Funestus were the most widely reported at 244 and 265 spatially unique sites respectively with the former showing the most ubiquitous distribution nationally. Anopheles gambiae, An. arabiensis, An. funestus and An. pharoensis were reported at sites located in all the transmission intensity classes with more reports of An. gambiae in the highest transmission intensity areas than the very low transmission areas.

Kaya 2002 explored the Use of Radar Remote Sensing for Identifying Environmental Factors Associated with Malaria Risk in Coastal Kenya. Image analysis was performed using eCognition software - a classification analysis package that uses an object-based approach rather than the traditional pixel-based routine. Image data is classified based on parcels of pixels known as 'objects' that are created using a segmentation routine, which separates significantly contrasted adjacent regions in an image based on image brightness values, and extracts the homogeneous regions as individual objects. Following segmentation, a classification was performed using the multi-temporal filtered and texture analysis images as input. A standard nearest neighbor classification was performed based on user-specified training objects. The resulting classification was validated with test sites. Classified polygons were extracted as GIS layers for use in the malaria risk map generation procedure. The premise for assessing areas at risk of malaria infection is based on the maximum distance a malaria-carrying mosquito can travel from its breeding ground to infect human hosts. The town of Mombasa (island in south part of image) is clearly identified as populated, smaller villages found in the middle part of the image. Forest type 1 (mangrove forests), were characterized by flooded areas with emergent vegetation. For this reason, backscattering characteristics, as well as textural information are similar to wetlands. Due to the similarities in environmental conditions, both landscape variables may be considered as high risk in terms of malaria breeding sites.

Omumbo, *et al.,* 2005, used discriminant analysis to model environmental and human settlement predictor variables to distinguish between four classes of parasite ratio (PR) risk shown to relate to disease outcomes in the region. The data search identified 330 parasite survey data points that fulfilled the inclusion criteria. Discriminant analysis was performed initially without controlling for ecological zone or urbanization and the accuracy of the prediction tested. OA was 72.4% (j ¼ 0.502, s ¼ 0.494). On visual comparison with historical (Government of Tanganyika 1956) and contemporary Modeled malaria risk maps, is significantly anomalous in southern Tanzania. The results were improved by stratifying the analysis according to two ecozone classes and by forcing the inclusion of urbanization as a predictor. These modifications marginally reduced OA in both ecozone 1 (OA ¼ 64.0%; j ¼ 0.483; s ¼ 0.478; and ecozone 2 (OA ¼ 61.4%; j ¼ 0.45; s ¼ 0.308;

but provided an output with fewer large-area anomalies when compared with historical and more recent climate-driven maps. The OA for the combined ecozone/urban adjusted was 62.1% (j ¼ 0.477, s ¼ 0.495).

Noor, *et al.,* 2009, in his study on the risks of malaria infection in Kenya used carried out a Bayesian space-time models using the Kenya *Pf*PR$_{2\text{-}10}$ data and the selected covariates, a spatial-temporal Bayesian generalized linear geostatistical mode. This model was implemented to predict a malaria map of Kenya for 2009. The underlying assumption of the Kenya *Pf*PR$_{2\text{-}10}$ model was that the probability of prevalence at any survey location was the product of two factors. First, a continuous function of the time and location of the survey, modified by a set of covariates, and modelled as a transformation of a space-time Gaussian random field. Second, a factor depending on the age range of individuals sampled in each survey. The distribution of the second factor was based on the procedure described by Smith, *et al., 2007.* The Bayesian spatial-temporal model was implemented in two parts starting with an inference stage in which a Markov Chain Monte Carlo (MCMC) algorithm was used to generate samples from the joint posterior distribution of the parameter set and the space-time random field at the data locations. This was followed by a prediction stage in which samples were generated from the posterior distribution of *Pf*PR$_{2\text{-}10}$ at each prediction location on a 1 × 1 km grid.

In this study of the risks of malaria infection in Kenya, P*lasmodium falciparum* parasite rate data *(Pf*PR) survey data was used as response variable which were identified using basic search principles and the following Predictor Variables: categorical forms of urbanization, rainfall, vegetation coverage, aridity, distance to water bodies, altitude and temperature. The relationships of the covariates in their continuous and categorical forms were first visually examined against *Pf*PR$_{2-10}$ data using scatter and box plots. These were used to aggregate the covariates into suitable categories that corresponded to biologically appropriate definitions, previous applications of remotely sensed variables and retention of effective sample sizes. A univariate non-spatial binomial logistic regression model was then implemented for each covariate with *Pf*PR$_{2-10}$ as the dependent variable in Stata/SE Version 10. The results of the univariate analyses were used to determine the relative strength of each candidate covariate as a predictor of *Pf*PR$_{2-10}$ and identify those which qualified for inclusion in the Bayesian geostatistical model. First, where there was more than one plausible way of categorizing a covariate, the size of the odds ratio, the Wald's p-value and the value of Akaike Information Criterion (AIC), were used to determine which approach resulted in categories with the strongest association with *Pf*PR$_{2-10}$. Once the best categorizations were determined, a collinearity test of all the covariates was undertaken and if a pair had a correlation coefficient > 0.9, the variable with the highest value of AIC was dropped from subsequent analysis. The selected covariates were then analyzed in a binomial multivariate logistic regression with *Pf*PR$_{2-10}$ as the dependent variable. Using backwards variable elimination, covariates with Wald's P > 0.2 were removed step-wise until a fully reduced model was achieved. Using the Kenya *Pf*PR$_{2-10}$ data and the selected covariates, a spatial-temporal Bayesian generalized linear geostatistical model was implemented to predict a malaria map of Kenya for 2009.

2.4 The Basis of the Study

This study follows the methodology that was used by Kazembe, *et al.*, 2008 in his study of applications of Bayesian approach in modeling risk of malaria-related hospital mortality. The studies response variables were distributed as a Bernoulli random variable. However the differences between the two studies were the area of application and number of covariate. Kazembe, *et al.*, 2008 study used data from Malawi while this study uses data from Kenya. The number of covariates in Kazembe, *et al.*, 2008 was six while this study used 13 covariates of which some were eliminated in the initial descriptive analysis.

Kazembe, *et al.*, 2008 analyzed and compared the following four logistic models; M0, M1, M2 and M3. Model M0 was a basic regression model of fixed covariates only. Model M1 assumed nonlinear functions for the continuous factors and tried to assess the gains of fitting a semi-parametric model. Model M2 considered all possible risk factors, i.e., simultaneously analysed nonlinear effects of age, time trend of calendar time, structured spatial effects, v, for 21 residential wards, unstructured spatial effects, u, heterogeneity effects, h, for 23 health facilities, and fixed effects, w'°, for categorical variables. In model M3, model M2 was extended to consider further temporal effects, whereby the effect of calendar time is decomposed into a time trend and seasonal component. The models were implemented in Bayes X version 1.4. For the four models, 40,000 iterations were carried out after a burn-in sample of 10,000, thinning every 20th iteration, yielding 2,000 samples for parameter estimation. It was observed that the risk of dying in hospital was lower in the dry season, and for children who travel a distance of less than 5 kms to the hospital, but increased for those who are referred to the hospital. The results also indicated significant differences in both structured and unstructured spatial effects, and the health facility effects revealed considerable differences by type of facility or practice. More importantly, the approach shows non-linearities in the effect of metrical covariates on the probability of dying in hospital. The study emphasized that the methodological framework used provided a useful tool for analyzing the data at hand and of similar structure.

CHAPTER THREE

3 STATEMENT OF RESEARCH QUESTION

3.1 Research Problem and Justification

Current malaria control initiatives are aimed at reducing malaria burden by half by the year 2011. Effective control requires evidence-based utilization of resources. Characterizing spatial patterns of risk, through maps, is an important tool to guide control programmes. Maps of malaria infection and disease risks can help to select appropriate suites of interventions. Advances in model based geo-statistics and assembly of malaria parasite prevalence data have led to the development of the most comprehensive malaria risk map in Kenya (Noor et al 2009).

However, due to low level of availability of malaria diagnostic tools in most health facilities and because of the long standing recommendation of presumptive treatment, most of the patients in the health facilities that present fever symptoms are recorded and treated as malaria cases (MoH 2010). The advantage of this approach of clinical diagnosis of malaria is that it has a high sensitivity i.e. the likelihood of missing a malaria case is minimal. However recent downward transition in the epidemiology of malaria across Kenya would suggest that the proportion of fevers that are malaria has also reduced (Okiro & Snow 2010). In return this has led to overestimation of malaria cases in the health facilities. Not only does result in wastage of antimalarial resources on treating non-malaria fevers but also decreases the chances of diagnosing other diseases that patients may be suffering thereby increasing the likelihood of severe and/or fatal outcomes.

To this end an analysis was carried out to predict and map risk of self-reported overall fever and malaria fever in Kenya using secondary community survey data. Here a Bayesian model-based geostatistical method was carried out to predict the risk of fever in Kenya. A comparison was carried out between the prevalence of fever and prevalence of self-reported Malaria in Kenya to determine if there is any correlation between the two.

As a new phase of malaria control in Kenya begins, the implications of the resulting malaria risk map in comparison to map of fever will inform the decision makers on the future case management of fever and malaria cases and also inform the prospects for the future of malaria control nationwide.

3.2 Broad Objective

To model the risk of self-reported fever and self-reported malaria fever in Kenya and determine their correlation with *P. falciparum* parasite prevalence modeled using empirical parasite rate data.

3.3 Specific Objectives

1. To determine the current distribution of self-reported fever in Kenya using Bayesian hierarchical modelling approaches.
2. To determine the current distribution of self-reported malaria fever in Kenya Bayesian hierarchical modelling approaches.
3. To determine the relationship of the distribution of self-reported fever and malaria fever against the *P. falciparum* parasite prevalence modeled using empirical parasite rate data.

3.4 Research Question

Is there a correlation between Bayesian geostatistical model of all reported fevers and self-reported malaria fever with *P. falciparum* parasite prevalence modeled using empirical parasite rate data?

CHAPTER FOUR

4 METHODOLOGY

In this chapter, the methodology of this study is explained in details. It is divided into seven subsections. Section 4.1 which introduces the type of study, while Sections 4.2 – 4.7 focus on more specific matters such as the analysis of the sample, estimation of Monte Carlo variability measures, and convergence of the algorithm and section 4.8 outlines the limitations of the study.

4.1 Study Design

This was a study that sought to model the risk of all reported fevers and self-reported malaria in Kenya against selected covariates for the study and to determine if the two models correlated with *P. falciparum* parasite prevalence modeled using empirical parasite rate data. The data used was from a cross-sectional national survey called Kenya Integrated Household Budget Survey (KIHBS) undertaken by the Kenya National Bureau of Statistics (KNBS) from May 2005 to May 2006.

4.2 Study Area and Population

Kenya is situated in the eastern part of the African continent. The country lies between 5 degrees north and 5 degrees south latitude and between 24 and 31 degrees east longitude. It is almost bisected by the equator. Kenya is bordered by Ethiopia (north), Somalia (northeast), Tanzania (south), Uganda and Lake Victoria (west), and Sudan (northwest). It is bordered on the east by the Indian Ocean. The 536-kilometre coastline, which contains swamps of East African mangroves and the port in Mombasa, enables the country to trade easily with other countries. The country is divided into 8 provinces and 158 districts (as of the 2009 Population and Housing Census). It has a total area of 582,646 square kilometres of which 571,466 square kilometres form the land area. Approximately 80 percent of the land area of the country is arid or semiarid, and only 20 percent is arable. The country has diverse physical features, including the Great Rift Valley, which runs from north to south; Mount Kenya, the second highest mountain in Africa; Lake Victoria, the largest freshwater lake on the continent; Lake Nakuru, a major tourist attraction because of its flamingos; Lake Magadi, famous for its soda ash; a number of rivers, including Tana, Athi, Yala, Nzoia, and Mara; and numerous wildlife reserves containing thousands of different animal species. The country falls into two regions: lowlands, including the coastal and Lake Basin lowlands, and highlands, which extend on both sides of the Great Rift Valley. Rainfall and temperatures are influenced by altitude and proximity to lakes or the ocean. The climate along the coast is tropical with rainfall and temperatures being higher throughout the year. There are four seasons in a year: a dry period from January to March, the long rainy season from March to May, followed by a long dry spell from May to October, and then the short rains between October and December.

Kenya's population was 10.9 million in 1969, and by 1999 it had almost tripled to 28.7 million (Central Bureau of Statistics, 1994, 2001a. The crude birth rate increased from 50 births per 1,000 populations in 1969 to 54 per 1,000 in 1979 but thereafter declined to 48 and 41 per 1,000 in 1989 and 1999, respectively. The crude death rate increased from 11 per 1,000 population in 1979-1989 to 12 per 1,000 for the 1989-1999 period. The infant mortality rate, which had steadily decreased from 119 deaths per 1,000 live births in 1969 to 88 deaths per 1,000 live births in 1979, and then to 66 deaths per 1,000 live births in 1989, increased briefly in 1999 to 77 per 1,000 but then resumed its decline in 2009.

Table 4-1: Basic Demographic Indicators

Selected demographic indicators for Kenya 1969, 1979, 1989, 1999 and 2009					
Indicator	**1969**	**1979**	**1989**	**1999**	**2009**
Population (millions)	10.9	16.2	23.2	28.7	39.4[a]
Density (pop/km^2)	19.0	27.0	37.0	49.0	67.7[a]
Percent urban	9.9	15.1	18.1	19.4	21.0[a]
Crude birth rate	50.0	54.0	48.0	41.3	34.8[b]
Crude rate death rate	17.0	14.0	11.0	11.7	U
Inter-censal growth rate	3.3	3.8	3.4	2.9	2.8[a]
Total fertility rate	7.6	7.8	6.7	5.0	4.6[b]
Infant mortality rate (per 1,000births)	119	88	66	77.3	52.0[b]
Life expectancy at birth	50	54	60	56.6	58.9[a]

[a] Revised projection figures
[b] KDHS results
[u] unknown
Sources: CBS, 1970; CBS, 1981; CBS, 1994; CBS, 2002a

Malaria is the leading cause of morbidity and mortality in Kenya, with close to 70 percent (24 million) of the population at risk of infection (Hamel, *et al.*, 2010). Although malaria affects people of all age groups, children under five years of age and pregnant women living in malaria endemic regions are most vulnerable. The human toll that malaria exacts and the economic and social impacts are devastating: sick children miss school, working days are lost, and tourism suffers. Malaria becomes puts communities in vicious cycle of poverty, where the disease prevents growth of the human and economic capital necessary to bring the disease under control. Moreover, malaria disproportionately affects the rural poor, who can neither afford insecticide-treated bed nets for prevention nor access appropriate treatment when they fall sick.

The Kenya Vision 2030 goal for the health sector is to provide equitable, affordable, quality health services to all Kenyans. The goal also aims to restructure the health care delivery system to shift the emphasis from curative to preventive health care. The goal of the second National Health Sector

Strategic Plan (NHSSP II 2005–2010) is to 'reduce health inequalities and to reverse the downward trend in health-related outcome and impact indicators' (Ministry of Health, 2004).

Malaria prevention and control activities in Kenya are guided by the National Malaria Strategy (NMS) 2009-2017 and the National Health Sector Strategic Plan 2005-2010. The NMS outlines malaria control activities based on the epidemiology of malaria in Kenya. The strategy aims to achieve national and international malaria control targets. The core interventions adopted in Kenya are the following:

- Vector control—using insecticide-treated nets (ITNs) and indoor residual spraying (IRS)
- Case management (using Artemisinin-based combination therapies (ACTs) and improved laboratory diagnosis)
- Management of malaria in pregnancy
- Epidemic preparedness and response
- Cross-cutting strategies including information, education, and communication (IEC) for behaviour change, as well as effective monitoring and evaluation

One of the objective of the Kenya National Malaria Strategy 2009-2017 is aimed to have 80 per cent of all self-managed fever cases receive prompt and effective treatment and 100 per cent of all fever cases who present to health facilities receive parasitological diagnosis and effective treatment by 2013. This is by strengthening capacity for malaria diagnosis and treatment; increasing access to affordable malaria medicines through the private sector; and strengthening home management of malaria. This initiative has not yet been scaled up in all the health facilities in Kenya.

4.3 Sampling and Sample Size

The data used in this study was taken from the health section of the Kenya Integrated Household Budget Survey (KIHBS). The following is how the sampling was carried out. A total of 13,430 households were randomly selected to comprise the KIHBS sample, which was designed to generate representative statistics at the national, provincial and district levels. The sampling design involved a number of stages.

Cluster selection: In the first stage, 1,343 clusters were stratified by district (and by both urban and rural areas within each district). The objective was to make the total sample representative and descriptive of the unequal distribution of the population across districts. In the KIHBS sample, 10 households were randomly selected with equal probability in each cluster to give a total sample of 13,430 households.

Strata: the urban and rural areas of all districts except Nairobi and Mombasa, which are entirely urban. However, in the six districts that contain municipalities, clusters in the urban sample were further stratified into six groups: five socio-economic classes in the municipality itself and other urban areas in the district. This ensured that different types of neighborhoods and social classes within municipal areas are all represented in the sample. The total sample sizes in rural and urban areas were 8,610 and 4,820 households respectively.

The KIHBS clusters are the Primary Sampling Units (PSUs) from the NASSEP IV sampling frame, which is designed to give nationally, and sub-nationally, representative household survey samples. The NASSEP IV sampling frame is composed of 1,800 clusters selected with probability proportional to size (pps) from a set of all Enumeration Areas (EA) used during the 1999 Population and Housing Census (a cluster is either an EA or an EA segment of about 100 households) The KIHBS clusters sampled in each district where selected with equal probability from the NASSEP IV frame. Therefore, the first stage consists of a defacto pps sub-sample of census EA segments. This sampling strategy produced an approximately self-weighted sample of households in each stratum.

With the basic sampling frame constructed, the next stage consisted of updating the NASSEP IV clusters through a cartographic and household listing exercise conducted in all urban and ASAL clusters as well as a portion of the rural clusters in which population was found to have changed significantly.

4.4 Data Collection

The data used for the project was taken from Kenya integrated household budget survey (KIHBS) 2005/06. Data collection for KIHBS 2005/06 was undertaken for a period of 12 months starting 16th May 2005. The Survey was conducted in 1,343 randomly selected clusters across all districts in Kenya and comprised 861 rural and 482 urban clusters. Following a listing exercise, 10 households were randomly selected with equal probability in each cluster resulting in a total sample size of 13,430 households.

The year-long survey was organized into 17 cycles of 21 days each, during which enumerators conducted household interviews in the clusters. Further, the districts were grouped into 22 zones that were logistically convenient for field teams to operate. Seasonal variation was captured by randomising visits to the selected clusters so that in each cycle at least one cluster was visited in each zone. See Appendix 1 for more information.

4.5 Variables

Dependent Variable

Proportion of self-reported fever for each district (model 1)
Proportion of self-reported malaria fever for each district (model 2)

Predictor Variable

Topographical Covariates

POLYID(spatial,map=m)*
POLYID(random) **

Socioeconomic and demographic covariates

Proportion of Male
Proportion of Under 5s
Proportion of ever attended school
Proportion Diagnosed by Health Worker
Proportion of Chronically ill
Proportion who slept under treated net
Proportion with protected source of drinking water
Proportion with the main cooking fuel as electricity/Gas LPG
Proportion with main cooking fuel as firewood
Proportion with main lighting as electricity/Gas LPG
Proportion with no Toilet
Access to Health Facility
Proportion using Pit latrine
Mean *P. falciparum* Parasite Prevalence by district (Noor et al 2009)

The spatial effect of the district was split up into a spatially correlated part (*) and an uncorrelated part (**) (Fahrmeir & Lang 2001b). The correlated part is modeled by a Markov random field prior, where the neighborhood matrix and possible weights associated with the neighbors are obtained from the map object m. The uncorrelated part is modeled by an i.i.d. Gaussian effect.

Table 4-2: Response Variable Description

No.	District	Total No. Interviewed	Total Number of Fevers	Number with self-reported Malaria	Proportion of Fevers	Proportion with reported malaria fever
	Baringo	1031	383	63	0.3715	0.10
	Bomet	802	261	46	0.3254	0.10
	Bondo	727	489	173	0.6726	0.25
	Bungoma	1399	668	218	0.4775	0.14
	Buret	898	328	88	0.3653	0.12
	Busia	917	620	216	0.6761	0.12
	Butere/Mumias	906	591	169	0.6523	0.24
	Embu	730	407	87	0.5575	0.04
	Garissa	975	381	85	0.3908	0.19
	Gucha	830	341	85	0.4108	0.14
	Homa Bay	751	567	204	0.755	0.12
	Isiolo	1019	428	118	0.42	0.04
	Kajiado	821	280	34	0.341	0.09
	Kakamega	1011	462	159	0.457	0.27
	Keiyo	844	161	19	0.1908	0.25
	Kericho	834	171	66	0.205	0.14
	Kiambu	1275	270	43	0.2118	0.29
	Kilifi	885	133	56	0.1503	0.06
	Kirinyaga	729	187	83	0.2565	0.09
	Kisii	782	201	102	0.257	0.07
	Kisumu	936	436	189	0.4658	0.02
	Kitui	968	278	128	0.2872	0.13
	Koibatek	950	196	32	0.2063	0.13
	Kuria	899	209	112	0.2325	0.09
	Kwale	1100	259	84	0.2355	0.12
	Laikipia	702	127	19	0.1809	0.11
	Lamu	843	220	119	0.261	0.24
	Lugari	1004	398	125	0.3964	0.13
	Machakos	1274	543	179	0.4262	0.03
	Makueni	1263	573	224	0.4537	0.16
	Malindi	1012	147	58	0.1453	0.12
	Mandera	1110	153	48	0.1378	0.03
	Maragua	783	196	70	0.2503	0.15
	Marakwet	803	172	33	0.2142	0.06
	Marsabit	807	180	77	0.223	0.13
	Mbeere	827	287	99	0.347	0.12

Meru Central	808	292	119	0.3614	0.12
Meru South	758	322	117	0.4248	0.27
Migori	923	436	240	0.4724	0.09
Mombasa	1066	228	93	0.2139	0.05
Moyale	1223	301	108	0.2461	0.09
Mt. Elgon	1186	481	155	0.4056	0.10
Muranga	614	195	53	0.3176	0.18
Mwingi	999	258	102	0.2583	0.06
Nairobi	2554	599	182	0.2345	0.02
Nakuru	1570	309	81	0.1968	0.10
Nandi	978	293	119	0.2996	0.09
Narok	894	115	60	0.1286	0.07
Nyambene	946	367	130	0.3879	0.18
Nyamira	889	230	89	0.2587	0.10
Nyandarua	881	109	16	0.1237	0.08
Nyando	838	445	180	0.531	0.26
Nyeri	924	137	22	0.1483	0.03
Rachuonyo	893	525	260	0.5879	0.21
Samburu	904	270	93	0.2987	0.20
Siaya	836	425	206	0.5084	0.09
Suba	771	467	207	0.6057	0.15
Taita Taveta	716	248	100	0.3464	0.07
Tana River	1147	252	141	0.2197	0.07
Teso	965	467	170	0.4839	0.10
Tharaka	859	276	96	0.3213	0.08
Thika	980	270	68	0.2755	0.11
Trans Mara	866	131	85	0.1513	0.16
Trans Nzoia	1214	432	156	0.3558	0.02
Turkana	1100	640	274	0.5818	0.14
Uasin Gishu	1042	186	72	0.1785	0.10
Vihiga	945	482	148	0.5101	0.16
Wajir	1095	222	99	0.2027	0.06
West Pokot	894	151	84	0.1689	0.04

4.6 Data Processing and Analysis

4.6.1 Data Preparation and Processing

Data processing included a number of important steps to prepare the raw data for analysis. The initial steps in data processing included: selecting the variables and data that was needed from a larger database of KIHBS. The data was then transferred to MS Excel and all the entries were double checked to minimize human error. Once all the data needed for the study was compiled, data cleaning began. The first step was to ensure 100 percent verification using KIHBS database to resolve any discrepancies. Next, a series of consistency and range checks were used to identify any unreasonable responses. Out of the covariates that were available from the database, new covariates were formed from classifying the original covariates. Some of the covariates were dropped from the study depending on the previous knowledge of their effects on the response variable. Therefore, 14 covariates were used for this study.

4.6.2 Statistical Analysis

4.6.2.1 Screening of Variables

Data analysis started with exploratory analysis to screen the variables to be used for the Bayesian analysis. The following exploratory analyses were carried out: collinearity analysis to assess correlation between the covariates for the study; Exploring linear relationship using scatterplots to assess the relationship of the covariates with the response variable; univariate non-spatial binomial logistic regression to determine the covariates that have a statistical significant association with the response variable; Multivariate logistic regression to determine the covariates that have statistically significant association with the response variable collectively.

4.6.2.2 Model Description

Consider a set of binomial data y, which expresses the number of successes over $N_i, i = 1, \dots, n$.

Hence $y \sim \text{Binomial}(\pi, N_i)$, resulting to a likelihood given by

$$f(y|\pi) = \prod_{i=1}^{n} \left\{ \begin{bmatrix} N_i \\ y_i \end{bmatrix} \pi^{y_i} (1-\pi)^{N_i - y_i} \right\} = f(y|\pi) = \prod_{i=1}^{n} \left\{ \begin{bmatrix} N_i \\ y_i \end{bmatrix} \pi^{n y_i} (1-\pi)^{N_i - n y_i} \right\},$$

where $N = \sum_{i=1}^{n} N_i$ is the total number of the Bernoulli experiments in the sample. For a beta prior distribution with parameters $\theta = (a,b)'$, denoted by Beta(a,b) and density function

$$f(\pi) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}.$$

The resulting posterior is also a beta distribution since

$$\begin{aligned} & f(\pi \mid y) \propto f(y \mid \pi) f(\pi) \\ & \propto \prod_{i=1}^{n} \binom{N_i}{y_i} \pi^{n\bar{y}} (1-\pi)^{N-n\bar{y}} \times \frac{\tau(\alpha)\tau(\beta)}{\tau(\alpha+\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \\ & \propto \pi^{n\bar{y}+\alpha--1}(1-\pi)^{N--n\bar{y}+\beta-1} \end{aligned}$$

Thus $\pi \mid y \sim beta(n\bar{y}+\alpha, N-n\bar{y}+\beta)$, with the posterior parameter $\tilde{\alpha} = (n\bar{y} + a, N - n\bar{y} + b)^T$.

The posterior mean and variance are respectively:

$$E(\pi \mid y) = \bar{\mu}_\pi = \frac{n\bar{y}+\alpha}{N+\alpha+\beta}$$

and

$$V(\pi \mid y) = \bar{\delta}_\pi^2 = \frac{(n\bar{y}+\alpha)(N-n\bar{y}+\beta)}{(N+\alpha+\beta)^2(N+\alpha+\beta+1)}.$$

4.6.2.3 Semi-parametric Bayesian Regression Model

Estimation of the model parameters was carried out through the Markov Chain Monte Carlo (MCMC) simulation techniques as implemented in BayesX version 2.0.1 with 100,000 iterations and discarded the initial 5,000 samples, and subsequently stored every 10th iteration, giving 9,500 samples which were summarized for assessing convergence and parameter estimation.

Given a set of observations $(y_i, w_i), i = 1......n$, where y_i is a binary response such that $y_i = 1$ if a person had fever and $y_i = 0$ a person did not have fever, and $w_i = (w_{i1}...., w_{ip})$ are covariates. A logistic model to estimate the probability of getting fever, $y_i = 1$ versus the probability of not getting fever, $y_i = 0$ was implemented. The response is distributed as a Bernoulli random variable such that:

$$f(y_i \mid n_i) = p_i^{y_i}(1 - p_i)^{1-y_i} = \exp\left[y_i n_i - \log(1 + \exp(n_i))\right],$$

where $p_i = p(y_i = 1)$ p and $n_i = \log it(p_i)$ is a canonical parameter linked to the linear predictor $n_i = w_i^{'} y$. Here y is a p-dimensional vector of unknown regression coefficients.

Since the observations are associated with district of residence, it was desirable to account for geographical differences. District level effects were incorporated in the model to allow expected spatial correlation and any unstructured area heterogeneity of fever, using a convolution prior. Mean distance to health facility per district was specified to permit variations to occur by the distance. An assumption of additional flexibility in the predictor was made to allow for nonlinear covariate effects.

Therefore some predictor variables were extended to a more general semi parametric predictor.

$$n_i = v_i + h_i + f_i(x_i) + w'$$

where v_1, v_2 {1, · · ·, V} are spatially structured effects and h_1, h_2 {1, · · ·, H} model unstructured heterogeneity at district level. f_i are unknown functions for nonlinear effects of continuous covariate xi. Note that the spatially structured effects and unobserved heterogeneity tries to capture all sources of unmeasured influential factors, some that occur locally or at large scale, or those that may vary with time.

4.6.2.4 Prior Distributions for Covariate Effects and Assumptions

Modelling and inference uses the fully Bayesian approach. In the Bayesian formulation, the specification of the proposed model is complete by assigning priors to all unknown parameters. For the fixed regression parameters, a suitable choice is the diffuse prior, but a weakly informative Gaussian prior is also possible.

In a Bayesian approach, unknown functions $f_i, j = 1....p,$, f_{str}, f_{unstr} and parameters g as well as the variance parameter δ^2 are considered as random variables and have to be supplemented with appropriate prior assumptions. In the absence of any prior knowledge independent diffuse priors $y_j \propto const, j = 1...r,$ are assumed for the parameters of fixed effects.

The basic assumption behind the P-splines approach was that an unknown smooth function f of a particular covariate x could be approximated by a spline of degree l defined on a set of equally spaced knots $\xi_o = x_{\min} < \xi_1 < .... < \xi_{r-1} < \xi_r = x_{\min}$ within the domain of x. It is well known that such a spline can be written in terms of a linear combination of $m = r+l$ B-spline basis functions B_t, i.e.

$$f(x) = \sum_{t-1}^{m} \beta_t \beta_t(x)$$

The basic functions B_t are defined locally in the sense that they are nonzero only on a domain spanned by $2+l$ knots. The vector $b = (b_1,...., b_m)$ is unknown and must be estimated from the data. In a simple regression spline approach the unknown regression coefficients are estimated using standard methods for fixed effects parameters. However, a crucial point with simple regression splines is the choice of the number and the position of knots. For a small number of knots the resulting spline spacc may be not flexible enough to capture the variability of the data. For a large number of knots estimated curves may tend to over-fit the data. As a remedy to these problems Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent regression coefficients to guarantee sufficient smoothness of the fitted curves. In a Bayesian approach, we replace difference penalties by their stochastic analogues, i.e. first or second order random walk models for the regression coefficients

$$\beta_t = \beta_{t-1} + \mu_t$$
$$\beta_t = 2\beta_{t-1} - \beta_{t-2} - \mu_t,$$

with Gaussian errors $\mu_t \sim N(0,\tau^2)$ and diffuse priors $\beta_1 \propto Const,$or β_1 and $\beta_2 \propto Const,$ for initial values, respectively. A first order random walk penalizes abrupt jumps $\beta_t - \beta_{t-1}$ between successive states and a second order random walk penalizes deviations from the linear trend $2\beta_{t-1} - \beta_{t-2}$. Random walk priors may be equivalently defined in a more symmetric form by specifying the conditional distributions of parameters β_t given its left *and* right neighbors, e.g. β_{t-1} and β_{t+1} in the case of a first order random walk. Then, random walk priors may be interpreted in terms of locally polynomial fits. A first order random walk corresponds to a locally linear and a second order random walk to a locally quadratic fit to the nearest neighbors, see e.g. (Besag, *et al.*, 1995). The amount of smoothness is controlled by the additional variance parameter τ^2, which corresponds to the smoothing parameter in a frequentist approach. The larger (smaller) the variance, the rougher (smoother) is the estimated functions.

For the spatially correlated effect *$f_{str(s)}$, $s=1,\ldots,S$,* Markov random field priors are chosen common in spatial statistics (Besag, *et al.*, 1991). These priors reflect spatial neighborhood relationships. For geographical data one usually assumes that two sites or regions *s* and *r are* neighbors if they share a common boundary. Then a spatial extension of random walk models leads to the conditional, spatially autoregressive specification

$$f_{unstr}(s) \mid \tau^2_{unstr} \sim N(0, \tau^2_{unstr})$$

where *Ns* is the number of adjacent regions, and $r \in \delta_s$ denotes that region *r* is a neighbor of region *s*. Thus the (conditional) mean of *fstr(s)* is an average of function evaluations *fstr(s)* of neighboring regions. Again the variance τ^2_{str} controls the degree of smoothness. For a spatially uncorrelated (unstructured) effect f_{unstr} a common assumption is that the parameters *$f_{unstr}(s)$* are i.i.d. Gaussian

$$f_{unstr}(s) \mid \tau^2_{unstr} \sim N(0, \tau^2_{unstr}).$$

For a fully Bayesian analysis, variance or smoothness parameters $\tau^2_j, j = 1\ldots, p, str, unstr$, are also considered as unknown and estimated simultaneously with corresponding unknown functions *fj*. Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by highly dispersed inverse gamma distributions $p(\tau^2_j) \sim IG(\alpha_j, \beta_j)$ with known hyperparameters α_j, β_j.

4.6.2.5 Bayesian Geo-statistical Prediction

Spatial autocorrelation was estimated within a Bayesian framework based on a geostatistical model. The individual fever status is considered a binary outcome variable Y_i with $Y_i = 1$ for individuals with fever and 0 for non-fever individuals. The model assumed a conditional Bernoulli model for the binary outcome variable where the probability *p* of an individual *i* being infected, given the location *j* of the individual was:

$$Y_{i,j} \sim Bernoulli(p_{i,j})$$

$$logit(p_{i,j}) = \alpha + \sum_{k=1}^{p} \beta_k \times x_{i,j} + u_i$$

where $Y_{i,j}$ is the infectious status of an individual in location j, $p_{i,j}$ is the probability of an individual being a case in location j, α is the intercept, $x_{i,j}$ is a matrix of covariates, β is a vector of coefficients and u_i is a geostatistical random effect defined by an isotropic exponential spatial correlation function:

$$f(d_{ab}; \phi) = \exp\left[-(\phi d_{ab})\right],$$

where d_{ab} are the distances between pairs of points a and b, and ϕ is the rate of decline in the spatial correlation per unit distance. Non-informative priors were used for α (uniform prior with bounds $-\infty$ and ∞) and the coefficients (normal prior with mean = 0 and precision = 1×10^{-4}). The prior distribution of ϕ had a minimum of 1 and a maximum of 600, $\phi \sim dunif(1,600)$. The precision of μ_i was given a non-informative gamma distribution $(\tau \sim dgamma(1,0.05))$.

The prediction of the prevalence of infection was performed by *kriging* the geostatistical random effect and adding it to the sum of the products of the coefficients for the fixed effects and the values of the fixed effects at each prediction location. A burn-in of 5,000 iterations was used, followed by 100,000 iterations where values for the intercept, coefficients and predicted probability of fever at the prediction locations were stored. Diagnostic tests for convergence of the stored variables were undertaken by use of autocorrelation plots; convergence was successfully achieved after 100,000 iterations. The outputs of Bayesian models including parameter estimates and spatial prediction are termed posterior distributions.

Maps of the posterior distributions of predicted fever prevalence were developed in BayesX version 2.0.1. Samples of the posterior distributions of the coefficients from the model were used to produce prediction maps.

Effects of the covariates on fever were accessed using plots posterior distributions of the predicted values. The following covariates were plotted: Access to health facility, Parasite prevalence and Proportion of who slept under Treated Nets.

4.7 Limitations and Validity of the Study

Given that the study was based on secondary data from KIHBS, there were several limitations that were bound to arise inherent in variable selection biasness. Assuming districts are homogenous is one limitation but it was not possible to conduct cluster level analysis cluster coordinates were lacking in the data.

The study in based on the self-reported fevers and self-reported malaria fever data from KIHBS. This is likely to introduce recall error from the respondents and therefore may not represent the exact figures on the ground.

CHAPTER: FIVE

5 RESULTS AND DISCUSSION

5.1 Results

5.1.1 Exploratory Analyses

The following exploratory analyses were carried out.

1. A collinearity analysis to assess correlation between the covariates for the study.
2. Exploring linear relationship using scatterplots to assess the relationship of the covariates with the response variable.
3. A univariate non-spatial binomial logistic regression to determine the covariates that have a statistical significant association with the response variable.
4. A multivariate logistic regression to collectively determine the covariates that have statistically significant association with the response variable.

5.1.2 Collinearity Analysis between Covariates

Collinearity was assessed between all possible pairs of predictor variables, and if a correlation coefficient of greater than 0.9 was observed, the variable with the lowest AIC score was selected from the correlated covariates. The table below shows the correlation coefficients and AIC score of the correlated variables.

Table 5-1: Correlation Coefficients AIC Score of the Correlated Variables

Variables	Correlation coefficients	AIC
Proportion with no Toilet	0.88 = = 0.9	-58.20442
Proportion using Pit latrine		-58.56771

5.1.3 Modeling Linear Relationships

Logit transformation was carried out on the response variable then linear relationships with the predictor variables were assessed using scatterplots in R project. Univariate analysis was carried out for the predictor variables that showed a linear relationship with the response variable. See Appendix 1 for scatter plots.

5.1.4 Univariate Analysis

A univariate non-spatial binomial logistic regression model was carried out for each covariate with risk of fever *as* the dependent variable in R project. The results of the univariate analyses were used to determine the relative strength of each candidate covariate as a predictor of fever and identify those which qualified for inclusion in the Bayesian geostatistical model. All covariates significant at $P < 0.05$ were included in the multivariate analysis and the subsequent analysis.

The following variables showed a significant association with the response variable.

Table 5-2: Univariate Non-Spatial Binomial Logistic Regression Model Output

Variable	Descriptive Statistics				P value
	Minimum	Maximum	Mean	Std. Dev.	
Proportion of Male	0.45	0.53	0.49	0.02	<0.001
Proportion of Under 5s	0.09	0.20	0.15	0.02	
Proportion of ever attended school	0.27	0.87	0.71	0.14	
Proportion Diagnosed by Health Worker	0.02	0.26	0.12	0.05	
Proportion of Chronically ill	0.01	0.16	0.06	0.03	
Proportion who slept under treated net	0.00	0.47	0.21	0.12	
Proportion with protected source of drinking water	0.00	0.23	0.06	0.05	
Proportion with main cooking fuel as electricity/Gas	0.00	0.11	0.01	0.01	
Proportion with main cooking fuel as firewood	0.13	0.26	0.19	0.03	
Proportion with main lighting as electricity/Gas LPG	0.00	0.20	0.03	0.03	
Proportion with no Toilet	0.00	0.14	0.04	0.04	
Access to Health Facility	1.41	24.54	6.69	6.13	
Parasite Prevalence	0.00	0.49	0.09	0.13	

5.1.5 Multivariate Analysis

The selected covariates were then analyzed in a binomial multivariate logistic regression with response variable as the dependent variable. Using both backward and forward variable elimination, covariates with Wald's P > 0.2 were removed step-wise until a fully reduced model was achieved.

Table 5-3: Multivariate Analysis Non-Spatial Binomial Logistic Regression Model Output

	Odds	LL	UL	P Value
Intercept	0.03	0.01	0.07	< 0.001
Proportion of Male	4.70	1.40	15.82	0.01
Proportion of Under 5s	13.77	4.02	47.22	0.00
Proportion of ever attended school	1.23	0.96	1.58	0.10
Proportion Diagnosed by Health Worker	8.60	5.43	13.62	< 0.001
Proportion who slept under treated net	1.72	1.42	2.09	0.00
Proportion with main cooking fuel as firewood	19.18	6.90	53.31	0.00
Proportion with no Toilet	0.63	0.33	1.18	0.15
Access to Health Facility	1.01	1.01	1.02	0.00
Parasite Prevalence	3.27	2.72	3.94	< 0.001

5.1.6 Spatial-temporal Bayesian Generalized Linear Geostatistical Model

Estimation of the model parameters was carried out through the Markov Chain Monte Carlo (MCMC) simulation techniques as implemented in BayesX version 2.0.1 with 100,000 iterations and discarded the initial 5,000 samples, and subsequently stored every 10th iteration, giving 9,500 samples which were summarized for assessing convergence and parameter estimation.

Given a set of observations $(y_i, w_i), i = 1....,n,$ where y_i is a binary response such that $y_i = 1$ if a person had fever and $y_i = 0$ a person did not have fever, and $w_i = (w_{i1},.....,w_{ip})$ are covariates. A logistic model to estimate the probability of getting fever, $y_i = 1$ versus the probability of not getting fever, $y_i = 0$ was implemented. The response is distributed as a Bernoulli random variable such that: $f(y_i \mid n_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \exp\left[y_i n_i - \log(1+\exp(n_i))\right]$ where $p_i = p(y_i = 1)$ and $n_i = \log it(p_i)$ is a canonical parameter linked to the linear predictor $n_i = w'_i y$. Here y is a p-dimensional vector of unknown regression coefficients.

Since the observations are associated with district of residence, it was desirable to account for geographical differences. District level effects were incorporated in the model to allow expected spatial correlation and any unstructured area heterogeneity of fever, using a convolution prior. Mean distance to health facility per district was specified to permit variations to occur by the distance.

5.1.6.1 Autocorrelation Plots

Convergence of the algorithm is a term that refers to whether the algorithm has reached its equilibrium (target) distribution. If this is true, then the generated sample comes from the correct target distribution. Hence, monitoring the convergence of the algorithm is essential for producing results from the posterior distribution of interest. Monitoring autocorrelations is also very useful since low or high values indicate fast or slow convergence, respectively. If all values are within a zone without strong periodicities and (especially) tendencies, then it is assumed convergence. There are many ways to monitor convergence.

Autocorrelation plots and time trace plots were used to determine if the MCMC algorithm converged. Convergence was monitored by plotting autocorrelation plots of the samples. Quantiles, median, mean and standard deviation for all parameters, estimated from the posterior distributions, were used to assess model fit. In particular, credible intervals were used to assess the significance of parameters. From the autocorrelation plots in figure 5-1 below and specific autocorrelation plots for the covariates are in appendix 3 shows that that the convergence was achieved. Trace plots are plots of the iterations versus the generated values. If all values are within a zone without strong periodicities and (especially) tendencies, then convergence is assumed. Trace plots were also used to confirm convergence of MCMC algorithm and as shown in Appendix 4, the convergence was achieved.

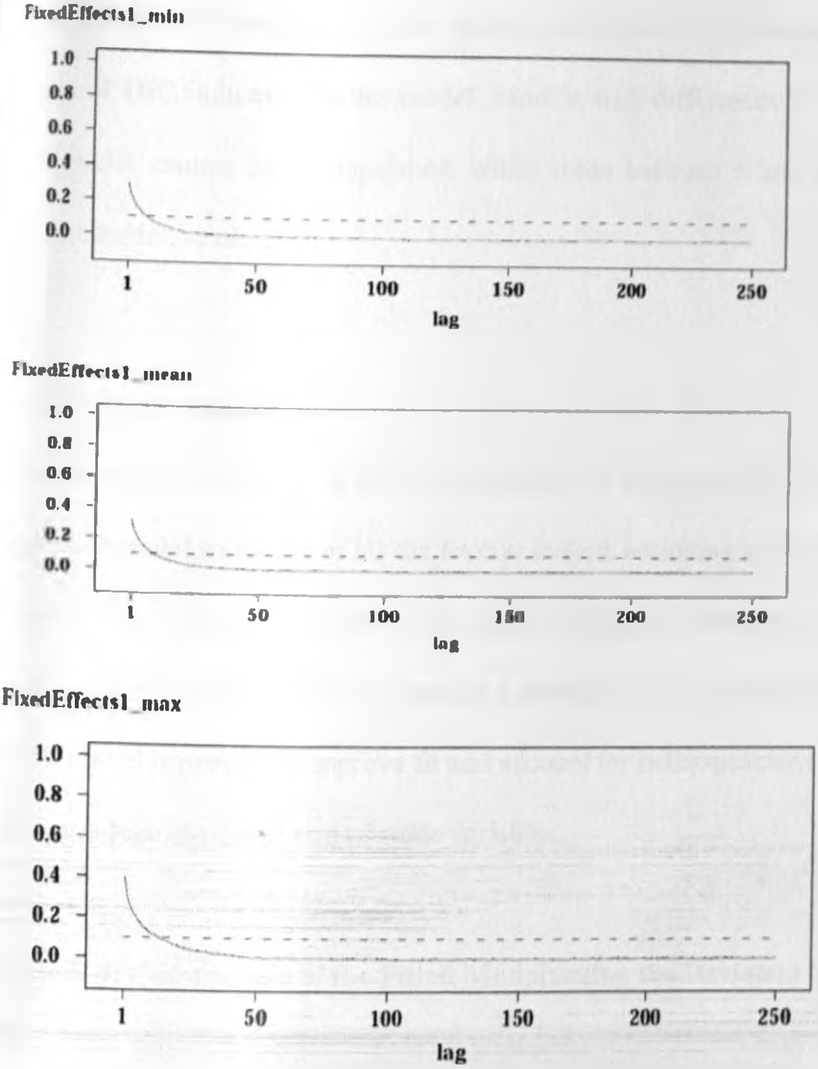


Figure 5-1: Autocorrelation Plots to test MCMC Algorithm Converged

Posterior deviance was monitored and a set of plausible models were compared using the Deviance Information Criterion (DIC) (Spiegelhalter, *et al.,* 2002). Specifically, model with all fever as a response variable and model with self-reported malaria fever as the response variable were compared. The two models had similar covariates. The DIC is given by $DIC = \bar{D} + pD$, where $\bar{D}$ is the posterior mean of the deviance, which is a measure of goodness of fit, and pD is the effective number of parameters, which is a measure of model complexity and penalizes over fitting. Since small values of $\bar{D}$ indicate good fit while small values of pD indicate a parsimonious model. small values of DIC indicate a better model. Models with differences in DIC of < 3 compared with the best model cannot be distinguished, while those between 3 and 7 can be weakly differentiated (Spiegelhalter, *et al.*, 2003).

5.1.6.2 Model Assessment:

Comparing the goodness of fit of two models, it was noted that the model of malaria was more preferred model to model of all the fevers. Indeed, assuming a semi-parametric model and random effects of the districts improved the models fitness. Evidently, modeling the impact of known factors alone is not sufficient to produce a satisfactory fit to the observations, and random effects at district level is needed to improve fit and account for heterogeneity and that the inclusion of random effects reduce the effect size of some variable .

Table 5-4: Comparison of the Fitted Models using the Deviance Information Criteria

	Models	
	Model of All Fevers	**Model of Malaria**
Model fit		
$\bar{D}$	21.124229	21.537995
pD	47.641519	47.20772
DIC	124.40727	115.95343
ΔDIC[$]	8.87297	

[$] Difference of the best fitting model against the other

5.1.6.3 Fixed Effects

Tables 5-5 and 5-6 below give posterior means standard deviation and 2.5%, 50% and 97.55 quintiles of the covariates. According to tables 5-5 and 5-6 below, shows that the increase of proportion of males decreases the fevers. Increase of proportion of under-fives increases the fever. Increase in the distance to the health facility and prevalence of the malaria parasites also increases the risk of fever. The other covariates that are proxy of the socio-economic status show that there increase of risk of fever with poor socio-economic status.

Table 5-5: Effect of Covariates on Fever

ariable	Mean	Std. Dev.	2.5% quant.	Median	97.5% quant.
onstant	0.31	0.64	-0.95	0.31	1.57
oportion of Male	-0.12	1.07	-2.25	-0.12	1.97
oportion of Under 5s	0.09	0.20	0.15	0.02	0.09
oportion of ever attended school	0.26	0.23	-0.19	0.26	0.70
oportion Diagnosed by Health Worker	0.75	0.42	-0.08	0.75	1.59
oportion with protected source of drinking ater	-0.09	0.66	-1.21	-0.09	1.38
oportion with firewood as main cooking fuel	0.60	2.83	-4.97	0.60	6.16
oportion who slept under treated net	0.76	0.69	-0.58	0.77	2.08
oportion with no Toilet	0.01	0.02	0.00	0.00	0.05
lean distance to Health Facility	0.01	0.03	0.00	0.00	0.00
lean Parasite Prevalence	0.01	0.03	0.00	0.00	0.05

Table 5-6: Effect of Covariates on Malaria Fever

ariable	Mean	Std. Dev.	2.5% quant.	Median	97.5% quant.
Constant	0.17	0.24	-0.29	0.17	0.64
Proportion of Male	-0.27	0.40	-1.06	-0.27	0.52
Proportion of Under 5s	0.02	0.41	-0.82	0.02	0.80
Proportion of ever attended school	0.11	0.08	-0.05	0.11	0.27
Proportion Diagnosed by Health Worker	0.55	0.16	0.24	0.55	0.87
Proportion with protected source of drinking water	-0.17	0.25	-0.67	-0.17	0.32
Proportion with firewood as main cooking fuel	-0.15	0.31	-0.76	-0.15	0.45
Proportion with no Toilet	0.55	0.23	0.09	0.56	1.01
Proportion who slept under treated net	0.0001	0.0002	0.0000	0.0000	0.0004
Mean distance to Health Facility	0.0000	0.0001	0.0000	0.0000	0.0002
Mean Parasite Prevalence	0.0001	0.0003	0.0000	0.0000	0.0003

5.1.6.4 Linear Effects

Figures 5-2 and 5-3 display linear effects of distance to the health facility on the risk is fever and self-reported malaria fever respectively. The effect of both all fevers and self-reported malaria fever is estimated to be almost linear, with the posterior means increasing with increasing all fevers and self-reported malaria fever. In other words the risk is lower for the people who are near the health facilities but increases for those who are further to the health facilities.

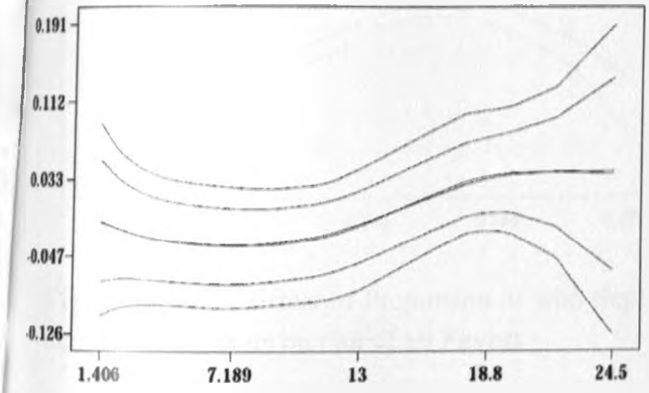


Figure 5-2: Effect of Access to Health Facility on the Risk of All Fevers in Kenya

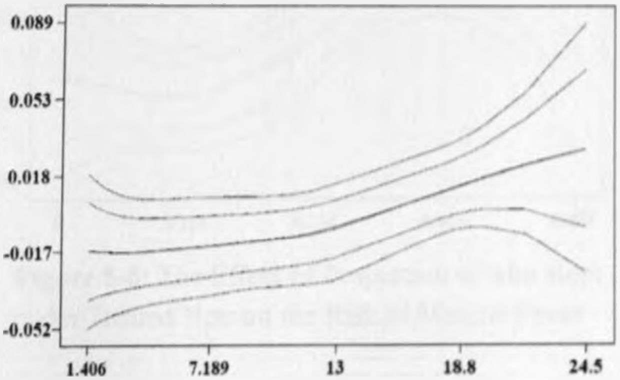


Figure 5-3: Effect of Access to Health Facility on the risk of Self-reported Malaria Fever in Kenya

Figures 5-4 and 5-5 also display linear effect malaria parasite prevalence on the risk of all fevers and self-reported malaria fever. This means that every unit increase in parasite prevalence increases the risks of all fevers and self-reported malaria fever by 0.01.

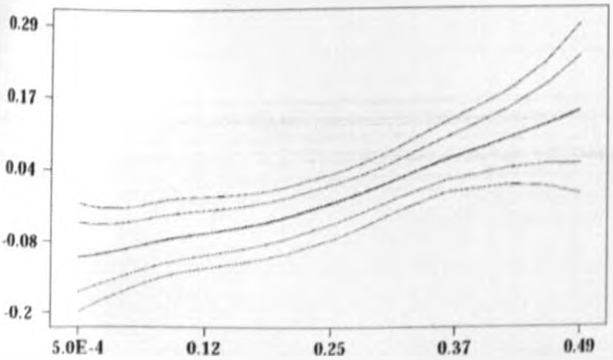


Figure 5-4: Effect of Parasite Prevalence on the Risk of All Fever in Kenya

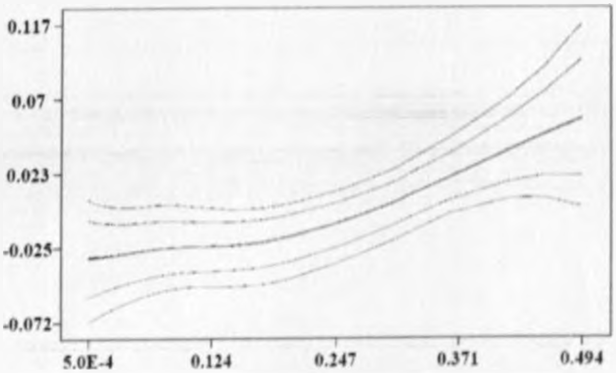


Figure 5-5: Effect of Parasite Prevalence on the Risk of Malaria Fever in Kenya

Figures 5-6 and 5-7, display the effect of proportion of people who slept under the net on the risk of all fever and malaria fever. In contrast to what is expected the figures below shows that increase of proportion of people who have treated nets increases the risk of all fevers and malaria fever by 0.76.

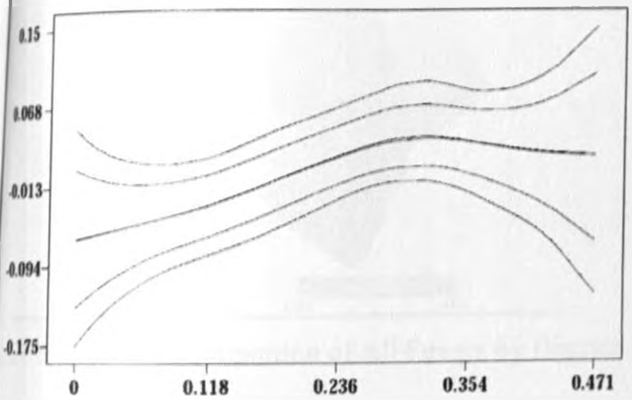


Figure 5-7: The Effect of Proportion of who slept under Treated Nets on risk of all Fevers

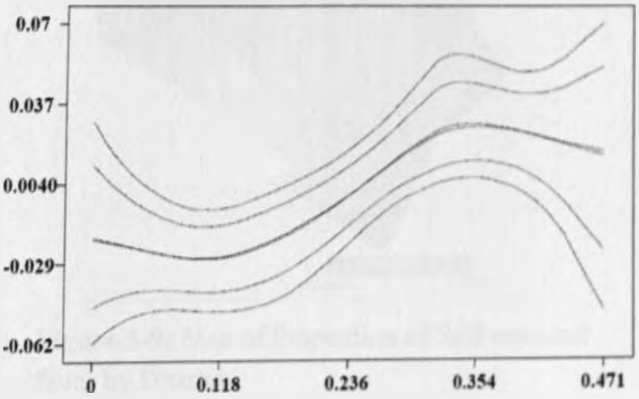
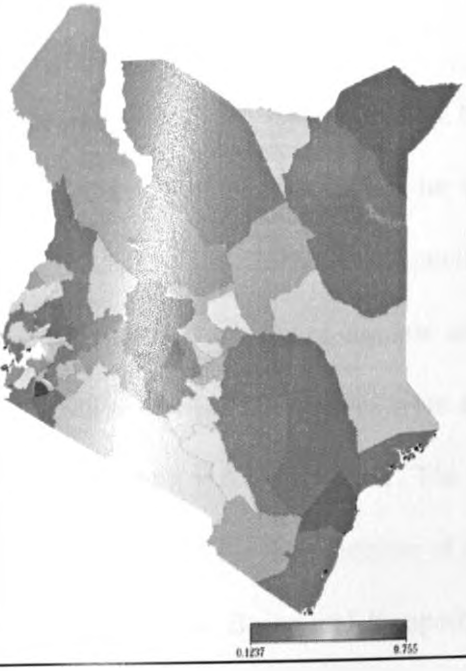


Figure 5-6: The Effect of Proportion of who slept under Treated Nets on the Risk of Malaria Fever

5.1.6.5 Spatial Effects

Figures 5-8 and 5-9 show the spatial effects on the risk of all fevers and malaria fever. There is evidence of spatial variation in risk of fever and self-reported malaria fever. It is clear those areas on the low lands like Nyanza province, which have high temperatures, report increased risk,



gure 5-8: Map of Proportion of All Fevers by District

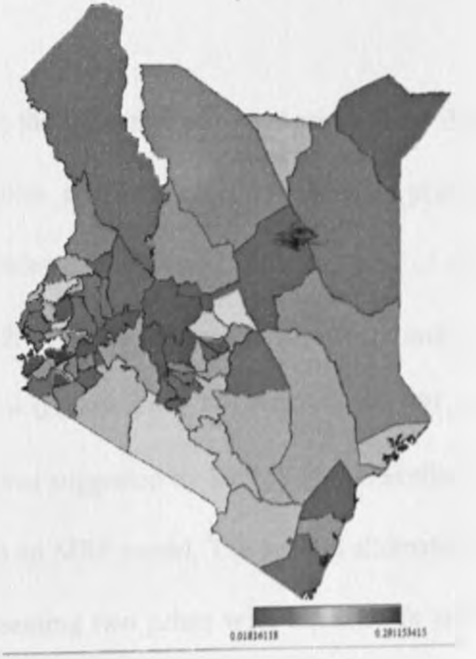


Figure 5-9: Map of Proportion of Self-reported fever by District

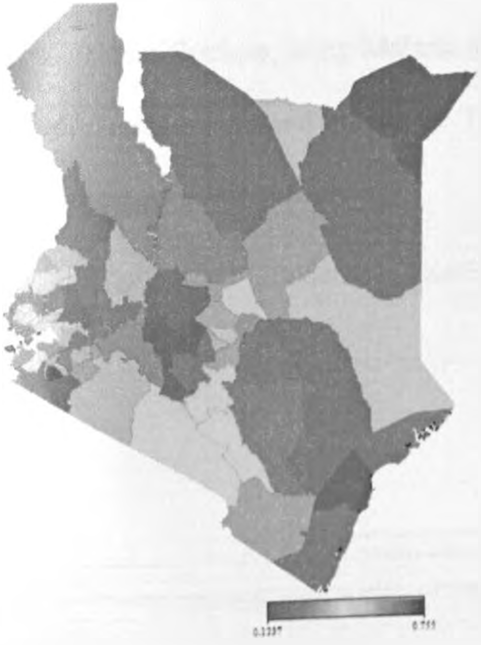


Figure 5-11: Predicted Mean of All Fevers by District

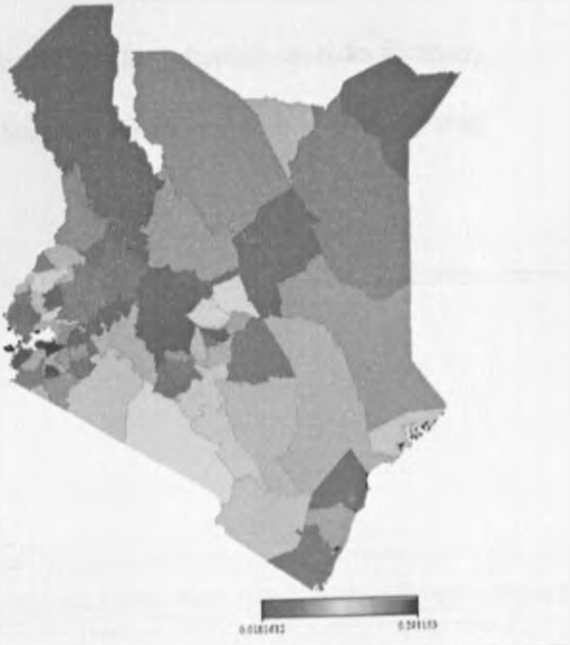


Figure 5-10: Predicted Mean of by Self-reported Malaria Fever District

5.1.6.6 Sensitivity Analysis

Tables 5-7 and 5-8 reports on the results investigating the influence of hyper-priors since the performance of the model can be sensitive to the choice of the variance components priors (Gelman A 2006). Alternative specifications were considered, and carried out sensitivity of all fevers and Malaria fever models assuming an IG with scale and shape parameters a and b respectively. Four alternatives were assumed: a = 0.5, b = 0.0005; a = 1, b = 0.005; a = 0.001, b = 0.001 and a = 0.01, b = 0.01. The first specification was suggested by Kelsall and Wakefield 1999, for modelling the precision of the spatial effects in an MRF model. The second alternative was proposed in Besag and Kooperberg 1995. The remaining two priors with equal scale and shape parameters, especially a = b = 0.001, have often been used as standard choice on the variances of random effects (Spiegelhalter, *et al*., 2003). Re-running MCMC simulations based on these specifications, using Malaria model, yield relatively similar inference on risks of fever, variance components and model fit. Therefore final choice of IG (a = 0.5, b = 0.0005) was appropriate for all the analyses.

Table 5-7: Sensitivity Analysis of all Fevers Model: Relative change of fixed effects, deviance information criterion, and variance component for different choices of hyper-parameter for $\tau_v^2 \tau_u^2 \tau_h^2$

	Hyperparameters for $\tau_v^2\tau_u^2\tau_h^2$			
Model fit	**a=0.5, b=0.0005**	**a =1, b =0.005**	**a = 0.001, b = 0.001**	**a = 0.01, b = 0.01**
$\overline{D}$	-128.52254	-263.172	-203.622	-230.705
pD	7.3640865	8.304392	-25.9086	-26.9373
DIC	-113.79436	-246.563	-255.44	-284.58
Fixed Effects				
Intercept	0.38 (-0.84,1.60)	0.19 (-0.27,0.66)	0.35 (-0.87,1.57)	0.31 (-0.94,1.56)
Proportion of Male	-0.40 (-2.41,1.62)	-0.35 (-1.12,0.42)	-0.27 (-2.31,1.76)	-0.15 (-2.23,1.93)
Proportion of Under 5s	-0.02 (-2.11,2.07)	0.05 (-0.75,0.84)	-0.20 (-2.31,1.91)	-0.34 (-2.49,1.81)
Proportion of ever attended school	0.22 (-0.19,0.64)	0.10 (-0.06,0.26)	0.24 (-0.19,0.66)	0.24 (-0.20,0.68)
Proportion Diagnosed by Health Worker	0.73 (-0.07,1.53)	0.52 (0.21,0.83)	0.72 (-0.11,1.55)	0.73 (-0.13,1.59)
Proportion with protected source of drinking water	0.02 (-1.24,1.27)	-0.20 (-0.69,0.28)	-0.01 (-1.30,1.29)	-0.05 (-1.37,1.27)
Proportion with firewood as main cooking fuel	0.07 (-1.42,1.55)	-0.06 (-0.64,0.52)	-0.03 (-1.59,1.52)	-0.09 (-1.70,1.53)
Proportion with main lighting as electricity/GAS	-0.95 (-2.81,0.91)	-0.18 (-0.90,0.53)	-0.94 (-2.83,0.96)	-0.85 (-2.79,1.08)
Proportion with no Toilet	0.56 (-0.58,1.70)	0.54 (0.11,0.97)	0.60 (-0.63,1.83)	0.68 (-0.60,1.96)
Proportion who slept under treated net	7.85E-05(-3E-04,5E-04)	2.17E-05 (-8E-05,1E-04)	2.72E-04 (-0.001,0.002)	7.28E-04(-0.002,0.003)
Mean distance to Health Facility	6.84E-05(-4E-04,5E-04)	1.03E-05 (-5E-05,7E-05)	2.53E-04 (-0.001,0.002)	7.70E-04(-0.003,0.004)
Mean Parasite Prevalence	6.88E-05(-5E-04,6E-04)	1.32E-05 (-8E-05,1E-04)	3.98E-04 (-0.004,0.005)	1.25E-03(-0.007,0.009)
Random Effects				
District: Structured	1.08E-03 (-0.01,0.01)	7.26E-04 (-0.01,0.01)	8.01E-03 (-0.02,0.036)	0.010149 (-0.017,0.038)
District: Unstructured	1.42E-03 (-0.01,0.01)	1.01E-03 (-0.00,0.01)	6.21E-03 (-0.01,0.02)	0.007186 (-0.003,0.018)

Table 5-8: Sensitivity Analysis of self-reported Malaria Fever Model: Relative change of fixed effects, deviance information criterion, and variance component for different choices of hyper-parameter for $\tau_v^2 \tau_u^2 \tau_h^2$

	Hyperparameters for $\tau_v^2\tau_u^2\tau_h^2$			
Model fit	**a=0.5, b=0.0005**	**a=0.5, b=0.0005**	**a=0.5, b=0.0005**	**a=0.5, b=0.0005**
$\overline{D}$	-263.172	-264.03381	-332.719	-365.169
pD	8.304392	13.726812	-18.5211	-20.7275
DIC	-246.563	-236.58019	-369.761	-406.624
Fixed Effects				
Intercept	0.19 (-0.27,0.66)	0.19 (-0.28,0.65)	0.17 (-0.29,0.64)	0.16 (-0.31,0.63)
Proportion of Male	-0.35 (-1.12,0.42)	-0.33 (-1.09,0.44)	-0.28 (-1.06,0.51)	-0.22 (-1.02,0.58)
Proportion of Under 5s	0.05 (-0.75,0.84)	0.01 (-0.79,0.81)	-0.01 (-0.82,0.79)	-0.07 (-0.88,0.74)
Proportion of ever attended school	0.10 (-0.06,0.26)	0.11 (-0.05,0.27)	0.11 (-0.05,0.27)	0.12 (-0.05,0.28)
Proportion Diagnosed by Health Worker	0.52 (0.21,0.83)	0.56 (0.25,0.88)	0.56 (0.24,0.88)	0.58 (0.26,0.91)
Proportion with protected source of drinking water	-0.20 (-0.69,0.28)	-0.18 (-0.67,0.31)	-0.17 (-0.67,0.32)	-0.14 (-0.65,0.36)
Proportion with firewood as main cooking fuel	-0.06 (-0.64,0.52)	-0.12 (-0.71,0.46)	-0.15 (-0.75,0.46)	-0.23 (-0.85,0.39)
Proportion with main lighting as electricity/GAS	-0.18 (-0.90,0.53)	-0.20 (-0.91,0.52)	-0.22 (-0.93,0.50)	-0.24 (-0.98,0.49)
Proportion with no Toilet	0.54 (0.11,0.97)	0.57 (0.13,1.01)	0.56 (0.10,1.02)	0.57 (0.09,1.05)
Proportion who slept under treated net	2.17E-05 (-8E-05,1E-04)	4.07E-05 (-7E-05,2E-04)	7.72E-05 (-3E-04,4E-04)	1.7E-04 (-4E-04,7E-04)
Mean distance to Health Facility	1.03E-05 (-5E-05,7E-05)	2.72E-05 (-6E-05,2E-04)	3.78E-05 (-2E-04,3E-04)	1.3E-04 (-6E-04,8E-04)
Mean Parasite Prevalence	1.32E-05 (-8E-05,1E-04)	2.99E-05 (-8E-05,2E-04)	5.48E-05 (-4E-04,5E-04)	1.59E-04(-6E-04,1E-03)
Random Effects				
District: Structured	1.09E-04 (-9E-04,0.001)	1.13E-04 (-8E-04,0.002)	8.26E-04 (-0.002,0.004)	1.30E-03 (-0.002,0.005)
District: Unstructured	0.00023 (-8E-04,0.001)	1.88E-04 (-7E-04,0.002)	9.08E-04 (-7E-04,0.003)	1.05E-03 (-4E-04,0.003)

5.2 Discussion

This study applied Bayesian techniques to analyze patterns and risk factors of fever. Logistic regression models was developed to have an in-depth understanding of factors associated with the probability of having fever, building on the existing methodological contributions by (Fahrmeir and Lang 2001) and (Fahrmeir, *et al.*, 2004).

A number of variables were used to explain the variation in the response and included spatial, continuous and heterogeneity terms. The spatially structured variation and unstructured heterogeneity were modeled using convolution prior and zero mean Gaussian heterogeneity priors as proposed by Besag, *et al*. The continuous variables are estimated non-parametrically by applying second order binomial random walk prior, which permits enough flexibility while avoiding over-fitting the data. The proposed methodology allowed all these factors to be estimated in a single framework. Because the models were highly parameterized and analytically intractable, the maximum likelihood approach was not feasible. Thus, the Bayesian inference, making use of MCMC simulation techniques, offered a viable alternative.

In this study it was found out that the risk of getting fever increased with increase of the distance to the health facility, parasite prevalence, proportion of under-fives, proportion diagnosed by health professional, chronically ill. However there was risk of fever increased with increase of the number of people who slept under the net until the relationship reaches a threshold where further increase of number of people who slept under the risk of fever decreases.

These results seem to suggest that when health care is accessible or available risk of fever goes down. Fever is a preventable disease, but delayed treatment or lack of effective treatment can lead severe cases or complications.

Children who are under the age of 5 are particularly vulnerable because of lack of immunity against the disease (Breman, *et al.,* 2004). The risk decreases with age. The increase in risk for those aged 6-14 years, although these are supposed to be protected through acquired immunity, may reflect some aspects of health seeking behavior, and emphasize the need for prompt and effective management of fever for all children including those aged over five years even if such cases may not frequently occur in the general population (Kazembe, *et al.,* 2008).

It is evident that treated bed net ownership alone is a poor indicator of fever control, and despite good distribution points in the country, does not translate into use and retreatment. Yet, usage and re-treatment are important indicators in the RBM campaign because these prevent contact with biting mosquitoes, and hence are critical to reducing infection and interruption of transmission.

The lower risk in the dry season should be interpreted with care. While the risk of infection is reduced during this period, this effect is directly linked to few cases being hospitalized, hence low number of fever. Another possible explanation is that during the dry season access to the hospital is easier than during rainy season, leading to early treatment, and therefore fewer avoidable fevers.

The spatial effects are often a surrogate of underlying unobserved information, and may give leads for further epidemiological research or assist in designing fever interventions. For example, the increased risk in rural areas may be an influence of different factors, such as unavailability or inaccessibility of health facilities resulting in increased risk for such children. These effects may also reflect health seeking behavior, which plays a critical role in accessing prompt and effective care. Scaling-up of interventions such as insecticide-treated nets or health promotions on appropriate and effective treatment in home or community based care should be emphasized in rural areas (World Health Organization 2004).

The data-driven approach we have taken in this analysis has a greater advantage in that the nonlinear effects of continuous variables are estimated, and avoids ad hoc categorizations. Indeed, the methodological framework applied provides useful tools for handling this type of data, and in similar conditions.

The application demonstrates that spatial and temporal analysis may reveal some salient features of the data, which may be overstepped by the classical regression. Flexible modeling, via nonparametric or semi-parametric model enabled to establish a better epidemiological relationship existing between the response and continuous explanatory variables.

A model diagnostic tool based on the posterior predictive distribution can be used to assess model adequacy by comparing the observed data with the samples drawn from the posterior predictive distribution.

In most African countries, most malaria cases occur at home, and the pattern may be biased towards urban areas that are well covered by health facilities. Moreover, one may argue that much of this data represent severe forms of fever, because studies on health seeking behavior for fever report that biomedical care is sought when the disease is nearly fatal (De Savigny, *et al.*, 2004). Health facility data can best be described as providing proxies for prevalence or morbidity and hence health need. A more representative data is through cross-sectional household surveys, e.g. the demographic and health surveys (DHS), however, these are often carried out every four years, thus the periodicity is not frequent enough for surveillance and to inform immediate decision making (De Savigny, *et al.*, 2004).

CHAPTER SIX

6 CONCLUSIONS AND RECOMMENDATIONS

In many resource-poor African countries, collection of population-based health data is a challenge and hospital data provide a critical source of information for decision making. This study set out to analyze risk factors of self-reported malaria and all fevers, using data from KHBS. This model, using the Bayesian approach, shows that risk of all self-reported fever and self-reported malaria fever is varied among gender, age, and socio-economic status. Fevers exhibit spatial variation.

From a public health perspective, with a goal of prevention and control, our results highlight that reducing malaria burden may require integrated strategies encompassing improved availability and access to health facilities; improving economic and social status and management of malaria parasite levels. Methodologically, this model can easily be adapted to analyze and compare other health indicator of similar structure and in like settings.

The model showed that there is a difference in the prevalence and spatial distribution of all self-reported fevers and self-reported malaria fever. This therefore translates to inefficient use of government funds in management of all fevers as Malaria fever.

The maps in this study provided a description of the geographic variation of self-reported malaria risk in Kenya,, and might help in the choice and design of interventions, which is crucial for reducing the burden of malaria in Kenya.

REFERENCES

Arnold, N., Thomas, A., Waller, L., & Conlon, E. (1999). Bayesian Models for Spatially Correlated Disease and Exposure Data. *Oxford University Press, USA*, *6*, p. 131.

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

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

Bodker, R., Akida, J., Shayo, D., Kisinza, W., Msangeni, H., Pedersen, E., et al. (2003). Relationship between altitude and intensity of malaria transmission in the Usambara Mountains, Tanzania. *Journal of medical entomology, 40*(5), 706-717.

Bornstein, D., Bredenberg, C., & Wood, W. (1963). Studies on the Pathogenesis of Fever. *The Journal of experimental medicine, 117*(3), 349.

Boyd, H., Flanders, W., Addiss, D., & Waller, L. (2005). Residual spatial correlation between geographically referenced observations: a Bayesian hierarchical modeling approach. *Epidemiology, 16*(4), 532.

Breman, J., Alilio, M., & Mills, A. (2004). Conquering the intolerable burden of malaria: what's new, what's needed: a summary. *The American journal of tropical medicine and hygiene, 71*(2 suppl), 1.

Brooker, S., Clements, A., & Bundy, D. (2006). Global epidemiology, ecology and control of soil-transmitted helminth infections. *Advances in parasitology, 62*, 221-261.

Carter, R., Mendis, K., & Roberts, D. (2000). Spatial targeting of interventions against malaria. *BULLETIN-WORLD HEALTH ORGANIZATION, 78*(12), 1401-1411.

Christensen, O., & Jr, P. R. (2002). geoRglm: A package for generalised linear spatial models. *R News, 2*(2), 26-28.

Clarke, S., Jukes, M., Njagi, J., Khasakhala, L., Cundill, B., Otido, J., et al. (2008). Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster-randomised, double-blind, placebo-controlled trial. *The Lancet, 372*(9633), 127-138.

Clements, A., Moyeed, R., & Brooker, S. (2006). Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa. *Parasitology, 133*(06), 711-719.

Corbett, J., Collis, S., Bush, B., Jeske, R., Martinez, R., Zermoglio, M., et al. (n.d.). *Almanac characterization tool. A resource base for characterizing the agricultural, natural, and human environments for select African countries. Texas Agricultural Experiment Station, Texas A\&M University System, Blackland Research and Extension Center.* Tech. rep., Report.

Cox, J., Craig, M., Sueur, D. L., & Sharp, B. (1999). Mapping malaria risk in the highlands of Africa.

Craig, M., Snow, R., & Sueur, D. L. (1999). A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitology Today, 15*(3), 105-110.

Cressie, N. (1993). Statistics for Spatial Data (Wiley Series in Probability and Statistics).

Cressie, N. (1988). Spatial prediction and ordinary kriging *Mathematical Geology, Springer, 20*, 405-421

Diggle, P., Jr, P. R., & Christensen, O. (2003). An introduction to model-based geostatistics. *Spatial statistics and computational methods, 173*, 43-86.

Diggle, P., Moyeed, R., Rowlingson, B., & 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.

Diggle, P., Tawn, J., Moyeed, R., WEBSTER, R., LAWSON, A., GLASBEY, C., et al. (1998). Model-based geostatistics. Discussion. Authors' reply. *Applied Statistics, 47*(3), 299-350.

Dinarello, C. (1996). Thermoregulation and the pathogenesis of fever. *Infectious disease clinics of North America, 10*(2), 433.

Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 89-102.

Emelda, O., & Robert, S. (n.d.). The relationship between reported fever and Plasmodium falciparum infection in African children. *Malaria Journal, 9.*

Fahrmeir, L., & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 50*(2), 201-220.

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica, 14*(3), 731-762.

Gelman, A. (2004). *Bayesian data analysis.* CRC press.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis, 1*(3), 515-533.

Gemperli, A., Vounatsou, P., Kleinschmidt, I., Bagayoko, M., Lengeler, C., & Smith, T. (2004). Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. *American journal of epidemiology, 159*(1), 64.

Gosoniu, L., Veta, A., & Vounatsou, P. (2010). Bayesian geostatistical modeling of malaria indicator survey data in Angola. *PloS one, 5*(3), e9322.

Gove, S. (1997). Integrated management of childhood illness by outpatient health workers: technical basis and overview. The WHO Working Group on Guidelines for Integrated Management of the Sick Child. *Bulletin of the World Health Organization, 75*(Suppl 1), 7.

Haining, R. (1993). *Spatial data analysis in the social and environmental sciences.* Cambridge Univ Pr.

Halstead, S. (1981). The Alexander D. Langmuir Lecture THE PATHOGENESIS OF DENGUE. *American journal of epidemiology, 114*(5), 632.

Hay, S., Guerra, C., Gething, P., Patil, A., Tatem, A., Noor, A., et al. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Medicine, 6*(3), e1000048.

Hamel, M., Odhacha, A., Roberts, J. & Deming, M. (2001). Malaria control in Bungoma District, Kenya: a survey of home treatment of children with fever, bedner use and attendance at antenatal clinics *Bulletin-World Health Organization, World Health Organisation, 79,* 1014-1023

Holtz, T., Marum, L., Mkandala, C., Chizani, N., Roberts, J., Macheso, A., et al. (2002). Insecticide-treated bednet use, anaemia, and malaria parasitaemia in Blantyre District, Malawi. *Tropical Medicine \& International Health, 7*(3), 220-230.

Kabatereine, N., Brooker, S., Tukahebwa, E., Kazibwe, F., & Onapa, A. (2004). Epidemiology and geography of Schistosoma mansoni in Uganda: implications for planning control. *Tropical Medicine \& International Health, 9*(3), 372-380.

Kallander, K., Nsungwa-Sabiiti, J., & Peterson, S. (2004). Symptom overlap for malaria and pneumonia--policy implications for home management strategies. *Acta Tropica, 90*(2), 211-214.

Kaya, S., Pultz, T., Mbogo, C., Beier, J., & Mushinzimana, E. (2002). The use of radar remote sensing for identifying environmental factors associated with malaria risk in coastal Kenya. *Citeseer*, (pp. 24-28).

Kazembe, L., Appleton, C., & Kleinschmidt, I. (2007). Choice of treatment for fever at household level in Malawi: examining spatial patterns. *Malaria Journal, 6*(1), 40.

Kazembe, L., Kleinschmidt, I., Holtz, T., & Sharp, B. (2006). International Journal of Health Geographics. *International Journal of Health Geographics, 5*, 41.

Kelsall, J., & Wakefield, J. (2002). Modeling spatial variation in disease risk. *Journal of the American Statistical Association, 97*(459), 692-701.

Kleinschmidt, I., Omumbo, J., Briet, O., De, N. V., Sogoba, N., Mensah, N., et al. (2001). An empirical malaria distribution map for West Africa. *Tropical Medicine \& International Health, 6*(10), 779-786.

Kleinschmidt, I., Sharp, B., Mueller, I., & Vounatsou, P. (2002). Rise in malaria incidence rates in South Africa: a small-area spatial analysis of variation in time trends. *American journal of epidemiology, 155*(3), 257.

Knorr-Held, L., & Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in medicine, 17*(18), 2045-2060.

Lawrence, K., Tobias, C., Jupiter, S., & Jimmy, N. (n.d.). Applications of Bayesian approach in modelling risk of malaria-related hospital mortality. *BMC Medical Research Methodology, 8*.

Meteopolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association, 44*(247), 335-341.

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

Neath, A., & Cavanaugh, J., (2010). Bayesian Estimation of Prediction Error and Variable Selection in Linear Regression *International Statistical Review, Wiley Online Library. 78*, 257-270

Noor, A., Gething, P., Alegana, V., Patil, A., Hay, S., Muchiri, E., et al. (2009). The risks of malaria infection in Kenya in 2009. *BMC infectious diseases, 9*(1), 180.

Now, S. (n.d.). STOP MALARIA NOW!

O'Meara, W.; Bejon, P.; Mwangi, T.; Okiro, E.; Peshu, N.; Snow, R.; Newton, C. & Marsh, K. (2008). Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya *The Lancet, Elsevier, 372*, 1555-1562

Omumbo, J., Hay, S., Snow, R., Tatem, A., & Rogers, D. (2005). Modelling malaria risk in East Africa at high-spatial resolution. *Tropical Medicine \& International Health, 10*(6), 557-566.

Omumbo, J., Ouma, J., Rapuoda, B., Craig, M., Lesueur, D., & Snow, R. (1998). Mapping malaria transmission intensity using geographical information systems GIS: an example from Kenya. *Annals of Tropical Medicine and Parasitology, 92*(1), 7-21.

Organization, W. H. (2006). *Guidelines for the treatment of malaria.* World Health Organization.

Orlando, Z., & Mikael, A. (n.d.). Mapping malaria incidence distribution that accounts for environmental factors in Maputo Province-Mozambique. *Malaria Journal, 9.*

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods.* Springer Verlag.

Robi, O., Noboru, M., Charles, M., Simon, H., & Robert, S. (n.d.). Distribution of the main malaria vectors in Kenya. *Malaria Journal, 9.*

Savigny, D. D., & Binka, F. (2004). Monitoring future impact on malaria burden in sub-Saharan Africa. *The American journal of tropical medicine and hygiene, 71*(2 suppl), 224.

Savigny, D. D., Mayombana, C., Mwageni, E., Masanja, H., Minhaj, A., Mkilindi, Y., et al. (2004). Care-seeking patterns for fatal malaria in Tanzania. *Malaria Journal, 3*(1), 27.

Schaffner, A. (2006). Fever--useful or noxious symptom that should be treated?]. *Therapeutische Umschau. Revue th{\'e}rapeutique, 63*(3), 185.

Snow, R., Gouws, E., Omumbo, J., Rapuoda, B., Craig, M., Tanser, F., et al. (1998). Models to predict the intensity of Plasmodium falciparum transmission: applications to the burden of disease in Kenya. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 92*(6), 601-606.

Soszynski, D. (n.d.). The pathogenesis and the adaptive value of fever. *order.*

Spiegelhalter, D., Best, N., Carlin, B., & Der, A. V. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 583-639.

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (1996). BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.*

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2004). WinBUGS user manual. *MRC Biostatistics Unit, Cambridge, UK, 2.*

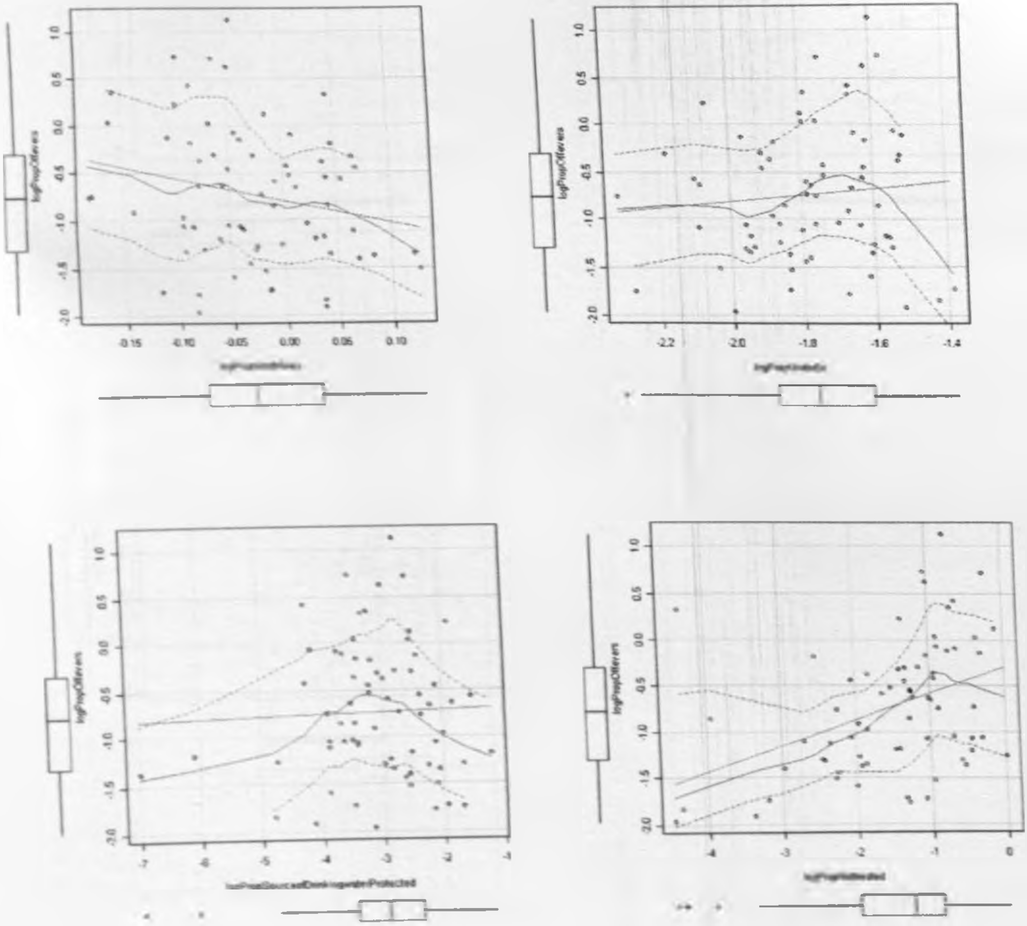
Team, R. (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria ISBN, 3*(10).

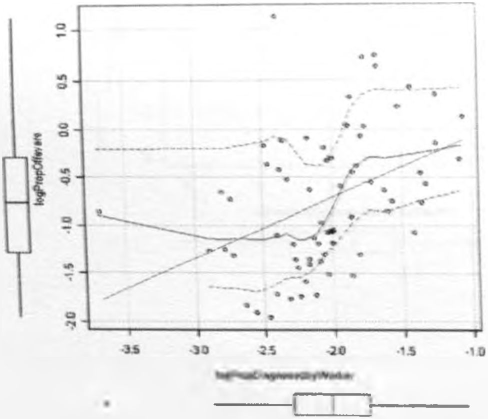
Thomson, M., Connor, S., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., et al. (1999). Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results. *The American journal of tropical medicine and hygiene, 61*(1), 2.

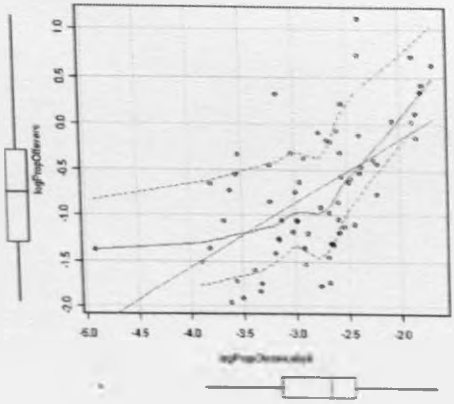
WHO, U. (2003). The Africa malaria report 2003. *Geneva: WHO, UNICEF.*

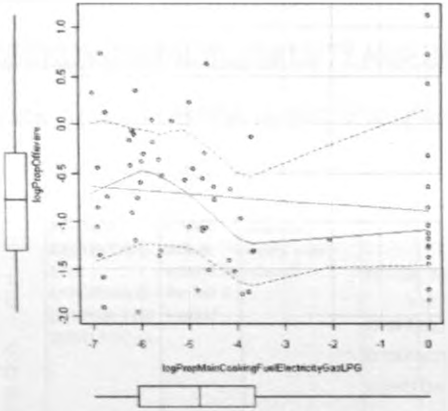
Zacarias, O., & Andersson, M. (2010). Mapping malaria incidence distribution that accounts for environmental factors in Maputo Province-Mozambique. *Malaria Journal, 9*(1), 79.

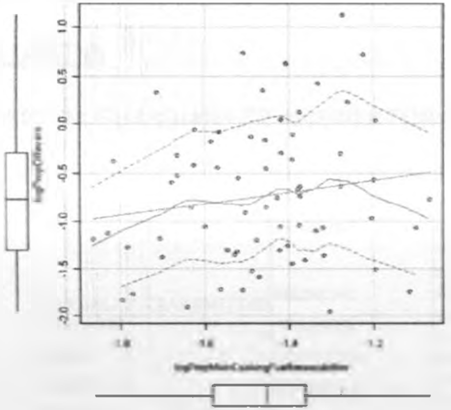
Appendix 1: Scatterplots for the Linear Relationships

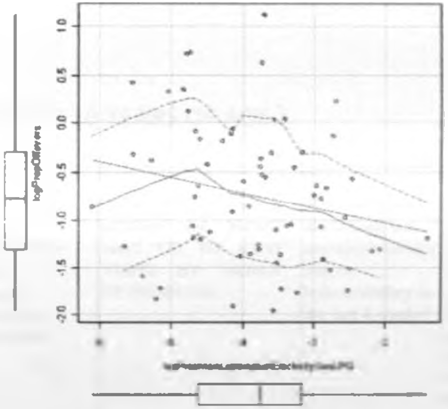


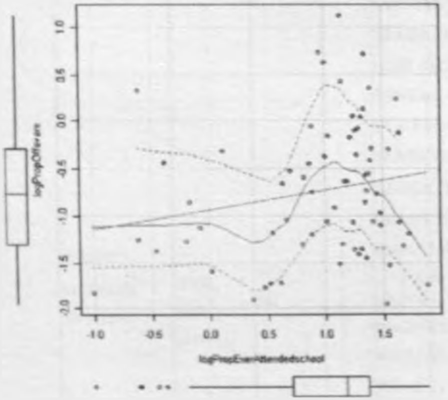

logPropOfferare
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
logPropDiagnosedbyWorker


logPropOfferare
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-5.0
-4.5
-4.0
-3.5
-3.0
-2.5
-2.0


logPropOffevers
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-7
-6
-5
-4
-3
-2
-1
0
logPropMainCookingFuelElectricityGasLPG


logPropOffevers
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-1.8
-1.6
-1.4
-1.2


logPropOffevers
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0


logPropOffevers
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-1.0
-0.5
0.0
0.5
1.0
1.5
logPropEverAttendedschool

Appendix 2: Kenya Integrated Household Budget Survey-Health Section Questionnaire

SECTION D: HEALTH, FERTILITY AND HOUSEHOLD DEATHS

[ASK OF ALL PERSONS IN THE HOUSEHOLD. MOTHERS OR GUARDIANS TO ANSWER FOR CHILDREN UNDER 10 YEARS OF AGE.]

D01	D02	D03	D04	D05		D06	D07	D08	D09	D10	D11
I D C O D E	ID NO. OF PERSON REPORTING THE INFORMATION FOR THE INDIVIDUAL	Was NAME sick or injured in the last 4 weeks?	Was NAME's sickness / injury work related?	What sort of sickness/injury did NAME suffer from?		Who diagnosed the illness?	How many days of work/school did NAME miss due to illness/injury in the last 4 weeks?	Did NAME consult a health provider on these sicknesses /injury in the last 4 weeks?	What kind of health provider did NAME visit? UP TO TWO VISITS BY ORDER OF PROBLEM.	How many times did NAME use any health service due to sickness/injury in the last 4 weeks?	Did NAME visit a health provider for any other health related reason (not sick) in the last 4 weeks?
	WRITE MEMBER ID CODE	YES... 1 NO ... 2 (→D11)	Yes. .....1 No 2	FEVER, MALARIA...................... 01 DIARRHEA STOMACH ACHE VOMITING UPPER RESPIRATORY(SINUSES) LOWER RESPIRATORY (CHEST, LUNGS 06 FLU ASTHMA HEADACHE SKIN PROBLEM DENTAL PROBLEM EYE PROBLEM EAR/NOSE/THROAT BACKACHE HEART PROBLEM BLOOD PRESSURE PAIN WHEN PASSING URINE DIABETES MENTAL DISORDER TB ....20 .... PROBLEM 1	SEXUALLY TRANSMITTED DISEASE21 BURN FRACTURE WOUND POISONING PREGNANCY RELATED UNSPECIFIED LONG-TERM ILLNESS27 HIV/AIDS TYPHOID OTHER (SPECIFY) PROBLEM 2	MEDICAL WORKER (DOCTOR, CLINICAL OFFICER, NURSE) AT HOSPITAL MEDICAL WORKER AT OTHER HEALTH FACILITY TRADITIONAL HEALER NON-HH MEMBER (NOT MEDICAL) HH MEMBER SELF HERBALIST FAITH HEALER OTHERS (specify) 9	FOR PERSONS 3YRS AND ABOVE N/A. 99 DAYS	YES. .... 1 NO.......... 2 [IF NO→D11)	REFERAL HOSPITAL DISTRICT/PROVINCIAL /HOSPITAL PUBLIC DISPENSARY PUBLIC HEALTH CENTER PRIVATE DISPENSARY/ HOSPITAL PRIVATE CLINIC TRADITIONAL HEALER MISSIONARY HOSP./DISP PHARMACY/CHEMIST KIOSK FAITH HEALER HERBALIST OTHER (SPECIFY) PROBLEM 1	NUMBER PROBLEM 2	YES. ..... 1. NO ....2. [IF NO →D13)

HEALTH

D01 ID CODE	D12 What kind of health provider did NAME visit? UP TO TWO PROVIDERS BY ORDER OF VISITS. REFERAL HOSPITAL DISTRICT/PROVINCIAL /HOSPIT02AL PUBLIC DISPENSARY PUBLIC 03 HEALTH CENTER PRIVATE DISPENSARY/ HOSPI 05 PRIVATE CLINIC TRADITIONAL HEALER MISSIONARY HOSP./DISP PHARMACY/CHEMIST KIOSK FAITH HEALER		D13 During the last 12 months, was NAME hospitalized or had an overnight stay(s) in a medical facility? YES. . . . . .1. NO . . . . .2 (IF NO»D16)	D14 Did NAME or other members of household have to borrow money inorder to pay for hospitalization ? YES.......... 1 NO.......... 2	D15 Did NAME or other members of household have to sell assets in order to pay for hospital-ization?	D16 During the last 12 months, did NAME stay over-night at a traditional healer's , herbalist or faith healer's dwelling?	D17 Did NAME or other members of household have to borrow money in order to pay for traditional healer, herbalist or faith healer?	D18 Did NAME or other members of household have to sell assets in order to pay for traditional healer, herbalist or faith healer?	D19 Is NAME physically handicapped in any way which limits or prevents activities or work? YES. 1. NO . . . . 2 (IF NO »D25)	D20 Was NAME's handicap work related? YES____________1 NO____________2 (IF NO »D24)	D21 Was NAME compensated for handicap? YES. . . . . .1 NO . . . . .2 (IF NO »D24)	D22 Was NAME compensated under any of the following? WORKMAN'S COMPENSATION OWN INSURANCE COVER.......2 OTHER COMPENSATION...... 3 EMPLOYER ARRANGEMENT...4	D23 How much did NAME receive in compensation for the handicap? KSHS.
	PROVIDER	PROVI											
1													
2													
3													

D01 ID CODE	D24 In what way(s) is NAME handicapped? MISSING HAND 1 MISSING FOOT 2 LAME 3 BLIND 4 DEAF 5 UNABLE TO SPEAK (DUMB) 6 MENTALLY DISABLED		D25 If NAME had to sweep the floor of the house, could he/she do so easily, with difficulty, or not at all? EASILY 1 WITH DIFFICULTY 2 NOT AT ALL 3	D26 If NAME had to walk for 2 kilometers on a flat path, could he/she do so easily, with difficulty, or not at all? EASILY 1 WITH DIFFICULTY 2 NOT AT ALL 3	D27 Does NAME suffer from a chronic illness? YES. . . . . .1. NO .2.	D28 What chronic illness does NAME suffer from? LIST UP TO 2. CHRONIC MALARIA/FEVER01 TUBERCULOSIS02 HIV/AIDS 03 STDs 04 DIABETES 05 ASTHMA 06 BILHARZIA/SCHISTOSOMIASIS 07 ARTHRITIS/RHEUMATISM 08 NERVE DISORDER 09 STOMACH DISORDER 10 SORES THAT DO NOT HEAL . . . . .1.1. CANCER. . . . . .12 PNEUMONIA	D29 How long has NAME suffered from this illness (these illnesses)? DO NOT KNOW 98 NOT STATED 99 . . . . 11 3 9		D30 Who diagnosed NAME's chronic illness? MEDICAL WORKER (DOCTOR, CLINICAL OFFICER, NURSE) AT HOSPITAL 1 MEDICAL WORKER AT OTHER HEALTH FACILITY 2 TRADITIONAL HEALER 3 NON-HH MEMBER (NOT MEDICAL)	D31 Did NAME sleep under a bed net to protect against mosquitos last night?	D32 Have the bed nets(s) ever been treated with insecticide to protect against mosquitos in the past six months?
	FIRST	SECOND			(»D31) ILLNESS 1	ILLNESS 2	YEARS	MONTHS			
1											
2											
3											

FERTILITY

D01	D33	D34	D35		D36		D37		D38		D39		D40	D41
ID CODE	PUT A '1' FOR ALL FEMALES WHO ARE AGED LESS THAN 12 YRS AND MORE THAN 49YRS AND ALL MALES, OTHERWISE CODE 2. DO NOT ADMINISTER THIS MODULE TO ALL INDIVIDUALS CODED 1.	Has NAME ever given birth to live births? YES 1 NO 2 (IF NO →D42)	How many children have you borne alive?		How many children has NAME borne alive who usually live in the household		How many children has NAME borne alive who usually live elsewhere		How many children has NAME borne alive who have died?		When was NAME's last child born?		Sex of last child(ren) born MALE 1 FEMALE 2 MALE TWINS 3 FEMALE TWINS 4 MULTIPLE BIRTHS 5 MALE - FEMALE TWINS 6	Is this last born child(ren) still alive? YES 1 NO 2 ONE OF THE TWINS 3 TWO OF THE MULTIPLE BIRTHS 4 ONE OF THE MULTIPLE BIRTHS 5 DK 8
			MALES	FEMALES	MALES	FEMALES	MALES	FEMALES	MALES	FEMALES	MONTH	YEAR		
1														
2														
3														

Sneak

Snick

DEATHS IN THE HOUSEHOLD

D01	D42	D43	D44		D45	D46
I D CODE	In the last 24 months has any household member died? (ask HH head or any other responsible member)	Sex of person who died	Age of person who died		Cause of Death	Where did NAME die ?
	YES 1 NO 2 (NEXT SECTION)	MALE 1 FEMALE 2	OVER 97 YEARS 97 DON'T KNOW 98 NOT STATED 99		MALARIA 01 PNEUMONIA 02 AIDS 03 TETANUS 04 TUBERCULOSIS 05 MALNUTRITION 06 ANAEMIA 07 CHILD BIRTH/PREGNANCY 08 SUDDEN DEATH 09 ASTHMA 10 CANCER 11 URINARY OBSTRUCTION 12 POISONING 13 SUICIDE 14 ACCIDENT 15 MEASELS 16 OTHERS SPECIFY 17	HOME 1 HEALTH FACILITY 2 OTHERS SPECIFY 3
			YEARS	MONTHS		
1						
2						
3						

SECTION D: HEALTH, FERTILITY AND HOUSEHOLD DEATHS

DEATHS IN THE HOUSEHOLD

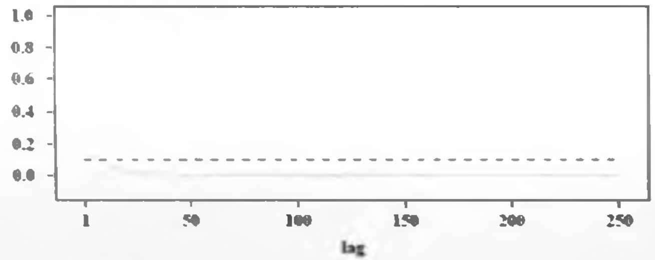
D42	D43	D44		D45	D46
In the last 24 months has any household member died? (ask HH head or any other responsible member)	Sex of person who died	Age of person who died		Cause of Death	Where did NAME die ?
YES 1 NO 2 (NEXT SECTION)	MALE 1 FEMALE 2	OVER 97 YEARS 97 DON'T KNOW 98 NOT STATED 99		MALARIA 01 PNEUMONIA 02 AIDS 03 TETANUS 04 TUBERCULOSIS 05 MALNUTRITION 06 ANAEMIA 07 CHILD BIRTH/PREGNANCY 08 SUDDEN DEATH 09 ASTHMA 10 CANCER 11 URINARY OBSTRUCTION 12 POISONING 13 SUICIDE 14 ACCIDENT 15 MEASELS 16 OTHERS SPECIFY 17	HOME 1 HEALTH FACILITY 2 OTHERS SPECIFY 3
		YEARS	MONTHS		

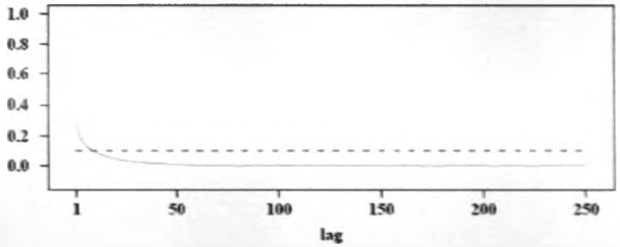
SECTION E: LABOUR

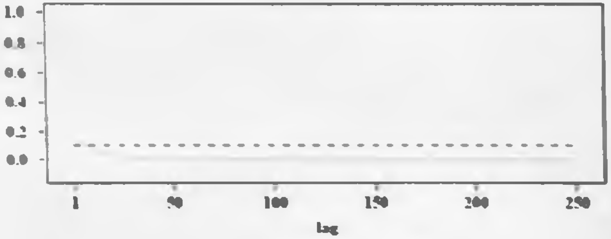
[ASK ALL HOUSEHOLD MEMBERS AGED 5 YEARS AND OLDER.] IF DID NOT DO TASK, WRITE ZERO; LESS THAN 1/2 HOUR, WRITE '0.5';

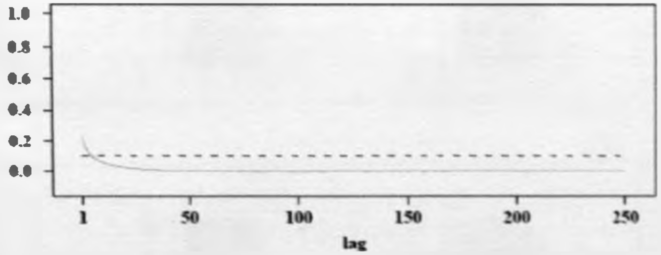
E01	E 0 2	E03	E 0 4	E05	E06	E07	E08	E09	E10	E11
I D C O D E	PUT CODE '1' FOR ALL INDIVIDUALS WHO ARE AGED LESS THAN 5 YEARS OTHERWISE CODE 2 DO NOT ADMINISTER THIS MODULE TO THE INDIVIDUAL CODED 1.	What was NAME mainly doing in the past 7 days? WORKED FOR PAY 01 ON LEAVE 02 SICK LEAVE 03 WORKED ON OWN/FAMILY BUSINESS 04 WORKED ON OWN/FAMILY AGRI.HOLDI 05 SEEKING WORK 06 (» E09) DOING NOTHING 07 (» E09) RETIRED 08 (» E09) HOMEMAKER 09 (» E09) FULL-TIME STUDENT 10 (» E14) IN CAPACITATED 11 (» E14) OTHER (SPECIFY) 12	Status in employment (Main) PAID EMPLOYEE WORKING EMPLOYER OWN-ACCOUNT WORKE 3 UNPAID FAMILY WORKER 4 APPRENTICE OTHER (SPECIFY	During the past 7 days, how many hours was NAME employed for a wage, salary, commission or any payment in kind? HOURS	During the past 7 days, how many hours did NAME work on the household farm, in a field or herding livestock? HOURS	During the past 7 days, how many hours did NAME work on any enterprise belonging to a member of household, including helping for no pay? HOURS	REVIEW QUESTIONS E05, E06 E07: DID THE RESPONDENT WORK FOR ANY HOURS IN ANY OF THESE TASKS DURING THE LAST 7 DAYS? YES (»E13) NO	Even though NAME did not do any income earning activities in the last 7 days, does NAME have a job, business, or other economic or farming activity to return to? YES 1 (»E13) NO 2	What is the main reason NAME was not working during the last 7 days? SICK 01 RETIRED 02 LOOKING FOR WORK 03 OUT OF SEASON 04 RETRENCHMENT/REDUNDANCY 05 TEMPORARY LAY OFF 06 DONT NEED WORK 07 BUSINESS CLOSED 08 TOO YOUNG/TOO OLD 09 OTHERS	In the past 4 weeks has NAME taken any action to look for any kind of work or start any kind of business/income generating activity? YES 1 NO 2
1										
2										
3										

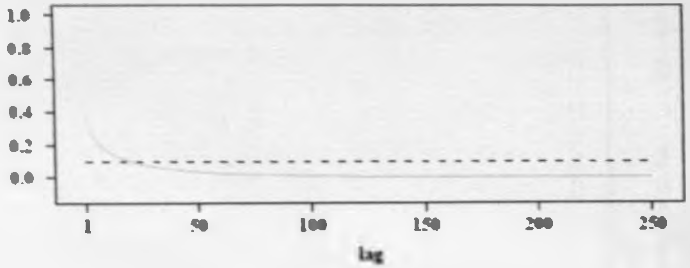
Appendix 3: Autocorrelation Plots

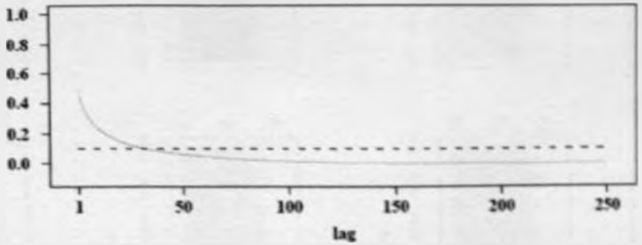

f_PropNettreated_pspline_mean


MaintoiletfacilityNotoilet_pspline_mean


ccessToHealthFacilityMean_pspline_mean

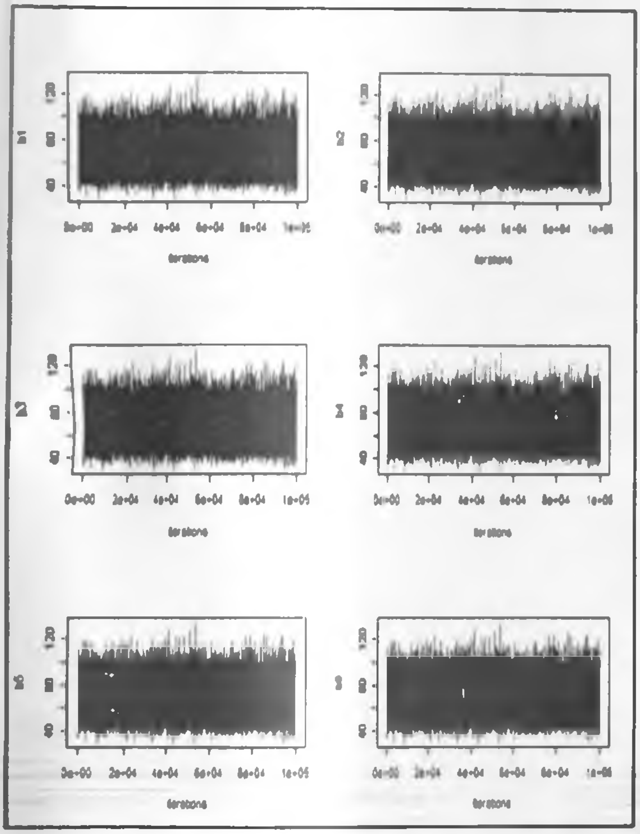

ParasitePrevalenceMean_pspline_mean


f_POLYID_spatial_mean

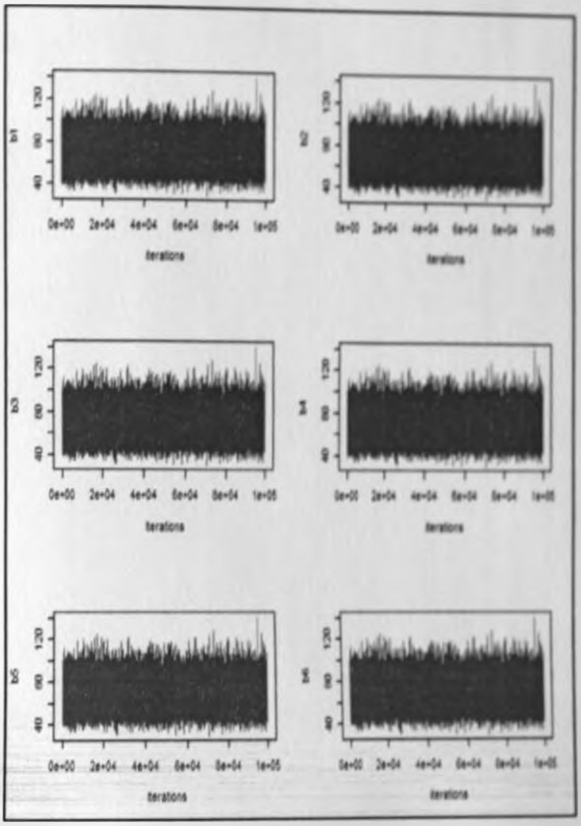

f_POLYID_random_mean

Appendix 4: Trace Plots MCMC Algorithm

Trace Plot for All Self-reported Fever Model



Trace Plot for Self-reported Malaria Model


b1
b2
b3
b4
b5
b6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
iterations