UNIVERSITY OF KWAZULU-NATAL

USE OF STATISTICAL MODELLING AND ANALYSES OF MALARIA RAPID DIAGNOSTIC TEST OUTCOME IN ETHIOPIA

2013

DAWIT GETNET AYELE

USE OF STATISTICAL MODELLING AND ANALYSES OF MALARIA RAPID DIAGNOSTIC TEST OUTCOME IN ETHIOPIA

By

DAWIT GETNET AYELE

(M.Sc. in Statistics)

Submitted in fulfilment of the academic
requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

in the

School of Mathematics, Statistics and Computer Science
University of KwaZulu–Natal
Pietermaritzburg
2013

Dedication

This thesis is dedicated to

My mother W/ro Ejigayehu Bereded
and
Grand mother Mulunesh Yeshaw

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Professor Temesgen Zewotir and Professor Henry Mwambi.

I, Dawit Getnet Ayele, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

August, 2013

Mr Dawit Getnet Ayele Date

Professor Temesgen Zewotir Date

Professor Henry Mwambi Date

Note

The following papers have been published from this thesis.

1. Ayele DG, Zewotir T, Mwambi H: Prevalence and risk factors of malaria in Ethiopia. *Malaria Journal* 2012, 11:195 doi:10.1186/1475-2875-1111-1195.

2. Ayele DG, Zewotir T, Mwambi H: The risk factor indicators of malaria in Ethiopia *International Journal of Medicine and Medical Sciences* 2013, 5(7):doi:10.5897/IJMMS2013.0956 5335-5347

3. Ayele DG, Zewotir T, Mwambi H: Spatial distribution of malaria problem in three regions of Ethiopia. *Malaria Journal* 2013, 12:207:doi:10.1186/ 1475-2875-1112-1207.

Acknowledgement

"BLESSED BE GOD, WHICH HATH NOT TURNED AWAY MY PRAYER, NOR HIS MERCY FROM ME." PSALMS 66:20

Before everything, I would like to give thanks to the Almighty God, who makes everything possible.

First of all I would like to thank my supervisors, Professor Temesgen Zewotir and Professor Henry Mwambi for their direction, advice, patience and encouragement throughout the research work. I would like to thank the staff members and postgraduate students in the School of Mathematics, Statistics and Computer Sciences for the hospitable environment they provided during my study. I would like to thank Dr. Solomon Hailu for editing the final manuscript.

I would, with deep appreciation, like to thank College of Agriculture, Engineering and Science for providing me the financial support over the years of my study. I would like to thank Amhara Regional Health Branch and The Carter Center, for providing me the data for this study.

I want to thank my family for their constant support all through my life; no achievement in my life would have been possible without them. I am forever indebted to my mother Ejigayehu Berededed. Her love, encouragement, and sacrifice at every step of the way have made me what I am today. I can only hope to one day to be as honest, loving, and hardworking parent as she is. I also express a feeling of bliss for having a loving sister like Lulit Getnet. She has always been sister who is genuinely concerned and ever encouraging.

Finally, everyone who contributed to the success of my study either directly or indirectly though not mentioned by name deserves my thanks.

Abstract

The transmission of malaria is among the leading public health problems in Ethiopia. From the total area of Ethiopia, more than 75% is malarious. Identifying the infectiousness of malaria by socio-economic, demographic and geographic risk factors based on the malaria rapid diagnosis test (RDT) survey results has several advantages for planning, monitoring and controlling, and eventual malaria eradication effort. Such a study requires thorough understanding of the diseases process and associated factors. However such studies are limited. Therefore, the aim of this study was to use different statistical tools suitable to identify socio-economic, demographic and geographic risk factors of malaria based on the malaria rapid diagnosis test (RDT) survey results in Ethiopia. A total of 224 clusters of about 25 households were selected from the Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia. Accordingly, a number of binary response statistical analysis models were used. Multiple correspondence analysis was carried out to identify the association among socio-economic, demographic and geographic factors. Moreover a number of binary response models such as survey logistic, GLMM, GLMM with spatial correlation, joint models and semi-parametric models were applied. To test and investigate how well the observed malaria RDT result, use of mosquito nets and use of indoor residual spray data fit the expectations of the model, Rasch model was used. The fitted models have their own strengths and weaknesses. Application of these models was carried out by analysing data on malaria RDT result. The data used in this study, which was conducted from December 2006 to January 2007 by The Carter Center, is from baseline malaria indicator survey in Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia.

The correspondence analysis and survey logistic regression model was used to identify predictors which affect malaria RDT results. The effect of identified socio-economic, demographic and geographic factors were subsequently explored by fitting a generalized linear mixed model (GLMM), i.e., to assess the covariance structures of the random components (to assess the association structure of the data). To examine whether the data displayed any spatial autocorrelation, i.e., whether surveys that are near in space have malaria prevalence or incidence that is similar to the surveys that are far apart, spatial statistics analysis was performed. This was done by introducing spatial autocorrelation structure in GLMM. Moreover, the customary two variables joint modelling approach was extended to three variables joint effect by exploring the joint effect of malaria RDT result, use of mosquito nets and indoor residual spray in the last twelve months. Assessing the association between these outcomes was also of interest. Furthermore, the relationships between the response and some confounding covariates may have unknown functional form. This led to proposing the use of semiparametric additive models which are less restrictive in their specification. Therefore, generalized additive mixed models were used to model the effect of age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed nonparametrically. The result from the study suggests that with the correct use of mosquito nets, indoor residual spraying and other preventative measures, coupled with factors such as the number of rooms in a house, are associated with a decrease in the incidence of malaria as determined by the RDT. However, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria. In order to determine whether or not the limited number of respondents had undue influence on the malaria RDT result, a Rasch model was used. The result shows that none of the responses had such influences. Therefore, application of the Rasch model has supported the viability of the total sixteen (socio-economic, demographic and geographic) items for measuring malaria RDT result, use of indoor residual spray and use of mosquito nets. From the analysis it can be seen that the scale shows high reliability. Hence, the result from Rasch model supports the analysis carried out in previous models.

Table of contents

List of tables

List of Figures

Acronyms

AM	Additive model
BRR	Balanced repeated replication
CA	correspondence analysis
CDTI	community-directed treatment with ivermectin
CML	conditional maximum likelihood
CRS	Cubic regression spline
CSA	Central Statistical Agency
DIF	Differential item functioning
DPQL	Double penalized quasi-likelihood
DIF	Differential item functioning
EA	Enumeration area
FMOH	Federal Ministry of Ethiopia
GAM	Generalized additive model
GAMM	Generalized additive mixed model
GCV	Generalized cross validation
GHQ	Gauss-Hermite Quadrature
GLM	Generalized linear model
GLMM	Generalized linear mixed model
GLS	Generalized least square
GML	Generalized maximum likelihood
IC	Class interval
ICC	Item characteristics curve
IRF	Item response functions
IRLS	Iteratively reweighted least square
IRM	Item response models
IRS	Indoor residual spraying
IRT	Item response theory
LLIN	long-lasting insecticidal net

LOWESS	Locally weighted scatter smoothing
MAD	Median absolute deviance
MCA	Multiple correspondence analysis
MCEM	Monte Carlo EM
MCNR	Monte Carlo Newton-Raphson
MGLMM	Marginal generalized linear mixed model
MINQE	Method of minimum norm quadratic equation
MIS	Malaria indicator survey
ML	Maximum likelihood
MML	Marginal maximum likelihood
MoM	Method of moments
MQL	Marginal quasi-likelihood
MSE	Mean square error
OCV	Ordinary cross validation
OLS	Ordinary least square
OR	Odd Ratio
PQL	Penalized quasi-likelihood
PSI	Person Separation Index
PSU	Primary Sampling Unit
RBM-MERG	Roll back malaria monitoring and evaluation group
RDT	Rapid diagnosis test
REML	Restricted maximum likelihood
RG	Random group
RM	Rasch model
SML	Simulated maximum likelihood
SNNPR	Southern Nations, Nationalities and People's Region
TCC	The Carter Center
UBRE	Un-biased risk estimation
WHO	World Health Organization
WLS	Weighted least square

Chapter 1

Introduction

Malaria is the most deadly, life-threatening problem caused by *Plasmodium* parasite infection affecting the world's most under developed countries, and regions lacking basic healthcare infrastructure (WHO, 2011). Through the disease predominates in Africa, also affects some of the well developed countries (Adhanom et al., 2006). The problem is extremely severe in Ethiopia where it has been the major cause of illness and death for many years (Adhanom et al., 2006, Federal Ministry of Health (FMH), 1999). According to records from the Ethiopian Federal Ministry of Health, 75% of the country is malarious with about 68% of the total population living in areas at risk of malaria (Adhanom et al., 2006, Federal Ministry of Health (FMH), 1999). That is, more than 50 million people are at risk of malaria (Lesaffre and Spiessens, 2001), and four to five million people are affected by malaria annually (FMH, 2004, WHO, 2006b). The transmission of malaria in Ethiopia depends on altitude and rainfall with a lag time varying from a few weeks before the beginning of the rainy season to more than a month after the end of the rainy season (Deressa et al., 2003, Tulu, 1993). Epidemics of malaria are relatively frequent (WHO, 2006c, Zhou et al., 2004) involving highland or highland fringe areas of Ethiopia, mainly areas 1,000-2,000 meters above sea level (Tulu, 1993, Adhanom et al., 2006, FMH, 2006b). Malaria transmission peaks bi-annually from September to December and April to May, coinciding with the major harvesting seasons. Therefore, this has serious consequences for Ethiopia's subsistence economy and for the nation in general. Major epidemics occur every five to eight years with focal epidemics as the commonest form. Early diagnosis and prompt treatment is one of the key strategies in controlling malaria.

Malaria diagnosis frequently relies on the patient's symptoms. Symptoms like fever, chills, sweats, headaches, muscle pains, nausea, and vomiting are not specific to malaria. *ApprClinical* diagnosis is inexpensive and can be effective. Clinicians often misdiagnose malarial infection. Misdiagnosis often leads to the unnecessary prescription of malaria medications which are becoming increasingly expensive as drug resistance grows globally and new medicines are required for effective treatment. Thus, increasing the accuracy of malaria diagnosis is becoming more important and will continue to be so in the future. There are broadly three different malaria diagnosis methods. These methods are microscopy, nucleic acid amplification tests and Rapid diagnostic tests (WHO, 2006a).

Microscopy diagnosis method is the most popular means of detecting malaria infection. But, this diagnosis method is available in better-equipped clinics. The malaria parasite can easily be confirmed using this technique. Therefore, important treatment information can be provided by identifying which of the multiple parasite species are in circulation and which drug treatment to initiate (WHO, 2006a).

The Nucleic Acid Amplification Tests (NAAT) detect parasite DNA circulating in the bloodstream and they are very sensitive. NAATs are currently not widely available in malaria endemic areas because of the expensive reagents and equipment as well as specialized training they require. Interpreting NAAT results can be challenging due to the fact that parasite DNA can remain in the bloodstream long after the infection has been cleared. Thus, differentiating an active infection from a recently cleared infection is difficult (LaBarre et al., 2010, Mens et al., 2006).

The Rapid Diagnostic Test (RDT) for malaria offers the potential to extend accurate malaria diagnosis to areas where microscopy services are not available in remote locations or after regular laboratory hours. RDTs have been developed in the lateral flow format and use finger-stick blood, taken only ten to fifteen minutes, and do not require a laboratory. Even non-clinical staff can easily learn to perform the test and interpret the results. However, these tests have limitations in that they lack the ability to detect mixed infections, all species of *Plasmodium*, and infections at low concentrations of parasites, including the inability to monitor response to therapy (Moody, 2002, Murray and Bennett, 2009). Malaria RDTs rely on the detection of parasite specific antigens (proteins) circulating in the bloodstream. The most common of these antigens are *Plasmodium Falciparum histidine-rich Protein2 (pfHRP2*) and *Plasmodium spp. lactose dehydrogenase* (pLDH) (WHO, 2009). Tests based on the *pfHRP2* antigen are specific to *Plasmodium falciparum*, the most dangerous species of malaria, and are more readily available and less expensive. *pLDH* based tests come in two varieties: pan-malarial tests which detect all malaria species or species specific tests that detect malaria species other than *Plasmodium falciparum*; and Pan-malarial tests, which are also available which detect the *Aldolase antigen* (Kakkilaya, 2003).

Among the three methods discussed above, microscopy remains the standard for diagnosing malaria. But, it is not accessible and affordable in most peripheral health facilities. The recent introduction of rapid diagnostic tests (RDT) for malaria has become a significant step forward in case detection, management and reduction of unnecessary treatment. RDT could be used in malaria diagnosis during population-based surveys and to provide immediate treatment based on the results (Reyburn et al., 2007, Tekola et al., 2008, Wongsrichanalai et al., 2007).

Demographic and Health Surveys (DHS) were carried out in Ethiopia in 2000, 2005 and 2011, and included a malaria module (CSA, 2000, CSA, 2006, CSA, 2012). From these surveys, it was recognized that the coverage and use of malaria intervention in the country was very low. In 2005, the Government of

Ethiopia's Federal Ministry of Health (FMH) developed a 5-year National Malaria Prevention and Control Strategy (FMH, 2006a). According to the strategy, areas less than 2,000 meters in altitude were considered 'malarious' and targeted to receive key malaria control interventions, including insecticide-treated nets (ITNs), indoor residual spraying of households with insecticide (IRS), and rapid diagnostic tests (RDTs) for malaria coupled with prompt and effective case management with *artemisinin*-based combination therapy (ACT) (Shargie et al., 2008).

Besides the demographic and health survey, various surveys were conducted to find malaria indicators. In 2007 Malaria Indicator Survey (MIS) was conducted in Ethiopia between September and December 2007 by Ministry of Health of Ethiopia in collaboration with CDC and USAID. The protocol for the MIS followed Roll Back Malaria Monitoring and Evaluation group (RBM MERG) guidelines (RBMM, 2005) with a few local modifications.

This survey was nationally representative. The objective was to determine parasite and anemia prevalence in the population at risk and to assess coverage, use and access to scaled-up malaria prevention and control interventions. In the survey, a two-stage random cluster sample of 7,621 households in 319 census enumeration areas (EAs; comprising approximately 200 households) was selected as primary sampling units, stratified by several domains, including altitude (i.e. less than 1,500 meters vs. between 1,500 and 2,500 meters) and degree of urbanization. The MIS household and women's questionnaires were adapted to the local context, and two types of questionnaires were used. The questionnaires included two structured, pre-coded ones with both closed and open ended questions: (i) a household questionnaire and (ii) a women's questionnaire. Both were based on Roll Back Malaria Monitoring and Evaluation group (RBM MERG) MIS Questionnaires (RBMM, 2005), modified to local conditions. The questionnaires were translated and printed in Amharic, Afaan Oromoo and Tigrigna languages and field-tested in non-survey EAs to determine the validity of the pre-coded answers (FMH, 2008).

The household questionnaire was administered to the household head or another adult if the household head was absent or unable to respond for any reason, and it elicited the following data: socio-demographic information and listing of household members; house construction materials and design; ownership of durable assets; availability, source of origin, type, condition and use of household mosquito net(s); and reported status of indoor residual spraying (IRS). The purpose of the household questionnaire was to identify children less than six years of age for specimen collection as well as women aged 15-49 years who were eligible to answer the women's questionnaire. The women's questionnaire was administered to these women as identified from the household questionnaire and it helped collected the following data: educational level; reproduction, birth history, and current pregnancy status; knowledge, attitudes and practices (KAP) on malaria preventive and curative aspects; reported history of fever among children less than five years of age in the previous two weeks; and reported treatment seeking behaviour for children less than five years of age with fever. In addition to the household and women questionnaires, blood samples were taken from all children less than five years old and from all household members in every fourth household. All children less than five years of age were included to ensure that no children under that age were missed during the survey, and only data for children under the same age are presented. The malaria diagnostic tests included rapid diagnostic tests (RDTs), blood slides for microscopic examination and haemoglobin level testing. RDTs were used in the survey to offer immediate treatment to individuals with a positive test. The RDT used (ParaScreen®, Zephyr Biomedical Systems, India) is a HRP2/pLDH-based antigen test detecting both *Plasmodium falciparum* and other *Plasmodium spp* (Graves et al., 2009).

After the collection of data and the release of the preliminary report, additional analyses were done based on the 2007 MIS survey by different researchers. (Jima et al., 2010) studied the coverage and use of major malaria prevention and control interventions of the Malaria Indicator Survey 2007 in Ethiopia. In their study, they found that since mid-2005, the Ethiopian national malaria control programme has considerably scaled-up its malaria prevention and control interventions, demonstrating the impact of strong political will and a committed partnership. Further, survey showed that efforts will have to be made to increase intervention access and use malaria intervention methods. To achieve the targets of coverage and use of malaria interventions, efforts have to be made to sustain and expand malaria intervention coverage and increase intervention access and use, and with strong involvement of the community. Based on these actions, Ethiopia expects to achieve its targets in terms of coverage and uptake of interventions in the coming years and move towards eliminating malaria (Shargie et al., 2010).

Besides the 2007 Malaria Indicator Survey (MIS) of Ethiopia, The Carter Center conducted a baseline household cluster survey in the Amhara, Oromiya and Southern Nations, Nationalities and Peoples' (SNNP) regions of Ethiopia between December 2006 and January 2007 during the end of the malaria season. The purpose was to obtain baseline information before large scale distribution of long-lasting insecticidal nets (LLINs) in early to mid-2007 and implementation of other integrated programs for prevention of malaria (Shargie et al., 2008). A questionnaire was developed as a modification of the survey household questionnaire which had two parts; the household interview, and malaria parasite form. The MIS was modified to survey each room in the house separately to ensure that all nets were in place, and to ascertain the density of sleepers per room as well as the number of sleeping rooms in (or outside) each house. This survey included peripheral blood microscopy and rapid diagnostic tests (RDT). The persons sleeping under each net were listed.

Based on the survey, routine surveillance data on malaria for the survey time period was obtained for comparison with prevalence survey results (Shargie *et al*., 2008). Shargie *et al.* (2008) found out that based on the ownership of nets, there were nearly a ten-fold increase as compared to the results of the 2005 Ethiopia Demographic and Health Survey (CSA, 2006) which was fewer than 5% of households in the Oromiya and SNNP regions. The results of the survey as well as the routine surveillance data demonstrated that malaria continues to be a significant public health challenge in these regions. However, the problem is more prevalent in SNNP than in Oromiya region. On the other hand, a study was conducted to estimate the prevalence of malaria parasites in Amhara, Oromiya, and Southern Nations, Nationalities and Peoples' (SNNP) regions of Ethiopia using the base line survey. Microscopy and RDT were used to investigate agreement between microscopy and RDT under field conditions. The samples were collected by taking fingerpick blood samples from all persons living in even-numbered households. The blood samples were tested using two methods: light microscopy of *Giemsa*-stained blood slides; and RDT (Tekola et al., 2008). From this study, they found that well conducted blood slide microscopy for malaria diagnosis for population based surveys remains the preferred option. The level of the agreement between RDT and light microscopy for malaria diagnosis warrants further investigations in clinical facilities in the Ethiopian context.

In addition to the two malaria indicator surveys, different surveys were conducted in different parts of Ethiopia. In 2003 (Peterson et al., 2009) studied the individual and household level factors associated with malaria incidence in a highland region of Ethiopia. The study was conducted in an area of the city of Adama (formerly Nazareth) located 120 km southeast of Addis Ababa. Data on incident malaria infections were obtained by assigning a unique study identification number to study households from August 1 to November 30, 2003. The cards were given to the heads of the households who instructed to present the card on all visits to the Adama Malaria Laboratory. Using this method, the data was collected and analysed first by examining the univariate associations between malaria incidence and other factors by regressing a single factor against individual malaria counts. Moreover, multivariate modelling was also used based on the statistical performance of factors in univariate analysis, and correlations among the factors.

The above study's strengths lie in its assessment of a wide range of both individual and household factors with regard to malaria risk, and the use of multilevel modelling. The study furthermore identified important malaria risk factors in a highland urban setting in Africa under epidemic conditions. The result showed that house distance to the major vector-breeding site was important in determining malaria risk. It suggests that vector control strategies targeted at such sites could greatly reduce the malaria burden in urban communities.

Other research on malaria epidemics and interventions from 1999 - 2004 was conducted in Kenya, Brundi, Southern Sudan and Ethiopia (Checchi et al., 2006). The researchers reviewed Medecins Sans Frontiers (MSF) program reports and used the available morbidity, mortality, diagnostic and treatment data from five interventions. These studies found that all four countries are moving to *Artemisinin-based* combination therapy (ACT) for outpatient treatment. They also suggested that further research is needed on methods to estimate needs (incidence) and coverage rapidly; and on strategies to expand treatment access efficiently.

To introduce the most advanced level of care for people with malaria infections in the health care system, it is important to scale up the malaria treatment programmes. This process requires continuous monitoring and counselling of patients in order to optimize medication benefits. A recent upsurge of malaria in endemic-disease areas with explosive epidemics in many parts of Africa is probably caused by many factors, including rapidly spreading resistance to antimalarial drugs, climatic changes, and population movements. Control efforts have been piecemeal and not coordinated. Strategies for control should have a solid research base both for developing antimalarial drugs and vaccines and for better understanding the pathogenesis, vector dynamics, epidemiology, and socio-economic aspects of the disease. Furthermore, for most countries in Africa, the costs of treatment programmes are enormous. Therefore, the outcome of this study not only will provide clinicians with the factors associated with malaria infections, but also provide between high risk patient differences on malaria prevention methods over time. That is, understanding specific barriers to medication and prevention of malaria for individual patients will be valuable in the development and implementation of evidence based interventions targeted at individual patients. The results can provide governmental and non-governmental organizations appropriate statistical models to analyze malaria indicator data in order to monitor malaria problems overtime. In general, after identifying a good-fitting, realistic model, the findings can be used to project the short-term future of the malaria epidemic, with the assumption that all parameter values and conditions remain constant.

In conclusion, the results of the studies conducted so far demonstrated that malaria continues to be a major cause of ill-health in Ethiopia. In addition, population movements contributed to the reappearance of the disease because most of population movements are from malaria free to highly malarious areas (Nathaniel, 2003). Therefore, the review section of this thesis identified the need for an in-depth study to identify the socio- economic, demographic and geographic factors thus leading to the reduction of the risk of malaria.

The current study will analyze the malaria indicator survey data by employing different statistical modelling approaches in order to determine the levels of malaria overtime. Factors that affect malaria treatment at the overall level, as well as individual level, will be sought. In general, a good-fitting, realistic model will be identified to project the short term future of the malaria epidemics. Hence, the findings will be valuable in tracking malaria patients and epidemics, identifying and testing different statistical methodologies which could be very helpful to critically understand binary response analyses and make recommendations on the appropriate techniques for further use.

To achieve this objective, the following steps were used

- The explanatory analysis was initially performed to identify the behaviour of the data.
- The relationship among malaria RDT result, socio-economic, demographic and geographic variables was investigated using multiple correspondence analysis.
- Malaria RDT result was obtained from complex sample survey. Therefore, to account for the survey design effect, the survey logistics method was used to investigate the effect of socio-economic, demographic and geographic factors on RDT result.
- To account for variability between the Probabilistic Sampling Units (PSU) which is *kebele* (smallest administrative unit in Ethiopia), generalized linear mixed model was used to fit the malaria RDT result data.
- The distribution of malaria is non-random across a landscape in areas of higher or lower transmission intensity and malaria risk. Spatial statistics analysis was performed to account for spatial autocorrelation and to check whether surveys that are near in space have similar malaria prevalence with the surveys that are far apart.
- The joint model under the generalized linear mixed model was used to investigate the joint effect of three predictor variables on malaria RDT result, use of mosquito nets and use of indoor residual spray (IRS) in the

last twelve months with socio-economic, demographic and geographic factors.

- Semiparametric model (GAMM) was applied to model the effect of age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed with indoor residual spray in the last twelve months nonparametrically while the other covariates (socio-economic, demographic and geographic factors) remain parametric.
- Rasch model was employed to test and investigate how well the observed malaria RDT result, use of mosquito nets and use of indoor residual spray data fit the expectations of the model.

In general, the study aims to investigate the different statistical approaches that are appropriate to model Malaria Indicator Survey (MIS) data. This is with the view of determining the levels of malaria across socio-economic, demographic and geographic factors that influence the malaria RDT result. Specifically, the purpose of this research is to assess the risk of malaria through the collection of household level baseline data, including housing construction, social-economic status, availability of latrines and water, altitude, coverage of spraying anti-mosquito and use of nets so as to establish a model which estimates the prevalence of malaria in all age groups through a malaria parasite prevalence survey. In addition, this study looks at the factors such as a change in socio-demographic characteristics, use of nets, public awareness or government-sponsored campaigns etc.

The thesis is organized as follows. The first chapter presents the introduction. In Chapter 2, a full description of the malaria Rapid Diagnosis Test (RDT) data is given with a further exploratory analysis of the data. The theory of correspondence analysis and its application to investigate the association between malaria RDT result, socio-economic, demographic and geographic factors are described in Chapter 3. Chapter 4 explores the socio-economic, demographic and geographic factors affecting malaria RDT using the generalized linear models, specifically, survey logistic method. Chapter 5 provides a comprehensive review of the generalized linear mixed models (GLMMs) including random effects models. Moreover, GLMMs are fitted to the malaria RDT data to explore socio-economic, demographic and geographic factors. In Chapter 6 a review and fitting of spatial statistics models to malaria RDT data are presented. Review and fitting of joint modelling of malaria RDT result and use of mosquito nets; and malaria RDT result and use of indoor residual spray (IRS) in the last twelve months are examined in Chapter 7. Chapter 8 looks at the semiparametric approaches, specifically generalized additive mixed models (GAMMs) while Chapter 9 presents the Rasch model analysis to malaria RDT result, use of indoor residual spray and use of mosquito nets. Finally, in Chapter 10 the discussions and conclusions as well as comparison of different models and possibilities for future research are presented.

Chapter 2

The data

Before getting into complex data analysis, it is of great importance to examine and get a general understanding of the data under consideration. It is this initial examination of the data that helps in determining the possible statistical techniques that could be applied to the data. The data used in this study is secondary data from The Carter Center (TCC) for the Malaria programme in Ethiopia. The Center is working in Ethiopia on two integrated disease control projects. These projects are malaria and onchocerciasis (MAONCHO) program; and Malaria and trachoma programmes. The Carter Center has committed itself to provide sufficient long-lasting insecticidal nets to most part of the country. In addition to the purchase and procurement of the requested nets, TCC is also helping to distribute them within and outside its current areas of operation in the regions of Amhara, Oromiya and the Southern Nations, Nationalities and Peoples Region (SNNPR). In order for TCC to assist the Federal Ministry of Health of Ethiopia in the assessment and evaluation of its malaria control, the Center needed to conduct a baseline household survey of net coverage and use as well as malaria prevalence within these three regions. The objective of the cluster survey was to assess the risk of malaria through the collection of household level baseline data for malaria risk indicators. The data included, housing construction, socio-economic status, availability of latrines and water, altitude, coverage of spraying and use of nets, indoor residual spraying, and estimation the prevalence of malaria in all age groups through a malaria parasite prevalence survey.

In order to achieve the above objective, TCC conducted a baseline household cluster malaria survey in Amhara, Oromiya and the Southern Nations and Nationalities People’s (SNNP) regions of Ethiopia from December 2006 to January 2007. A questionnaire was developed as a modification of the Malaria

Indicator Survey (MIS) Household Questionnaire. The questionnaire had two parts: the household interview and malaria parasite form. The MIS was modified to survey each room. Furthermore, each room in the house was listed separately.

For the baseline household cluster malaria survey, which was conducted by TCC, a multi-stage cluster random sampling was used. By assuming the lowest measurement of prevalence malaria indicator, the sample size was estimated. Assuming prevalence of malaria to be the lowest indicator measured, the prevalence in the population was estimated to be 8%. In Amhara region, each zone was regarded as a separate domain; while in Oromiya and SNNPR, the community-directed treatment with *ivermectin* (CDTI) areas combined were taken as one domain. Furthermore, to estimate the required sample size, the following formula was used.

$$n = Z^2 \, p \, (1 - p)/e^2$$

where p is the expected malaria prevalence (p = 0.08), Z is level of significance 95% = total value = 1.96 (value in the standard normal distribution) and e is acceptance error (0.02).

In addition to these values, a 10% non-response rate was factored into the calculation of the sample size. For TCC baseline household cluster malaria survey in Amhara, Oromiya and the Southern Nations and Nationalities People’s (SNNP) regions of Ethiopia, which was conducted in 2007, the design was a population-based household cluster survey. Based on these clusters, Zoneal-level estimates of indicators were obtained for Amhara region, and sub-regional estimates were taken for Oromiya and SNNPR. All ten Amhara zones were surveyed as separate domains, with sixteen clusters in each zone (total 160 clusters). Bahir Dar town and two *woredas* with less than 10% of the population living in malarious areas were excluded. In Oromiya and SNNPR, sampling was done directly at the *kebele* level. From the total number of individuals who participated in the survey, 7,745 in Amhara, 1,996 in Oromiya and 1,860 in SNNP from all age groups were tested using RDT (The Carter Center (TCC), 2007).

Further studies on the sampling procedure for the survey were conducted by different researchers (Emerson et al., 2008, Shargie et al., 2008). The sampling design was employed in order to select households within each first-stage cluster, or *Kebele* (smallest administrative unit in Ethiopia). From the 224 selected *Kebele*s, 25 households were chosen, from which even-numbered households were selected for the malaria Rapid Diagnostic Tests (RDT). All individuals in these twelve households were eligible for individual interviews. Furthermore, each room in the house was listed separately. By using the mosquito nets as a guide, it was possible to determine the number of persons sleeping in each room. This information was useful in determining the number of sleeping rooms both within and outside the house. In addition to the number of rooms and number of nets, the persons sleeping under each net were listed. The sampled areas and domains as well as the survey sites are presented in Figure 2.1.

Malaria parasite testing was performed on consenting residents. The blood sample subjected to the malaria Rapid Diagnostic Test was collected by taking finger prick blood samples from participants. The Rapid Diagnostic Test used was *ParaScreen* which is capable of detecting malaria infection with high degree of sensitivity. The test uses approximately *5 μl* of blood and is readable after fifteen minutes following the manufacturer's guidelines. Participants with positive rapid tests were immediately offered treatment according to national guidelines.

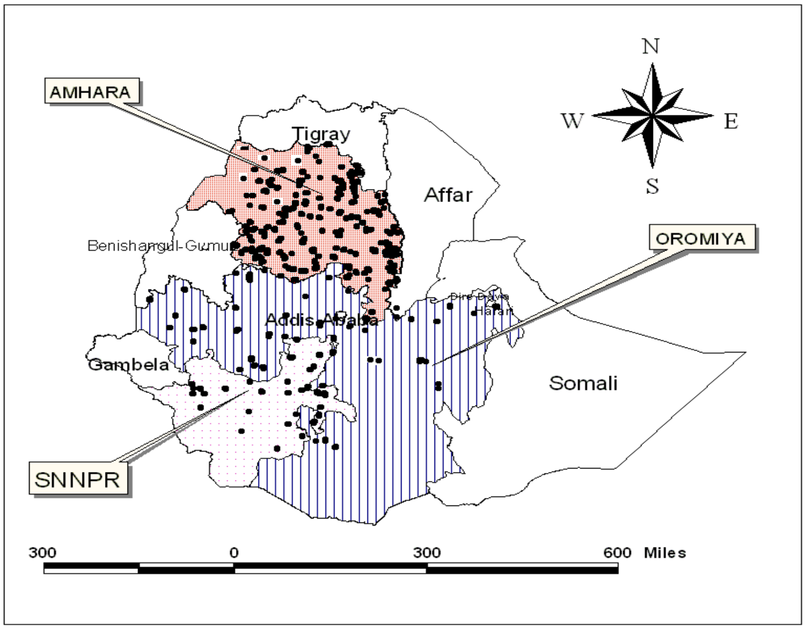


Figure 2. 1: Map of Ethiopia showing the surveyed households

2.1 Variables of interest

The variables used for the analyses in this study included malaria rapid diagnosis test, socio-economic, demographic and geographic variables. Malaria rapid diagnosis test was collected from consenting household members. The response variable and the covariates are given as follows.

Response variable: The outcome of interest is the RDT result. RDTs assist in the diagnosis of malaria by detecting evidence of malaria parasites in human blood and are an alternative to diagnosis based on clinical grounds or microscopy, particularly where good quality microscopy services cannot be readily provided. Thus, the response variable was binary, indicating that either a person was positive or not positive.

Independent variables: The independent (predictor) variables consisted of baseline socio-economic, demographic and geographic variables, which were collected from each household. The socio-economic variables were the following: main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, access to radio and television, total number of rooms, main construction material of the rooms' walls, main construction material of the room's roof and main construction material of the room's floor, incidence in the past twelve months of indoor residual spray; use of mosquito nets and total number of nets. Geographic variables were region and altitude, and demographic variables were gender, age and family size. Of these variables, age and sex were collected at the individual level, while altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms, main construction material of walls, roof and floor, incidence of indoor residual spraying and use of mosquito nets were all collected at the household level. The levels and coding of the categorical variables are given in Table 2.1.

Table 2. 1: Table of variables

Variables	Levels and coding
Region	1 = Amhara , 2 = Oromiya, 3 = SNNP
Main source of drinking water	1= Unprotected, 2 = protected, 3 = Tap water
Time to collect water	1=<30 minutes, 2 = 30 to 40 minutes, 3 = 40 – 90 minutes, 4 = >90 minutes
Toilet facilities	1 = No facility, 2 = pit latrine, 3 = toilet with flush
Availability of electricity	1 = yes, 2 = no
Availability of radio	1 = yes, 2 = no
Availability of television	1 = yes, 2 = no
Main material of the room's wall	1 = cement block, 2 = mud block/stick/wood, 3 = corrugated metal
Main material of the room's roof	1 = thatch, 2 = stick and mud, 3 = corrugate
Main material of the room's floor	1 = earth/Local dung plaster, 2 = wood, 3 = cement
Spraying of indoor residual spray in the past twelve months	1 = yes, 2 = no
Use of mosquito nets	1 = yes, 2 = no
Rapid Diagnosis test (RDT)	0 = Negative, 1= Positive

2.2 Baseline characteristics of the study population

The data analyzed consisted of malaria rapid diagnosis tests of respondents in the rural parts of Amhara, Oromiya and SNNP regions of Ethiopia. During the study period, 5,708 households that were located in 224 clusters, covered in the survey. From the total 5,708 households, Amhara, Oromiya and SNNP regions covered 4,101 (71.85%), 809 (14.17%) and 798 (13.98%) households respectively. The distribution of toilet facility, source of drinking water and time to collect water is presented in Table 2.2. The table shows that in Amhara and Oromiya regions, the majority of people most frequently used unprotected water supplies with percentage equal to 66.30% and 79.70% respectively. In contrast to these regions, in SNNP, the use of unprotected water was found to be slightly over half (56%). On the other hand, 25.4% of the people in SNNP region used protected water followed by those in Amhara (17.8%) and Oromiya (8.8%) regions. From the total households, 18.6% in SNNP, 16% in Amhara and 11.5% in Oromiya regions use tap water for drinking. Unprotected water includes, unprotected spring, unprotected dug well (use bucket and rope) and surface water (river/dam/lake/pond/stream). Similarly, protected water includes, capped spring, protected dug well (use hand pump), tube well or borehole and cart with small Tank. Furthermore, the tap water also includes public tap or standpipe, piped into yard and piped into dwelling.

The total time taken to collect water is also presented in Table 2.2. Based on the result, more households in Amhara region (72.3%) than in the other two regions (62.6 – 64.6%) had to travel less than 30 minutes on average to get their water. Furthermore, 8.6% of the households in Oromiya region travel more than 90 minutes to collect water. But, in Amhara and SNNP regions 2.9% and 3.6% of their residents respectively took more than 90 minutes to collect water.

Table 2. 2: Distribution of toilet facility, source of drinking water and time to collect water by region

Socio-economic variables	Region		
	Amhara	Oromiya	SNNPR
Toilet facility			
No facility	73.40%	72.60%	40.90%
Pit latrine	26.60%	26.00%	59.10%
Toilet with flush	0.00%	1.40%	0.00%
Source of drinking water			
Unprotected water	66.30%	79.70%	56.00%
Protected water	17.80%	8.80%	25.40%
Tap water	16.00%	11.50%	18.60%
Time to get water			
Less than 30 minute	72.30%	62.60%	64.60%
Between 30 - 90 minutes	24.90%	28.80%	31.80%
Greater than 90 minutes	2.90%	8.60%	3.60%

The great majority of households, namely 73.4% in Amhara, 72.6% in Oromiya and 40.9% in SNNP regions had no access to toilet facility (Table 2.2). Again, Amhara and Oromiya lagged behind SNNP in the use of pit latrine includes pit latrine with no cement slab, pit latrine with slab and pit latrine with cement slab and vent pipe toilet. Table 2.2 shows that more than half of houses in SNNP (59.1%) having a latrine toilet.

Furthermore, the distribution of positive RDT results by toilet facility, source of drinking water and distance to get water is presented in Figure 2.2. In the figure, it is clear that respondents with no toilet facility (11%) had more positive RDT results, followed by pit latrine (5.9%) and toilet with flush (4.3%). Similarly, households who travelled long distance (5.8%) have a high percentage of positive RDT results than those travelling shorter distances. Persons who have unprotected water (6.5%) as source of drinking water have greater chance to be RDT positive compared to those using protected and tap waters.

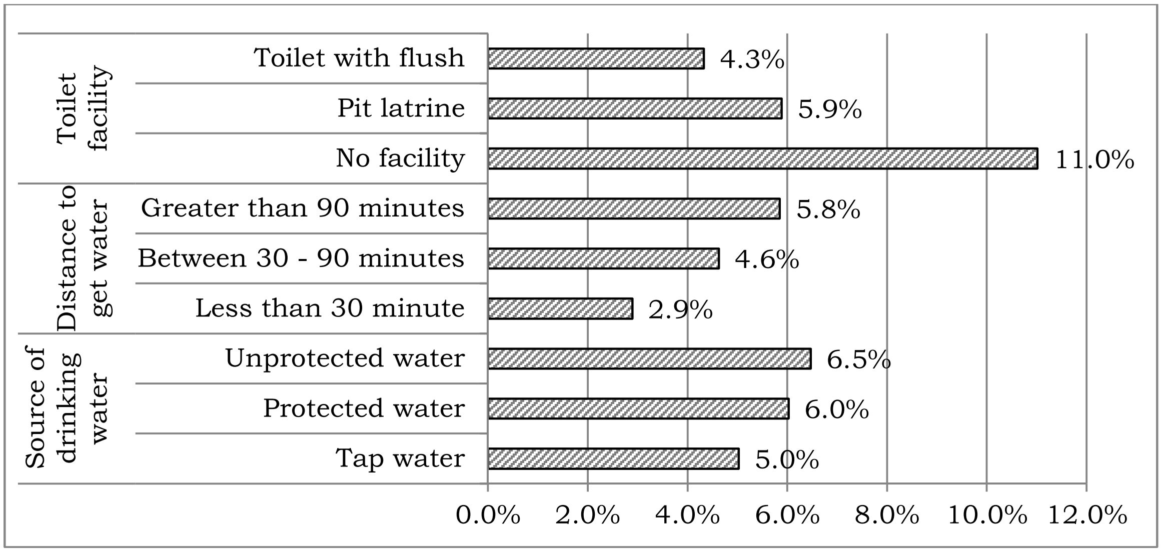


Figure 2. 2: Distribution of positive RDT result by toilet facility, source of drinking water and distance to get water

More than 90% of households in all regions have house walls made of wood. Similarly, more than 90% of the households for all regions have earth or local dung floor. However, the roof material varied across regions, with the majority of houses in Oromiya (67.3%) and SNNP (74%) regions having corrugated iron roofs compared to Amhara (39.8%). On the other hand, where 59.3% had thatch roofs in Amhara, followed by 23.1% in Oromiya and 0.5% in the SNNP region.

Table 2. 3: Ditribution of material for house construction by region

Socio-economic variables	Region		
	Amhara	**Oromiya**	**SNNP**
Wall material			
Cement block	0.3%	2.9%	0.2%
Mud block/stick/wood	99.6%	90.1%	99.8%
Corrugated metal	.1%	7.0%	0.0%
Roof material			
Thatch	59.3%	23.1%	0.5%
Stick and mud	0.9%	9.6%	25.5%
Corrugate	39.8%	67.3%	74.0%
Floor material			
Earth/Local dung plaster	96.2%	92.6%	96.98%
Wood	2.0%	6.9%	1.02%
Cement	1.8%	.5%	2.0%

The distribution of positive RDT result by wall, roof and floor materials of the house is presented in Figure 2.3. The figure shows that the percentage of positive RDT results in cemented floors was 0.9%, 7.7% in wooden floor and 5.9% in earth or local dung plastered. On the other hand, the percentage of RDT result in corrugate roof was 5.4%, 17.3% in stick and mud roof and 5.8% in thatch roofs. The percentage of positive RDT in corrugated metal wall was found to be 5.9%, 7.3% in mud/stick/wood wall and 3.2% in cement block (Figure 2.3).

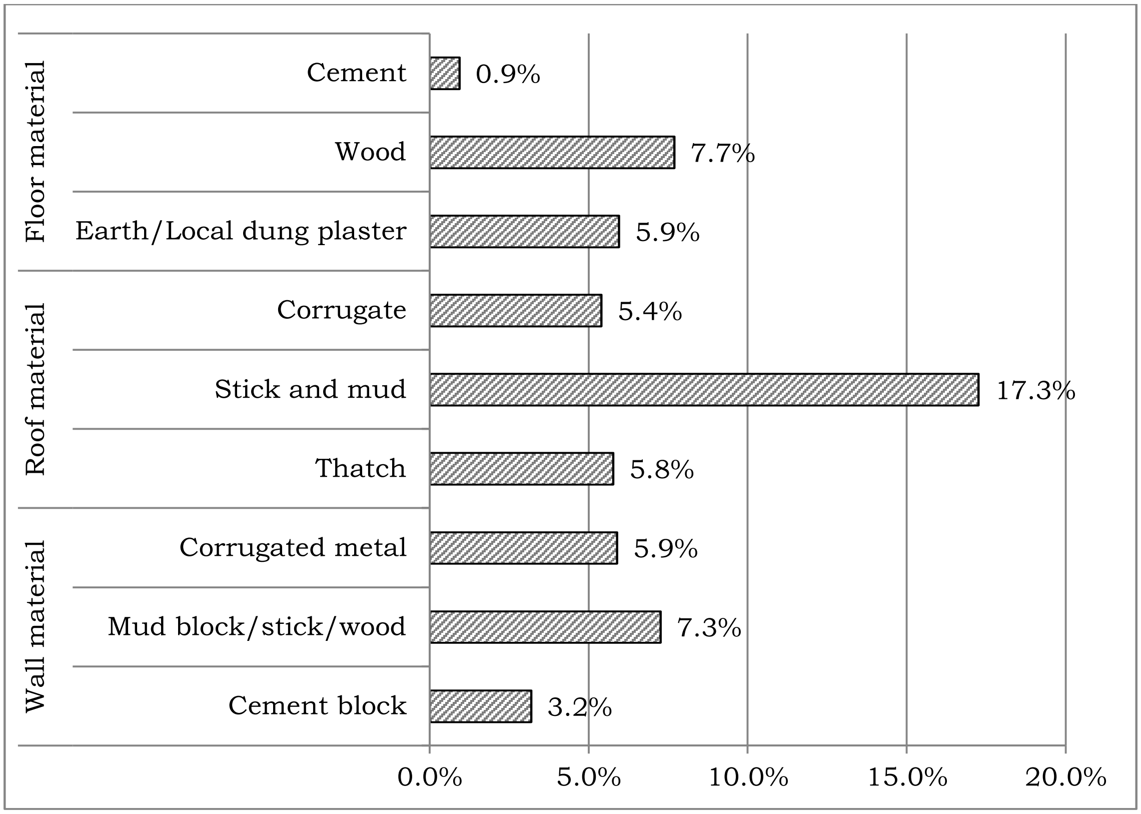


Figure 2. 3: Distribution of positive RDT result by wall, roof and floor materials of the house

In the survey, representative household heads were asked if they had access to radio, television and electricity. From the result it was found that electricity and televisions were very rare in the surveyed households. In Amhara, Oromiya and SNNP regions 94.3%, 97.9% and 94.1% of the households did not have access to electricity respectively. Similarly, more than 97% of the households in the three regions have no television. Unlike television access, radios were more common. The data show that 75.2%, 62.9% and 58.7% of the households in Amhara, Oromiya and SNNP regions have access to radio respectively (Table 2.4).

Table 2. 4: Ditribution of availability of radio, television and electricity by region

Socio-economic variables	Region		
	Amhara	**Oromiya**	**SNNP**
Availability of radio			
Yes	24.80	37.10	41.30
No	75.20	62.90	58.70
Availability of Television			
Yes	1.20	2.47	0.74
No	98.80	97.53	99.26
Availability of electricity			
Yes	5.70	2.10	5.90
No	94.30	97.90	94.10

Use of mosquito nets and indoor residual spraying drugs in the last twelve months were included in the survey. The use of mosquito nets was derived by direct questioning about who slept in each net in the household, and who slept without a net. The results show that 38.3% in Amhara, 43.7% in oromiya and 48.2% in SNNP regions use mosquito nets. Besides the use of mosquito nets, information on the use of indoor residual spraying in the last twelve months was collected. The result revels that those households who live in SNNP region use more indoor residual spraying (30.9%) compared to Amhara (29.6%) and Oromiya (27.4%) regions (Figure 2.4).

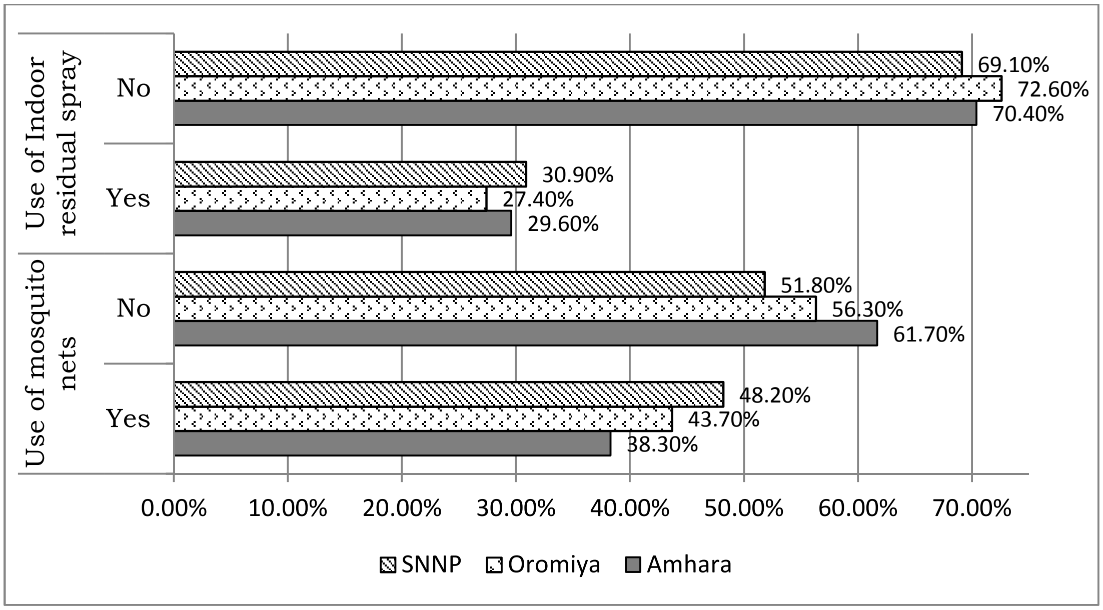


Figure 2. 4: Distribution of use of mosquito nets and indoor residual spraying by RDT result

Figure 2.5 shows the distribution of age group and family size by malaria RDT result. Most houses, i.e., age group 31-45 accounts for 72.9% of all positive malaria RDTs and 72.6% of all negative malaria RDTs. Similarly, family size 5–10 persons accounts for 58.7% of all positive malaria RDTs and 53.2% of all negative malaria RDTs.

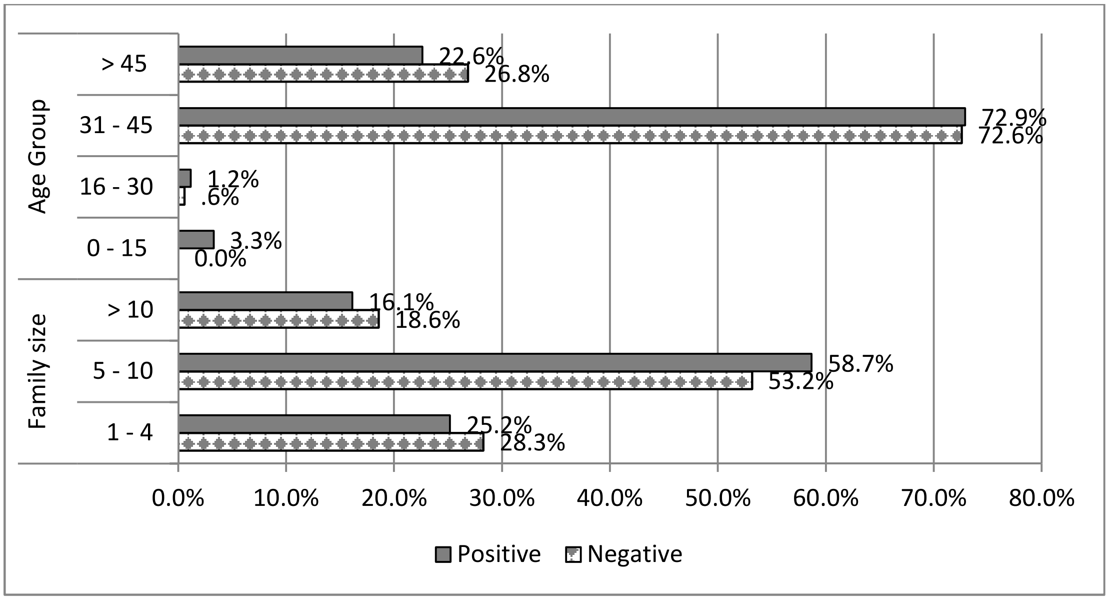


Figure 2. 5: Distribution of age group and family size by RDT result

Table 2.5 shows descriptive characteristics of total rooms and total number of persons in the household. Most houses in Amhara (90.25%), Oromiya (71.2%) and SNNP (74.56%) regions had only one sleeping room. A very small proportion of people (<1%) reported having more than three sleeping rooms. Furthermore, the average number of rooms in Amhara, Oromiya and SNNP regions was found to be 1.15, 1.3 and 1.3 respectively. Furthermore, the average number of persons per household ranged from 4.7 to 5.6 by region and was 4.9 overall. In Amhara the median household size was five whereas in both Oromiya and SNNPR it was six persons.

Table 2. 5: Distribution of total number of rooms and total number of members of the household by region

	Region		
	Amhara	**Oromiya**	**SNNP**
Total Number of Rooms			
1	90.25%	71.20%	74.56%
2	8.83%	22.62%	21.93%
3	0.85%	4.70%	3.01%
4	0.05%	0.99%	0.25%
5+	0.02%	0.49%	0.25%
Family size			
1	2.34%	1.36%	0.75%
2	10.59%	6.67%	4.01%
3	16.90%	13.10%	12.91%
4	17.32%	14.46%	15.91%
5	17.88%	16.69%	20.05%
6	15.22%	20.40%	20.18%
7	9.93%	10.75%	11.15%
8	5.27%	6.67%	5.64%
9+	4.56%	9.89%	9.40%

The age and gender-specific malaria prevalence, by region is shown in Table 2.6. This table demonstrates that there is no significant difference in prevalence by age group as well as by region. Moreover, the pattern of malaria prevalence by age is not homogeneous across the study regions. In addition, Table 2.6 shows that there is no difference in prevalence between males 4.05% and females 4.55%.

Table 2. 6: Malaria prevalence by region, age group and gender

Age group	Male			Female			Total		
	Tested	+ive	%	Tested	+ive	%	Tested	+ive	%
Amhara									
<5	643	28	4.35	603	31	5.14	1,246	59	4.74
5-14	1,144	43	3.76	1,240	49	3.95	2,384	92	3.86
15-49	1,316	55	4.18	1,998	94	4.70	3,314	149	4.50
>=50	426	17	3.99	375	18	4.80	801	35	4.37
Total	**3529**	**143**	**4.05**	**4216**	**192**	**4.55**	**7745**	**335**	**4.33**
Oromiya									
<5	225	1	0.44	213	2	0.94	438	3	0.68
5-14	293	1	0.34	368	4	1.09	661	5	0.76
15-49	342	2	0.58	420	4	0.95	762	6	0.79
>=50	66	2	3.03	69	0	0.00	135	2	1.48
Total	**926**	**6**	**0.65**	**1,070**	**10**	**0.93**	**1,996**	**16**	**0.80**
SNNPR									
<5	142	11	7.75	134	6	4.48	276	17	6.16
5-14	346	23	6.65	326	20	6.13	672	43	6.40
15-49	332	16	4.82	443	20	4.51	775	36	4.65
>=50	78	5	6.41	59	4	6.78	137	9	6.57
Total	**898**	**55**	**6.12**	**962**	**50**	**5.20**	**1,860**	**105**	**5.65**
Three regions									
<5	1,010	40	3.96	950	39	4.11	1,960	79	4.03
5-14	1,783	67	3.76	1,934	73	3.77	3,717	140	3.77
15-49	1,990	73	3.67	2,861	118	4.12	4,851	191	3.94
>=50	570	24	4.21	503	22	4.37	1,073	46	4.29
Total	**5353**	**204**	**3.81**	**6,248**	**252**	**4.03**	**11,601**	**456**	**3.93**

The prevalence of malaria by altitude is given in Tables 2.7. For each surveyed household the altitude was determined at the time of the survey. Based on the values of altitude for the households, Amhara had the greatest range of altitudes. For Oromiya and SNNPR the altitude for households is below 1000 meters or above 2500 meters. The majority of households (93.4%) in all regions were found at altitudes between 1000 to 2500 meters. Moreover, there were a significant number of malaria cases detected at altitudes above 2000 meters.

Unlike Oromiya and SNNP regions, there was an expected decline in prevalence by altitude up to 3000 meters in Amhara (Table 2.7). But, the prevalence of malaria above 2500 meters was found to be 8.3%. For persons who lived above 3000 meters, the prevalence of malaria was 1.33%. No positive malaria cases were detected above 2000 meters for Oromiya region (Table 2.7), but in SNNPR there was a high prevalence of 72.6% for households who lived at 1500-2000 meters.

Table 2. 7: Malaria prevalence by altitude and region

Altitude class	Amhara		Oromiya		SNNP		Total	
	Tested	%	Tested	%	Tested	%	Tested	%
<=1000m	125	1.61	0	0.00	0	0.00	125	1.08
1000-1500m	859	11.09	343	17.18	327	17.58	1529	13.18
1500-2000m	2973	38.39	1316	65.93	1351	72.63	5640	48.62
2000-2500m	3142	40.57	337	16.88	182	9.78	3661	31.56
2500-3000m	543	7.01	0	0.00	0	0.00	543	4.68
>3000m	103	1.33	0	0.00	0	0.00	103	0.89
Total	**7745**		**1996**		**1860**		**11601**	

According to the result in Figure 2.6, there was a declining trend in percentage of malaria prevalence from 48.6% at 1500-2000m to 31.6% at 2000-2500m. Highland or highland fringe areas, mainly those at 1000 – 2000 meters are often described as the limit of transmission of malaria in Ethiopia. But, there are some cases found above 2000 meters. These cases may have resulted from local epidemics or from movement of people from lower altitudes.

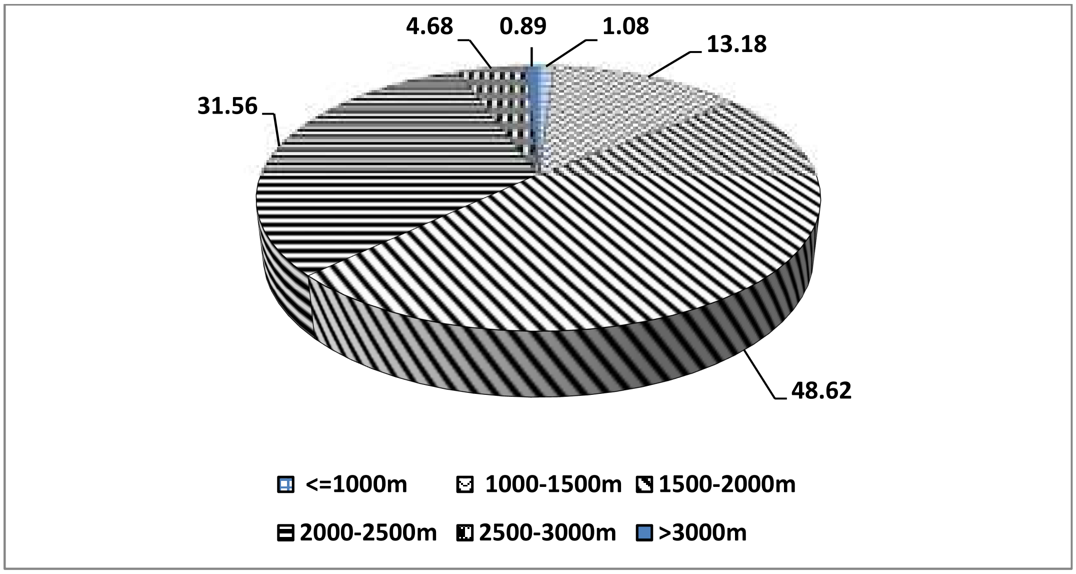


Figure 2. 6: Distribution of altitude by positive malaria RDT result

2.3 Summary

An integrated malaria survey was conducted in 224 clusters covering 5,708 households in three regions of Ethiopia between December 2006 and early February 2007, at the end of the peak malaria season. Blood slides from 9,352 people of all ages living in even numbered households were examined for malaria parasites. Net usage was assessed from all households included in the survey. The maximum number of nets owned was five and the median was zero. Moreover, the maximum rooms in the house were found to be five. Furthermore, it can be seen that there was no difference in net use by gender. There was a declining trend of prevalence of malaria by altitude. The majority of the households used unprotected water. More than half of the households in the survey areas had no access to toilet facility and the majority of the households were constructed with wood or stick wall, and their floors were mainly earth. Roofs were mainly made of thatch in Amhara, but of corrugated iron in Oromiya and SNNP regions. Very low percentage of households had electricity and television, while quarter (25%) of the households had a radio.

The study of assessment of variables by multiple correspondence analysis technique allowed the analysis of the relationship between the socio-economic, demographic, geographic and malaria RDT result factors. The use of the multiple correspondence techniques in comparison to other advanced statistical results was made both analytically and empirically across the geographic regions. The advantage of applying multiple correspondence analysis is that it gives more detailed information about the relationship between different variables. Moreover, the results will be easier to interpret. The application of multiple correspondence analysis with detailed theoretical background will be discussed in the next chapter.

Chapter 3

Correspondence analysis

3.1 Introduction

The cross-tabulation of categorical data is perhaps the most commonly encountered and simple form of analysis in research. Therefore, ordering things in time has been the interest of many researchers. Based on this fact, correspondence analysis (CA) is one of a statistical visualization methods used to analyzing data in contingency tables. This method first developed in France (Benzécri, 1973, Greenacre, 1984). Different authors proposed this method under various names. These method names include the Dutch Homeneity Analysis (Gifi, 1990), the Japanese Qualification Method (Hayashi, 1954), the Canadian Dual Scaling (Nishisato, 1980). These analogous have different theoretical foundations but all methods lead to equivalent solutions (Greenacre and Blasius, 2006, Tenenhaus and Young, 1985). Correspondence analysis is thought of as a principal component method for normal, contingency table data. It can be used to analyze cases-by-variable-categories matrices of non-negative data. Correspondence analysis is also a multivariate descriptive data analytic technique. Even the most commonly used statistics for simplification of data may not be adequate for description or understanding of the data. The correspondence analysis results provide information which are similar to those produced by principal component or factor analysis (Hill, 1974). Using this result, it is possible to explore the structure of the categorical variables included in the table. The simplified form data provides useful information about the data (Van der Heijden and de Leeuw, 1985, Hair et al., 1995). The relationship of the categories of rows and columns of the data can be represented using correspondence analysis graphs. The graphical representation of the relationships between the row and column categories is in the same space which is also produced using correspondence analysis. In general, correspondence analysis simplifies complex data and provides a detailed description of practically every bit of information in the data, yielding a simple, yet exhaustive analysis (Greenacre and Blasius, 2006, Johnson and Wichern, 2007).

Correspondence analysis has several features that distinguish it from other techniques of data analysis. The multivariate treatment of the data through multiple categorical variables is an important feature of correspondence analysis. This multivariate nature has advantage to reveal relationships which could occur during a series of pair wise comparisons of variable (Tian et al., 1993). Correspondence analysis works effectively for a large data matrix, if the variables are homogeneous, and the data matrix structure is either unknown or poorly understood. There are some advantages of correspondence analysis over other methods. This advantage is related to joint graphical displays. This graphical display produces two dual displays whose row and column geometries have similar interpretations. This facilitates the analysis to detect different relationships. In other multivariate approaches for graphical data representation, this duality is not present (Askell-Williams and Lawson, 2004).

Multiple correspondence analysis (MCA) which is part of a family of descriptive methods is an extension of correspondence analysis (CA) and allows investigating the pattern of relationships of several categorical dependent variables. It is the multivariate extension of CA to analyze tables containing three or more variables. In addition to this, MCA can considered as a generalization of principal component analysis for categorical variables which reveal patterning in complex data sets.

MCA helps to describe patterns of relationships distinctively using geometrical methods by locating each variable/unit of analysis as a point in a low-dimensional space. MCA is useful to map both variables and individuals, so allowing the construction of complex visual maps whose structuring can be interpreted. Moreover, this technique offers the potential of linking both variable centred and case centred approaches.

The rest of the chapter is organized as follows. An overview of the theory of MCA is presented in sections 3.2. Multiple correspondence analysis (MCA) is fitted to malaria RDT result data in section 3.3. Summary and discussion of this chapter is given in section 3.4.

3.2 Review of Multiple Correspondence Analysis (MCA)

Suppose there are n observations on p categorical variables. Assume q_j different values for variable j. Next define a matrix, $\boldsymbol{G}_j$ which is $n \times q_j$ matrix. This matrix is known as indicator matrix. The $n \times q$ matrix $\boldsymbol{G}$, with q the sum of q_j can be obtained by concatenating the $\boldsymbol{G}_j$'s (Greenacre, 1984). In general, MCA is defined as the application of weighted PCA to the indicator matrix $\boldsymbol{G}$ (Benzécri, 1973). Furthermore, $\boldsymbol{G}$ is divided by its grand total np to obtain the correspondence matrix $\boldsymbol{F} = \frac{1}{np}\boldsymbol{G}$, i.e., $\boldsymbol{1}_n^t \boldsymbol{F} \boldsymbol{1}_q = 1$, where $\boldsymbol{1}_i$ is an $i \times 1$ vector of ones. The vectors $\boldsymbol{r} = \boldsymbol{F}\boldsymbol{1}_q$ and $\boldsymbol{c} = \boldsymbol{F}^t \boldsymbol{1}_n$ are the row and column marginals respectively. These marginals are the vectors of row and column masses. Suppose the diagonal matrices of the masses are defined as $\boldsymbol{D}_r = diag(\boldsymbol{r})$ and $\boldsymbol{D}_c = diag(\boldsymbol{c})$. Note that, the i^{th} element of $\boldsymbol{r}$ is $f_{i.} = \frac{1}{n}$ and the s^{th} element of $\boldsymbol{c}$ is $f_{.s} = \frac{n_s}{np}$ where n_s is the frequency of category s (Greenacre and Blasius, 2006).

MCA can be defined as the application of PCA to the centered matrix $\boldsymbol{D}_r^{-1}(\boldsymbol{F} - \boldsymbol{r}\boldsymbol{c}^t)$ with distances between profiles given by the chi-squared metric defined by $\boldsymbol{D}_c^{-1}$. The n projected coordinate of the row profiles on the principal axes are called row principal coordinates. The $n \times k$ matrix $\mathbf{X}$ of row principal coordinates is defined by

$$\boldsymbol{X} = \boldsymbol{D}_r^{-1/2} \widetilde{\boldsymbol{F}} \boldsymbol{V}_k, \tag{3.1}$$

where $\widetilde{\boldsymbol{F}} = \boldsymbol{D}_r^{-1/2}(\boldsymbol{F} - \boldsymbol{rc}^t)\boldsymbol{D}_c^{-1/2}$ and $\boldsymbol{V}_k$ is the $q \times k$ matrix of eigenvectors corresponding to the k largest eigenvalues $\lambda_1, \dots, \lambda_k$ of the matrix $\widetilde{\boldsymbol{F}}^t\widetilde{\boldsymbol{F}}$. The projected row profiles can be plotted in the different planes defined by these principal axes called row principal planes (Greenacre and Blasius, 2006).

The categories for column profile can be described by the column profiles. The value can be calculated by dividing the columns of $\boldsymbol{F}$ by their column marginals. Interchanging rows with columns and all associated entities can be used for the dual analysis of columns profiles. This is done by transposing the matrix $\boldsymbol{F}$ and repeating all the steps. The metrics used to define the principal axes (weighted PCA) of the centered profiles matrix $\boldsymbol{D}_c^{-1/2}(F - \boldsymbol{rc}^t)^t$ are $\boldsymbol{D}^c$ and $\boldsymbol{D}_r^{-1}$.

The $q \times k$ matrix $\boldsymbol{Y}$ of columns principal coordinates is now defined by

$$\boldsymbol{Y} = \boldsymbol{D}_c^{-1/2}\widetilde{\boldsymbol{F}}^t\boldsymbol{U}_k, \tag{3.2}$$

where $\boldsymbol{U}_k$ is the $n \times k$ matrix of eigenvectors corresponding to the k largest eigenvalues $\lambda_1, \dots, \lambda_k$ of the matrix $\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}^t$. To aid visualization and interpretation of the projected column profiles in the planes defined by principal axes, which are called column principal planes, can be plotted (Johnson and Wichern, 2007).

The absolute contribution of the variable j to the inertia of the column principal component α in the α^{th} column of $\boldsymbol{Y}$ is given by

$$c_{j\alpha} = \sum_{s \in M_j} s\epsilon M_j\, f_{.s} y_{s\alpha}^2$$

where M_j is the set of categories of variable j. The relation between the absolute contribution $c_{j\alpha}$ and the correlation ratio between the variable j and the row standard component α is given by

$$\eta_{j\alpha}^2 = \sum_{s\in M_j} \frac{n_s}{n}(\bar{x}_{s\alpha}^* - 0)^2 = p \times c_{j\alpha}. \tag{3.3}$$

Note that factor loadings for PCA are correlations between the variables and the components (the correlation ratios) are known as discrimination measures. These values can be interpreted in MCA as squared loadings.

Suppose $\bar{\boldsymbol{X}}^* = \boldsymbol{X}^*\boldsymbol{T}$ and $\bar{\boldsymbol{Y}} = \boldsymbol{YT}$, where $\boldsymbol{TT^t} = \boldsymbol{T^tT} = \mathbb{I}_{\boldsymbol{k}}$. Let $\boldsymbol{X}^*\boldsymbol{Y^t} = \bar{\boldsymbol{X}}^*\bar{\boldsymbol{Y}}^{\boldsymbol{t}}$**.** Then, these relations show that the lower rank approximation is not unique. Furthermore, the MCA solutions $\boldsymbol{X}^*$ and $\boldsymbol{Y}$, are not unique over orthogonal rotations. The non-uniqueness can be explored to improve the interpretability of the original solution by means of rotation. Rotation of the column principal coordinates matrix $\boldsymbol{Y}$ to simple structure must be followed by the same rotation of the row standard coordinates matrix $\boldsymbol{X}^*$. The interpretation of the correlation ratios can be simplified for the matrices $\boldsymbol{Y}$ and $\boldsymbol{X}^*$ by rotation (Greenacre, 2000).

For the method of rotation, the Varimax based function can be used. After rotation of $\boldsymbol{X}^*$ and $\boldsymbol{Y}$, the relation (3.3) becomes

$$\tilde{\eta}_{j\alpha}^2 = p \sum_{s\in M_j} f_{.s}\tilde{y}_{s\alpha}^2, \tag{3.4}$$

where $\tilde{\eta}_{j\alpha}^2$ is the correlation ratio between the variable j and α^{th} column of $\tilde{X}^*$.

The graphical approach to represent the correspondence approach is the *biplot* representation. Therefore, *biplot* information is represented by $n \times p$ data matrix. As the name indicates, it refers to the two kinds of information contained in a data matrix. The information in the rows pertains to samples or sampling units and that in the columns pertains to variables. The scatter plot can represent the information on both the sampling units and the variables in a single diagram. This representation is useful to visualize the position of one sampling unit relative to another (Dray et al., 2003). In addition to this, it helps to visualize the relative importance of each of the two variables to the position of any variables. Matrix array can be constructed with several variables using scatter plots. The idea behind *biplots* is to add the information about the variables to the graph. Therefore, the construction of a *biplot* leads the sample principal components and the best two-dimensional approximation to the data matrix $\boldsymbol{X}$ approximates the j^{th} observation $\boldsymbol{x}_j$ in terms of the sample values of the first two principal components. Specifically,

$$\boldsymbol{x}_j = \overline{\boldsymbol{X}} + \hat{y}_{j1}\hat{\boldsymbol{e}}_1 + \hat{y}_{j2}\hat{\boldsymbol{e}}_2 \tag{3.5}$$

where $\hat{\boldsymbol{e}}_1$ and $\hat{\boldsymbol{e}}_2$ are the first two eigenvectors of $\boldsymbol{S}$ and equivalent to $\boldsymbol{X}_c'\boldsymbol{X}_c = (n-1)\boldsymbol{S}$ and $\boldsymbol{X_c}$ denotes the mean correlated data matrix with rows $\left(\boldsymbol{x_j} - \overline{\boldsymbol{X}}\right)'$. The eigenvectors determine a plane and the coordinates of the j^{th} unit are the pair of values of the principal components $(\hat{y}_{j1}, \hat{y}_{j2})$. The pair of eigenvectors has to be considered in order to include the information on the variables in the plot. These eigenvectors are coefficient vectors for the first two sample principal components. Thus, each row of the matrix positions ($\widehat{\boldsymbol{E}} = [\hat{e}_1, \hat{e}_2]$) a variable in the graph and the magnitudes of the coordinates of the variables show the weightings of the variables. The weightings represent each principal component of the variables. The plots of the variable with corresponding position are indicated by a vector. Singular value decomposition is the direct approach to obtain a *biplot*. Then, the singular decomposition expresses the $n \times p$ mean correlated $\boldsymbol{X}_c$ as

$$\underset{(n\times p)}{\boldsymbol{X}_c} = \underset{(n\times p)}{\boldsymbol{U}}\ \underset{(p\times p)}{\Lambda}\ \underset{(p\times p)}{\boldsymbol{V}'}$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\boldsymbol{V} = \widehat{\boldsymbol{E}} = \left[\hat{\boldsymbol{e}}_\mathbf{1}, \ldots, \hat{\boldsymbol{e}}_{\boldsymbol{p}}\right]$ is an orthogonal matrix whose columns are the eigenvector of $\boldsymbol{X}_c'\boldsymbol{X}_c = (n-1)\boldsymbol{S}$. The best rank two approximation to $\boldsymbol{X}_c$ is obtained by replacing Λ by $\Lambda^* = diag\ (\lambda_1, \lambda_2, 0, \ldots, 0)$. Therefore, this result is known as *Eckart-Young* theorem. The approximation is given as

$$\boldsymbol{X}_c = \boldsymbol{U}\Lambda^{*}\boldsymbol{V}' = [\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2] \begin{bmatrix} \hat{\boldsymbol{e}}_1' \\ \hat{\boldsymbol{e}}_2' \end{bmatrix} \quad (3.6)$$

where $\hat{\boldsymbol{y}}_1$ and $\hat{\boldsymbol{y}}_2$ are the $n \times 1$ vector of values for the first and second principal components respectively.

The *biplot* represents each row of the data matrix by the point located by the pair of values of the principal components. The i^{th} column of the data matrix is represented as an arrow from the origin to the point with coordinates (e_{1i}, e_{2i}), the entries in the i^{th} column of the second matrix $[\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_p]'$ approximations. Furthermore, the idea of a *biplot* extends to canonical correlation analysis, multidimensional scaling and even more complicated nonlinear techniques.

3.3 Application of multiple correspondence analysis

The application of multiple correspondence analysis is used to visualize the associations between the socio-economic, demographic and geographic parameters and the malaria RDT result. Multiple correspondence analysis helps to track the impact of socio-economic, demographic and geographic parameters and the malaria RDT result. Therefore, applying correspondence analysis helps to summarize important effects including interactions in effect reducing the dimensionality of the problem. Beyond the better understanding of the structure of the data the computational time may be significantly reduced. Furthermore, the graphical interpretation of the data is a useful tool in an exploratory research and the reduction of the level of the associations between the investigated parameters.

When applying MCA method, variables are divided into distinct subgroups that contain variables of similar types such as socio-economic, demographic and geographic variables. Variables analyzed with MCA generally are assumed to be categorical. This technique is described by (Guitonneau and Roux, 1977). To apply MCA to both continuous and discrete data, continuous variables could be categorized through a process of mutually exclusive and exhaustive discretization or coding (Greenacre, 1984). Multiple correspondence analysis locates all the categories in a Euclidean space. To examine the associations among the categories, it is important to plot the first two dimensions of the Euclidean space. For the multiple correspondence analysis, malaria RDT result and the other socio-economic, demographic and geographic variables are considered. The demographic variables are sex, age and family size. For the multiple correspondence analysis, the continuous age and family size variables were recoded to be appropriate for the analysis. The socio-economic variables are source of drinking water, time to collect water, toilet facility, availability of radio, television and electricity, construction material for room's floor, wall and roof, use of indoor residual spray, use of mosquito nets, total number of rooms in the house and total number of nets in the house. Besides the socio-economic and demographic variable, there were geographic variables included in the analysis. These variables are region and altitude. Therefore, to perform the MCA analysis all socio-economic, demographic and geographic variables were included to the multiple correspondence analysis.

For MCA analysis, each principal inertia values expressed as a percentage of the total inertia, which quantifies the amount of variation accounted for by the corresponding principal dimension. In addition to this the principal inertia is decomposed into components for each of the rows and columns. The decomposed rows and columns provide the numerical contributions used to interpret the dimensions and the quality of display of each point in the reduced space. The parts which are expressed as percentages are useful to explain the method of determination of the dimensions. The same parts of the dimensions can be expressed relative to the inertia of the corresponding points in the full space and this helps to assess how close the individual points are to the dimension.

Table 3.1 presents inertia and Chi-Square decomposition for multiple correspondence analysis. Correspondence analysis employs chi-square distances to calculate the dissimilarity between the frequencies in each cell of a contingency table. The calculation of the chi-square distances is cell-independent. Table 3.1 suggested that the two dimensions accounts for 19.4% of the total association. The total chi-square statistic in Table 3.1, which is a measure of the association between the rows and columns in the full dimensions of the table, is 2169476 with degrees of freedom 2050. This chi-square represents all pairwise interactions among the factors. The maximum number of dimensions (or axes) is the minimum of the number of rows and columns, minus one.

From Table 3.1, the singular value indicates the relative importance of each dimension in explaining of the inertia, or proportion of variation, in the participant and variable profiles. The singular values can be interpreted as the correlation between the rows and columns of the contingency table. As in principal components analysis, the first dimension explains as much variance as possible, the second dimension is orthogonal to the first and displays as much of the remaining variance as possible, and so on. Singular values of greater than 0.2 indicate that the dimension should be included in the analysis (Hair et al., 1995). However, the proportion of variance explained by each dimension must be balanced with the cut-off point. The singular value and the inertia are directly related i.e., the inertia is an indicator of how much of the variation in the original data is retained in the reduced dimensional solution (Bendixen, 1996). Furthermore, the percentages of inertia accounted for by the first twelve axes are 10.7 per cent and 8.7 per cent, 5.73 per cent, 5.12 per cent, 4.61 per cent, 4.1 per cent, 4.02 per cent, 3.81 per cent, 3.61 per cent, 3.55 per cent, 3.54 per cent and 3.45 per cent, respectively (Table 3.1).

Table 3.1: Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	2 4 6 8 10 ----+----+----+----+----+---
0.42757	0.18282	232503	10.72	10.72	**************************
0.38438	0.14775	187901	8.66	19.38	**********************
0.3126	0.09772	124277	5.73	25.11	**************
0.29555	0.08735	111089	5.12	30.23	*************
0.28047	0.07866	100043	4.61	34.84	************
0.26462	0.07002	89054	4.1	38.94	**********
0.26193	0.06861	87250	4.02	42.97	**********
0.25503	0.06504	82716	3.81	46.78	**********
0.24806	0.06154	78259	3.61	50.39	*********
0.24591	0.06047	76909	3.55	53.93	*********
0.24557	0.06031	76696	3.54	57.47	*********
0.24356	0.05932	75444	3.48	60.94	*********
0.23959	0.0574	73005	3.37	64.31	********
0.23772	0.05651	71869	3.31	67.62	********
0.23474	0.0551	70079	3.23	70.85	********
0.23154	0.05361	68179	3.14	73.99	********
0.22675	0.05142	65388	3.01	77.01	********
0.22274	0.04961	63094	2.91	79.92	*******
0.21997	0.04839	61539	2.84	82.75	*******
0.21788	0.04747	60370	2.78	85.54	*******
0.2095	0.04389	55817	2.57	88.11	******
0.2031	0.04125	52458	2.42	90.53	******
0.1965	0.03861	49106	2.26	92.79	******
0.18357	0.0337	42856	1.98	94.76	*****
0.17417	0.03033	38578	1.78	96.54	****
0.16618	0.02761	35119	1.62	98.16	****
0.14744	0.02174	27646	1.27	99.44	***
0.08754	0.00766	9745	0.45	99.89	*
0.04423	0.00196	2488	0.11	100	
Total	1.70588	2169476	100		
Degrees of Freedom = 2025					

Based on this result, the first twelve axes accounting for 60.9 per cent of the amounts of variance and would expect 39.1 per cent of the inertia to be accounted by the remaining axes. As can be seen from the table, 93 per cent of the association can be represented well in twenty three dimensions. However, these data can be considered in just two dimensions. The first axis accounting for approximately 10.72 per cent of the inertia and the second axis accounts approximately 8.66 per cent. The percentages of inertia in MCA are low and tend to be close to one another and this latter fact might lead to an assumption that individual axes might be unstable.

Figure 3.1 presents the scree plot of singular values. One method to assess most appropriate number of dimensions for interpretation is using scree plot. The scree plot presents the proportions of variance explained (Hair et al., 1995). As can be seen from the figure, the scree plot suggests that the proportion of variance explained drops faster up to 7th dimension and less rapidly up to dimension 26. As discussed by (Hair et al., 1995), 0.2 can be considered as a cut-off point as a first step. But, this cut-off point suggests that only 90.5 per cent variation can be explained with 22 dimensions.

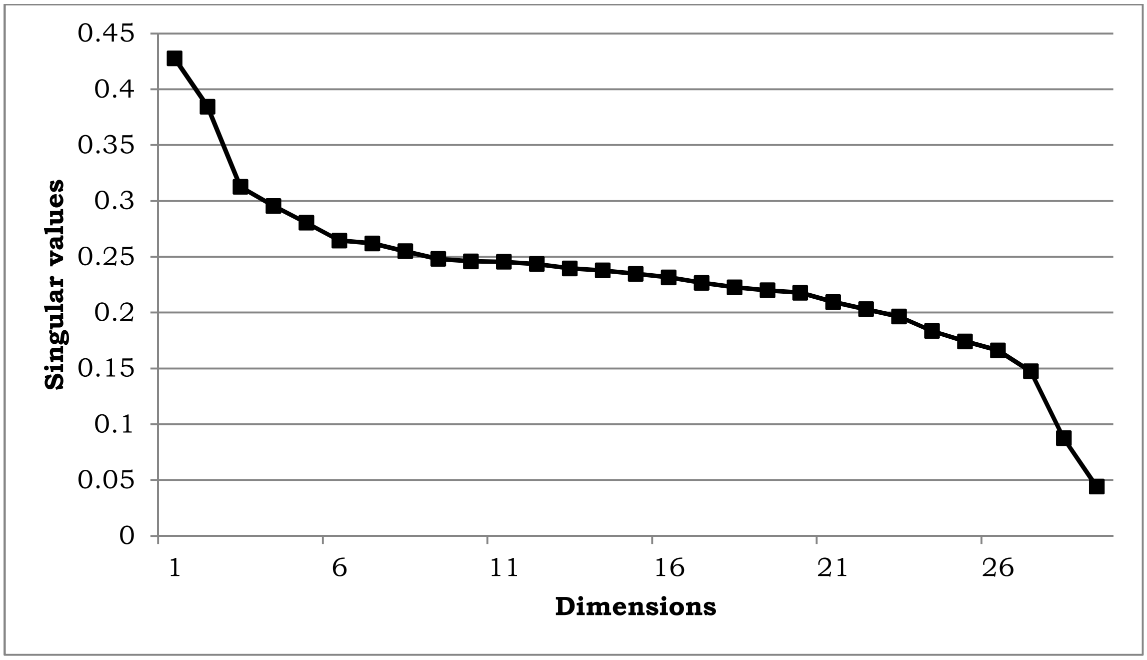
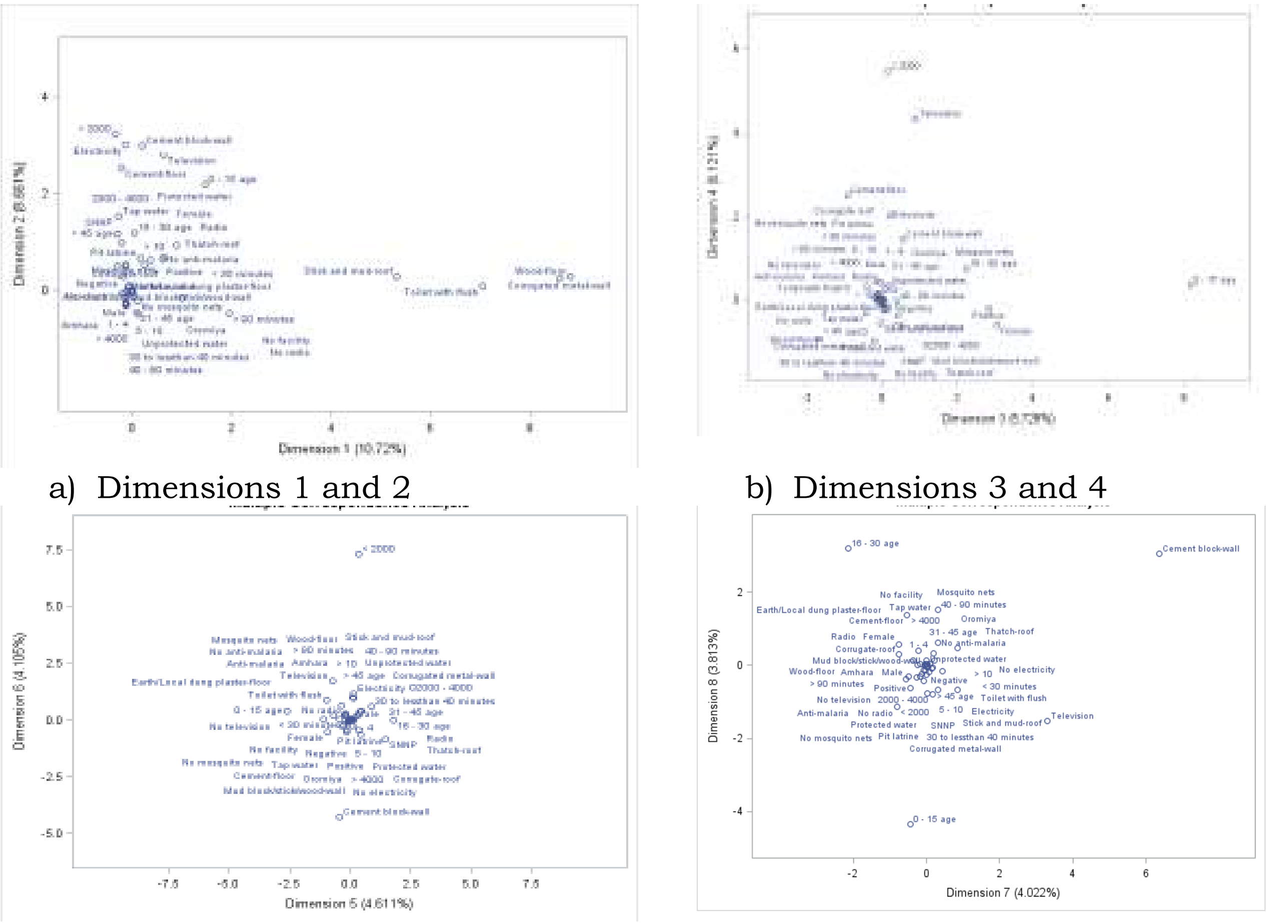


Figure 3.1: Scree plot of singular values

Figure 3.2 contains the multiple correspondence analysis scaling solution coordinates for the variables for twelve dimensions, with Dimension 1 on the horizontal axis and Dimension 2 on the vertical axis and so on. Multiple correspondence analysis locates all the categories in a Euclidean space. The first two dimensions of this space are plotted to examine the associations among the categories. Dimension 1 accounts for 10.72 per cent of the variance in the data and Dimension 2 accounts for 8.66 per cent of the variance. The twelve dimensions totally accounts for 60.9 per cent of the variations. It can be seen that variable like stick and mud roof, toilet with flush, wood floor and corrugated metal wall appears separately in the right hand side of the chart. Therefore, these variables have to be included in the interpretation of dimension 1 and similarly for other dimensions.

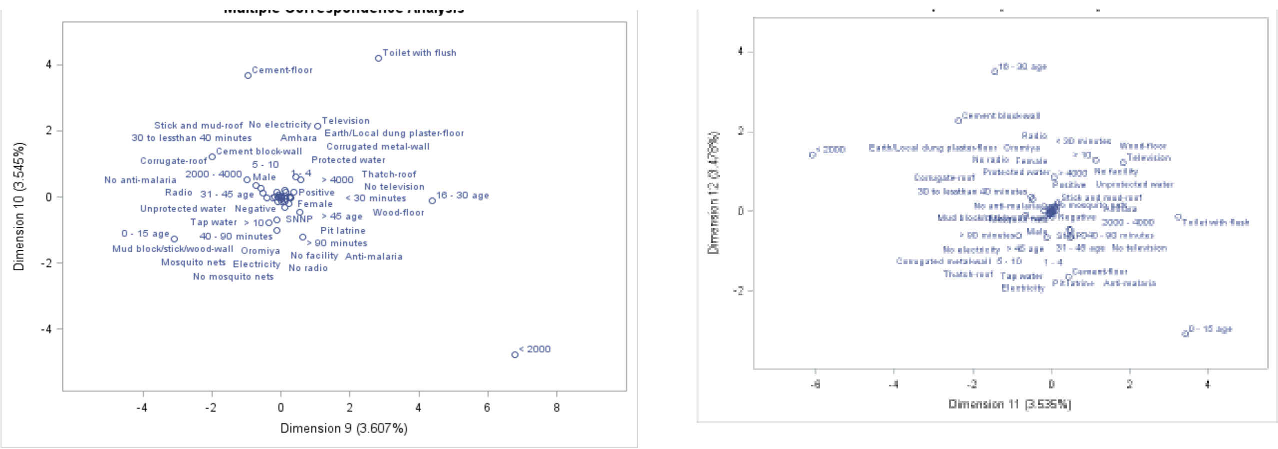


a) Dimensions 1 and 2

b) Dimensions 3 and 4

c) Dimensions 5 and 6

d) Dimensions 7 and 8



e) Dimensions 9 and 10

f) Dimensions 11 and 12

Figure 3.2: Multiple correspondence analysis plot for twelve dimensions

It is important to note that this two-dimensional chart is part of the twenty two dimensional solutions. Interpreting of each dimension is considered as the contribution of variables to that dimension (Clausen, 1998). This is because a variable that appears on the two-dimensional chart might be a major contributor to another dimension but might not be located in the existing two-dimensional plane (Nishisato, 1994). As can be seen in Figure 3.2, the right quadrant of the plot (dimensions 1 and 2) shows that the categories stick and mud roof, toilet with flush, wood floor and corrugated metal wall are associated. To the top of the plot, altitude less than 2000 meter, use of electricity, cement block wall, cement floor, use of television, protected water, altitude between 2000 – 4000 meters are associated. On the other hand, positive malaria RDT result, not using indoor residual spray, thatch roof, earth or dung plaster floor are grouped together. Furthermore, negative malaria RDT result, use of indoor residual spray, use of malaria nets, pit latrine toilet and corrugated floor are associated. Similarly, unprotected water, 30 – 40 minutes to get water, no toilet facility and no radio are associated together. This interpretation of the plot is based on points found in approximately the same direction and in approximately the same region of the space.

So far, the association between socio-economic, demographic, geographic variables and malaria RDT result was assessed based on dimension 1 and 2. Therefore, the contribution of dimension 1 and dimension 2 has been interpreted. As can be seen from Table 3.1, dimension 1 and 2 constitute 19.4 per cent of the variation. But, the other 20 dimensions all together constitute 71.2 per cent of the variation. Except the relationships between dimension 4 and 3, dimension 5 and 2, dimension 5 and 3 and dimension 7 and 1, the relationship between the variables for other combination of dimensions show that they are located at the center of the graphs. The relationships between variables show similar relationships as of dimension 1 and 2.

3.4 Summary and Discussion

In this study, multiple correspondence analysis was used as a way to graphically represent and interpret the relations between primary meanings in different malaria RDT result, socio-economic, demographic and geographic variables. Multiple correspondence analysis provides useful interpretative tools that can further the understanding of the conceptual context in which socio-economic, demographic and geographic variables by malaria RDT result occurs.

As it was discussed above, multiple correspondence analysis is a method for exploring associations between sets of categorical variables. Mathematically, it is a method for breaking down the value of the goodness-of-fit statistic into components due to the rows and columns of the contingency table. It can also be considered as a technique for assigned order to unordered categories. Therefore, the MCA approach involves defining a set of points, with associated masses, in a multidimensional space structured by Euclidean distance. Furthermore, the display is also thought of as a framework for reconstructing the original data as closely as possible. To display the relationship, the coordinate positions of the row and column points are used.

The association using MCA gives the relationship among coded variables and their associations. The technique allows the analysis of the relationships between the variables and different levels of one variable. Furthermore, the results of the analysis can be seen analytically and visually. This method of display gives detailed information of the relationship between variables and their associations. Therefore, the result from multiple correspondence analysis shows that there is association between malaria RDT result and different socio-economic, demographic and geographic variables. Moreover, there is an indication that some socio-economic, demographic and geographic factors have joint effects. It is important to confirm the association between socio-economic, demographic and geographic factors using advanced statistical techniques. Therefore, future investigations need to be done to identify those variables that show significant relationships. By identifying those variables which could have joint effect, it is important to determine the principal axes and the identification of selection of variables to take forward for further analysis. Furthermore, the interaction effects between socio-economic, demographic and geographic variables will be included in the further analysis for this study.

The commonly used methods for discrete (e.g binary) data are a direct extension of generalized linear models for independent observations to the context of correlated data. Therefore, a review of these models is provided in the next chapter. The survey conducted in the Amhara, Oromiya and SNNP regions involves the complex survey method. Detailed review of survey logistic model is also provided in the next chapter. In addition to this, these models will be fitted to the malaria rapid diagnosis test result data to identify socio-economic, demographic and geographic factors that affect malaria rapid diagnosis test result.

Chapter 4

Prevalence and risk factors of Malaria in Ethiopia using Generalized Linear Models

4.1 Introduction

The class of generalized linear models includes many well-known statistical models such as: multiple regression for normal responses; logistic and probit regression for binary responses; binomial counts, or proportions; Poisson and negative binomial regression; log-linear categorical data analysis models; gamma regression for variance models; and exponential and gamma models for survival time models.

The literature on generalized linear models and their extensions are vast (Berridge and Crouchley, 2011, Zuur et al., 2009, Zurr et al., 2007, Fox, 2008, Madsen and Thyregod, 2010). Generalized linear models have been extended in many ways, such as accommodating random and mixed effects, accommodating correlated data, relaxing distributional assumptions, allowing semiparametric linear predictors, etc (Schimek, 1997, Smith et al., 2004).

In statistics, the flexible generalization of ordinary least squares regression is generalized linear model (GLM). The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Generalized linear models were formulated by John Nelder and Robert Wedderburn in 1972 as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression (Nelder and Wedderburn, 1972). John Nelder and Robert Wedderburn proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters.

In summary the current chapter is organized as follows. An overview of the theory of GLM and survey logistic is presented in sections 4.2 – 4.5. The survey logistic model is fitted to malaria RDT result data in section 4.6. Summary and discussion of this chapter is given in section 4.7.

4.2 Generalized Linear Model

Generalized Linear Model (GLM) is an extension of the linear modelling process that allows models to be fitted to data that follow probability distributions other than the Normal distribution. GLM helps to include response variables that follow any probability distribution in the exponential family of distributions. The exponential family includes such useful distributions as the Normal, Binomial, Poisson, Multinomial, Gamma, Negative Binomial, and others. Hypothesis tests applied to the Generalized Linear Model do not require normality of the response variable, nor do they require homogeneity of variances. Hence, Generalized Linear Models can be used when response variables follow distributions other than the Normal distribution.

Let $y_1,\ldots,y_n$ denote n independent observations on a response variable $\mathbf{y}$. We treat y_i as a realization of a random variable Y_i. In the general linear model formulation we assume that Y_i has a normal distribution with mean μ_i and variance σ^2

$$Y_i \sim N(\mu_i, \sigma^2),$$

and further assumed that the expected value μ_i is a linear function of p predictors that take values $x_i' = (x_{i1}, \ldots, x_{ip})$ for the i^{th} observation, so that

$$\mu_i = \boldsymbol{X}_i'\beta,$$

where β is a vector of unknown parameters.

The Exponential Family of Generalized Linear Models

Nelder and Wedderburn introduced the generalized linear model (GLIM) in 1972. The GLM models consist of independent responses Y_i, $i = 1, 2, \dots, n$, with an exponential family distribution as follows

$$f(y) = f(y|\theta, \psi) = \ exp\left[y^{\theta}\left(g^{-1}(X_i'\beta_G)\right), \psi\right], \qquad (4.1)$$

where *exp* represents an exponential family member with parameters $\theta$and ψ; and ψ may be known (Nelder and Wedderburn, 1972). The parameter θ is a function of the mean and can be written as $g^{-1}(X_i'\beta_G)$ (a function of a linear combination of the regressors). In general, generalized linear models have three features (McCullagh and Nelder, 1989). These features are the random component, systematic component and the link function. These feature are explained as follows.

- A *random component* consists of a response variable Y from the exponential family with independent observations $(y_1, y_2, \dots, y_n)$. The density function for exponential family is given by

$$f_Y(y_i) = \exp\left[\frac{y^{\theta} - \ b(\theta)}{\psi} + \ c(y_i, \psi)\right], \qquad i = 1, 2, \dots n$$

where, $Y_1, Y_2, \dots, Y_n$, are assumed to be independent. θ and ψ are parameters while $b(\theta)$ and $c(y, \psi)$ are known functions. The parameter θ is termed the canonical parameter and is related to $E[Y_i]$ through $b(\bullet)$. Therefore, $\mu = E[Y_i] = \ b'(\theta)$.

The variance of Y is a function of the mean and the scale parameter or dispersion parameter ψ,

$$V\,ar[Y_i] = \ \psi\frac{\partial\mu}{\partial\theta} = \ \psi b''\,(\theta).$$

where, $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$ with respect to θ. In general, the mean and variance of Y can be derived by using the property $\int f(y|\theta,\psi)dy = 1$. Taking the first and second derivatives with respect to θ from both sides of the equation gives

$$\int \left(y - b'(\theta)\right) f(y/\theta,\psi)dy = 0 \text{ and}$$

$$\int \left[\psi^{-1}\left(y - b'(\theta)\right)^2 - b''(\theta)\right] f(y/\theta,\psi)dy = 0.$$

Therefore, $E(y) = \mu = b'(\theta)$ and $var(y) = b''(\theta)\psi$. Unlike multiple regression and other normal distribution based models, the variance of generalized linear models can depend on the mean. If $b''(\theta)$ is expressed as a function a of the mean, $b''(\theta) = V(\mu)$, then V is called the variance function. The parameter ψ is a scale parameter. When it is unknown it must be estimated along with θ.

- The *systematic component* of a GLM relates a vector $(\eta_1, \dots, \eta_n)$ to the regressor variables through a linear model. Associated with each response y_i is a vector $\boldsymbol{X}_i$ denote the value of predictor $\boldsymbol{X}_i = (x_{i1}, x_{i2} \dots, x_{ip})'$ of values of p explanatory variables, then the distribution of the response variable y_i depends on $\boldsymbol{X}_i$ through the linear predictor η_i where

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} .$$

The systematic component of the linear form places the regressors on an additive scale. Therefore, this scale makes the interpretation of their effects simple. Moreover, the significance of each regressor can be tested with a linear hypothesis $H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$ for $i = 1, 2, \dots, k$.

- The function $g(\mu_i)$ is called a link function which connects the linear predictor to the mean $E[Y]$. This is done through a monotonic, differentiable function

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{ip}\, x_p.$$

Here, the link is a linearizing transformation of the mean which is a function that maps the mean onto a scale where regressor effects are linear. The link is used to allow η_i to range freely while restricting the range of μ_i. For example, the inverse logit link $\mu = 1/(1 + e^{-\eta})$ maps $(-\infty, \infty)$ onto $(0, 1)$, which is an appropriate range if μ_i is a probability. The monotonicity of the link function guarantees that this mapping is one-to-one. Therefore, the generalized linear model can be expressed in terms of the inverse link function,

$$E[Y_i] = g^{-1}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{ip}\, x_p).$$

For a linear predictor which is equal to canonical parameter θ, the canonical link is given by $\theta(\mu)$. The canonical link is useful and reasonable link function. The canonical link does the estimation method, but it is necessary to restrict generalized linear modelling to canonical link functions (Agresti, 2002).

The notation $F(x_i' \beta_G)$ can be used for $g^{-1}(x_i' \beta_G)$, $i = 1, 2, \ldots, n$, stacked in a vector for the generalized linear mean model. Therefore, generalized linear model for the entire dataset can be expressed as additive form as follows

$$\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{X}\beta_G) + \epsilon, \; where\, \epsilon \sim (\boldsymbol{0}, \langle \psi a(\mu_i) \rangle). \tag{4.2}$$

The application of iteratively reweighted least squares was extended to obtain maximum likelihood estimates (Finney, 1952, Nelder and Wedderburn, 1972). The term deviance was introduced as a measure of model fit. Moreover, generalized analysis of variance was considered as the change in deviance of a sequential fit of nested models (Good, 1967). (McCullagh and Nelder, 1983) first introduced generalized linear models and their second edition in 1989 (McCullagh and Nelder, 1989) serves as the standard monograph on generalized linear models. The literature on generalized linear models and their extensions are voluminous. Generalized linear models have been extended in many ways, such as accommodating random and mixed effects, accommodating correlated data, relaxing distributional assumptions, allowing semiparametric linear predictors, etc (Schimek, 1997, Smith et al., 2004).

4.3 Estimation in Generalized Linear Models

The method of maximum likelihood (ML) can be used to estimate the parameters in the linear predictor η_i. Assume $Y_i, i = 1, \ldots n$ be independent, the joint likelihood is the product of the likelihoods for each Y_i. The log likelihood for β_G, as a function of an arbitrary β, is then

$$l(\beta|\mathbf{y}) = \sum_{i=1}^{n}\left\{\frac{[y_i\theta_i - b(\theta_i)]}{\psi} + c(y_i, \psi)\right\}. \tag{4.3}$$

The likelihood problem can be solved by taking the derivative of the log likelihood $l(\beta|\mathbf{y})$ under the properties of the exponential family and the fact that the link g is monotonic. The score equations obtained from equating the first order derivatives of the log likelihood to zero gives

$$S(\beta) = \sum_{i}^{n}\frac{\partial\theta_i}{\partial\beta}\,[y_i - b(\theta_i)] = 0. \tag{4.4}$$

Since $\mu_i = b'(\theta_i)$ and $V_i = V(\mu_i) = b''(\theta_i)$, then

$$\frac{\partial\mu_i}{\partial\beta} = b''(\theta_i)\frac{\partial\theta_i}{\partial\beta} = V_i\frac{\partial\theta_i}{\partial\beta},$$

and the result implies the following equations

$$S(\beta) = \sum_{i}^{n}\frac{\partial\theta_i}{\partial\beta}\,V_i^{-1}[y_i - \mu_i] = 0. \tag{4.5}$$

Solving the score equation (4.5) gives the ML estimates of β.

The score equations can be solved iteratively. Initial solution of the equations is guessed and then updated until iterative algorithm converges to the solution $\hat{\beta}$, called the maximum likelihood estimate of β. The methods of Fisher's scoring and Newton-Raphson are the two most popular and widely used iterative algorithms for the maximum likelihood estimation. The Fisher's scoring method is equivalent to the iterative reweighted least squares. The Newton-Raphson method solves maximum likelihood estimates iteratively using the standard least-squares methods (Agresti, 1990, McCullagh, 2008). Classical inferences based on asymptotic likelihood theory become available, including Wald-type tests, likelihood ratio tests and the score tests, all asymptotically equivalent once the maximum likelihood estimates have been obtained. Moreover, with some models such as the logistic regression model, ø is a known constant. For models, like the linear normal model, estimation of ø may be required to estimate the standard errors of the elements in β. There are several ways of estimating ø, one of which is given by

$$\hat{ø} = \frac{1}{N - p} \sum_{i} (y_i - \hat{\mu}_i)^2 lV_i(\hat{\mu}_i)$$

where n is the total number of observations and p is the number of parameters in the model.

Detailed discussion of Fisher's scoring and Newton-Raphson can be found in different literatures (Agresti, 1990, Kutner et al., 2005, McCullagh, 2008, McCullagh and Nelder, 1989, Schabenberger and Pierce, 2002).

4.4 Survey logistic regression for binary data

The logistic regression model is classified under generalized linear models. This model is used to model binary data. But, the standard statistical methods are inappropriate for analyzing survey data due to clustering and stratification used in the survey design. Therefore, some adjustments to the classical methods that take account of the survey design are necessary in order to make valid inferences (Chen and Mantel, 2009). Therefore, the logistic regression model used to analyze data from complex sampling designs is referred to as survey logistic regression models. Survey logistic regression models have the same theory as ordinary logistic regression models. The difference between ordinary and survey logistic is that survey logistic accounts for the complexity of survey designs, i.e. sampling techniques, such as stratified random or cluster sampling including multi-stage sampling. But, for data from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical. To apply survey logistic to the current problem, the first stage primary sampling unit (PSU), was a *Kebele* (the smallest administrative unit in Ethiopia). In the second stage, households with in a *kebele* sampled. The response of the i^{th} person in the j^{th} household and h^{th} *Kebele* can be specified as y_{ijh} ($i = 1,2,\ldots,mh_j$; $j = 1,2,\ldots,n_h$; and $h = 1,2,\ldots,H$) where y_{ijh} equals 1 if there is positive malaria rapid diagnosis test result in the j^{th} household within h^{th} *Kebele* (PSU), and 0 otherwise. Thus, the log-likelihood function in this case is given by

$$l(\beta;\boldsymbol{y}) = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}}\left\{y_{ijh}log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right) - log\left(\frac{1}{1-\pi_{ijh}}\right)\right\}$$

and the survey logistic regression model is given by

$$logit(\pi_{ijh}) = x'_{ijh}\beta\ , \quad i = 1,2,\ldots,m_{hj};\ j = 1,2,\ldots,n_h;\ \text{and}\ h = 1,2,\ldots,H$$

where x_{ijh} is the row of the design matrix corresponding to the characteristics of the i^{th} person in the j^{th} PSU within h^{th} stratum, and β is a vector of unknown parameters of the model. To obtain reliable inference about the effects of factors from the fitted model, it is important to include all design variables in the model as explanatory variables (Pfeffermann, 1993).

Estimation of Parameters

For ordinary logistic regression, a method of maximum likelihood estimation is used to estimate parameters of the model. But, estimation of the standard errors of the parameter estimates is very complicated for data which comes from complex designs. The complexities in variance estimation arise partly from the complicated sample design and the weighting procedure imposed. So a rough estimate for the variance of a statistic based on a complicated sample can be obtained either by ignoring the actual complicated sample design used and proceeding to the estimation process using the straightforward formulae of the simple random sampling or another similarly simple design (Park, 2008, Rabe-Hesketh and Skrondal, 2006, Biewen and Jenkins, 2006). But, the incorporation of sampling information is important for the proper assessment of the variance of a statistic (Park, 2008). Since weighting and specific sample designs are particularly implemented for increasing the efficiency of a statistic, their incorporation in the variance estimation methodology is of major importance (Schaefer et al., 2003). Thus, the bias induced under this simplifying approach depends on the particular sampling design and should be investigated circumstantially (Lehtonen and Pahkinen, 2004). Therefore, there are several methods to obtain the covariance matrix. These methods include the Taylor expansion approximation procedure, jackknife estimator, bootstrap estimator, balanced repeated replication method and random groups method (Wolter, 1985, Lee and Forthofer, 2006).

Taylor expansion approximation procedure

The Taylor series approximation method relies on the simplicity associated with estimating the variance of a linear statistic, even with a complex sample design. By applying the Taylor linearization method, nonlinear statistics are approximated by linear forms of the observations (by taking the first-order terms in an appropriate Taylor-series expansion). But, it has to be noted that Taylor series linearization is, essentially used in elementary cases, while influence function can be deployed in complex cases.

The estimation of variance of the general estimator is adapted from the Taylor-series expansion. To use the Taylor series expansion, consider a finite population N. Let $p-$dimensional parameter vector be denoted by $\boldsymbol{Y} = (Y_1, \dots, Y_p)'$ where, Y_j are population totals or means. The corresponding estimator vector is denoted by $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_p)'$ based on a sample size s of $n(s)$. Therefore, the estimators $\widehat{Y}_j$, $j = 1, \dots, p$ depends on the sampling design generating the sample s. Let us consider a nonlinear parameter $\theta = f(\boldsymbol{Y})$ with a consistent estimator denoted by $\hat{\theta} = f(\widehat{\boldsymbol{Y}})$. Therefore, the interest here is to find an appropriate expression for the design variance of $\hat{\theta}$ and constructing a suitable estimator of the variance of $\hat{\theta}$ (Wolter, 1985).

Suppose that continuous second–order derivative exists for the function $f(\boldsymbol{Y})$. Therefore, using the linear terms of the Taylor-series expression, the approximate linearized expression is

$$\hat{\theta} - \theta = \sum_{j=1}^{s} \frac{\partial f(\boldsymbol{Y})}{\partial Y_j} \left(\hat{y}_j - y\right), \tag{4.6}$$

where, $\partial f(\boldsymbol{Y}) / \partial Y_j$ refers to partial derivation. Using equation (4.6), the variance approximation of $\hat{\theta}$ is given by

$$V(\tilde{\theta}) = V\left(\sum_{j=1}^{s} \frac{\partial f(\boldsymbol{Y})}{\partial Y_j}(\hat{y}_j - y)\right) = \sum_{j=1}^{s} \frac{\partial f(\boldsymbol{Y})}{\partial Y_j} . \frac{\partial f(\boldsymbol{Y})}{\partial Y_k} V((\hat{y}_j, \hat{y}_k). \qquad (4.7)$$

Here, the variance of nonlinear estimator $\hat{\theta}$ has been reduced to a function of variances and covariances of s linear estimators $\hat{Y}_j$ (Wolter, 1985). Therefore, the variance estimator $\hat{V}(\hat{\theta})$ is obtained from (4.7) (Skinner et al., 1989).

The resulting variance estimator in equation (4.7) is referred to as the first order approximation. Extending the Taylor series expansion could develop second or even higher-order approximations. However, in practice, the first-order approximation usually yields satisfactory results, with the exception of highly skewed populations (Wolter, 1985). Standard variance estimation techniques can then be applied to the linearized statistic. This implies that Taylor linearization is not a *'per se'* method for variance estimation, it simply provides approximate linear forms of the statistics of interest and then other methods should be deployed for the estimation of variance itself. The Taylor linearization method is a widely applied method, quite straightforward for any case where an estimator already exists for totals. However, the Taylor linearization variance estimator is a biased estimator. Its bias stems from its tendency to under estimate the true value and it depends on the size of the sample and the complexity of the estimated statistic. Though, if the statistic is fairly simple, like the weighted sample mean, then the bias is negligible even for small samples, while it becomes nil for large samples (Särndal et al., 1992). On the other hand for a complex estimator like the variance, large samples are needed before the bias becomes small. In any case, however, it is a consistent estimator.

Jackknife estimator

The jackknife technique is developed by (Quenouille, 1949, Quenouille, 1956). The main idea of jackknife is to divide the sample into disjoint parts, dropping one part and recalculating the statistic of interest based on incomplete sample. The dropped part is re-entered in the sample and the process is repeated successively until all parts have been removed once from the original sample. These replicated statistics are used in order to calculate the corresponding variance. Disjoint parts mentioned above can be either single observation in a simple random sampling or clusters of units in multistage cluster sampling schemes. The choice of the way that sampling units are entered, re-entered in the sample leads to a number of different expressions of jackknife variance.

It should also be noted that the jackknife method for variance estimation is more applicable in with replacement designs, though it can also be used in without replacement surveys when the sampling fraction is small (Wolter, 1985). However, this is rarely the case when we are dealing with business surveys. The impact of its use in surveys with relatively large sampling fraction is illustrated, via simulation in (Smith et al., 1998), while, as mentioned in (Shao and Tu, 1995) the application of jackknife requires a modification to account for the sampling fractions only when the first stage sampling is without replacement. In any case, due to their nature, jackknife variance estimation methods seem to be more appropriate for (single or multistage) cluster designs, where in each replicate a single cluster is left out of the estimation.

If the number of disjoint parts (e.g. clusters) is large, the calculation of replicate estimates is time consuming, making the whole process rather than time-demanding in the case of large-scale surveys (Yung and Rao, 2000). So, alternative jackknife techniques have been developed (Efron, 1982).

Jackknife linearized variance estimation is a modification of the standard jackknife estimator based on its linearization. Its essence is that repeated recalculations of a statistic are replaced by analytic differentiation. The result is a formula that it is easy to calculate. For example for stratified cluster sample the bias adjusted variance formula, presupposing sampling with replacement, is (Canty and Devison, 1999):

$$\hat{v} = \sum_{h=1}^{H}(1-f_h).\frac{1}{n_h.(n_h-1)_-}\sum_{j=1}^{n_h} l_{hj}^2.$$

The factor l_{hj} is the 'empirical influence value' for the j^{th} cluster in stratum h (Canty and Devison, 1999). The effort required for calculating l_{hj} is based on the complexity of the statistic. For the linear estimator in stratified cluster sampling:

$$\hat{\theta} = \textstyle\sum_{j,h} y'_{hj}$$

where,

$$y'_{hj} = \textstyle\sum_k \omega_{hjk}.y_{hjk}$$

is the sum of $y's$ in every cluster j in each stratum h, and ω_{hjk} is the design weights then

$$l_{hj} = n_h \,. y'_{hj}.$$

For the ratio of two calibrated estimators, l_{hj} is:

$$l_{hj} = \frac{l_{hj}^{y} - \hat{\theta}.l_{hj}^{z}}{l^T W_z}$$

where

$$\hat{\theta} = \frac{l^T W_y}{l^T W_Z}$$

while y and z are the vectors of the observations in the dataset and l_{hj}^{y} , l_{hj}^{z} and W are calculated from the data analytically.

Therefore, the main advantage of jackknife estimator is that it is less computationally demanding, while it generally retains the good properties of the original jackknife method. However, in case of non-linear statistics, it requires the derivation of separate formulae, as is the case with all linearised estimators. Therefore, its usefulness for complex analyses of survey data or elaborate sample designs is somewhat limited. More details can be found at (Canty and Devison, 1999, Rao, 1997), while an insightful application is made by (Holmes and Skinner, 2000).

Bootstrap estimator

Similar to jackknife method, bootstrap method was introduced outside survey sampling which was originated by (Efron, 1979, Efron, 1981, Efron, 1982). Bootstrap was introduced for samples of independent and identically distributed observations. Since then, there has been much theoretical and empirical research examining properties of the bootstrap estimator. Moreover, bootstrapping has become a popular tool for classical statistical analysis (Shao and Tu, 1995). The bootstrap involves drawing a series of independent samples from the sampled observations, using the same sampling design as the one by which the initial sample was drawn from the population and calculating an estimate for each of the bootstrap samples (Rao and Wu, 1988).

Balanced repeated replication (BRR) method

The balanced repeated replication method (BRR) (or balanced half samples, or pseudoreplication) developed for the case with a large number of strata. This method has a very specific application in cluster designs where each cluster has exactly two final stage units or in cases with a large number of strata and with only two elements per stratum. The aim of this method is to select a set of samples from the family of $2k$ samples, compute an estimate for each one and then use them for the variance estimator in a way that the selection satisfies the balance property (Särndal et al., 1992).

In the BRR technique, the formation of pseudo samples starts H strata and *r=2* sample clusters per stratum. If there are no PSUs per stratum, these form the replication. Therefore, the total sample can be split into 2^H overlapping half-samples each with H sample clusters. Therefore, the estimate $\hat{\theta}_i$ can be constructed for each half-samples and be used to estimate $V(\hat{\theta})$. But, it is computationally expensive to evaluate all 2^H possible $\hat{\theta}_i$. Therefore, it is possible to select a balanced set of only k half-samples where k is minimum multiple of 4 greater than H. Therefore, the estimator can be given as follows.

$$V(\hat{\theta}) = \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2/k. \qquad (4.8)$$

The estimator (4.8) has equal asymptotic precision to the same estimator evaluated over all *2H* half-samples. The gain in precision of the variance estimate compared to simple replication needs to be balanced against the increased computation required (Rao and Wu, 1985). The recent research result of (Rao and Shao, 1996) shows that any asymptotically correct estimator can only be obtained by using repeated division, i.e. repeatedly grouped balanced half samples. Therefore, the use of BRR with business surveys is typically difficult, as stratification is regularly used and the manipulation of both data and software becomes very difficult. According to (Rao, 1997) the main advantage of BRR method over the jackknife is that it leads to asymptotically valid inferences for both smooth and non-smooth functions. However, it is not easily applicable for arbitrary sample sizes like the bootstrap and the jackknife.

Random groups method

For complex surveys, the random group method is one of the first methods developed in order to simplify variance estimation. To estimate the parameters using random group method, drawing sub-samples from the population is required. Then the variance will be assessed based on deviances from the union of sub-samples (Wolter, 1985). This technique is described as follows. To estimate the variance, the design of the survey should involve *r* independent replications of the same basic design. This process gives a final sample consisting of *r* replicates (Skinner et al., 1989). Let $\hat{\theta}$ denotes the estimator of θ from the whole sample. Hence, any statistic $\hat{\theta}$ for the parent sample can be recomputed for each of *r* replicates giving $\hat{\theta}_1, \dots, \hat{\theta}_r$. $\hat{\theta}_i$ is the estimator obtained from the r^{th} random group and $\hat{\hat{\theta}} = \sum_{i=1}^{r} \hat{\theta}_i / r$. Therefore, the variance estimator $V(\hat{\hat{\theta}})$ can be estimated by

$$V\left(\hat{\hat{\theta}}\right) = \frac{1}{r(r-1)} \sum_{i=1}^{r} (\hat{\theta}_i - \hat{\hat{\theta}})^2.$$

Hence, $\hat{\theta}$ can be estimated by $V(\hat{\hat{\theta}})$ (Wolter, 1985), where

$$V(\hat{\theta}) = \frac{1}{r(r-1)} \sum_{i=1}^{r} (\hat{\theta}_i - \hat{\theta})^2.$$

Random groups method can be distinguished into two main variations, based on whether the sub-samples are independent or not. But, in practice, survey sample is drawn at once and random groups technique is applied in the sequel by drawing, essentially, sub-samples of the original sample. In such cases, we have to deal with dependent random groups. For the case of independent random groups, random groups method provides unbiased linear estimators, though small biases may occur in the estimation of non-linear statistics. In case of dependent random groups, a bias is introduced in the results, which, however, tends to be negligible for large-scale surveys with small sampling fraction. In such circumstances the uniformity of the underlying sampling design of each sub-sample is a prerequisite for safeguarding the acceptable statistical properties of the random groups variance estimator.

Comparison of the methods

The applicability of variance estimation methods depends on the sampling design and the adjustments. Obviously, the best approach to estimate the variance is exact formulae, but the exact methods for many practical cases of complex surveys are too difficult to be derived. There are many theoretical studies conducted to compare replication methods with Taylor linearization. These theoretical studies to compare the estimation methods were conducted by (Krewski and Rao, 1981, Rao and Shao, 1992). These studies showed that linearization and replication approaches are asymptotically equivalent and both methods lead to consistent variance estimators. Among the replication methods, jackknife methods have similar properties with linearization approach. But, the properties of balanced repeated replications and bootstrap techniques are comparable. In general, in the case of simple situations of sample designs and estimation features, linearization may be simpler to interpret and less time demanding. However, in case of complex survey design and estimation strategies, replication methods are equivalently flexible.

The summarized findings for the comparison of the variance estimation methods are presented in (Wolter, 1985). After reviewing and summarizing from five different studies (Bean, 1975, Deng and Wu, 1987, Dippo and Wolter, 1984, Frankel, 1971, Mulry and Wolter, 1981), (Wolter, 1985) concludes that '… we feel that it may be warranted to conclude that the TS [Taylor series] method is good, perhaps best in some circumstances, in terms of the mean square error (MSE) and bias criteria, but the BHS [balanced half-samples] method in particular, and secondarily the RG [random groups] and J [jackknife] methods are preferable from the point of view of confidence interval coverage probabilities' (Wolter, 1985, pp. 361).

Furthermore, the advantages of flexibility and cost compared among the variance estimation methods by (Wolter, 1985). Based on the comparison, Taylor linearization method, Jackknife estimator, Balanced repeated replications and Random groups methods are equally flexible. But, based on costs, Jackknife is more expensive than the others. Moreover, the random group method is slight edge in the terms of flexibility. In the stratified sampling setting with a fixed number of strata, bootstrap procedures are available that provides improvements over classical approaches for constructing confidence intervals based on the normal approximation. However, the improvements are of second order and are generally only noticeable when the sample sizes are small. Moreover, in the case where there are an increasing number of strata, replication methods are likely to lose their appealing features as they provide minor asymptotic improvement over the standard normal approximation.

Model Selection and Model Checking for survey logistic

The same selection procedure which can be used for logistic regression could be applied for survey logistic regression models. However, the selection procedures (i.e. forward, backward, and stepwise) are not yet included in SAS 9.2 for PROC SURVEYLOGISTIC procedure. Therefore, the best alternative to select the best model is to start with the saturated model and observe the contribution of each effect to deviance reduction given by type III analysis of effects, then exclude one variable with insignificant effect (one at a time) and observe the contribution of the remaining effects to deviance reduction. This process will continue until the model has only significant effects.

In addition, the Akaike's information criterion (AIC) introduced by (Akaike, 1974), and the Schwarz Criterion (SC) (also known as Bayesian Information criterion (BIC)) introduced by (Schwarz, 1978) can also be used to compare the goodness-of-fit of two nested models. These methods are used to adjust the likelihood ratio statistic $-2logL$ which measures the deviation of the log-likelihood of the fitted model from the log-likelihood of the maximal possible model (Vittinghoff et al., 2005). It is necessary to adjust $-2logL$. The reason for the adjustment is that, $-2logL$ will always decrease as a new explanatory variable enters the model even if it is insignificant. Therefore, the AIC is given by

$$AIC = -2logL + 2p$$

where p is the number of parameters used in the model. This technique, which tolerates violation of parametric model assumptions, can be used to compare multiple nested models, and it does not rely entirely on p-values for determining significance of explanatory variables. In addition to AIC, another criterion, i.e. SC, adjusts the $-2logL$ statistic for the number of parameters and is given by

$$SC = -2logL + p\,log(n)$$

where p is as explained above and n is the overall sample size. Therefore, the smaller the value of the criteria, the better the goodness-of-fit of the model.

The AIC and SC criteria will be used to test for the goodness-of-fit of the model. Since the criteria involve $-2logL$ is only used for variable selection in the case of ungrouped binary data, they are used as approximations. The Hosmer-Lemeshow goodness-of-fit statistic which is used in the case of ungrouped binary data, is not yet implemented in the PROC SURVEYLOGISTIC.

Model checking

For all types of statistical models, assessing model fit is important. Assessing the model includes OLS linear regression models. For such models, assessing of the model is typically examined by statistics like the coefficient of determination (or R^2) and the F-ratio. But, for other members of the generalized linear model, these cannot be applied. Therefore, assessing the model relies on a more general set of criteria for assessing model fit. Furthermore, to assess the goodness of fit, two different statistical methods can be used. These methods are the deviance and Pearson X^2. These methods are approximates for small samples. But, for large samples, the two methods are statistically equivalent. These methods measure the discrepancy of fit between the maximum log-likelihood achievable and the achieved log-likelihood by the fitted model (Jiang, 2001, Kutner et al., 2005).

Table 4. 1: Fit range of models

Model	Link function	Fitted values
Null model	$g(\mu_i) = \alpha$	$\hat{\mu}_i = \hat{\mu}^{(n)}$
Intermediate model	$g(\mu_i) = \boldsymbol{X}'\beta$	$\hat{\mu}_i = g^{-1}(\boldsymbol{X}'\beta)$
Saturated model	$g(\mu_i) = \alpha_i$	$\hat{\mu}_i = \hat{\mu}_i^{(s)}$

Suppose there are n observations, the fit range of models can be given as follows (Table 4.4). The most widely used statistic, log-likelihood whose idea is similarly to sum of squares for linear models for constructing criteria for assessing goodness of fit for generalized linear models, is Deviance. But, the question is what Deviance means for goodness of fit. If the deviance is huge, then the model "*doesn't fit very well*". And if deviance is small, it "*fits well*". But, it is not possible to be specific. Therefore, the scaled deviance of the intermediate model is given by

$$D(\boldsymbol{y};\hat{\mu}) = 2\left[l\left(\hat{\mu}^{(s)},\psi;\boldsymbol{y}\right) - l(\hat{\mu},\psi;\boldsymbol{y})\right]$$

$$= \sum_{i=1}^{n} 2w_i \left\{y_i\left(\theta\left(\hat{\mu}^{(s)}\right) - \theta\left(\hat{\mu}_j\right)\right) - b\left(\hat{\mu}^{(s)}\right) + b\left(\hat{\mu}_j\right)\right\}/\psi$$

$$= \frac{D^*(\boldsymbol{y};\hat{\mu})}{\psi} \geq 0,$$

where $\mathrm{l}(\hat{\mu}, \psi; y)$ is the log-likelihood under the current model, $\mathrm{l}(\hat{\mu}^{(s)}, \psi; y)$ is the log-likelihood under the maximum achievable (saturated) model, $\hat{\mu}$ is the MLE in the intermediate model, and $D^{*}(y; \hat{\mu})$ is called the deviance of the intermediate model. The general aim of the deviance is to minimize D ($D(\boldsymbol{y}; \hat{\mu})$) by maximizing $l(\hat{\mu}, \psi; \boldsymbol{y})$. Furthermore, the deviance is used to compare two nested models having P_1 and P_2 parameters respectively, where $P_1 < P_2$. Let $\hat{\mu}^1$ and $\hat{\mu}^2$ denote the corresponding MLEs.

Therefore, the test statistic is

$$\frac{D^{*}(\boldsymbol{y}; \hat{\mu}^1) - D^{*}(\boldsymbol{y}; \hat{\mu}^2)}{\psi} = -2[l(\hat{\mu}^1, \psi; \boldsymbol{y}) - l(\hat{\mu}^2, \psi; \boldsymbol{y})] \sim X^2_{P_2 - P_1}.$$

If ψ is unknown, it is normally estimated from the large model:

$$\hat{\psi} = \frac{1}{n - P_2} \sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i^2)^2}{V(\hat{\mu}_i^2)},$$

where

$$V(\hat{\mu}_i^2) = \frac{var(y_i)}{a(\psi)} = w_i var(y_i)/\psi.$$

For unknown ψ, it can be estimated by $\psi = \frac{D}{n-p}$, where n is the number of observations and p is the number of parameters. D (or D^{*}) has an asymptotic chi-square distribution with $n - p$ degrees of freedom. To use this statistical methods, asymptotic properties of the goodness-of-fit test of the current model should be satisfied (Schabenberger and Pierce, 2002, Der and Everitt, 2002).

For the measure of goodness-of-fit, Pearson X^2 is used. For the categorical dependent variable, this indicator is quite indicative of the X^2 statistics. Furthermore, Pearson's X^2 test examines the sum of the squared differences between the observed and expected number of cases per covariate pattern divided by its standard error. For ordinary logistic regression, let n

observations are independently sampled, a covariate pattern is defined to be a unique set of the $\boldsymbol{x}_i$'s, where $i = 1, \dots, n$, and m_k will represent the number of subjects with the same covariate pattern where $k = 1, \dots, K$ represents the number of unique covariate patterns. For the estimated probabilities $\hat{\pi}_i$, the values are the same for all m_k subjects in the same covariate pattern. Let y_i represents the outcome for all i^{th}subject, and y_k represents the sum of the observed outcomes in the k^{th} covariate pattern. The Pearson's X^2 goodness-of-fit for logistic regression is expressed as the sum of the squared Pearson's residuals, that is

$$X^2 = \sum_{k=1}^{K} \frac{(y_k - m_k\hat{\pi}_k)^2}{m_k\hat{\pi}_k(1 - \hat{\pi}_k)}$$

is distributed approximately chi-square with $K - (p + 1)$ degree of freedom, $m_k\hat{\pi}_k$ is large for every k, K is the number of covariate patterns and p is the number of independent covariates model.

In 1980, Hosmer and Lemeshow developed a set of goodness-of-fit tests to avoid problems associated with the asymptotic distribution of X^2 –test. Using (Hosmoer and Lemeshow, 1980) suggestion, subjects have to be grouped into g groups and X^2 – test is estimated using the amalgamated cells. Therefore, to use Hosmer and Lemeshow recommended method, observations have to be partitioned into $g = 10$ equal-sized groups based on their ordered estimated probabilities. Then,

$$G_{HL}^2 = \sum_{j=1}^{10} \frac{\left(O_j - E_j\right)^2}{E_j\left(1 - E_j/n_j\right)} \sim X_8^2$$

where

n_j = number of observations in the j^{th} group

O_j = $\sum_j y_{ij}$ = observed number of cases in the j^{th} group

E_j = $\sum_j \hat{p}_{ij}$ = expected number of cases in the j^{th} group

For fitting logistic regression models using complex survey data, the sampling weight can be calculated as the inverse of the product of the conditional inclusion probabilities at each stage of sampling. This represents the number of units that the given sampled observations represented in the total population. Expanding each observations by its sampling weight produces a dataset for the N units in the total population. Therefore, for complex survey Hosmer-Lemeshow goodness-of-fit test, the observed and expected cell counts are the total population size (Archera et al., 2007).

4.5 Data analysis using survey logistic model

The data analysis for this study was done using SAS version 9.2. The deviance was used to compare alternative models during model selection. Change in the deviance was used to measure the extent to which the fit of the model improves when additional variables were included. To avoid confounding effects, the model was fitted in two steps. The model was fitted to each predictor variables one at a time. In stage two the significant predictors were retained in a multivariate logistic regression model. In addition to the main effects, possible combinations of up to three-way interaction terms were added and assessed to further avoid and mitigate the problem of confounding. Therefore, the main effects and the possible combinations of up to three-way interaction terms were fitted. The selected model was the one with the smallest change in deviance compared to all possible models.

Let the response $y_{ijh} = 1$ if the i^{th} person has been positive for malaria rapid diagnosis test and $y_i = 0$ otherwise. Therefore, the fitted survey logistic model is given as

$$logit(\pi_{ijh}) = \log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right) = x'_{ijh}\beta$$

where, $\pi_{ijh} = E(y_i) = \mathrm{P}(\mathrm{y_i} = 1)$, x'_{ijh} is a vector of appropriately coded values of the explanatory variables and β is a vector of unknown parameters.

The objective of the analysis is to identify the individual characteristics that could be associated with the malaria rapid diagnosis test outcome. On the other hand, this study focused on identifying the household characteristics which could be associated with the increase/decrease of the number of malaria infected household members. These household characteristics which were included in the model are main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, number of persons per room, main material of the room's wall, main material of the room's roof, main material of the room's floor, use of indoor residual spray in the past twelve months, use of mosquito nets, number of nets per person, family size, region and altitude. The individual characteristics are gender and age.

To make statistically valid inferences, the analysis of the data from the study accounted for design effects of the study. The SAS procedure (PROC SURVEYLOGISTIC) which performs logistic regression for categorical responses in sample survey data was used (SAS, 9.2). The maximal model with significant effects is given in Table 4.2. These models have the smallest deviance ($-2logL$) amongst all the nested models with the three-way interaction effects. Based on the final model, six interactions reduced the deviance ($-2logL$). Therefore, the final model includes all the main effects and the six interaction effects.

Table 4. 2: Type 3 analysis of effects for the survey logistic model

Effect	DF	Chi-Square	Pr > ChiSq
Age	1	14.6585	0.0001
Gender	1	24.3933	<.0001
Family size	1	1.9782	0.1596
Region	2	1.7835	0.4099
Altitude	1	0.1126	0.7372
Main source of drinking water	2	56.4991	<.0001
Time to collect water	1	851.0891	<.0001
Toilet facilities	2	4.7555	0.0928
Availability of electricity	1	0.6455	0.4217
Availability radio	1	1.3791	0.2403
Availability television	1	0.7465	0.3876
Total number of rooms	1	52.2942	<.0001
Main material of the room's wall	2	28.571	<.0001
Main material of the room's roof	2	38.0472	<.0001
Main material of the room's floor	2	32.909	<.0001
use of indoor residual spray	1	24.7274	<.0001
Number of months room sprayed	1	38.2539	<.0001
Use of mosquito nets	1	15.1781	<.0001
Total number of nets	1	4.1535	0.0458
Main source of drinking water and main material of the room's roof	4	56.5889	<.0001
use of indoor residual spray and use of mosquito nets	1	21.7258	<.0001
Time to collect water and main material of the room's floor	2	10.3219	0.0013
Gender & main source of drinking water	1	160.2781	<.0001
Gender & main material of the room's floor	2	18.9357	<.0001
Gender, Main source of drinking water and electricity	2	5.7837	0.0162

Toilet facilities, availability of television, number of rooms per person, main material for walls, number of months the room was sprayed, number of mosquito nets per person, age and family size were found to be significant main effects. In addition to the main effects, five significant two-way interaction terms and one three-way interaction terms was obtained. The two-way interaction terms were: the interaction between main source of drinking water and main material of the room's roof; use of indoor residual spray and use of mosquito nets; time taken to collect water and floor material; gender and main source of drinking water; gender and main material of the room's floor; and gender and use of indoor residual spray. Three-way interaction between gender, main source of drinking water and availability of electricity was also significant. Age, family size, toilet facilities, availability of television, number of persons per room, wall material and number of months indoor residual sprayed in the room were the significant main effects, which were not involved in significant interaction terms (Table 4.2). Accordingly, the effect of these variables can be directly interpreted using the odds ratio (OR).

Tables 4.3 and 4.4 present estimates of socio-economic, demographic and geographic factors on RDT. Based on the result for a unit increase in age, implies a reduction of the odds of a positive malaria test by 3.0% (OR = 0.970, p - value = 0.0001). Furthermore, for a unit increase in family size, the odds of a positive RDT increased by 5.7% (OR = 1.057, p - value < .0001). Furthermore, compared to households which had no toilet facilities, those with a pit latrine were at lower risk of malaria diagnosis (OR = 0.725, p-value = <.0001) as well as households with flush toilets (OR = 0.552, p - value = <.0001). Households who were using mosquito nets were found to be at a lower risk of malaria compared to the households who were not using mosquito nets (OR = 0.91, p - value = <.0001). Furthermore, for a unit increase in the number of nets, the odds of positive malaria diagnosis test decreases by 54% (OR = 0.46, p - value = <0.0001) for the household.

Table 4. 3: Estimates and odds ratios of socio-economic, demographic and geographic factors on RDT

Effects	Estimate	OR	95% C.I.		P -value
			Lower	Upper	
Intercept	-3.030	0.048	0.016	0.125	0.001
Age	-0.031	0.970	0.319	0.995	0.0001
Sex (ref. male)					
Female	-1.820	0.162	0.053	0.418	<.0001
Family size	0.049	1.057	1.014	1.124	<.0001
Region (ref. SNNP)					
Amhara	-0.099	0.906	0.178	16.374	0.521
Oromiya	-0.184	0.832	0.238	8.581	0.183
Toilet facility (Ref. No facility)					
Pit latrine	-0.3213	0.725	0.575	0.943	<.0001
Toilet with flush	-0.5935	0.552	0.432	0.909	<.0001
Main source of drinking water (ref. protected water)					
Tap water	-0.038	0.963	0.316	0.973	<.0001
Unprotected water	0.717	2.048	0.673	5.289	0.007
Availability of television (ref. no)					
Yes	0.304	1.356	0.446	3.500	0.024
Number of rooms/person	-0.473	0.623	0.205	1.001	0.044
Main material of room's wall (ref. cement block)					
Mud block/stick/wood	-2.326	0.098	0.032	0.252	0.048
Corrugated metal	-0.620	0.538	0.471	0.826	0.001
Main material of room's roof (ref. corrugate)					
Thatch	1.325	3.761	1.236	9.712	<.0001
Stick and mud	-1.960	0.141	0.046	0.364	<.0001
Main material of room's floor (ref. earth/Local dung plaster)					
Wood	-1.701	0.183	0.149	0.443	<.0001
Cement	-3.927	0.014	0.011	0.876	0.018
use of indoor residual spray (ref. yes)					
No	1.857	6.405	2.105	16.539	0.046
Use of mosquito nets (ref. no)					
Yes	-0.095	0.910	0.299	0.949	<.0001
Number of nets/person	-0.782	0.457	0.150	0.981	<.0001

Table 4. 4: Estimates and odds ratios of socio-economic, demographic and geographic factors on RDT for interaction effects

	Estimate	OR	95% CI		P -value
			Lower	Upper	
Main source of drinking water and main material of the room's roof (ref. Protected water & cement block)					
Tap water and Mud block/stick/wood	-3.339	0.035	0.007	0.177	<.0001
Tap water and Corrugated metal	-3.377	0.034	0.007	0.184	<.0001
Unprotected water and Mud block/stick/wood	-4.008	0.018	0.003	0.130	<.0001
Unprotected water and Cement block	-1.857	0.156	0.022	1.119	<.0001
Time to collect water and material of room's floor (ref. Less than 30 minutes and earth/local dung plaster)					
Greater than 90 minutes and Cement	-0.423	0.655	0.066	1.478	<.0001
Greater than 90 minutes and Wood	-0.721	0.486	0.160	1.478	0.0013
Between 30 - 40 minutes and Cement	-1.901	0.149	0.049	1.478	<.0001
Between 30 - 40 minutes and Wood	1.554	4.729	0.821	9.220	<.0001
Between 40 - 90 minutes and Cement	-0.739	0.933	0.129	1.258	0.0011
Between 40 - 90 minutes and Wood	0.554	3.769	1.835	7.232	<.0001
Gender and main source of drinking water and main material of the room's roof (ref. Male & protected water)					
Female and Tap water	-0.069	0.933	0.624	1.397	0.0972
Female and Unprotected water	1.327	3.769	1.948	7.293	<.0001
Gender and material of room's floor (ref. Male and earth/Local dung plaster)					
Female and Cement	-0.372	0.689	0.158	1.254	<.0001
Female and Wood	-4.893	0.008	0.003	0.017	<.0001
use of indoor residual spray and use of mosquito nets (ref. Yes &no)					
No and Yes	0.104	1.110	0.898	1.372	0.0319
Gender, main source of drinking water and electricity (ref. Male, protected water & yes)					
Female, tap water and no	0.550	1.734	1.137	2.643	0.0172
Female, unprotected water and no	-1.319	0.267	0.132	0.542	0.0049

Interaction effects

The relationship between gender, main source of drinking water and availability of electricity is presented in Figure 4.1. The risk of positive malaria RDT is higher for unprotected water use by female respondents. However, for both males and females, positive RDT is low for households using tap water and electricity.

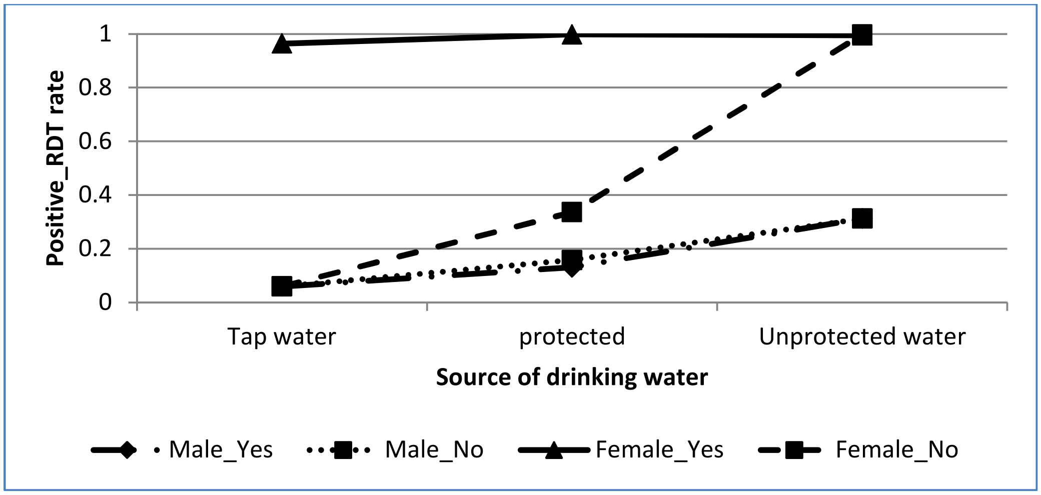


Figure 4. 1: Log odds associated with rapid diagnosis test and gender, source of drinking water with availability of electricity

With reference to households that have tap water for drinking and corrugated iron-roofed houses, the risk of positive malaria RDT was significantly lower than for households living in stick and mud-roofed houses and drinking unprotected water. As Figure 4.2 indicates, higher positive malaria diagnosis test was found for households that reportedly used unprotected water for drinking.

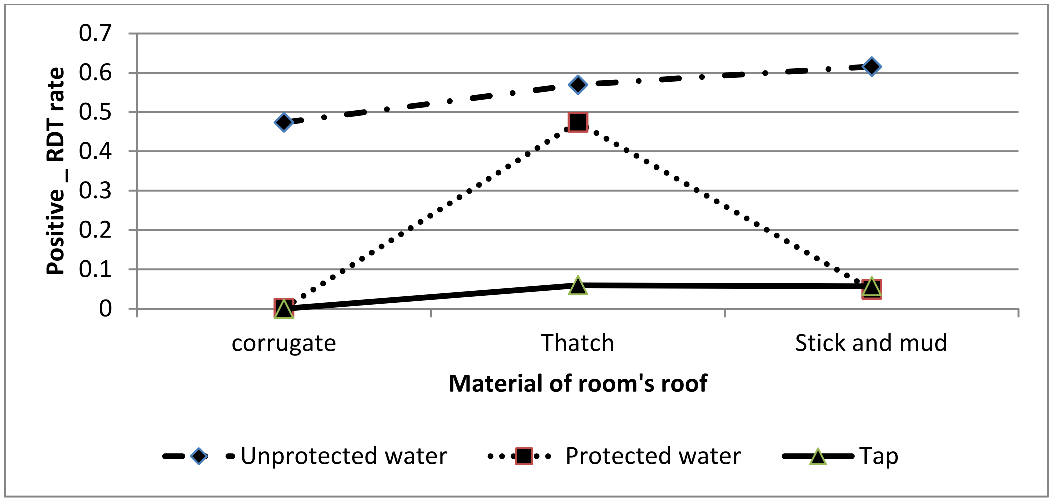


Figure 4. 2: Log odds associated with rapid diagnosis test and material of room's roof with main source of drinking water

The OR values for the interaction between gender and main material of the room's floor is given in Figure 4.3. Based on the result, positive malaria diagnosis test was significantly higher for females than for males who reported that the material of the room’s floor was earth/local dung as well as those who reported that the material of the room’s floor was wood. There was however, higher positive malaria diagnosis test found for both males and females who reported that the material of the room’s floor was wood.

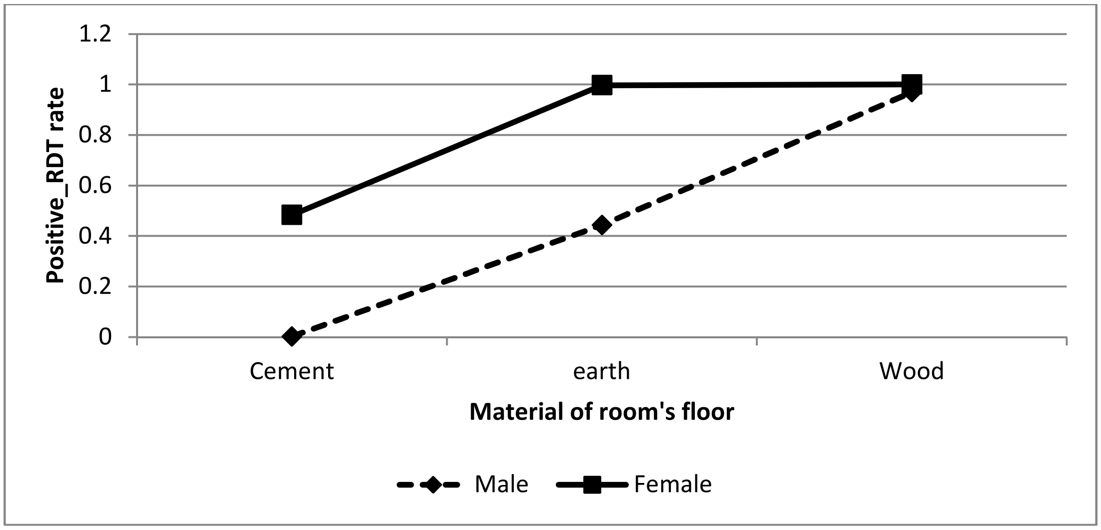


Figure 4. 3: Log odds associated with rapid diagnosis test and gender with material of room's floor

Positive RDT was significantly higher for respondents living in a room with a wooden or earth/local dung floor than for those living in a room with a cement floor for respondents who took 40-90 minutes to collect water. But, for respondents who took less than 40 minutes to collect water, positive RDT was low (refer Figure 4.4).

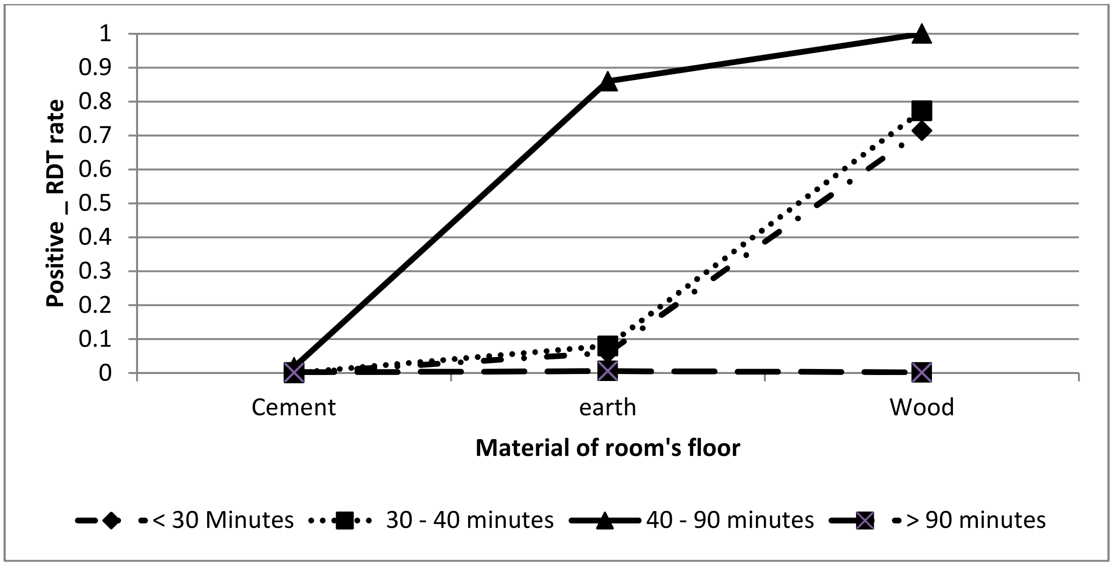


Figure 4. 4: Log odds associated with rapid diagnosis test and material of room's floor with time to collect water

Prevalence of malaria was significantly higher for male than for female respondents who were living in a house treated with indoor residual spray (refer Figure 4.5). For both males and females who were living in a house that had not been sprayed, the risk of positive malaria was significantly higher. On the other hand, for males living in a house that had not been treated with indoor residual spraying, the risk of malaria infection for males is more than that of females.

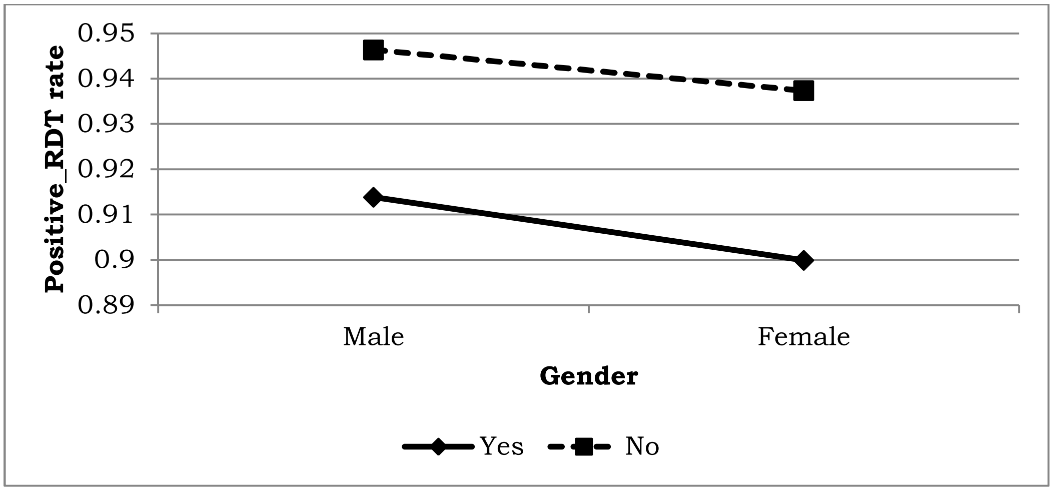


Figure 4. 5: Log odds associated with rapid diagnosis test and use of use of indoor residual spray with gender

The use of mosquito nets and applying indoor residual spray to the walls of the house altered the risk of malaria. The risk of malaria was low for individuals who lived in houses that had been sprayed and used malaria nets. It is shown in Figure 4.6 that the estimated risk of malaria was higher for individuals with no mosquito nets.

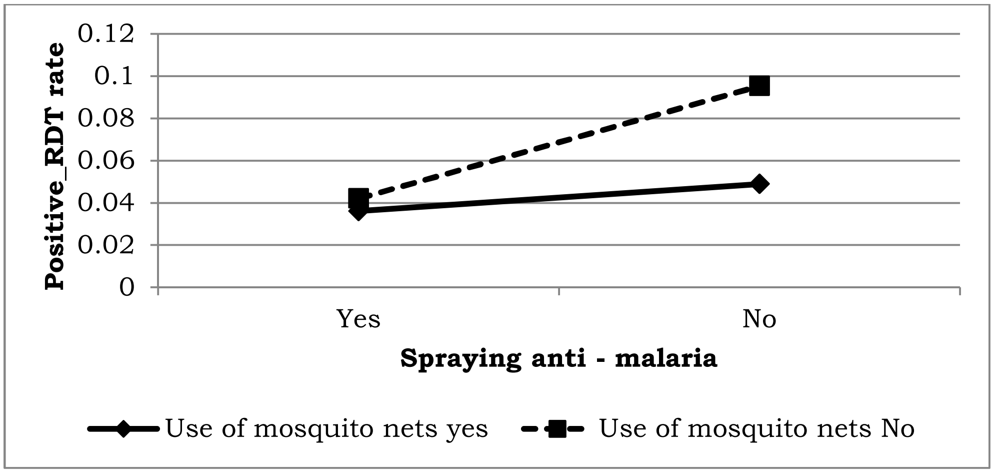


Figure 4. 6: Log odds associated with rapid diagnosis test and use of use of indoor residual spray with use of mosquito nets

The other result which is important to be discussed is the predictive accuracy/ability of the model. Therefore, the procedures used for fitting binary response models to data, produce statistics on the prediction ability of the model, such as c, Sommer's D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT). Using the SAS notation, these statistics are given by

$$
\begin{aligned}
\mathrm{c} &= (\mathrm{n_t} - 0.5(\mathrm{t} - \mathrm{n_c} - \mathrm{n_d})\mathrm{t}^{-1} \\
\mathrm{SD} &= (\mathrm{n_c} - \mathrm{n_d})\mathrm{t}^{-1} \\
\mathrm{GKG} &= (\mathrm{n_c} - \mathrm{n_d})(\mathrm{n_c} + \mathrm{n_d})^{-1} \\
\mathrm{KT} &= (\mathrm{n_c} - \mathrm{n_d})(0.5\mathrm{N}(\mathrm{N} - 1))^{-1}
\end{aligned}
$$

where n is the total number of individuals in the data set, t is a total number of pairs given by $n(n-1)/2$, n_c is a number of concordant pairs (a pair of observations is concordant if a response y is 1 and the predicted probability is high), n_d is a number of discordant pairs (a pair of observations is discordant if the response y is 1 and the predicted probability is low), and tied pairs are given by $\mathrm{t} - \mathrm{n_c} - \mathrm{n_d}$ (Agresti, 1984). The Predictive ability of the model is given under the association of predicted probabilities and observed responses. From the result it is observed that out of the 7,531,272 (informative) pairs, 86.5% were concordant and 0.8% were tied. The other rank correlation is the "c" value. This value ranges from 0 to 1. The value 0 implies there is no association. Moreover, c is equal to the area under the receiver operating characteristic (ROC) curve. Based on the values, the prediction accuracy is poor if c is between 0.5 to 0.6, moderate if between 0.6 to 0.7, acceptable if between 0.7 to 0.8 and excellent if greater than 0.8. Based on the values (c = 0.869) the model is excellent. Furthermore, the Somers' D (SD) statistic which is also related to concordance via D= 2*(c-0.5) = 0.738, is simply a rescaled version of concordance that takes values between -1 and 1, like a usual correlation coefficient instead of 0 and 1. The other value is Gamma. This value is the surplus of concordant pairs over discordant pairs. This value ignores percentage ties. Therefore, if tied pairs ignored and the ranking of two pairs guessed based on knowledge of the independent variable x, then it is possible to predict the second x. If the second value is more than the first, then the rank of the second y value will be greater than the rank of the first y value. From the result, the gamma value is 0.542. Therefore, knowing the independent variable reduces our errors in predicting the rank (not value) of the dependent variable by 54.2%.

4.6 Summary and discussion

The generalized linear models using survey logistic regression provided a tool for assessing factors that affect malaria rapid diagnosis test. The present study was conducted based on the 2006 baseline malaria indicator survey in Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia. This survey was a population-based household cluster survey. There were 224 clusters and each cluster consists of 25 households. For this survey, the sampling frame was the rural population of Amhara, Oromiya and SNNP regions. Therefore, the data used for this study was from complex survey. For the statistical analysis, the study used generalized linear model. For this study, gender, age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms per person, main material of the room's wall, main material of the room's roof, main material of the room's floor, incidence of indoor residual spray in the past twelve months, use of mosquito nets and total number of nets per person with up to three-way interaction effects were used for the analysis.

Based on these facts, the findings of this study show that the following socio-economic factors are related to malaria risk: construction material of walls, roof and floor of house; main source of drinking water; time taken to collect water; toilet facilities and availability of electricity. Besides socio-economic factors, there are demographic and geographic factors that also had an effect on the risk of malaria. These include gender, age, family size and the region where the respondents lived. In addition to the main effects, there were interactional effects between the socio-economic, demographic and geographic factors that also influenced the risk of malaria. Most notable of these were the interaction between the main source of drinking water and the main construction material of the room's roof; the time taken to collect water and the main construction material of the room's floor; gender and the main source of drinking water; gender and the availability of electricity; gender and the main construction material of the room's floor and finally, interaction between gender, main source of drinking water and the availability of electricity.

From the study, it was observed that residents living in the Amhara region were found to be more at risk of malaria than those living in the SNNP and the Oromiya regions. Similarly, houses that were treated with indoor residual spray were less likely to be affected by malaria. One of the most important finding to which may inform public health policy in the control of malaria infection was that households with no toilet facilities were more likely to be positive for malaria diagnosis test than those with good toilet facilities. From the results, it was observed that households with no toilet facilities were more likely to be positive for malaria diagnosis test. Furthermore, positive malaria diagnosis rate decreased with age. But, for household size, the risk of malaria increased per unit increase in family size. Generally, malaria parasite prevalence differed between age and gender with the highest prevalence occurring in children and females. The findings of the association between socio-economic factors and malaria prevalence are similar to some of the results from previous studies (Banguero, 1984, Koram et al., 1995, Sintasath et al., 2005). In addition to this in 1998 and 2000, studies were conducted by (Ghebreyesus et al., 2000, Snow et al., 1998) in Ethiopia and Kenya respectively. The objectives of the studies were to assess different types of materials used in the construction of walls, roofs and floors of a house. They used generalized linear models, poisson and logistic models, for their study. Based on their findings, they observed association between any roof, wall and floor material and risk of malaria.

Therefore, the finding of this study gives similar findings to those from previous studies.

This study suggests that having toilet facilities, access to clean drinking water and the use of electricity offers a greater chance of not being positive for malaria diagnosis. Using mosquito nets and indoor residual spray treatment on the walls of the house were also found to be a way of reducing the risk of malaria. In addition to this, having a cement floor and corrugated iron roof were found to be means of reducing the risk of malaria. Based on the study findings, different types of housing have an influence on the risk of malarial transmission with those houses constructed of poor quality materials having an increased risk. Moreover, the presence of particular structural features, such as bricks, that may limit contact with the mosquito vector, also reduces infection. Therefore, the risk of malaria is higher for households in a lower socio-economic bracket than for those that enjoy a higher status and who are able to afford to take measures to reduce the risk of transmission.

This study suggests that with the correct use of mosquito nets, indoor residual spray and other preventative measures, coupled with factors such as the number of rooms in a house, the incidence of disease is decreased. However, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria.

In this chapter, the analysis method of the study data was survey logistic model based and survey design effects were included. But, there are other variabilities in the model. These variabilities related to the errors which are correlated and also nonconstant variability of the error terms. Moreover, use of survey logistic cannot allow investigating more than one source of variation when modelling the explanatory variables. Furthermore, this variability was not included in the model. Therefore, in the next chapter, we will develop a model which includes the additional variabilities.

Chapter 5

The risk factor indicators of malaria in Ethiopia using generalized linear mixed models

5.1 Introduction

In the previous chapter, we adopted the survey logistic model approach which is under generalized linear model for malaria RDT data. This model is an alternative statistical methodology used to identify factors affecting the malaria risk (Ayele et al., 2012, Natarajan, 2008). But, this model is survey based, whereas the *kebeles* are chosen at random which could result in some variability between the sampling units. Generalized Linear Mixed Models (GLMM) explore the idea of statistical models that incorporate random factors into generalized linear models. GLMMs add random effects or correlations among observations to a model, where observations arise from a distribution in the exponential family. The generalized linear mixed model has many advantages. The use of GLMMs can allow random effects to be properly specified and computed and errors can also be correlated. In addition to this, GLMMs can allow the error terms to exhibit non constant variability while also allowing investigation into more than one source of variations. This ultimately leads to greater flexibility in modelling the dependent variable. In this chapter, the objective is to determine the socio-economic, demographic and geographic factors using generalized linear mixed model.

Classical linear models can be generalized using the Generalized Linear Models (GLMs) by exploring the exponential family of sampling distributions (McCullagh and Nelder, 1989). GLM models have an immense impact on both theoretical and practical aspects in statistics. To perform the analysis, there are a number of statistical software tools to fit the generalized linear mixed model. Diversified methodologies arise in the implementation and estimation in the GLMMs. But, there are still plenty of room within the GLMMs framework for further investigation and improvements. This can overcome the over-dispersion in the data and at the same time, accommodate the population heterogeneity. Therefore, the addition of random effects allows accommodating correlation in the context of a broad class of models for non-normally distributed data. These models become more applicable in practical situations. The generalized linear mixed model is applicable in a wide range of areas. For example in modelling problems in plant breading, modelling HIV infections in clinical trials (Jiang, 2007), for joint modelling of multivariate outcomes, etc (Molenberghs and Verbeke, 2005).

Therefore, this chapter is organized as follows. The theory behind GLMM is presented in sections 5.2 and 5.3. The fitted result of RDT malaria data is presented in section 5.4. Summary and discussion of the chapter is presented in section 5.5.

5.2 Generalized linear mixed models (GLMMs)

Generalized linear mixed models are extension of the GLMs. The term 'mixed' in the GLMMs means that the random effects together with the fixed effects are both contained in a model for an outcome of interest to get a modified model. The word “generalized” refers to nonnormal distributions, but the model can include normal distributed data as a special case. This model can overcome the over-dispersion in the data and at the same time, accommodate the population heterogeneity. The main difference in the structure of GLMMs as compared with GLMs is the incorporation of the random effects, term $\boldsymbol{b}_i$, into the linear predictor. But also the nature of the data may dictate the use of GLMMs rather than GLMs. Therefore, the addition of random effects allows accommodating correlation in the context of a broad class of models for non-normally distributed data. These models become more applicable in many practical situations. But, the calculation becomes very complicated because of the inclusion of random effects.

The structure of the generalized linear model involves three points. These points are the distribution of the data, the function of the mean to be modelled and the predictors.

For model formulation, let Y_{ij} be the j^{th} response measured for cluster $i, i = 1, \ldots, N, j = 1, \ldots, n_i$. In addition, let $\boldsymbol{Y}_i$ denote the n_i - dimensional vector of all measurements available for cluster i. Conditionally on random effects $\boldsymbol{b}_i$, it assumes that the elements Y_{ij} of $\boldsymbol{Y}_i$ are independent, following generalized linear mixed model, but the linear predictor extended with subject specific-regression parameters $\boldsymbol{b}_i$. Based on these facts, it is assumed that all Y_{ij} have densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \beta, \psi) = \exp\left\{\frac{y_{ij}(\theta_{ij}) - \Psi(\theta_{ij})}{\emptyset} + c(y_{ij}, \psi)\right\},$$

where the mean μ_{ij}, the conditional mean of y_{ij} for a specific set of unknown parameters θ and ψ, and for known functions $\psi(.)$ and $C(.)$ is modelled through a linear predictor. In the expression, θ and $\emptyset$ are the natural parameters. The linear predictor contains fixed parameters β as well as subject-specific parameters $\boldsymbol{b}_i$,

$$g(\mu_{ij}) = g[E(y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}'_{\boldsymbol{ij}}\beta + \boldsymbol{z}'_{\boldsymbol{ij}}\boldsymbol{b}_i \tag{5.1}$$

for a known link function $g(.)$, and $\boldsymbol{x_{ij}}$ and $\boldsymbol{z_{ij}}$ are the fixed and random effects vectors containing known covariate values, β and $\boldsymbol{b}_i$ are p-dimensional and q-dimensional vectors of known covariate values corresponding to fixed and random effect parameters respectively, as in the normal mixed models (McCullagh and Searle, 2001).

5.3 Estimation and prediction of the fixed and random effects

Estimation method for fixed effects in generalized linear models which is based on normality assumptions is standard for linear models. For many GLMs, maximum likelihood is a standard method of estimation. Parameter estimates of the model can be obtained by partially differentiating the log-likelihood of (5.1) with respect to β and $\boldsymbol{b}_{\mathrm{i}}$, and iteratively solving the resulting estimating equations. But, evaluating the likelihood method is difficult for GLMMs.

For a set of observations $\boldsymbol{y}_i$ where, $i = 1,2, \ldots N$, the interest is in the parameter estimates. The density function of $\boldsymbol{y}_i$ can be denoted as $f_i(y_{ij}|\boldsymbol{b}_i, \beta, \phi)$. The random effects model can be fitted by maximization of the likelihood. Therefore, the contribution of the i^{th} cluster to the likelihood is given by

$$f_i(y_{ij}|\boldsymbol{b}_i, G, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b}_i, \beta, \phi)\, f_i(\boldsymbol{b}_i|G) db_i.$$

where it is important to note that the random effects $\boldsymbol{b}_{\mathrm{i}}$ are integrated out to get the marginal likelihood equation for the parameters of interest. Moreover, the likelihood for β, ϕ and G can be derived from the likelihood function L. This function can be written as

$$L(\beta, D, \psi) = \prod_{i=1}^{N} f_i(y_{ij}|\boldsymbol{b}_i, G, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{b}_i, \beta, \psi)\, f_i(\boldsymbol{b}_i|G) db_i.$$

To find the estimates, there are two main approaches, Classical and Bayesian approaches. In classical inference, the concern is about the likelihood function $L(\beta, D, \psi)$. The parameter estimate is treated as fixed but unknown. By differentiating the log-likelihood function, the parameter estimate which maximizes the likelihood function of the observed data can be obtained. But, it is difficult to evaluate the marginal likelihood function when this likelihood involves high dimensional integral. Various methodologies were proposed in the computation of the likelihood function and hence the maximum likelihood estimates (Wu, 2010).

Maximum Likelihood method (ML)

For parameter estimation, maximum likelihood method is the traditional methodology. Estimation method of fixed effects in GLMs is based on the well-defined log-likelihood and is simple to construct an objective function based on the independence of the data. In linear mixed models, estimation of parameters is based on the marginal likelihood of the data and can be evaluated analytically (Jiang, 2007). With GLMMs, to obtain maximum likelihood estimates, one would maximize the marginal likelihood

$$L(\beta, \theta, \boldsymbol{y}) = \int f(y|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b} \tag{5.2}$$

where, $f(y|\boldsymbol{b})$ is the conditional distribution of the data and $f(\boldsymbol{b})$ is the distribution of random effects. Evaluation of the likelihood involves integration over the distribution of random effects. Because the random effects enter the model non-linearly, the integration is often complicated and even intractable (Littell et al., 2006, Molenberghs et al., 2001, Schall, 1991). The use of maximum likelihood approach in Generalized linear mixed models was studied by (Schall, 1991). Based on the findings of this research, the numerical integration method is found to be only appropriate for simple cases in which the likelihood function involves only integrals of low dimension where such integrals can be factorized into a product of low dimensional integral.

Restricted Maximum Likelihood method (REML)

An extension method of the ML method is Restricted Maximum Likelihood method. It is mainly used for estimating the variance component. This method maximizes the likelihood of linear combinations of elements y. Following similar procedures as in Maximum Likelihood method, the estimates can be obtained by differentiating the log-likelihood function with respect to the variance components, i.e.,

$$L(\beta, \theta, \boldsymbol{y}) = \int f(y|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b}.$$

This expression may be integrated as integrating the mean parameter β out of the likelihood function. The EM algorithm for REML estimation is given by (Laird, 1982). But, it is important to note that the bias of the MLE depends on the dimension of the mean parameter β (McCullagh and Searle, 2001, Schall, 1991).

Penalized quasi-likelihood, Laplace approximation and Guassi-Hermit quadrature methods

To approximate the likelihood to estimate GLMM parameters, different methods have been proposed by different researchers. These methods include pseudo and penalized quai-likelihood, Laplace approximation and Gauss-Hermite quadrature (Breslow and Clayton, 1993, Schall, 1991, Wolfinger and O'Connell, 1993 , Pinheiro and Chao, 2006).

The Pseudo-likelihood Approach

This approach is based on a decomposition of the data into the mean and an appropriate error term, based on a Taylor series expansion of the mean that is a non-linear function of the linear predictor (Molenberghs and Verbeke, 2005). This non-linear function arises after inverting the link function in order to express the conditional mean as a function of the linear predictor. The basic idea is to remove non-linearity by applying Taylor series (linearization) to $g^{-1}(X\beta + ZU)$ about the current estimates of β and U. Hence, this approach is referred to as the linearization method. SAS GLIMMIX procedure documentation (SAS, 9.2) summarizes this approach in the following way. Once the linearization of μ about $(\tilde{\beta})$ and $(\tilde{U})$ has been applied, the model

$$P = \boldsymbol{X}\beta + \boldsymbol{Z}U + \varepsilon$$

is a linear mixed model with the pseudo-response P, fixed effects β and random effects U as well as $var(\varepsilon)$ = $var(P|U)$ = $\widetilde{\Delta}^{-1} A^{\frac{1}{2}} R A^{\frac{1}{2}} \widetilde{\Delta}^{-1}$, where $\widetilde{\Delta} = (\frac{\partial g^{-1}(n)}{\partial \eta})_{\tilde{\beta},\tilde{U}}$. The matrix $\boldsymbol{A}$ is a diagonal matrix containing the variance function of the model and $\boldsymbol{R}$ is a diagonal matrix, i.e $\boldsymbol{R} = \psi\boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.

The marginal variance in the linear mixed pseudo model can be defined as

$$V(\theta) = \boldsymbol{ZGZ'} + \widetilde{\Delta}^{-1} A^{\frac{1}{2}} R A^{\frac{1}{2}} \widetilde{\Delta}^{-1} \tag{5.4}$$

where θ is $(q \times 1)$ vector containing all unknowns in G and R. Based on the linearized model, an objective function can then be defined assuming that the distribution of P is known. The maximum log pseudo-likelihood and restricted log pseudo-likelihood for P are given as follows respectively.

$$l(\theta, p) = -\frac{1}{2} log|V(\theta)| - \frac{1}{2} r'V(\theta)'r - \frac{f}{2} log(2\pi)$$

and

$$l_R(\theta, p) = -\frac{1}{2} log|V(\theta)| - \frac{1}{2} r'V(\theta)'r - \frac{1}{2} log|\boldsymbol{X}'V(\theta)'\boldsymbol{X}| - \frac{f-k}{2} log(2\pi)$$

where $r = \boldsymbol{p} - \boldsymbol{X}(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}p}$, f denotes the sum of frequencies used in the analysis and k denotes the rank of $\boldsymbol{X}$. At convergence the fixed effects parameters are estimated and the random effects are predicted as

$$\hat{\beta} = (\boldsymbol{X'V(\hat{\theta})^{-1}X})^{-1}\boldsymbol{X'V(\hat{\theta})^{-1}P}$$
$$\hat{U} = \hat{G}Z'V(\hat{\theta})^{-1}\hat{r}$$

The parameter estimates are then used to update the linearization, which results in a new linear mixed model. The process continues until the relative change between parameter estimates at two successive iterations is sufficiently small.

There are two commonly used approximations based on Taylor's expansion of the mean. A subject specific expansion referred to as the penalized quasi-likelihood (PQL) approximation uses $\tilde{\beta} = \hat{\beta}$ and $\widetilde{U} = \widehat{U}$, which are the current estimates of fixed effects and predictors of random effects. The population-average expansion referred to as the marginal quasi-likelihood (MQL) uses $\tilde{\beta} = \hat{\beta}$ and $\widetilde{U} = 0$, which are the same current estimates of fixed effects and the random effects are not incorporated in the linear predictor.

Penalized Quasi-Likelihood method (PQL)

The quasi-likelihood method was developed by (Wedderburn, 1974). The quasi-likelihood function is constructed with fewer assumptions than the likelihood function. However, the construction of the quasi-likelihood function requires the relationship between the mean and variance of the data.

Let $\theta = (\theta_1, \dots, \theta_c)'$ and $G(\theta) = diag\left(\theta_1 I_{q_1}, \dots, \theta_c I_{q_c}\right)'$ where I_{q_j} is a $q_j \times q_j$ identity matrix. Assume the random effects u_j are independent and distributed as $N(0, \theta_{jI_{q_1}})$ the integrated likelihood of (α, θ) is

$$L(\alpha, \theta) \propto (2\pi)^{-q/2}|G|^{1/2} \int exp\left\{-\frac{1}{2\varnothing}\sum_{i=1}^{n} d_i(y_i; \mu_i) - \frac{1}{2}b'G(\theta)u\right\}du$$

Therefore, a conditional algorithm of quasi-likelihood function Q_i is given by

$$Q_i = d_i(y_i; \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{y_i\, T^2 V(\mu_i)}\, dt$$

where, y_i are independent measurements from a distribution with density from the exponential family, T is the unspecified constant of proportionality relating $Var(y_i)$ to $var(\mu_i)$ (Goldstein, 2011, Lin, 2007).

Some researchers further included a term into the quasi-likelihood function to form the penalized quasi-likelihood (PQL) method. For random effects which follow a normal distribution with mean 0 and a variance-covariance matrix G, the penalized quasi likelihood function is given by

$$PQL = \sum Q_i - \frac{1}{2}\beta' G^{-1}\beta \qquad (5.3)$$

where, $\beta' G^{-1}\beta$ is the penalized term added into quasi-likelihood function. Moreover, arbitrary selection of the value of β can be prevented using the added term (Green, 1990, Wolfinger, 1993).

Therefore, the maximum quasi-likelihood equation can be obtained by differentiating equation 5.3.

But, the estimates which are obtained using PQL in GLMMs are biased towards zero for some variance components. Biased-corrected PQL was suggested by (Lin and Breslow, 1996). This study suggested a method which improves the asymptotic performance of PQL estimates. But, the suggested method inflates the variance.

Marginal quasi-likelihood (MQL)

The marginal quasi-likelihood method is similar to PQL method. The difference between the two methods is that the Taylor series expansion which is given by

$$Y_{ij} = \mu_{ij} + \epsilon_{ij} = h\left(x'_{ij}\beta + Z'_{ij}u_i\right) + \epsilon_{ij}$$

is considered for the mean around the current estimates $\hat{\beta}$ and $\hat{b}_i = 0$ for the fixed and random effects respectively. For MQL, the result is similar except for the current predictor of the mean $\hat{\mu}_{ij}$ is of the form $\left(x_{ij}, \hat{\beta}\right)$ instead of $h\left(x'_{ij}\hat{\beta} + Z'_{ij}\hat{b}_i\right)$. Therefore, the Pseudo data can be written as

$$Y_i^* = V_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta}$$

and satisfies the linear mixed model

$$Y_i^* \approx X_i\beta + Z_i b_i + \in_i^*.$$

The calculation between pseudo-data is used to fit the model iteratively. This estimate is known as marginal quasi-likelihood (MQL) estimate (Breslow and Clayton, 1993, Goldstein, 2011).

Approximation of the integrand using Laplace Approximation and Gausse-Hermite quadrature

Laplace Approximation

Laplacian approximations are frequently used and evaluate marginal likelihoods or posterior means functions (Barndorff-Nielsen and Cox, 1989, Breslow and Clayton, 1993, Tierney and Kadane, 1986, Wolfinger, 1993). To standard Laplace approximation can be described as follows.

Suppose that we want to evaluate integrals of the form (Molenberghs and Verbeke, 2005)

$$I = \int e^{-Q(b)} d\boldsymbol{b}. \tag{5.5}$$

Suppose $\widehat{\boldsymbol{b}}$ is the value of $\boldsymbol{b}$ for Q is minimized. Then, the second-order Taylor expansion of $Q(b)$ around $\widehat{\boldsymbol{b}}$ is of the form

$$Q(b) \approx Q\left(\widehat{\boldsymbol{b}}\right) + \frac{1}{2}(\boldsymbol{b} - \widehat{\boldsymbol{b}})'Q''(\boldsymbol{b})(\boldsymbol{b} - \widehat{\boldsymbol{b}})$$

where Q''(b) is equal to the Hessian of Q, i.e. the matrix of the second order derivatives of Q, evaluated at $\hat{\boldsymbol{b}}$**.** The integral I can be approximated by replacing $Q(b)$ in (5.5). Thus,

$$I \approx (2\pi)^{q/2} \left|Q''\left(\hat{\boldsymbol{b}}\right)\right|^{-1/2} e^{-Q(b)}. \tag{5.6}$$

The Laplace approximation to the integral uses many different estimates of $\boldsymbol{b}$ as necessary according to the different modes of the Q function. Each integral in (5.5) is proportional to an integral of the form (5.5), for a $Q(b)$ function given by

$$Q(b) = (a;\psi)^{-1} \textstyle\sum_{j=1}^{n_i} [y_{ij}(\boldsymbol{X}'_{ij}\beta + \boldsymbol{Z}'_{ij}\boldsymbol{b}) - \Psi(\boldsymbol{X}'_{ij}\beta + \boldsymbol{Z}'_{ij}\boldsymbol{b})] - \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b},$$

such that Laplace's method can be applied. Here, the model $\hat{\boldsymbol{b}}$ of Q depends on the unknown parameters β, ψ and $\boldsymbol{G}$.

Gauss-Hermite Quadrature

Gauss-Hermite Quadrature (GHQ) is often used for numerical approximation of integrals with Gaussian kernels. In generalized linear mixed models random effects are assumed to have Gaussian distributions, but often the marginal likelihood, which has the key role in parameter estimation and inference, is analytically intractable. Furthermore, Gauss-Hermite Quadrature is feasible tools for numerical evaluation of the integrals.

The likelihood function for two level logistic models can be written as follows

$$\int_{-\infty}^{\infty} \pi_i\{\left(\pi_{ij}\right)^{s_{ij}}\left(1 - \pi_{ij}\right)^{n_{ij} - s_{ij}}\} f(b_i;G)du_j$$

and

$$\pi_{ij} = \left\{1 + \exp(x_{ij}\beta_j\right\}^{-1} \ ; \ \beta_j = \beta + u_j$$

where $f(b_i;G)$ is assumed to be a multivariate normal density.

$$f(b_i;G) = \int_{-\infty}^{\infty} P\left(u_j\right) f\left(u_j\right) du_j.$$

Therefore, Gauss-Hermite quadrature approximations is

$$\int_{-\infty}^{\infty} P(v)\, e^{-v^2} dv \;\approx\; \sum_{q=1}^{Q} P(x_q) w_q \tag{5.7}$$

where $\sum_{q=1}^{Q} P(x_q) w_q$ is a Gauss-Hermite polynomial evaluated at a series of quadrature points indexed by q. A model with a single random intercept can be represented as

$$P(u_j) = \prod_i \frac{\exp(x_{ij}\beta + u_j)}{1 + \exp(x_{ij}\beta + u_j)}.$$

In general, quadrature methods can be applied to poisson, binomial, multinomial and ordered category models. But, Gauss-Hermite quadrature is effectively limited to the normal distribution because of the exponential term in equation 5.7.

Simulated Maximum Likelihood method (SML)

Simulated Maximum Likelihood (SLM) method was suggested by (Geyer and Thompson, 1992, Gelfand and Carlin, 1993). (McCulloch, 1997) studied the use of Simulated Maximum Likelihood method on the GLMMs. In SML method, the likelihood is estimated directly without considering the log-likelihood function by simulation. The simulation to estimate the value of the likelihood is given by

$$\begin{aligned} L(\beta, \phi, G|y) &= \int f_{y|u}(y|u, \beta, \psi) f_u(u|G) du \\ &= \int \frac{\int f_{y|u}(y|u, \beta, \psi) f_u(u|G)}{h_u(u)} h_u(u) du \\ &\cong \frac{1}{N} \sum_{k=1}^{N} \frac{f_{y|u}(y|u^{(k)}, \beta, \psi) f_u(u^{(k)}|G)}{h_u(u^{(k)}} \end{aligned}$$

where, N is the total number of simulated value, $h_u(\boldsymbol{u})$ is the importance sampling function and $\boldsymbol{u}$ is a vector of random effects simulated from this distribution by any sampling technique. In theory, the estimates of the parameter are independent of the choice of importance sampling function, $h_u(u)$ and calculated numerically based on the likelihood function approximated using simulations. The efficiency of estimates depends on the choice of importance sampling function. If the importance function in SML is far away from the density of the random effects, the resulting estimator may be inefficient. Therefore, it is important to be careful when implementing SML method (McCullagh and Searle, 2001).

Empirical Bayes Estimation

In practice, estimation of the marginal parameters (β, D and ψ) of the marginal distribution of $\boldsymbol{Y_i}$ is important. But, estimating the random effects $\boldsymbol{b}_i$ is also very important. To detect special profiles, the estimate of subject variability is very important. For the prediction of subject-specific evolutions, estimating the random effects are important. Therefore, Bayesian inference is based on the posterior function which is given by

$$f(\boldsymbol{b}_i \,\big|\, y, \beta, D, \psi) = \frac{f(y|\boldsymbol{b}_i,\beta,\psi)f(\boldsymbol{b}_i|G)}{\int f(y|\boldsymbol{b}_i,\beta,\psi)f(\boldsymbol{b}_i|G)d\boldsymbol{b}_i}$$

based on a density function on y, namely $f(y|\boldsymbol{b}_i, \beta, \psi)$ and a posterior distribution on the parameter (Wu, 2010). Therefore, prior information has to be collected on the parameter θ and assign a suitable prior density to the parameter θ in order to construct the posterior density. The parameter θ is treated as random variables in Bayesian approach (Lee, 2004). However, there exists arguments about the specification of the prior density, i.e., either conjugate prior is chosen just for convenience or the choice of prior can be subjective. In addition to this, the aim is also to evaluate the posterior density and obtain the posterior mean as the Bayesian estimate.

Monte Carlo Newton Raphson method (MCNR)

Newton Raphson method is a popular iterative method to find the maximum likelihood estimates. If the log-likelihood function on the data y and the parameter space can be denoted by $L(\beta, \phi, G|y)$. Then $L'(\beta, \phi, G|y)$ and $L''(\beta, \phi, G|y)$ are the first and second order derivatives respectively. In each Newton Raphson iteration, current parameter estimates can be updated to the next iteration and the procedure continues until convergence is achieved. For GLMMs, the likelihood function and its derivatives may be difficult to evaluate in the Newton Raphson procedures. The use of Monte Carlo Newton Raphson method for calculation of the estimates in the GLMMs was proposed by (Kuk and Cheng, 1997). The Monte Carlo algorithm requires the random effects being simulated from a conditional function given the observed y and the current estimate. As shown by (Kuk and Cheng, 1997) the convergent rate for MCNR was faster than that of Monte Carlo EM. So, it is computationally more efficient.

Monte Carlo EM method (MCEM)

An iterative method for the computation of maximizer on the posterior density is the EM algorithm. The algorithm includes an E-step in expectation and then follows an M-step in maximization. The EM algorithm is popular estimation for data with missing values, i.e., the basic idea of EM algorithm is that for a given observed data, it is assumed to have some missing data to the random effects (Dempster et al., 1997). In the E-step, the expectation can be computed over the missing data to approximate the likelihood function. Afterwards, a maximizer of the likelihood given the working values of the parameter estimates in the M-step can be found. The conditional distribution is updated using the new maximizer and the algorithm is iterated until convergence is reached. The implementation of the use of Monte Carlo EM algorithm where the E- step by a Monte Carlo method was suggested by (McCulloch, 1994, Wei and Tanner, 1990). Therefore, the random effects of the GLMMs can be treated as missing values and apply the EM algorithm. However, the expectation is too difficult when the density of the data cannot be written in a closed form.

Gibbs sampler

Bayesian approach is an alternative method to Classical approach. In order to obtain Bayesian estimates, the prior distribution on each parameter is specified. After specifying the prior distribution, get the posterior mean of each parameter from its conditional distribution. The value have to be specified. Markov chain Monte Carlo and in particular Gibbs sampling for fitting GLMM for point referenced data was suggest by (Diggle et al., 1998). The Gibbs sampler is a special case of the Metropolis-Hastings algorithm and has been found to be very useful in many multidimensional applications. Therefore, the standard implementation of the Gibbs algorithm requires sampling from the full conditional posterior distributions. This application has the following forms:

$$p(\beta_k|\beta_{-k},\boldsymbol{U},\boldsymbol{Y}) \propto \prod_{i=1}^{n}\prod_{j=1}^{n_i}\frac{\exp(\boldsymbol{X}_{ijk}\beta_k.Y_{ij})}{1+\exp(\boldsymbol{X}_{ij}^t\beta+U_i)} \tag{5.8}$$

$$p(\boldsymbol{U}_i|\boldsymbol{U}_{-i},\sigma^2,\psi,\boldsymbol{Y}) \propto \prod_{i=1}^{n}\prod_{j=1}^{n_i}\frac{\exp(\boldsymbol{X}_{ijk}\beta_k.Y_{ij})}{1+\exp(\boldsymbol{X}_{ij}^t\beta+U_i)}|\Sigma_i|^{-\frac{1}{2}} \times$$

$$\exp(-\tfrac{1}{2}(\boldsymbol{U}_i - \Sigma_{-i,i}\Sigma_{-i}^{-1}\boldsymbol{U}_i)^2(\Sigma_i)^{-1}) \tag{5.9}$$

$$p(\psi|\boldsymbol{U},\sigma^2) \propto |\Sigma|^{-\frac{1}{2}}\exp(-\tfrac{1}{2}(\boldsymbol{U}^t\Sigma^{-1}\boldsymbol{U}+\tfrac{b_1}{\psi})\psi^{-(a_1+1)} \tag{5.10}$$

$$p(\sigma^2|\boldsymbol{U},\psi)\sim Inverse\ Gamma\ (a_2+\tfrac{n}{2},b_2+\tfrac{1}{2}\ \boldsymbol{U}^t\boldsymbol{R}\boldsymbol{U}) \tag{5.11}$$

where, $\beta_{-k}=(\beta_1,\dots,\beta_{k-1},\beta_{k+1},\dots,\beta_k)^t$, $\boldsymbol{U}_{-i}=(U_1,\dots,U_{i-1},U_{i+1},\dots,U_n)^t$, $\Sigma_{-i,i}=\Sigma_{i,-i}^t=Cov(U_{-i},U_i), \Sigma_{-i}=Cov\ (U_{-i},U_i)$, $R_{kl}=\rho(\psi;d_{kl})$ and
$\Sigma_{-i}=\sigma^2-\Sigma_{i,-i}\Sigma_{-i}^{-1}\Sigma_{i,-i}$ (Wu, 2010).

Samples from $p(\sigma^2|\boldsymbol{U},\psi)$ can be drawn easily as this is a known distribution. The conditionals of the other parameters do not have standard forms and a random walk Metropolis algorithm with a Gaussian proposal density, having mean equal to the estimate from the previous iteration and variance derived from the inverse second derivative of the log-posterior, could be employed for simulation. The likelihood calculations in (5.8) and (5.10) require inversions of the $(n-1)\times(n-1)$ matrices, Σ_{-i}, i = 1,…,n and the $n \times n$ matrix Σ, respectively. Matrix inversion is an order three operations, which has to be repeated for evaluating the conditional distribution of all n random effects U_i and that of the ψ parameter, within each Gibbs sampling iteration. This leads to an enormous demand of computing capacity and makes implementation of the algorithm extremely slow (or possibly infeasible), especially for large number of locations (Jiang, 2007).

Inference for fixed and random effects

In a regression analysis, the objective is to see if an effect is associated with the outcome. After the analysis, if the covariates has no association with outcome, then $\beta_j = 0$ for $j = 1, \dots, p-1$. If the covariates associated with the outcome, then $\beta_j \neq 0$. For random effects, it can be concluded that there is no association with outcome when the effect has zero variability. Since GLMMs are based on maximum likelihood approach, the obtained estimates are asymptotically normally distributed; as a result, tests such as the Wald-type as well as likelihood ratio tests can be used as similar to linear mixed models. Inferences for linear mixed model are discussed below.

Inference for Fixed effects

(Verbeke and Molenberghs, 2000) show that inferences about the fixed effects can be done using the approximate Wald tests (also referred to as Z-test), the t-tests and F-tests. The Wald test as well as the associated confidence of β_j is obtained from approximating the distribution of $(\hat{\beta}_j - \beta_j)/s.e\,(\hat{\beta}_j)$ by a standard univariate normal distribution of $\hat{\beta}_j$, $j = 1, \dots, p$. More generally, it may be of interest to construct confidence intervals and tests of hypotheses about certain linear combinations of the components of β. For instance, given any known matrix L, a test for hypothesis

$$H_0 : \boldsymbol{L}\beta = 0 \text{ versus } H_A : \boldsymbol{L}\beta \neq 0, ,$$

follows from the fact that the distribution of

$$(\hat{\beta} - \beta)' L' \left[L(\sum_{i=1}^{N} \boldsymbol{X}_i' V_I^{-1} \boldsymbol{X})' \; L_i' \right]^{-1} L((\hat{\beta} - \beta)$$

follows asymptotically a chi-square distribution with rank (L) degrees of freedom. Alternatively, approximate t and F statistics can be used for testing hypothesis about the fixed effects. In fact, it is pointed out that the t- and F -statistics rectify the downward bias of the standard errors in the Wald test statistics due to failing to take into account the variability introduced by estimating the variance parameters. For large samples, large sample normality of estimators can be used to utilize Wald tests. This can be specified individual parameters as $\hat{\beta}_i - \beta_{i,0} / \sqrt{\widehat{var}_{\infty}(\hat{\beta}_i)} \sim AN(0,1)$ or for a set of linear combinations of the parameters, $L'\hat{\beta} - L'\beta_0 \sim AN(0, L'I^{-1}L)$ where I represents the observed or expected information matrix. An approximate F test can be carried out by dividing the Wald test by the numerator degrees of freedom and approximating the denominator degrees of freedom ($rank\ (L)$). There are several methods that are available for estimating the denominator degrees of freedom; one of which is the Satterthwaite approximation. All these tests are based on large sample approximation. It is worth noting that different methods lead to different results. This is due to the fact that different subjects contribute independent information, which results in numbers of degrees of freedom which are typically large enough (McCullagh and Searle, 2001). The presence of single random effects or multiple random effects can be tested. For this test, the score test can be used. This test was proposed by (Commenges and Jacqmin-Gadda, 1997, Commenges et al., 1994, Jacqmin-Gadda and Commenges, 1995, Lin, 1997). The advantage of this test is that, the maximum likelihood estimators under GLMM are not required for testing.

The likelihood ratio (LR) test can also be used for comparison of nested models with different mean structure. The likelihood ratio test for two nested models is constructed by comparing the maximized log-likelihoods, say $\hat{l}_{full}$ and $\hat{l}_{reduced}$ for the full and reduced models respectively. The two models are nested in the sense that the reduced model is the special case of the full model. To compare $\hat{l}_{full}$ and $\hat{l}_{reduced}$, minus twice the logarithm of the ratio of these maximized likelihoods can be used and the test statistic is given by

$$-2\,ln\,\lambda_N = -2ln\left(\frac{\hat{l}_{reduced}}{\hat{l}_{full}}\right)$$

comparing the statistic to a chi-square distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models. Small values of $-2ln\lambda_N$ are obtained when $\hat{l}_{reduced}$ is similar to $\hat{l}_{full}$, indicating that the reduced model is a good one. The LR test results for fixed effects are not valid if models are fitted using REML rather than ML. This is because REML log-likelihood functions are based on different observations, which makes them no longer comparable (Verbeke and Molenberghs, 2000).

Inference for random effects

With the asymptotic normality of parameter estimates, approximate Wald tests for random effects can be obtained in the same way as with the fixed effects. However, the normal approximation fails if the parameter to be tested takes values on the boundary of the parameter space. Likewise, the likelihood ratio test suffers from the same problems as the approximate Wald tests. For instance, suppose we have a random coefficient model with a random intercept and slope given by

$$y_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i})t_{ij} + e_{ij} \quad \text{and}$$
$$\mathrm{U_i} = \mathrm{var}\begin{pmatrix}\mathrm{U_{0i}}\\ \mathrm{U_{1i}}\end{pmatrix}, \mathrm{var}(\mathrm{U_i}) = \mathrm{G} = \begin{pmatrix}\mathrm{G_{11}} & \mathrm{G_{12}}\\ \mathrm{G_{21}} & \mathrm{G_{22}}\end{pmatrix}.$$

This model is referred to as the 'full' model. Here, consider the possibility that slopes, for example, do not vary across units. That is, consider slopes as being fixed rather than random such that there will be a 'reduced' model, which is given by

$$y_{ij} = \beta_0 + u_{0i} + \beta_1 t_{ij} + e_{ij} \quad \text{and}$$
$$\mathrm{U_i} = u_{0i}, \mathrm{var}(\mathrm{U_i}) = \mathrm{G_{11}}.$$

For both models, assume that the $var\,(e_i) = R_i = \sigma^2 I_n$. Both the full and reduced models have the same mean structure, $E(y_{ij}) = X_i\beta$. Both however have different covariance models, $V_i = Z_i G Z_i' + \sigma^2 I_n$. The full model has the usual form of Z_i given by

$$\mathrm{Z_i} = \begin{pmatrix}1 & \mathrm{t_{i1}}\\ \vdots & \vdots\\ 1 & \mathrm{t_{in}}\end{pmatrix} \quad \text{hence} \quad \mathrm{G} = \begin{pmatrix}\mathrm{G_{11}} & \mathrm{G_{12}}\\ \mathrm{G_{21}} & \mathrm{G_{22}}\end{pmatrix}$$

whereas the reduced model takes the form

$$Z_i = I_n \quad \text{hence} \quad G = G_{11}\ .$$

Considering the fact that the models are nested, hypothesis test of whether slopes vary across units seem to be applicable. However, testing whether slopes do not vary across units requires that the variance G_{22} in the full model to be equated to zero. This means that the null hypothesis involves checking whether G_{22} takes values on the boundary of the parameter space for G_{22}. The theory that underlies the use of the likelihood ratio test is no longer appropriate when the null hypothesis involves a parameter in the boundary space. This is because the likelihood ratio test does no longer have $\times^2$ distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models (Fitzmaurice et al., 2004, Verbeke and Molenberghs, 2000). It should also be noted that in contrast to the likelihood ratio test for the fixed effects, valid likelihood ratio tests are obtained under REML instead of ML.

Therefore, care should be taken when using output from linear mixed model that was fitted to the pseudo-data. For instance, when one compares nested models using the likelihood ratio (LR) test, the test should be based on the likelihood from the observed data rather than the likelihood corresponding to the linear mixed model for pseudo-data. With regard to inference on the variance components, approximate Wald tests and LR test can be used as long as parameters to be tested are not on the boundary of the parameter space.

Generalized linear mixed models for binary response

Binary data can be specified either as a series of zeros and ones (Bernolli form) or as a frequency of 'success' out of 'trials' (binomial form). Therefore, the development of GLMMs for dichotomous data has been an active area of statistical research. By adopting a logistic or probit regression model, various methods for incorporating and estimating the influence of the random effects, have been developed (Pendergast et al., 1996).

The logistic regression model, which includes the mixed effects, is a common choice for analysis of multilevel dichotomous data. In the GLMM, this model utilizes the logit link, namely

$$g(\mu_{ijk}) = logit\,(\mu_{ijk}) = log\left[\frac{\mu_{ijk}}{1-\ \mu_{ijk}}\right] =\ \eta_{ijk},$$

The conditional expectation $\mu_{ijk} = E(Y_{ijk}|v_i, x_i)$ equals $p(Y_{ijk} = 1|v_i, x_{ijk})$, i.e., the conditional probability of a response given the random effects. Here, Y_{ijk} corresponds to the i^{th} respondent in the j^{th} household with k^{th} probabilistic sampling unit (PSU).

Therefore, this model can also be written as

$$P\left(Y_{ijk} = 1\middle|v_i, x_{ijk}, z_{ijk}\right) = g^{-1}\left(\eta_{ijk}\right)$$

where, the inverse link function $g^{-1}\left(\eta_{ijk}\right)$ is the logistic cumulative distribution function (cdf), namely

$$g^{-1}\left(\eta_{ijk}\right) = [1 + exp(-\eta_{ijk})]^{-1}.$$

The logistic distribution simplifies parameter estimation, because the probability density function (pdf) is related to the cdf (Agresti, 2002).

5.4 Evaluation of malaria rapid diagnosis test using GLMMs

One of the main objectives of this study is to identify socio-economic, demographic and geographic factors affecting malaria rapid diagnosis test. In our previous discussion (Chapter 4), Generalized Linear Model (the survey logistics model approach) was used to identify factors affecting malaria rapid diagnosis test. But, this method is survey based, whereas the *Kebeles* are chosen at random which could result in some variability between the sampling units. Therefore, effect of *Kebeles* on malaria rapid diagnosis test was ignored. When the random effect (*Kebele*) is included in the analysis the model becomes generalized linear mixed models. Generalized Linear Mixed Models (GLMM) explore the idea of statistical models that incorporate random factors into generalized linear models. GLMMs add random effects or correlations among observations to a model, where observations arise from a distribution in the exponential family. The generalized linear mixed model has many advantages. The use of GLMMs can allow random effects to be properly specified and computed and errors can also be correlated. In addition to this, GLMMs can allow the error terms to exhibit non constant variability while also allowing investigation into more than one source of variation. This ultimately leads to greater flexibility in modelling the dependent variable.

To analyze the malaria rapid diagnosis test data PROC GLIMMIX in SAS was used. For this analysis, malaria rapid diagnosis test was considered as a response variable. Moreover, the socio - economic, geographic and demographic variables were considered as explanatory variables. The socio-economic variables are main source of drinking water, time to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms per person, main material of the room's wall, main material of the room's roof, main material of the room's floor, use of indoor residual spray in the past twelve months, number of months rooms are sprayed, use of mosquito nets, total number of nets per person and type of nets. Geographic variables are region and altitude, and demographic variables are gender, age and family size. The mean structure is examined first by evaluating whether factors that affect malaria rapid diagnosis test are still important. Different method of estimations, Pseudo-Likelihood, Maximum Likelihood with Laplace Approximation and Maximum Likelihood with Adaptive Quadrature methods were used.

To perform analysis using PROC GLIMMIX, it is important to assume that for the model which contains random effects, the distribution of the data conditional on the random effects is known. Therefore, the distribution is a member of the exponential family distributions. Moreover, the conditional expected value of the data takes the form of a linear mixed model after a monotonic transformation is applied. Using PROC GLIMMIX, for models containing random effects, parameter estimates could be obtained by applying pseudo-likelihood techniques as in (Breslow and Clayton, 1993, Wolfinger and O'Connell, 1993). This is the default method for PROC GLIMMIX. Pseudo-likelihood method for generalized linear mixed models uses Taylor series expansions of the GLMM. The expansion is either the vector of random effects solutions or the mean of the random effects. These expansions are also referred to as the subject specific and marginal expansions. The abbreviation identifies the method as a pseudo-likelihood technique. But, estimation using Pseudo-likelihood method did not converge. Furthermore, GLMMs estimation of model parameters can be obtained by using maximum likelihood where the marginal distribution is numerically approximated by the Laplace method (METHOD = LAPLACE) or by adaptive gaussian quadrature (METHOD = QUAD).

Therefore, the analysis was performed using classical Gaussian and adaptive Gaussian quadrature as well as Laplace approximations. As discussed earlier, the likelihood obtained is based on numerical integration. Different numbers of quadrature points were used to estimate the effect of socio-economic, demographic and geographic variables. To identify the impact of different number of quadrature points, different quadrature points were used. The use of different quadrature points, (Q = 3, 5, 10, 20), did not lead to considerable difference for parameter estimation. But, for quadrature points greater than 5, there were slight difference for the estimation of parameters. But, there is no difference between parameter estimates for quadrature points 10 and 20. As a result, for the analysis, classical gaussian quadrature with large number of quadrature point was used. After the estimation of parameters, appropriate statistical inferences for the fixed and covariance parameters of the model can be performed. Tests of hypotheses for the fixed effects are based on Wald-type tests and the estimated variance-covariance matrix. The COVTEST statement option in PROC GLIMMIX enables to perform inferences about covariance parameters based on likelihood ratio tests.

The assessment of the model fit was performed using the log pseudo-likelihood and the generalized chi-square test. The minus twice the residual log pseudo-likelihood of the model fit was found to be 10551.76, whereas the generalized chi-square was 60022.2. The ratio of the generalized chi-square statistics divided by the degree of freedom is given by 1.07. This ratio measures the residual variability in the margin distribution of the data. Since the value is close to 1 (1.07), this indicates that the variability in the data has been properly modelled and hence there was no residual over-dispersion. This indicates that there is no lack of fit when the random effect was introduced in the model.

For the analysis, statistical inferences for the covariance parameters were performed. Significance tests were based on the ratio of likelihoods. The GLIMMIX procedure distinguishes two types of random effects. Depending on the parameters of the covariance structure, the procedure distinguishes between "G-side" and "R-side" random effects. The associated covariance structures ($\boldsymbol{G}$ or $\boldsymbol{R}$) are similarly termed the G-side and R-side covariance structure, respectively. R-side effects are also called "residual" effects. Therefore, if a random effect is an element of γ, it is a G-side effect and the model is the G-side covariance structure. The likelihood ratio test was obtained by fitting the model subject to the constraints imposed by the test-specification. The test statistic was formed as twice the difference of the log (pseudo) likelihoods of the full and the reduced models. The dimension of the parameter space is one. The random effect specified through the Z matrix as G-side effect for the variance of the random effect is *Kebele* effect. The estimate of the variance of the *kebele* is significant, i.e., there is *kebele*-to-*kebele* heterogeneity in the RDT of malaria. Tests whether the $\boldsymbol{G}$ matrix reduced to a zero matrix or not was performed. This eliminates all G-side random effects from the model. But, the result shows that the $\boldsymbol{G}$ matrix (X^2 = 2849.63, P - Value = <.0001), cannot be reduce to zero matrix. Therefore, there is G-side effects for RDT of malaria.

Table 5. 1: Type 3 analysis of effects for the GLMM

Effect	Num DF	F Value	Pr > F
Age	1	10.16	0.0014
Gender	1	0.12	0.7257
Family size	1	75.32	<.0001
Region	2	0.02	0.9761
Altitude	1	215.47	<.0001
Main source of drinking water	2	6.59	0.0014
Time to collect water	3	7.46	<.0001
Toilet facilities	2	5.2	0.0055
Availability of electricity	1	17.61	<.0001
Availability radio	1	2.82	0.0732
Availability television	1	4.5	0.034
Number of rooms/person	1	38.49	<.0001
Main material of the room's wall	2	12.94	<.0001
Main material of the room's roof	2	12.27	0.0262
Main material of the room's floor	2	13.37	<.0001
Use of indoor residual spray	1	986.9	<.0001
Number of months room sprayed	1	944.72	<.0001
Use of mosquito nets	1	2.64	0.1127
Number of nets/person	1	13.48	0.0002
Age and gender	1	0.19	0.9918
Main source of drinking water and main material of the room's roof	4	4.57	0.0004
Gender and use of mosquito nets	1	11.59	<.0001
Time to collect water and main material of the room's floor	4	14.57	0.0024
Gender & main source of drinking water	1	33.46	<.0001
Gender and main material of the room's floor	2	5.67	0.0035
Gender and use of indoor residual spray	1	849.57	<.0001
Use of mosquito nets and number of nets per person	1	849.57	<.0001
Age, gender and source of drinking water	4	8.42	<.0001
Age, gender and availability of electricity	2	7.8	0.0004

Model selection was achieved by first including into the model all predictor variables and then evaluating whether or not any interaction terms needed to be incorporated. This was achieved by fitting model effects one at a time, each of the interaction terms formed from the predictor variables, and retaining in the model only those interaction terms which were significant. This process continued until the final maximal model was obtained. The final chosen model for the malaria Rapid Diagnosis Test contained all main effects as well as six two-way interaction terms, and two three-way interaction terms. The final model is presented in Table 5.1.

Age, family size, altitude, main source of drinking water, time taken to collect water, availability of toilet facilities, availability of electricity, access to radio or television, number of rooms per person, main construction material of the rooms' walls and floors, incidence of indoor residual spray in the past twelve months, number of months the room was sprayed and total number of nets per person were found to be significant main effects. From these main effects, the following were involved in the interaction effects: main source of drinking water; time to collect water; availability of electricity; main construction material of the wall room, roof and floor; incidence of indoor residual spray; and the use of mosquito nets. There are two three-way and eight two-way significant interaction terms. The three-way interaction term is between age, gender and main source of drinking water and between age, gender and availability of electricity. The two-way interaction terms are between source of water and roof material; between number of nets per person and use of mosquito nets; between gender and availability of electricity; between gender and floor material; between time to collect water and construction material of room's floor; between gender and use of indoor residual spray; and between gender and number of months the room was sprayed. The interpretation of the results is presented as follows.

Table 5. 2: Estimates of odds ratio for main effects

Effect	Estimate	OR	95% C.I.		P-value
			Lower	Upper	
Intercept	0.622	1.863	1.369	2.536	<.0001
Age	-0.009	0.992	0.987	0.996	0.0002
Gender (Ref. Male)					
Female	-0.027	0.973	0.637	1.487	0.8995
Family size	0.037	1.038	1.018	8.118	<.0001
Region (Ref. SNNP)					
Amhara	0.004	1.044	0.972	1.036	0.8271
Oromiya	0.002	1.072	0.963	1.043	0.9053
Altitude	-0.007	0.978	0.945	0.998	<.0001
Main source of drinking water (Ref. protected water)					
Tap water	1.591	4.909	1.892	7.751	<.0001
Unprotected water	0.725	2.065	1.066	3.902	0.031
Time to collect water (Ref. less than 30 minutes)					
30 - 40 minutes	0.721	2.056	1.066	3.900	0.031
40 - 90 minutes	1.470	4.349	2.284	8.373	<.0001
> 90 minutes	0.069	1.071	0.959	1.065	0.6932
Availability of toilet facility (Ref. No facility)					
Pit latrine	-0.130	0.878	0.694	0.940	0.005
Toilet with flush	-0.112	0.894	0.610	0.956	0.0141
Availability of electricity (ref. no)					
Yes	0.166	1.181	0.987	1.133	0.1098
Availability of radio (ref. yes)					
No	-0.022	0.978	0.980	1.009	0.4328
Availability of television (ref. yes)					
No	-0.104	0.901	0.845	0.960	0.0013
Number of rooms/person	-0.057	0.945	0.908	0.982	0.004
Main material of room's wall (Ref. cement block)					
Corrugated metal	-0.329	0.719	0.700	0.740	<.0001
Mud block/stick/wood	-0.322	0.725	0.570	0.922	0.0086
Main material of room's roof (Ref. Corrugate)					
Thatch	0.006	1.006	0.995	1.018	0.0269
Stick and mud	0.045	1.046	1.016	1.077	0.0024
Main material of room's floor (Ref. /Local dung plaster)					
Cement-floor	-0.174	0.840	0.624	1.132	0.2532
Wood-floor	-0.136	0.872	0.657	1.158	0.3456
Use of indoor residual spray (ref. No)					
Yes	-0.396	0.673	0.656	0.690	<.0001
Number of months the room sprayed	-0.053	0.949	0.945	0.953	<.0001
Use of mosquito nets (ref. No)					
Yes	-0.009	0.991	0.999	1.019	0.0778
Number of nets/person	-0.034	0.966	0.949	0.984	0.0002

Table 5.2 presents odds ratio estimates associated with age, gender, family size, region, altitude, toilet facilities, main source of drinking water, time to collect water, availability of electricity, radio and television, number of rooms per person, main construction material of room's roof, use of indoor residual spray, number of months the room sprayed, use of mosquito nets and number of nets per person. Based on the results, for a unit increase in family size, the odds of positive malaria RDT Test for individuals increases by 3.76% (OR = 1.0376, P-value < 0.0001). Furthermore, for a unit increase in altitude, the odds of positive malaria RDT decreases by 0.2% (OR = 0.998, P - value <0.0001).

With reference to individuals with no toilets facility, the odds of malaria RDT was seen to be positive for more individuals with toilet with flush facilities (OR = 0.894, P-value = 0.0141) followed by pit latrines (OR = 0.878, P-value = 0.005). Moreover, for a unit increase in the number of total rooms, the odds of malaria RDT for individuals decreased by 5.5% (OR = 0.945, P-value = 0.004).

Table 5. 3: estimates and odd ratios for interaction effects

Effect	Estimate	OR	95% C.I.		P-value
			Lower	Upper	
Main source of drinking water and main material of the room's roof (ref. Protected water and cement block)					
Tap water and Mud block/stick/wood	-0.034	0.967	0.944	0.991	0.006
Tap water and Corrugated metal	-0.264	0.768	0.626	0.829	0.019
Unprotected water and Mud block/stick/wood	-0.008	0.992	0.966	1.000	0.020
Unprotected water and Cement block	-0.032	0.968	0.906	1.035	0.549
Time to collect water and material of room's floor (ref. Less than 30 minutes and earth/local dung plaster)					
Greater than 90 minutes and Cement	-0.039	0.962	0.857	1.079	0.5048
Greater than 90 minutes and Wood	-0.294	0.745	1.201	1.500	<.0001
Between 30 - 40 minutes and Cement	-0.016	0.985	0.980	1.053	0.3901
Between 30 - 40 minutes and Wood	0.145	1.156	1.147	1.165	0.0048
Between 40 - 90 minutes and Cement	-0.172	0.842	1.226	1.151	<.0002
Between 40 - 90 minutes and Wood	0.200	1.221	1.312	1.137	0.3901
Gender and main source of drinking water (ref. Male and protected water)					
Female and tap water	0.0169	1.017	0.941	1.099	0.0488
Female and unprotected water	-0.0795	0.924	0.854	0.999	0.0467
Gender and material of room's floor (ref. Male and earth/Local dung plaster)					
Female and Cement	-0.0175	0.983	0.619	0.998	0.0408
Female and Wood	0.2741	1.315	0.859	2.014	0.0075
Gender and use of mosquito nets (ref. Male and yes)					
Female and no	-0.034	0.967	0.964	0.969	<.0001
Gender and use of indoor residual spray (ref. Male and no)					
Female and yes	0.0018	1.002	0.985	1.030	0.0055
Number of nets per person and use of mosquito nets (ref. No)					
Yes	0.00491	1.005	1.000	1.010	0.0467
Age and gender (ref. Male)					
Age and female	0.0336	1.034	0.992	1.002	0.4011
Age, gender, main source of drinking water (ref. Male and protected water)					
Female and tap water	-0.00098	0.999	0.998	1.000	0.0119
Female and unprotected water	0.00199	1.002	1.001	1.003	<.0001
Age, gender and electricity (ref. Male and yes)					
Female and no	0.00335	1.003	0.995	1.105	0.0003

Interaction effects

Figures 5.1 and 5.2 show the distribution of malaria RDT against the main source of drinking water for both males and females respectively. As age increased, positive malaria diagnosis was less likely for males than for females who were using protected, unprotected and tap water for drinking. Furthermore, as age of respondents increased, malaria RDT was less likely to be positive for individuals who used tap water for drinking (OR = 0.98, P - Value < 0.0001) for males and (OR = 1.077, P - Value < 0.0001) for females. More specifically, positive malaria diagnosis rates increased with age for females whereas it decreased for males as age increased (Figures 5.1 and 5.2). The Figures further show that the gap in the Rapid Diagnosis Test between respondents using unprotected, protected and tap water for drinking widens with increasing age.

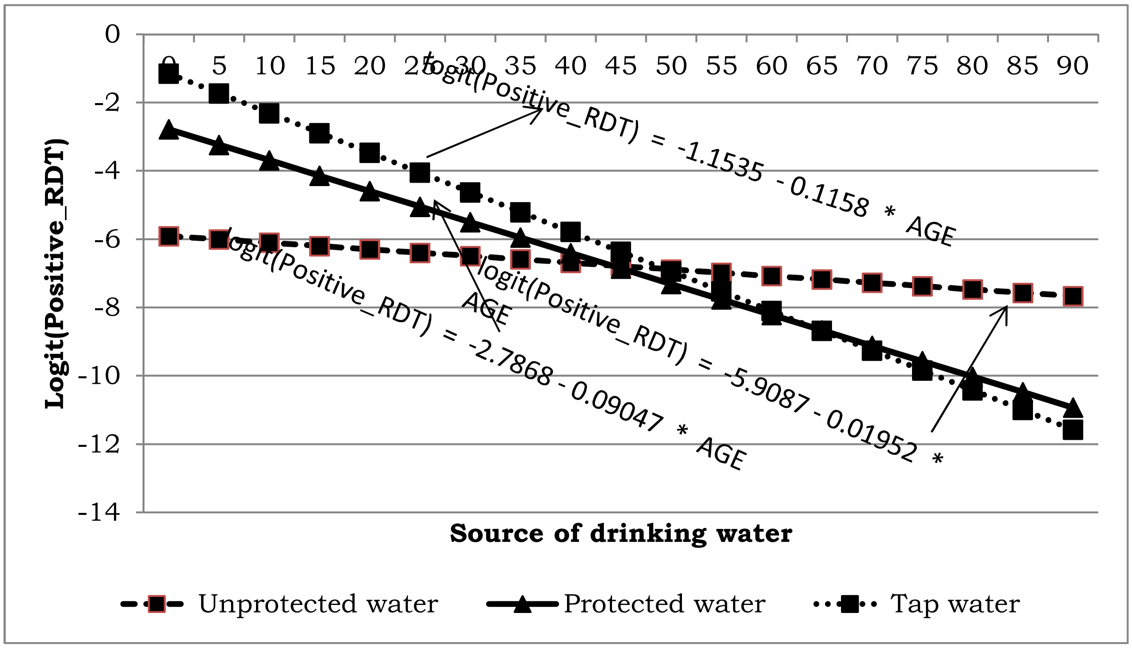


Figure 5. 1: Log odds associated with rapid diagnosis test and age for male respondents with source of drinking water

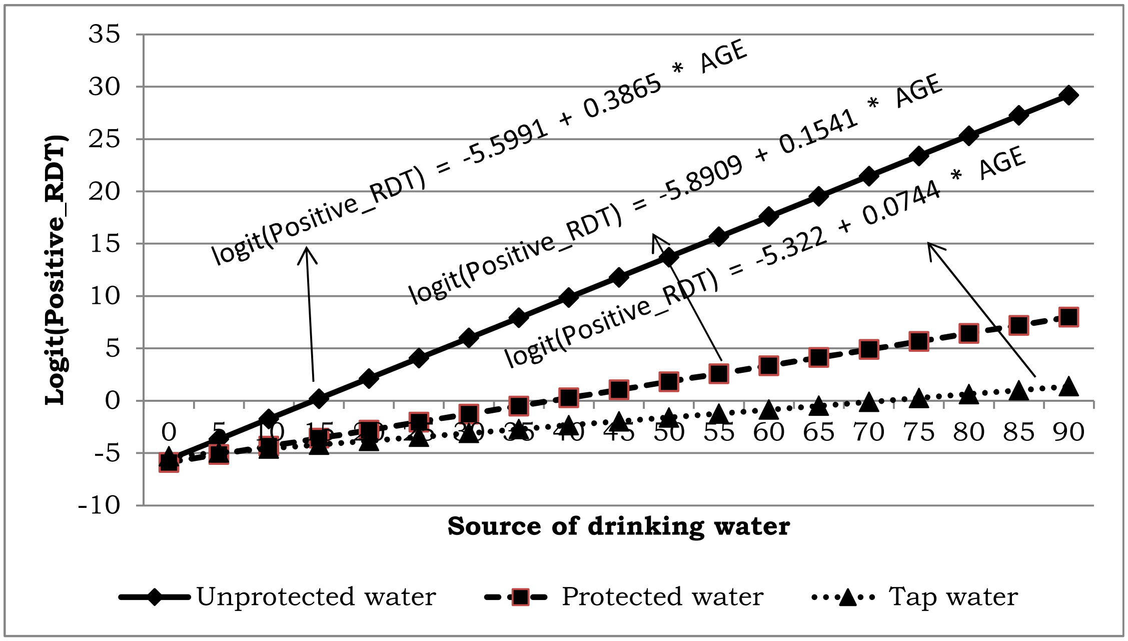


Figure 5. 2: Log odds associated with rapid diagnosis test and age for female respondents with source of drinking water

The relationship between age, gender and availability of electricity is presented in Figure 5.3. As the Figure indicates, positive malaria RDT decreases as age increases for both male and female respondents, whether or not they had access to electricity. However, the rate of decrease was not the same for males and females after controlling for other covariates in the model. The rate of increase for females who responded positively to having electricity was 9.14% higher than the other categories [OR = 1.0914, p-value < 0.001]. Probabilities for this interaction are presented in Figure 5.3.

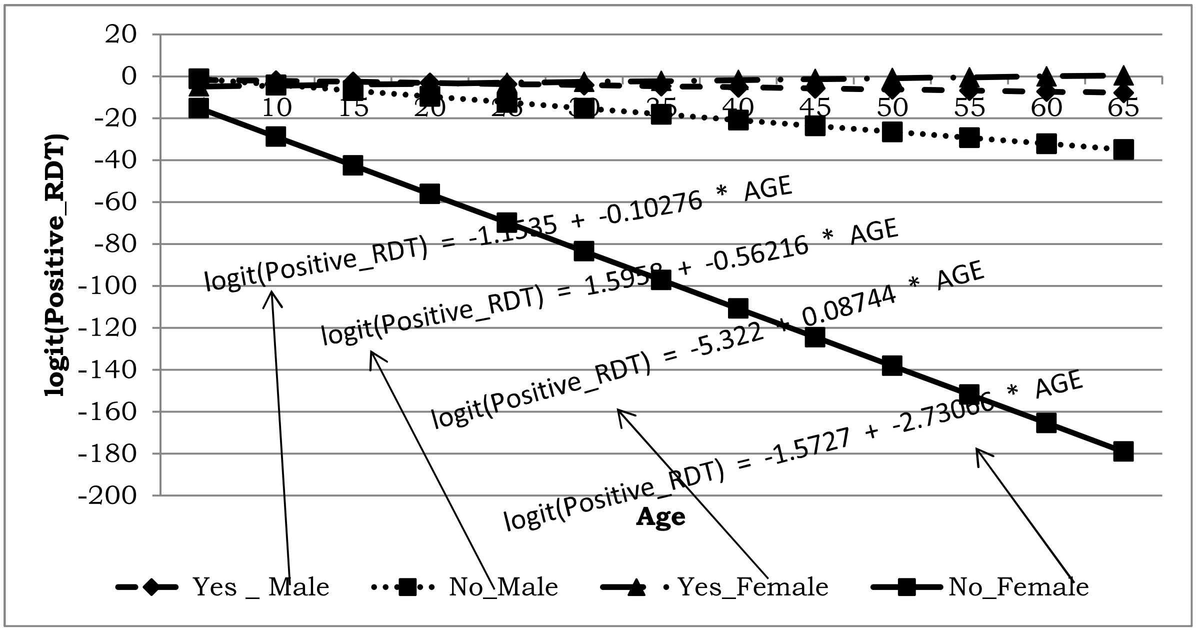


Figure 5. 3: Log odds associated with rapid diagnosis test with age for male and female respondents with availability of electricity

Interaction effects between main source of water and main construction material of the room's roof are presented in Figure 5.4. From the Figure, it is clearly seen that with respondents who reported using tap water as well as protected and unprotected water for drinking, positive rapid diagnosis of malaria was significantly higher when the roof of the house was thatched, followed by those who occupied a stick and mud roof and finally respondents living in a house with a corrugated iron roof. The difference in rapid diagnosis test between the respondents' use of tap, protected and unprotected sources of drinking water and having a thatch or stick/mud roof was particularly significant. It has also shown that for a corrugated iron roof, positive rapid diagnosis test was significantly lower for respondents who reported using tap water for drinking than for those who were using protected and unprotected water.

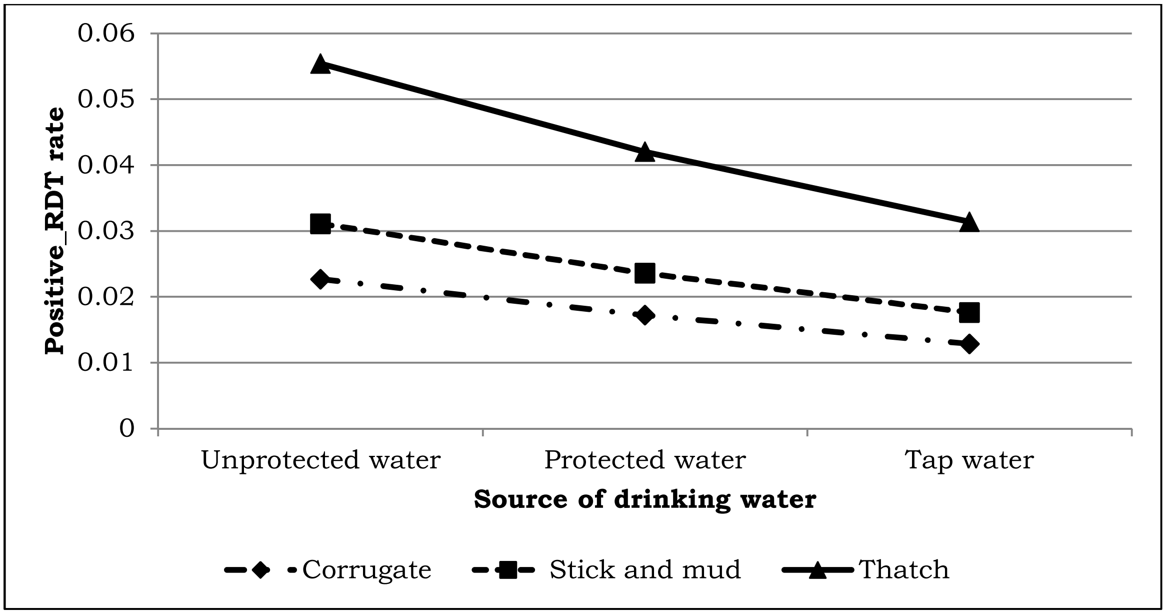


Figure 5. 4: Log odds associated with rapid diagnosis test and source of drinking water with material of the room's roof

The other two-way interaction effect which is significant is between the time taken to collect water and main construction material of the room's floor (Table 5.1). This result is presented graphically in Figure 5.5. Positive malaria rapid diagnosis test was significantly higher in a room with an earth or dung and plaster floor than in one with cement or wooden floors for respondents who took < 30 minutes and >90 minutes to collect water. But, for respondents who took less than 90 minutes to collect water and had a cement floor, positive rapid diagnosis test is low. Furthermore, for respondents who took between 30 to 40 minutes to collect water, there was lower positive rapid diagnosis test for respondents with an earth or dung and plaster floor and a wooden floor.

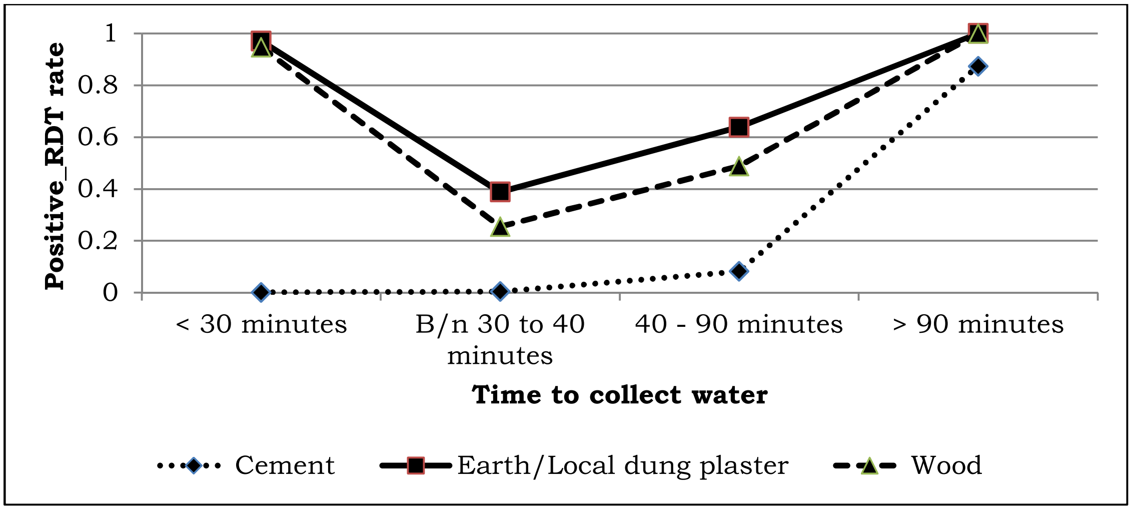


Figure 5. 5: Log odds associated with rapid diagnosis test and time to collect water with material of the room's floor

The relationship between the main construction material of the room's floor and gender for a household is presented in Figure 5.6. As the Figure indicates, positive Rapid Diagnosis Test was significantly higher for males than females with respondents who reported having an earth or dung and plaster floor. But, the result is higher for females for those who reported having a wooden floor in their house. There was however, no significant difference in positive rapid diagnosis test between females and males who reported having a room with a cement floor.

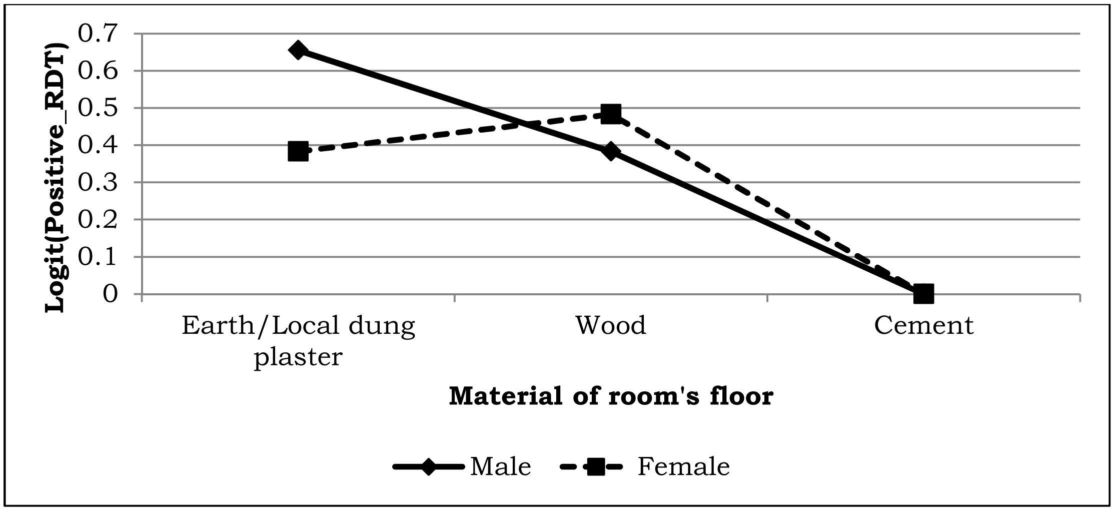


Figure 5. 6: Log odds associated with rapid diagnosis test and material of room's floor with gender

The interaction effect between gender and main source of drinking water is presented in Figure 5.7. The Figure shows that the risk of malaria for households using unprotected water is significantly higher than for those households who reported having protected and tap water for both males and females. Moreover, for female members of the household, the risk of malaria was higher for those households who reported having unprotected water.

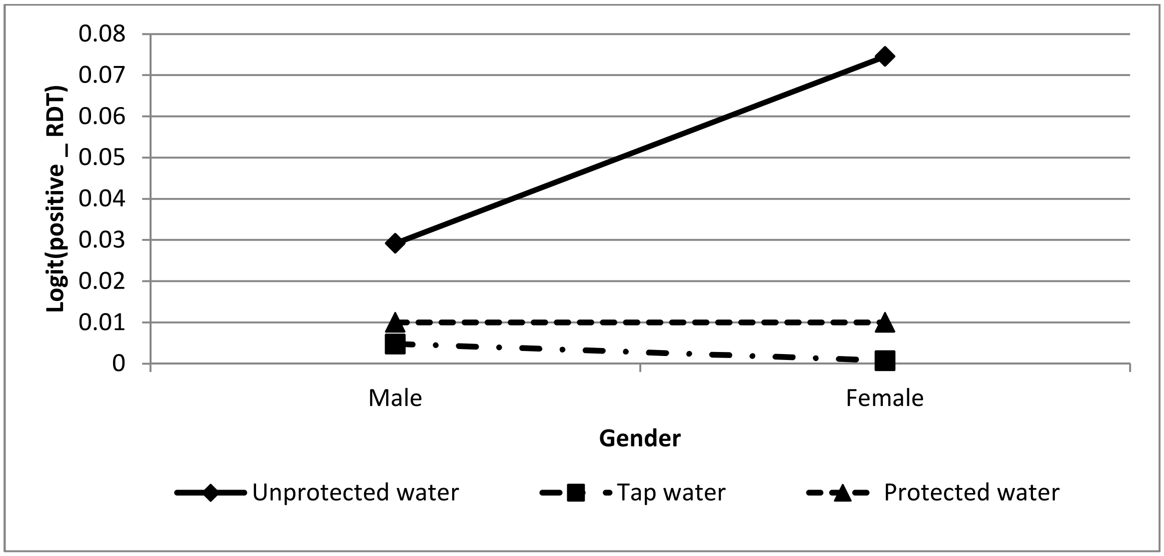


Figure 5. 7. Log odds associated with rapid diagnosis test and main source of drinking water with gender

Figure 5.8 presents the interaction effect between the use of indoor residual spray and gender for individuals. Prevalence of malaria was significantly higher for male than for female respondents who were living in a house treated with indoor residual spraying. For males living in a house, which was not treated with indoor residual spraying, the positive malaria result was significantly higher than it was for females.

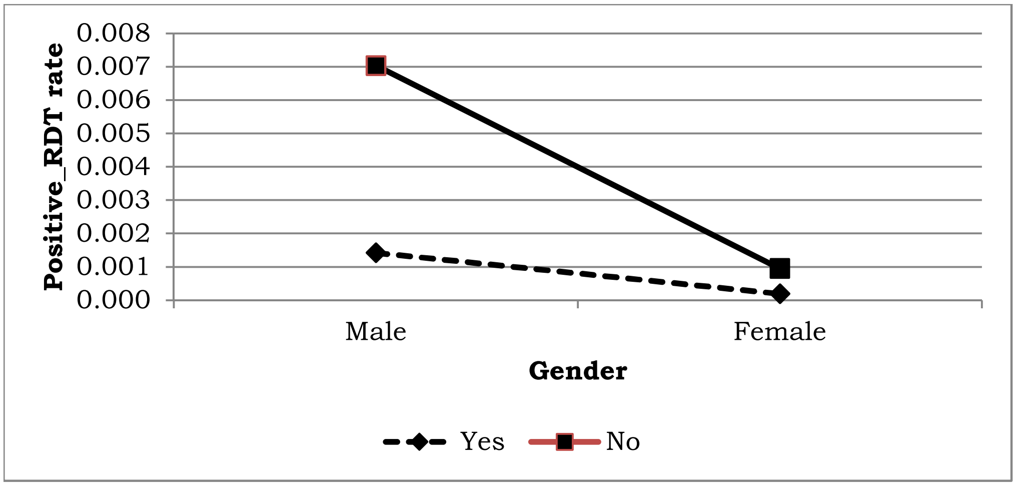


Figure 5. 8: Log odds associated with rapid diagnosis test and use of indoor residual spray of respondents with gender

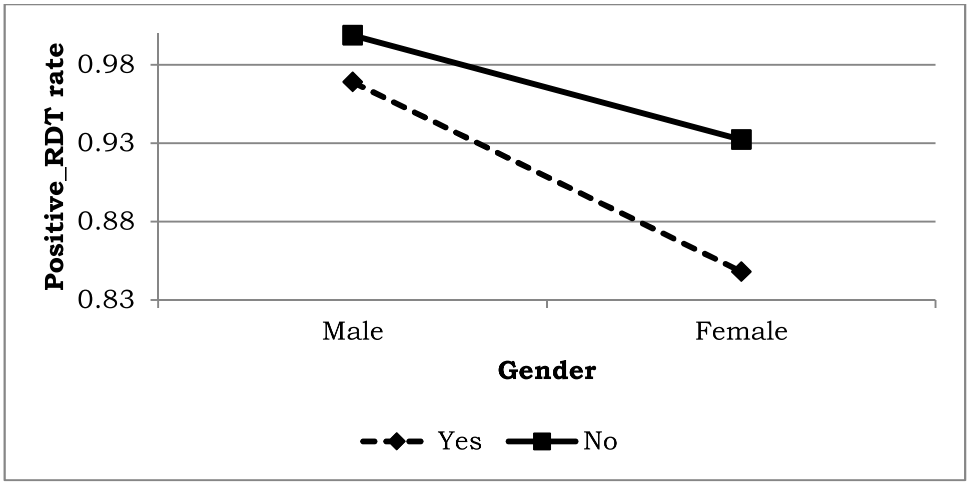


Figure 5. 9: Log odds associated with rapid diagnosis test and use of mosquito nets with gender

Similarly, the interaction effect between use of mosquito nets and gender is presented in Figure 5.9. As the Figure indicates, the risk of malaria is higher for males than for females using mosquito nets when sleeping.

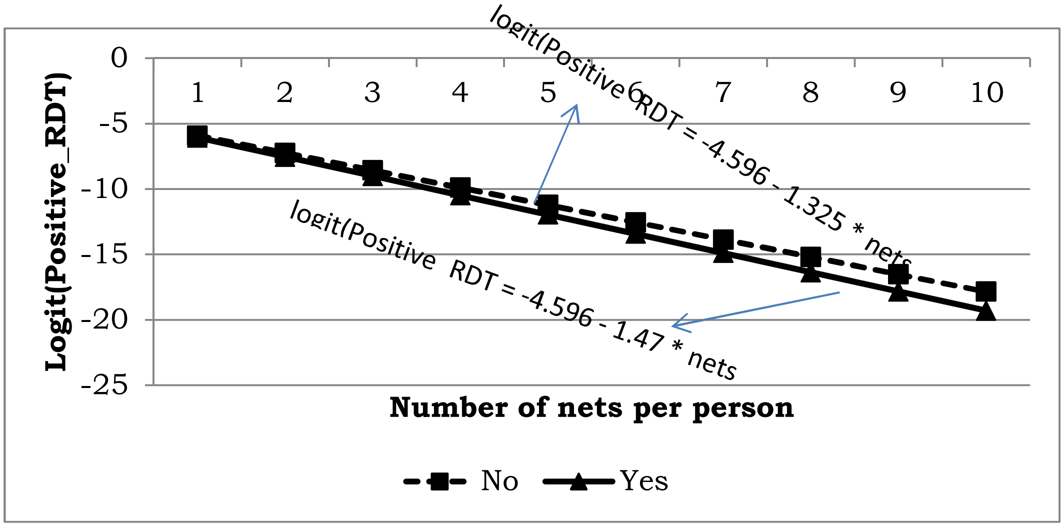


Figure 5. 10: Log odds associated with rapid diagnosis test and use of mosquito nets with number of nets per person

As the number of mosquito nets increased, the risk of malaria was less likely for household members with and without nets. However, the risk of malaria was found to be much lower for individuals as the number of nets increased (Figure 5.10). This Figure shows that for individuals with and without the use of mosquito nets, the risk of malaria decreased as the number of net ownerships in the household increased.

5.5 Summary and discussion

The study indicates that socio-economic, demographic and geographic factors are responsible for the transmission of malaria. These factors are age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, availability of radio, total number of rooms, main construction material of the room's walls, main construction material of the room's floor, use of indoor residual spray, use of mosquito nets and total number of nets were the major factors associated with malaria rapid diagnosis test results. In addition to the main effects, three-way and two-way interaction effects were identified. The three-way interactions were between age, gender and main source of drinking water and age, gender and availability of electricity. The two-way interaction effects were between main source of drinking water and main construction material of the room's roof, time taken to collect water and main construction material of the room's floor, age and gender, gender and main source of drinking water, gender and availability of electricity, and gender and main construction material of the room's floor.

In the present study, the effect of socio-economic factors shows that residents with no toilet facilities were found to be at more risk of malaria than those with toilet facilities. Additionally, malaria prevalence is low for households with a greater number of rooms in the house. On the other hand, having more mosquito nets over beds was found to be one way of reducing the risk of malaria. The prevalence of malaria for households with access to clean water was found to be less. Malaria rapid diagnosis was found to be higher for those respondents living in thatched houses, or ones with stick and mud roofs. Therefore, having a house with a corrugated iron roof was found to reduce the risk of malaria. Furthermore, the prevalence of malaria for households with earth and local dung and plaster floors was found to be higher. Moreover, the treatment of walls of houses with indoor residual spraying was found to be one means of reducing the risk of malaria.

Based on demographic factors associated with malaria, our findings showed that females and children are at a greater risk. Furthermore, the malaria prevalence rate was found to be less for households with fewer people in the house. Malaria prevalence was similarly associated with geographic factors. The association between malaria and altitude showed that malaria prevalence is higher for households who are living at lower altitudes.

In conclusion, the government of Ethiopia has adopted various strategies to control malaria. These include early diagnosis, prompt treatment, selective vector control, epidemic prevention and control. In addition to this, the government has supporting strategies such as human resource development, monitoring and evaluation. One of the government's key goals in the control of malaria is to achieve the complete elimination of malaria within those geographical areas with historically low malaria transmission and achieve near zero malaria transmission in the remaining malarious areas of the country. For this reason, evidence based strategies to prevent malaria is an attractive strategy for the country (Goovaerts, 1997). Therefore, the results from this study showed that malaria is associated with socio-economic, demographic and geographic factors, mainly influenced by poverty levels. Malaria is generally regarded as a disease of the poor or those living in poverty. The more wealthy households who can afford to have toilet facilities, a greater number of rooms in the house, clean drinking water, and well built houses were found to be less affected by malaria. Furthermore, it was found that women and children are more vulnerable to malaria. Lack of bed nets contributes to this vulnerability. Moreover, as our results indicate having more bed nets is one means of reducing malaria and evidence suggests that households who are unable to afford sufficient mosquito nets, due to large families and low incomes, are more affected by malaria. Women and children are also exposed to mosquito bites while they are travelling long distances to fetch water. As the wealthier households were found to be less vulnerable to malaria than the poor households, improving the living conditions of the communities could be one way of achieving the malaria control goals set by the health professionals.

The method used in this chapter investigated the variability between PSU's which is *kebele*. But, besides PSU's variability, there might be spatial variability between selected households. Therefore, this variability will be investigated in the next chapter.

Chapter 6

Spatial distribution of malaria problem in three regions of Ethiopia

6.1 Introduction

In the previous chapter (Chapter 5), the generalized mixed model was used. But, the distribution of malaria is non-random across a landscape in areas of higher or lower transmission intensity and malaria risk. The transmissions are separated by greater or lesser distances from each other. Based on geographical aggregation, there are two distinct levels. These are, the focal unit of malaria transmission, the area over which human malaria is actively transmitted originating from a specific aquatic breeding site and the household or other reasonably identified point of contact between a small group of humans and mosquito vectors. The baseline household cluster malaria survey which was conducted by The Carter Center from December 2006 to January 2007 includes the geographical locations of the reference of each household. Therefore, it is of interest to know whether the data display any spatial autocorrelation. Furthermore, it is important to check whether surveys that are near in space have malaria prevalence or incidence that is similar with the surveys that are far apart. This is important because spatially correlated data cannot be regarded as independent observations. If the analysis does not take account of the correlation structure of the data, the estimates obtained from the analysis may be inaccurate because of the underestimated standard errors. Therefore, the objective of this chapter is to undertake statistical analysis of malaria incidence to identify important socio-economic, demographic and geographic variables associated with the disease and to produce prevalence maps of the area illustrating the variation in malaria risk using spatial statistics analysis.

Spatial statistical analysis provides useful insights about the causes, patterns, and prevalence of malaria transmission. There are different methods available to display disease distributions and analyze spatial patterns. By considering a variety of linkages or looking at the patterns of clustering of a malaria distribution, it is possible to investigate the factors at large or small scale. Therefore, tools for spatial representation of events have recently improved, with the widespread availability of geographical information systems (GIS). To model spatial variation of disease as well as the relationship of malaria to socio-economic, demographic and geographic factors and the health care system, GIS technology is an important tool (Craig et al., 1999, Schellenberg et al., 1998).

Spatial statistics can be divided into three methods. These are: point pattern analysis, methods for lattice data and geostatistics (Schabenberger and Gotway, 2005, Cressie, 1993). *Point referenced data:* - is often called geocoded or geostatistical data. *Areal data:* - is often called lattice data. Some spatial data sets feature both point and areal-level data. *Point pattern data:* - The response occurrence of the event is often fixed and only the locations where it occurs are thought of at random. Of these, the geostatistical approach is most relevant to epidemiological analysis which is conducted at the landscape scale and based on remote sensing (Chiles and Delfiner, 1999, Goovaerts, 1997, Goovaerts et al., 2005).

A common approach to integrate spatially correlated data with the random effects and proceed with maximum likelihood based approaches for estimating the covariate and covariogram parameters, is based on the theory of generalized linear mixed models (GLMM). Using GLMM, numerical approximation can be implemented (Lesaffre and Spiessens, 2001, Schabenberger and Gotway, 2005). Therefore, the aim of this chapter is to review the theory of spatial statistics and then fit them to malaria RDT data.

Specifically, interest is to model the effect of socio-economic, demographic and geographic factors on malaria rapid diagnosis test status. This chapter is organized as follows. An overview of theory of spatial statistics is presented in Sections 6.2 – 6.4. The spatial statistics model is fitted to malaria RDT data in Section 6.5. Section 6.6 gives summary and discussion of the chapter.

6.2 Models for spatial correlations

Non-Gaussian spatial problems may be formally analyzed in the context of generalized linear mixed models (GLMM). Specification of the likelihood of the random variable $y(s)$ is required where s generally denotes the location of the observation is made. As in classical generalized linear models (GLMs), there is a canonical parameter corresponding to the distribution, which is normally a function of the location parameter via the link function $g(.)$ for the distribution. This function is assumed to be linear in the explanatory variables. In the classical formulation of GLMs containing only fixed effects, $g(\mu) = X\beta,$ where X is the matrix of explanatory variables (Berridge and Crouchley, 2011, Zuur et al., 2009, Zurr et al., 2007, Fox, 2008, Madsen and Thyregod, 2010). To incorporate a spatial process, we assume $y(s_i|\alpha)$ is conditionally independent for any location s_i with conditional mean $E[y(s_i)|\alpha] = \mu(s_i)$. The parameter α is used to define the distribution of s. Then, the spatially correlated random effect is incorporated into the linear predictor as

$$g(\mu) = \boldsymbol{X}\beta + \boldsymbol{Z}\alpha \qquad (6.1)$$

where X and Z are the design matrices. The error term accommodates over-dispersion relative to the mean-variance relationship implied by the distribution under consideration. The random effect at location s_i, $\alpha \sim Gau(0, \varSigma_\alpha(\theta))$ and $\varepsilon \sim Gau(0, \sigma_\varepsilon^2 I),$ with spatial correlation is parameterized by θ in $\varSigma_\alpha(\theta)$ (Schabenberger and Gotway, 2005). Note that s_i is just one location and $\boldsymbol{s} = (s_1,\ldots,s_k)'$ denotes a vector of k locations with variance-covariance matrix Σ.

Spatial dependence may be represented by a range of functions (Hengl, 2007). To describe spatial correlation of observations, there are three major functions used in geostatistics. These major functions are the correlogram, the covariance, and the semivariogram. Semivariogram is also more simply called the variogram. In geostatistics, the variogram is a key function and is used to fit a model for the spatial correlation in the data. The model which is obtained using the variogram is used in kriging estimation procedures, a method which was first used to minimize (Goovaerts, 1997). Moreover, variogram models are also used to understand maximum distances of spatial autocorrelation which can further be used in construction of search parameters for different interpolation techniques. A variogram represents both structural and random aspects of the data under consideration. The structural part of the variogram model is represented by the range of a variogram. Furthermore, the variogram values increase with increases in the distance of separation until it reaches the maximum at a distance known as the “range”. To develop the variogram, assume $\mu(s)$ is a constant, that is constant mean $\mu(s)$, and define

$$var\{Z(s_1) - Z(s_2)\} = 2\gamma(s_1 - s_2). \tag{6.2}$$

In statement (6.2), the variance of s_1 and s_2 is through their difference $s_1 - s_2$ and the process which satisfies this property is called intrinsically stationary. The function $2\gamma(.)$ is called the variogram and $\gamma(.)$ the semivariogram.

The other concept here is isotropy. Suppose the process is intrinsically stationary with semivariogram $\gamma(\boldsymbol{h})$, $\boldsymbol{h} \in \boldsymbol{R}^d$. If $\gamma(h) = \gamma_0(\|h\|)$ for some function γ_0, i.e. if the semivariogram depends on its vector argument $\boldsymbol{h}$ only through its length $\|h\|$, then the process is isotropic. Therefore, a process which is both intrinsically stationary and isotropic is also called homogeneous. Isotropic processes are convenient to deal with because there are a number of widely used parametric forms for $\gamma_0(h)$. Using semivariance $\gamma_0(t)$ for interval distance class t, lag distance interval t, c_0 (nugget variance) ≥ 0, c_1 (structural variance) $\geq c_0$ and R is the range parameter, some of the examples are:

1. Linear:

$$\gamma_0(t) = \left\{ \begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1 t & if\ \ t > 1. \end{array} \right\}$$

Here c_0 and c_1 are positive constants. The function tends to ∞ as $t \rightarrow \infty$ and so does not correspond to a stationary process.

2. Spherical:

$$\gamma_0(t) = \left\{ \begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1 t \left\{ \frac{3}{2} \frac{t}{R} - \frac{1}{2} \left(\frac{t}{R} \right)^3 \right\} & if\ \ 0 < t \leq R, \\ c_0 + \ c_1 t & if\ \ t \geq R. \end{array} \right\}$$

This is valid if d = 1; 2 or 3, but for higher dimensions it fails the non-positive-definiteness condition. It is a convenient form because it increases from a positive value c_0 when t is small, levelling at the constant $c_0 + c_1$ at $t = R$. This is of the "nugget/range/ sill" form which is often considered a realistic and interpretable form for a semivariogram.

3. Exponential:

$$\gamma_0(t) = \left\{ \begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1 (1 - e^{-t/R}) & if\ \ t > 1. \end{array} \right\}$$

This is simpler in functional form than the spherical case (and valid for all d) but without the finite range of the spherical form. The parameter ***R*** has a similar interpretation to the spherical model however, of fixing the scale of variability.

4. Gaussian:

$$\gamma_0(t) = \left\{\begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1(1 - e^{-t^2/R^2}) & if\ \ t > 1. \end{array}\right\}$$

5. Exponential-power form:

$$\gamma_0(t) = \left\{\begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1(1 - e^{-|t/R|^p}) & if\ \ t > 1. \end{array}\right\}$$

Here $0 < p \leq 2$. This form generalizes both the exponential and Gaussian forms, and forms the basis for the families of spatial covariance functions introduced by (Sacks et al., 1989). However, in generalizing the results from one dimension to higher dimensions, these authors used a product form of covariance function in preference to constructions based on isotropic processes (Gaetan and Guyon, 2010).

6. Relational quadratic:

$$\gamma_0(t) = \left\{\begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1 t^2(1 + e^{t^2/R}) & if\ \ t > 1. \end{array}\right\}$$

7. Wave:

$$\gamma_0(t) = \left\{\begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1\{1 - \frac{R}{t}\sin(\frac{t}{R})\} & if\ \ t > 1. \end{array}\right\}$$

8. Power law

$$\gamma_0(t) = \left\{\begin{array}{ll} 0 & if\ t = 0, \\ c_0 + \ c_1 t^{\lambda} & if\ \ t > 1. \end{array}\right\}$$

Non-positive-definiteness requires $0 \leq \lambda < 2$. This generalizes the linear case, and it is an example of a semivariogram that does not correspond to a stationary process (Gaetan and Guyon, 2010).

9. The Matérn class: This method which was given by (Matérn, 1960) neglected in favour of simpler analytic forms. (Handcock and Stein, 1993, Handcock and Wallis, 1994) demonstrated the flexibility of this method in

handling a variety of spatial data set. The class is best defined in terms of isotropic covariance. Therefore,

$$C_0\,(t) = \frac{1}{2^{\theta_2-1}\Gamma(\theta_2)}\left(\frac{2\sqrt{\theta_2 t}}{\theta_1}\right)^{\theta_2} K_{\theta_2}\left(\frac{2\sqrt{\theta_2 t}}{\theta_1}\right)$$

where $\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape of parameter, $\Gamma(.)$ is the gamma function, $K_{\theta_2}$is the modified Bessel function.

For most of the variograms, $\Upsilon_0(0) = 0$, but Υ_0 increases from a non-negative value near t = 0 (the nugget) to a limiting value (the sill) which is either attained at a finite value t = R (the range). The shape of the semi variograms have the form which is presented in Figure 6.1 (Clay and Shanahan, 2011).

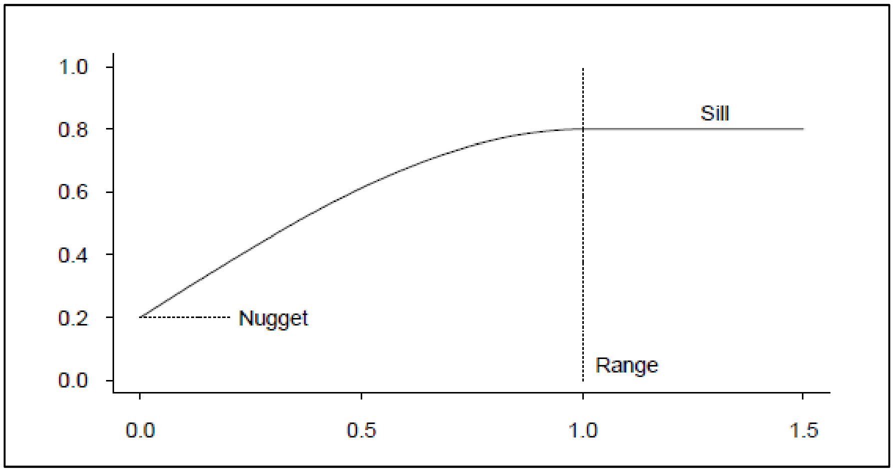


Figure 6. 1: Idealized form of variogram function, illustrating the nuggest, sill and range

For positive nugget, it is paradoxical because the positive nuggets imply discontinuity in the covariance function. This situation is a well-known feature of spatial data. Furthermore, these cases have various explanations. Among the possible explanations, the simplest explanation related to some residual white noise over and above any smooth spatial variation (Waller and Gotway, 2004).

To deal with anisotropic processes, there are a number of direct generalizations. From these methods, the simplest method is geometric anisotropy. A semivariogram with the form of geometric anisotropy is given by

$$\Upsilon(h) = \Upsilon_0(\|\boldsymbol{A}h\|)$$

where Υ_0 is an isotropic semivariogram and $\boldsymbol{A}$ is a $d \times d$ matrix, representing a linear transformation of R^d. If $\boldsymbol{A}$ is the identity this reduces to isotropic case, the process is isotropic in some linearly transformed space. Furthermore, for a positive definite matrix $\boldsymbol{A}$, the contours of equal covariance are ellipses instead of circles. To generalize the anisotropy, let the simple independent intrinsically stationary process be $Z_1,\dots,Z_p$. Then

$$Z = Z_1 + \dots + Z_p,$$

is also intrinsically stationary, with semivariogram given by

$$\Upsilon(h) = \Upsilon_1(h) + \dots + \Upsilon_p(h),$$

$\Upsilon_1,\dots,\Upsilon_p$ denoting the semivariograms of $Z_1,\dots,Z_p$ respectively. Thus

$$\Upsilon(h) = \sum_{i=1}^{p} \Upsilon_0(\boldsymbol{A}_i h),$$

where Υ_0 is an isotropic semivariogram and $\boldsymbol{A}_1,\dots,\boldsymbol{A}_p$ are matrices, is a valid semivariogram generalizing geometric anisotropy which is called zonal anisotropy (Gaetan and Guyon, 2010).

Moreover, for some nonlinear function $g(s)$, the process $Z(g(s))$, rather than $Z(s)$, is a stationary isotropic process. Therefore, non-stationarity as well as non-isotropic cases can be handled (Sampson and Guttorp, 1992). Spatial covariance or semivariogram function can be defined arbitrarily. To define the function, positive definiteness has to be satisfied. Generally, $cov\,\{Z(s_1), Z(s_2)\} = C(s_1,\ s_2)$. But, this equation does not support any form of stationary condition. Therefore, for positive definiteness, the relation

$$\sum_i \sum_j a_i a_j C(s_i, s_j) \geq 0.$$

This relation holds for any finite set of points $s_1, \ldots, s_n$ and arbitrary real coefficients $a_1, \ldots, a_n$. Furthermore, based on *Bochner's theorem*, the left hand side of the above relation is the variance of $\sum_j a_i Z(s_i)$. For d dimensional stationary process, *Bochner's theorem* implies that

$$C(h) = \int \ldots \int \cos(w^T h)\, g(w) dw$$

where $G(dw) = g(w)dw$ the integral is over R^d and G is a positive bounded spectral measure (Cliff and Ord, 1981). For

$$\int \ldots \int |C(h)| dh < \infty$$

G is automatically differentiable. For positive definiteness, $g(w) \geq 0$ for all w. Therefore, if the process is isotropic, $C(h) = C_0(\|h\|)$ for some function C_0 of univariate argument, then the spectral representation simplified to

$$C_0(t) = \int_{(0,\infty)} Y_d(wt) \Phi(dw),$$

where Φ is non-decreasing on $[0, \infty)$ with $\int \Phi(dw) < \infty$ and

$$Y_d(t) = \left(\frac{2}{t}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{\frac{d-2}{2}}(t)$$

and $J_v(.)$ denotes the *Bessel* function of first order v (Schabenberger and Gotway, 2005). Moreover, there is corresponding theory for the variogram. For second-order stationary process of semivariogram $\Upsilon(.)$, if $a_1, \ldots, a_n$ are constants with $\sum a_i = 0$, then

$$\sum_i \sum_j a_i a_j \Upsilon(s_i - s_j) \leq 0.$$

Therefore, this equation is a conditional non-positive definiteness condition (Cressie, 1993).

6.3 Estimation

After developing the main concepts of spatial covariance and variogram, the next question is estimation. To estimate the variograms, there are different methods. These methods are *Matheron's* (Method of moments) estimator, the *Cressie-Hawkins* robust estimator and estimators based on order statistics and quantiles. Therefore, the general scenario for estimation is that there is a process $\{Z(s), s \in D\}$ observed at a finite number of points $s_1, \ldots, s_N$.

For estimating the variogram, the simplest estimator is the method of moments (MoM) estimator. This method is also known as Matheron's estimator and proposed by (Matheron, 1962). One could plot the squared differences $\{Z(s_i) - Z(s_j)\}^2$ against the lag distance h (or $\|h\|$). Moreover, $\{Z(s_i) - Z(s_j)\}^2$ unbiasedly estimates the variogram at lag $s_i - s_j$. The semivariogram estimator with distance $s_i - s_j = h$ (averages of the squared differences of points) apart is known as the classical or Matheron. More useful estimator of the variogram is obtained by summarizing the squared difference. For this estimator, the sampling points $s_1, \ldots, s_N$ lie on a regular lattice (Schabenberger and Gotway, 2005) and defined by

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(s_i, s_j) \in N(h)} \{Z(s_i) - Z(s_j)\}^2 \qquad (6.3)$$

where $N(h)$ denotes all pairs (s_i, s_j) for $s_i - s_j = h$ and $|N(h)|$ denotes the cardinally of $N(h)$. Furthermore, $N(h)$ will either be empty or some reasonably sized subset of the set of all pairs of sampling points. In the case where the points do not lie on lattice, the same equation (6.3) is applied, but the definition of $N(h)$ to be changed as

$$N(h) = \{(s_i, s_j): s_i - s_j \in T(h)\},$$

where $T(h)$ being some small neighborhood or tolerance region around h. Moreover, (Jurnel and Huijbregts, 1978) recommended choosing *T(h*) large to contain at least 30 pairs of points, and this can still be recommended as a rule of thumb (Schabenberger and Gotway, 2005).

One of the shortcomings to use the MoM method is that it is not robust against outlying value of Z. This arises from the skewness of the distribution. If the process is assumed to be Gaussian, then for a specific s and h, the distribution of $\{Z(s_i) - Z(s_j)\}^2$ is of the form $2\gamma(h)X_1^2$, and the X_1^2 distribution is highly skewed. However, if $X \sim X_1^2$, then $X^{1/4}$ has a nearly symmetric distribution and sample averages of $|Z(s_i) - Z(s_j)|^{1/2}$ expected to be much better behaved than $\{Z(s_i) - Z(s_j)\}^2$ (Gaetan and Guyon, 2010). (Cressie and Hawkins, 1980) suggested another estimator. This estimator is known as Cressie-Hawkin's robust estimator. Based on (Cressie and Hawkins, 1980), the fourth root transformation of $\{Z(s_i) - Z(s_j)\}^2$yields an approximately Gaussian random variable with mean

$$E[|Z(s_i) - Z(s_j)|^{1/2}] \approx \frac{1}{2}\pi^{-1/2}\Gamma \times \gamma(s_i - s_j)^{1/4}.$$

Moreover, the expected value of the fourth power of

$$\frac{1}{|N(h)|}\sum_{|N(h)|}|Z(s_i) - Z(s_j)|^{1/2}$$

turns out to be

$$2\gamma(h)\left(0.457+\frac{0.494}{|N(h)|}+\frac{0.045}{|N(h)|^2}\right).$$

Therefore, the suggested estimator

$$2\gamma(h)=\frac{1}{0.457+0.494/|N(h)|}\left\{\frac{1}{|N(h)|}\sum_{(s_i,s_j)\in|N(h)|}\left|Z(s_i)-Z(s_j)\right|^{1/2}\right\}^4$$

is approximately unbiased estimator of $2\gamma(h)$. Therefore, from the equation, the first factor is a bias correction term.

The Cressie-Hawkin (CH) estimator is not resistance estimator under gross contamination of the data. Therefore, CH estimator has small amount of contribution in a Gaussian process, there is a breakdown point of 50% and unbounded influence function using CH and Matheron estimators (Schabenberger and Gotway, 2005). Here, the percentage of the data can be replaced by arbitrary values. For example, if the median absolute deviance (MAD) is an estimator of scale with 50% breaks down point and a smooth influence function, then, for a set of numbers $x_1,\ldots,x_n$, the MAD is

$$MAD=b\times median_i\{x_i-median_j(x_j)\}$$

where $medan_i(x_i)$ denotes the median of x_i and factor b is chosen to yield approximate unbiasedness and consistency. Suppose $x_1,\ldots,x_n$ are independent realization from a $G\,(\mu,\sigma^2)$. A robust estimator of scale for 50% breakdown point of a smooth influence function is suggested by (Rousseeuw and Croux, 1993). Furthermore, the Q_n estimator is given by the k^{th} order statistic of the $n(n+1)/2$ inter-point distance. For $h=[n/2]+1$ and $k=\binom{h}{2}$,

$Q_n = C\{|x_i - x_j|; i < j\}_{(k)}$. Therefore, Q_n estimator has positive small sample bias (Croux and Rousseeuw, 1992).

For observed spatial data $Z(s_i),\ldots,Z(s_j)$, let $N(h)$ denote the set of pairwise differences $T_i = Z(s_i) - Z(s_i + h), i = 1,\ldots,n(n+1)/2$. The calculation for $Q_{|N(h)|}$ for T_i gives the semivariogram estimator at lag h

$$\bar{\gamma}(h) = \frac{1}{2} Q^2_{|N(h)|}. \tag{6.4}$$

Therefore, Q_n has 50% breakdown point, $\bar{\gamma}(h)$ has a 50% breakdown point interms of the process of differences T_i. Equation (6.4) is resistance to roughly 30% of outliers among the $Z(s_i)$. This is established by (Genton, 2001) through simulation.

Use of quantiles of the distribution of $\{Z(s_i) - Z(s_j)\}^2$ or $|Z(s_i) - Z(s_j)|$ is one of the approaches to estimate the empirical semivariogram. Suppose $[Z(s_i) - Z(s_j)]'$ are bivariate Gaussian with common mean. Then,

$$\frac{1}{2}\{Z(s) - Z(s+h)\}^2 \sim \gamma(h)x_1^2$$

$$\frac{1}{2}|Z(s) - Z(s+h)| \sim \sqrt{\frac{1}{2}\gamma(h)|U|} \qquad U \sim G(0,1)$$

For $q^{(p)}_{|N(h)|}$ denotes the p^{th} quantile,

$$\hat{\gamma}_p(h) = q^{(p)}_{|N(h)|}\left\{\frac{1}{2}[Z(s) - Z(s+h)]^2\right\}.$$

A median-based estimator ($p = 05$) gives

$$\hat{\gamma}_p(h) = \frac{1}{2}\, median_{|N(h)|}\{[Z(s) - Z(s+h)]^2\}/0.455$$

$$= \frac{1}{2}\,(median_{|N(h)|}\{|Z(s) - Z(s+h)|^{1/2}\}^4/0.455$$

(Cressie, 1993, p. 75).

The sample variogram can be estimated using parametric models. Suppose there are samples from a homogeneous spatial process, in which the variogram has been estimated for a sequence of distance h. The empirical semivariogram $\hat{\gamma}(h)$ is an unbiased estimator of $\gamma(h)$, but it provides estimates only at a finite set of lags or lag classes. The properties of the semivariogram estimators $\hat{\gamma}(h)$, $\bar{\gamma}(h)$ and $\tilde{\gamma}(h)$ have been extensively investigated for a single value of h, as a function over all h. But, the estimators fail the condition of non-positive definiteness conditions and these estimators lack a very important property. Therefore, spatial predictions derived from such estimators might have negative variances. To avoid negative variances, the empirical $\gamma(h)$ has to be replaced by some parametric form which is known to be conditionally non-positive definite (Gaetan and Guyon, 2010). Hence, it is important to seek a parametric family which adequately models the observed data. In general, there are three methods to be considered. These methods are: least square estimation, maximum likelihood (ML), restricted maximum likelihood (REML), and Bayesian estimators.

Least square estimation

Suppose the semivariogram $\gamma(h)$ have been estimated at a finite set of values of h, and desire to fit a model specified by the parametric function $\gamma(h;\theta)$ in terms of a finite parameter vector θ. Suppose it is assumed that MoM estimator $\hat{\gamma}$ has been used and let $\hat{\gamma}$ denote the vector of estimates, $\gamma\,(\theta)$ the vector of model values at the same vector of h values. Therefore, there are three well-used version of non-linear least-squared estimators. These estimates are: ordinary least squares (OLS), in which θ can be minimized using $\{\hat{\gamma}-\gamma(\theta)\}'\{\hat{\gamma}-\gamma(\theta)\}$. The second one is generalized least square (GLS), θ can be minimized as $\{\hat{\gamma}-\gamma(\theta)\}'V(\theta)^{-1}\{\hat{\gamma}-\gamma(\theta)\}$ where $V(\theta)$ denotes the covariance matrix of $\hat{\gamma}$. This estimator depends on an unknown θ because the problem is non-linear. The other method is weighted least squares (WLS). Here, θ can be minimized using

$\{\hat{\gamma} - \gamma(\theta)\}'W(\theta)^{-1}\{\hat{\gamma} - \gamma(\theta)\}$ where, $W(\theta)$ is a diagonal matrix whose diagonal entries are the variances of the entries of $\hat{\gamma}$. WLS is used for the variance of $\hat{\gamma}$ but not covariance. Unlike WLS, GLS allows for both variance and covariance.

Furthermore, the three estimators (OLS, GLS, WLS) expected to be in increasing order of efficiency but in decreasing order of convenience to use. Here, OLS is immediately implementable by a nonlinear least square procedure, whereas GLS and WLS require specification of the matrices $V(\theta)$ and $W(\theta)$. For Gaussian process, the following expression is given

$$var[\{Z(s+h) - Z(s)\}^2] = 2\{2\gamma(h)\}^2, \tag{6.5}$$

$$corr[\{Z(s_1+h_1) - Z(s_1)\}^2, \{Z(s_2+h_2) - Z(s_2)\}^2] =$$

$$\frac{\{\gamma(s_1 - s_2 + h) + \gamma(s_1 - s_2 - h_2) - \gamma(s_1 - s_2 + h_1 + h_2) - \gamma(s_1 - s_2)\}^2}{4\gamma(h_1)\gamma(h_2)}. \tag{6.6}$$

This equation can be used to evaluate the matrices $V(\theta)$ and $W(\theta)$. As one of least square estimator, it is possible to use GLS estimator. But, it is complicated to implement this method. Because of this, there is no guarantee for the resulting minimization problem to have unique solution (Schabenberger and Gotway, 2005).

To solve the complication, the approximation of WLS criterion was proposed by (Cressie, 1985). Suppose $\hat{\gamma}$ is evaluated on a finite set $\{h_j\}$ and choose θ to minimize

$$\sum_j |N(h_j)| \left\{\frac{\hat{\gamma}(h)}{\gamma(h_j;\theta)} - 1\right\}^2. \tag{6.7}$$

WLS solution can be derived under the approximation of equation (6.7) and can be given as

$$var\{\hat{\gamma}(h)\} \approx \frac{8\gamma^2(h)}{|N(h)|}. \tag{6.8}$$

Equation 6.8 follows from (6.7). If $Z(s_i) - Z(s_j)$ is individual terms, then it is independent. This assumption is not exactly satisfied. But, it is a reasonable approximation. If the pairs (s_i, s_j) lying in $N(h)$ are widely spread over the sampling space, the assumption of independence can be a reasonable approximation. Therefore, (6.7) is not difficult to implement than OLS. Furthermore, this method expected to be substantially more efficient. In addition to MoM estimator of $\hat{\gamma}$, the robust estimator $\bar{\gamma}$ can be used. Therefore, by derivation of equations (6.5) and (6.6), and assuming normal distribution,

$$\frac{\{Z(s+h) - Z(s)\}^2}{2\gamma(h)} \sim X_1^2.$$

as mean 1 and variance 2. Therefore, $var[\{Z(s+h) - Z(s)\}^2] = 2\{2\gamma(h)\}^2$ follows from this relation. The literature for this approach can be found from (Genton, 2000, Genton et al., 2001 , Zimmerman and Zimmerman, 1991).

Maximum likelihood estimation

For sampling from Gaussian process, the estimation is straightforward principle to express likelihood function and maximize numerically. Estimation of the spatial process using likelihood function was first used by (Kitanidis, 1983, Mardia, 1984). The computation of inverse and determinant of the model covariance matrix are required for the evaluation of the likelihood function. As compared to Cressie's WLS procedure, the sampling properties of the maximum likelihood estimates are clear. Suppose deterministic linear regression terms with no essential change in the methodology are considered. Then the model to be considered is

$$Z \sim N(\boldsymbol{X}\beta, \Sigma)$$

where Z an n-dimensional vector of observations, X is an $n \times q$ matrix of known regressors, β is a q-vector of unknown regression parameters and Σ is the covariance matrix of the observations (Waller and Gotway, 2004). Therefore,

$$\Sigma = \alpha V(\theta)$$

where α is an unknown scale parameter and $V(\theta)$ is a vector of standard covariance determined by the unknown parameter vector θ. Suppose there is exponential variogram structure. The covariance function can be given as

$$Cov\{Z(s_1), Z(s_2)\} = \begin{Bmatrix} c_0 + c_1 & if\ s_1 = s_2 \\ c_1 \exp(-|s_1 - s_2|/R & if\ s_1 \neq s_2 \end{Bmatrix}$$

and define $\alpha = c_1$, $\emptyset = c_0/(c_0 + c_1)$, which is called the nugget ratio, $\theta = (\emptyset, R)$ and let $V(\theta)$ denote the matrix whose diagonal entries are all $1/(1 - \emptyset)$ and off-diagonal entries are of the form $v_{ij} = \exp(-d_{ij}/R)$. Moreover, d_{ij} is the distance between the i^{th} and j^{th} sampling points.

With Z defined by $Z \sim N(\boldsymbol{X}\beta, \Sigma)$, its density is

$$(2\pi)^{-n/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}\ (\boldsymbol{Z} - \boldsymbol{X}\beta)'\Sigma^{-1}(\boldsymbol{Z} - \boldsymbol{X}\beta)\right\}.$$

The negative log likelihood is given by

$$(\beta, \alpha, \theta) = \frac{n}{2}\log(2\pi) + \frac{n}{2}log\alpha + \frac{1}{2}log|V(\theta)| + \frac{1}{2\alpha}(\boldsymbol{Z} - \boldsymbol{X}\beta)'V(\theta)^{-1}(\boldsymbol{Z} - \boldsymbol{X}\beta). \quad (6.9)$$

Therefore, for given V, define $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}$ which is the GLS estimator of β based on covariance matrix $\boldsymbol{V}$, then

$$(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta})'V^{-1}X = 0,$$

and

$$(\boldsymbol{Z} - \boldsymbol{X}\beta)'V^{-1}(\boldsymbol{Z} - \boldsymbol{X}\beta) = \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta} + \boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta\right)'\boldsymbol{V}^{-1}(\boldsymbol{X} - \boldsymbol{X}\hat{\beta} + \boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta)$$

$$= \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta}\right)'V^{-1}\left(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta}\right) + \left(\hat{\beta} - \beta\right)'\boldsymbol{X}'V^{-1}\boldsymbol{X}\left(\hat{\beta} - \beta\right). \ (6.10)$$

This equation confirms that the choice of β minimizes the generalized sum of squares criterion and leads to a sum of squares of generalized residuals denoted by

$$G^2 = \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta}\right)' V^{-1}(\boldsymbol{Z} - \boldsymbol{X}\hat{\beta}). \tag{6.11}$$

For equation (6.9), define $\hat{\beta}(\theta) = (\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V(\theta)^{-1}Z$ and G^2 by $G^2(\theta)$. Using equation (6.11) leads to

$$l\left(\hat{\beta}(\theta), \alpha, \theta\right) = \frac{n}{2}\log(2\pi) + \frac{n}{2}log\ \alpha + \frac{1}{2}log|V(\theta)| + \frac{1}{2\alpha}G^2(\theta). \tag{6.12}$$

Minimizing this equation with respect to α gives

$$\hat{\alpha}(\theta) = \frac{G^2(\theta)}{n}. \tag{6.13}$$

Therefore, relation (6.12) and (6.13) is called a profile negative log likelihood (Kitanidis, 1983, Mardia, 1984).

Restricted maximum likelihood

(Patterson and Thompson, 1971) originally proposed the idea of restricted maximum likelihood (REML) estimation in connection with variance components in linear models. But, the situation proposed by Patterson and Thompson is similar to Gaussian models for spatial data. This idea was pointed out by different authors (Cressie, 1993). Suppose $Y_1, \ldots, Y_n$ are independent univariate random variables, each $N(\mu, \sigma^2)$ with unkown μ and σ^2. The maximum likelihood estimator of μ and σ^2 are $\hat{\mu} = \frac{1}{n}\sum_i Y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_i (Y_i - \bar{Y})^2$. But, this definition of $\hat{\sigma}^2$ is a biased estimator, whereas the more usual unbiased estimator of σ^2 is $\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$. Suppose instead of basing the maximum likelihood estimator on the full joint density of $Y_1, \ldots, Y_n$, it is based on the joint density of vectors of contrasts, i.e., $(Y_1 - \bar{Y},\ Y_2 - \bar{Y},\ \ldots, Y_{n-1} - \ \bar{Y})$. This distribution does not depend on μ. The maximum likelihood estimator of σ^2, turns out to be the unbiased estimator $\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$. Constructing an estimate of σ^2 based on an $(n-1)$ dimensional vector of contrasts gives usual maximum likelihood estimator based on the full n-dimensional data vector (Waller and Gotway, 2004).

To extend this idea, let $W = A'Z$ be a vector of $n-q$ linearly independent contrasts, i.e. the $n-q$ columns of A are linearly independent and $A'X = 0$, then

$$W \sim N(0, \boldsymbol{A}'\Sigma\boldsymbol{A}),$$

and the joint negative log likelihood function based on $\boldsymbol{W}$ is

$$l_w(\alpha,\theta) = \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2} log\alpha + \frac{1}{2} log|\boldsymbol{A}'V(\theta)\boldsymbol{A}| + \frac{1}{2\alpha} W'(\boldsymbol{A}'V(\theta)\boldsymbol{A})^{-1}W.$$

Here, it is possible to choose $\boldsymbol{A}$ to satisfy $\boldsymbol{AA}' = I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $A'A = I$. Further calculations given by (Harville, 1974, Patterson and Thompson, 1971) showed that the above equation can be simplified as follows.

$$l_w(\alpha,\theta) = \frac{n-q}{2}\log(2\pi) + \frac{n=q}{2} log\alpha - \frac{1}{2} log|\boldsymbol{X}'\boldsymbol{X}| + \frac{1}{2} log|\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X}| + \frac{1}{2} log|V(\theta)| + \frac{1}{2\alpha} G^2(\theta),$$

where $G^2(\theta) = n(\hat{\alpha}(\theta))$. To minimize with respect to α, set $\hat{\alpha} = \ G^2(\theta)/(n-q)$ and the above equation is reduced to

$$l_w^*(\theta) = \ l_w(\tilde{\alpha},\theta) = \ \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2} log\frac{G^2(\theta)}{n-q} - \frac{1}{2} log|\boldsymbol{X}'\boldsymbol{X}| + \frac{1}{2} log|\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X}| + \frac{1}{2} log|V(\theta)| + \frac{n-q}{2}.$$

Based on (Harville, 1974, Patterson and Thompson, 1971), $\boldsymbol{A}$ is $n \times (n-q)$ matrix and let G denote the $n \times q$ matrix $V^{-1}X(X'V^{-1}X)^{-1}$, so that $\hat{\beta} = G'Z$. Let $B = [A|G]$, i.e, the $n \times n$ matrix formed by placing the matrices $\boldsymbol{A}$ and G. Then,

$$|B| = |B'B|^{1/2} = \begin{vmatrix} \boldsymbol{A}'\boldsymbol{A} & \boldsymbol{A}'G \\ G'\boldsymbol{A} & G'G \end{vmatrix}^{1/2} = |\boldsymbol{A}'\boldsymbol{A}|^{1/2}\left|G'G - G'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'G\right|^{1/2}.$$

But, $\boldsymbol{A}'\boldsymbol{A} = I$, $\boldsymbol{A}\boldsymbol{A}' = I - \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ and $G'G - G'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'G = (\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Therefore, $|B| = |\boldsymbol{X}'\boldsymbol{X}|^{-1/2}$. It is known that the density of Z is given by

$$f_Z(z) = (2\pi)^{-n/2}\alpha^{-n/2}|V|^{-1/2}exp\left\{-\frac{1}{2\alpha}(\boldsymbol{Z}-\boldsymbol{X}\beta)'V^{-1}(\boldsymbol{Z}-\boldsymbol{X}\beta)\right\}. \tag{6.13}$$

Let $Z^* = B'Z = (\boldsymbol{Z}'\boldsymbol{A}, \boldsymbol{Z}'G)' = \left(W', \hat{\beta}'\right)'$. The Jacobian transformation from Z to Z^* is given by $|B|^{-1} = |X'X|^{1/2}$. Furthermore, equations (6.10) and (6.11) give

$$(\boldsymbol{Z}-\boldsymbol{X}\beta)'V^{-1}(\boldsymbol{Z}-\boldsymbol{X}\beta) = G^2(\theta) + \left(\hat{\beta} - \beta\right)'\boldsymbol{X}'V^{-1}\boldsymbol{X}(\hat{\beta} - \beta).$$

Here, $G^2(\theta)$ is a function of elements orthogonal to $\hat{\beta}$ and a function of W. Therefore, equation 6.14 leads to

$$f_{w,\hat{\beta}}\left(w,\hat{\beta}\right) = |\boldsymbol{X}'\boldsymbol{X}|^{1/2}(2\pi)^{-n/2}\alpha^{-n/2}|V|^{-1/2}exp\left\{-\frac{1}{2}G^2(\theta) - \frac{1}{2\alpha}\left(\hat{\beta}-\beta\right)'\boldsymbol{X}'V^{-1}\boldsymbol{X}(\hat{\beta}-\beta)\right\}.$$

Integrating the above equation with respect to $\hat{\beta}$ gives

$$f_w(w) = |\boldsymbol{X}'\boldsymbol{X}|^{1/2}(2\pi)^{-n/2}\alpha^{-n/2}|V|^{-1/2}|\boldsymbol{X}'V^{-1}\boldsymbol{X}|^{-1/2}exp\left\{-\frac{1}{2}G^2(\theta)\right\}.$$

Bayesian procedure

(Handcock and Stein, 1993, Le and Zidek, 1992) considered using Bayesian procedure to spatial statistics. Models which were defined as $Z \sim N(X\beta, \Sigma)$ and $\Sigma = \alpha V(\theta)$ are considered with the improper prior density by different authors

$$\pi(\beta, \alpha, \theta) \propto \frac{\pi(\theta)}{\alpha}$$

for some prior $\pi(\theta)$. Therefore, the posterior density has the form

$$\pi(\beta, \alpha, \theta|Z) \; \propto \frac{\pi(\theta)}{\alpha}(2\pi)^{-n/2}\alpha^{-n/2}|V(\theta)|^{-1/2}exp\left\{-\frac{1}{2\alpha}(\boldsymbol{Z} - \boldsymbol{X}\beta)'V(\theta)^{-1}(\boldsymbol{Z} - \boldsymbol{X}\beta)\right\}.$$

Define,

$$\hat{\beta}(\theta) = \; (\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V(\theta)^{-1}Z$$

and ignoring constants equation (6.10) gives

$$\pi(\beta, \alpha, \theta|Z) \; \propto \frac{\pi(\theta)}{\alpha}\alpha^{-n/2}|V(\theta)|^{-1/2}exp\left\{-\frac{G^2(\theta)}{2\alpha}\right\}.exp\left\{-\frac{1}{2\alpha}\left(\beta - \hat{\beta}\right)'\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X}(\beta - \hat{\beta})\right\}. \tag{6.14}$$

And integrating this equation with respect to β gives

$$\pi(\beta, \alpha, \theta|Z) \; \propto \frac{\pi(\theta)}{\alpha}\alpha^{-n/2}|V(\theta)|^{-1/2}exp\left\{-\frac{G^2(\theta)}{2\alpha}\right\}.\alpha^{q/2}|\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X}|^{-1/2}$$

and integrating with respect to α gives

$$\pi((\theta|Z) \; \alpha \; \pi(\theta)|V(\theta)|^{-1/2}G^2(\theta)^{-(n-q)/2}|\boldsymbol{X}'V(\theta)^{-1}\boldsymbol{X}|^{-1/2}$$

(Handcock and Stein, 1993).

If $\pi(\theta)$ is ignored in (6.14), the posterior density of θ is precisely the REML estimation. But, a fully Bayesian approach involves not maximizing (6.9), but integrating with respect to the components of θ, and the two methods are different. Integration with respect to θ must be performed numerically.

MINQE estimation

The other method of estimation is the method of minimum norm quadratic estimation (MINQE). This method was originally developed by C.R. Rao (Rao, 1979). When compared to the other estimation methods, MINQE is restricted in scope. Even though this method is restricted in scope, it is competitive with the other methods (Kitanidis, 1983).

Suppose the universal kriging model is given by

$$Z = \boldsymbol{X}\beta + \eta,$$

where the semivariogram of η is $\gamma\ (.\,;\theta)$ is of the form

$$\gamma\ (h;\theta) = \sum_{h=1}^{k} \theta_k \gamma_k(h),$$

where γ is a linear combination of k known semivariograms $\gamma_1,\dots,\gamma_k$, with unknown weights $\theta_1,\dots,\theta_k$. Similar to REML estimation, suppose $W = A'Z$ to be a vector of orthogonal contrasts to $\boldsymbol{X}$, where the columns of X include a constant term, so that the covariance of W is of the form $-A'\Gamma(\theta)A = \psi(\theta)$, where $\Gamma(\theta)$ is the matrix with entries $\gamma\left(s_i - s_j;\theta\right)$, $i,j,\dots,n$ being the sampling points. Let $\psi_1,\dots,\psi_k$ denote the corresponding ψ matrix where $\gamma = \gamma_k$, for each of $k = 1,\dots,k$. The problem is therefore to estimate the coefficient $\{\theta_k\}$ when observed data have the covariance matrix

$$\psi\ (\theta) = \sum_{k=1}^{k} \theta_k \psi_k.$$

Suppose, $p' = (p_1,\dots,p_k)$ is a vector, $p'\theta$ estimated by the quadratic form $Y'HY$. To find the unbiased estimate, $E\{Y'HY\} = E\{tr(HYY')\} = \sum \theta_k tr(H\psi_k)$ is required. Thus, $tr(H\psi_k) = p_k$. In general, the idea behind MINQE is to choose minimum variance unbiased estimator. The variance of $Y'HY$ is of the form $tr(HVHV)$ for some matrix V. In the case of Gaussian process, $var\{Y'HY\} = 2tr(H\psi(\theta)H\psi(\theta))$ (Schabenberger and Gotway, 2005).

But, in practice, $V = \psi(\alpha)$ for some prior guess α of θ. A Lagrange multiplier solution to the resulting contained optimization problem leads us to $A_i = \psi(\alpha)^{-1}\psi_i\psi(\alpha)^{-1}$. The estimator $\hat{\theta}$ will be unbiased estimator of θ if

$$Y'A_iY = tr\left(A_i\psi(\hat{\theta})\right) = \sum_j \hat{\theta}_j tr(A_i\psi_i)$$

for all $i = 1, \ldots, n$. Let B denote the matrix entries $b_{ij} = tr(A_i\psi_i)$, and C with entries c_{ij} denotes the inverse of B, then

$$\hat{\theta}_i = \sum c_{ij}\boldsymbol{Y}'A_j\boldsymbol{Y}.$$

If $\psi(\theta)$ cannot be written as a linear function of θ, then we replace the distribution A_i with

$$A_i = \psi(\alpha)^{-1}\left\{\frac{\partial}{\partial\alpha_i}\psi(\alpha)\right\}\psi(\alpha)^{-1}.$$

But, this method is less motivated compared with general procedures such as maximum likelihood and REML (Stein, 1987).

Measures of Spatial Autocorrelation

There are two types of spatial autocorrelation measures. These are *Moran's I* and *Geary's C*.

Moran's I

This method can be used to tests for global spatial autocorrelation for continuous (Moran, 1950). The test is based on cross-products of the deviations from the mean. The deviation is calculated for n observations on a variable x at locations i,j as:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

where $\bar{x}$ is the mean of the x variable, $\boldsymbol{w_{ij}}$ are the elements of the weight matrix, and $\boldsymbol{S_0}$ is the sum of the elements of the weight matrix: $S_0 = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} w_{\boldsymbol{ij}}$. When compared to correlation coefficient, *Moran's I* has similarity but not equivalent. The values vary from -1 to +1. In the absence of autocorrelation and regardless of the specified weight matrix, the expectation of *Moran's I* statistic is $-1/(n-1)$. This value tends to zero as the sample size increases. For a row-standardized spatial weight matrix, the normalizing factor S_0 equals n, and the statistic simplifies to a ratio of a spatial cross product to a variance. A *Moran's I* coefficient larger than $-1/(\mathrm{n}-1)$ indicates positive spatial autocorrelation, and a *Moran's I* less than $-1/(\mathrm{n}-1)$ indicates negative spatial autocorrelation. Thus, the variance is given by

$$\mathrm{Var(I)} = \frac{\mathrm{n}\{\mathrm{n}^2 - 3\mathrm{n} + 3)\mathrm{S}_1 - \mathrm{nS}_2 + 3\mathrm{S}_0^2\} - \mathrm{k}\{\mathrm{n}(\mathrm{n}-1)\mathrm{S}_1 - 2\mathrm{nS}_2 + 6\mathrm{S}_0^2\}}{(\mathrm{n}-1)(\mathrm{n}-2)(\mathrm{n}-3)\mathrm{S}_0^2} - \frac{1}{(\mathrm{n}-1)^2}$$

where $\mathrm{S}_1 = \frac{1}{2}\sum\sum_{\mathrm{i}\neq\mathrm{j}}(\mathrm{W}_{\mathrm{ij}} + \mathrm{W}_{\mathrm{ij}})^2 = 2\mathrm{S}_0$ for symmetric W containing 0's and 1's

$$\mathrm{S}_2 = \sum_{\mathrm{i}}(\mathrm{W}_{\mathrm{ij}} + \mathrm{W}_{\mathrm{ij}})^2 \text{ where } \mathrm{W}_{\mathrm{i0}} = \sum_{\mathrm{j}} \mathrm{W}_{\mathrm{ij}} \text{ and } \mathrm{W}_{0\mathrm{i}} = \sum_{\mathrm{j}} \mathrm{W}_{\mathrm{ij}}$$

Geary's C

The other measure of Spatial Autocorrelation is Geary's C statistic. This method is proposed by (Geary, 1954). It is based on the deviations in responses of each observation with one another:

$$\mathrm{C} = \frac{\mathrm{n}-1}{2\mathrm{S}_0} \frac{\sum_{\mathrm{i}} \sum_{\mathrm{j}} \mathrm{w}_{\mathrm{ij}}(\mathrm{x}_{\mathrm{i}} - \mathrm{x}_{\mathrm{j}})^2}{\sum_{\mathrm{i}}(\mathrm{x}_{\mathrm{i}} - \overline{\mathrm{X}})^2}.$$

Geary's C ranges from 0 (maximal positive autocorrelation) to a positive value for high negative autocorrelation. Its expectation is 1 in the absence of autocorrelation and regardless of the specified weight matrix (Sokal and Oden, 1978). For *Geary's C* value which is less than 1, it indicates positive spatial autocorrelation. The variance is estimate is given by

$$\mathrm{Var(c)} = \frac{1}{\mathrm{n(n-2)(n-3)S_0^2}}\Big\{\mathrm{S_0^2[(n^2-3)-k(n-1)^2]} + \mathrm{S_1(n-1)[n^2-3n+3-k(n-1)]} + \frac{1}{4}\mathrm{S_2(n-1)[k(n^2-n+2)} - \mathrm{(n^2+3n-6)]}\Big\}$$

where S_0, S_1 and S_2 are defined similarly as in Moran's I.

Comparison of *Moran's I* and *Geary's C* suggested that Moran's I is a more global measurement and sensitive to extreme values of x, whereas Geary's C is more sensitive to differences in small neighbourhoods. In general, Moran's I and Geary's C result in similar conclusions. However, Moran's I is preferred in most cases. Cliff and Ord (1975, 1981) have shown that Moran's I is consistently more powerful than Geary's C.

6.4 Spatial Prediction

Modelling spatial data is not only useful for identifying significant covariates but for producing smooth maps of the outcome by predicting at unsampled locations. Spatial prediction is usually referred to as kriging. Kriging is an optimal interpolation based on regression against observed values of surrounding data points, weighted according to spatial covariance values. Interpolation refers to an estimation of a variable at an unmeasured location from observed values at surrounding locations (Bivand et al., 2008). Kriging has some advantages. These advantages are

- helps to compensate for the effects of data clustering, assigning individual points within a cluster less weight than isolated data points,
- gives an estimate of estimation error (kriging variance), along with estimate of the variable,
- ensures availability of estimation error which provides a basis for stochasticity,
- allows simulation of possible realization.

The spatial prediction which is called kriging can statistically be defined as follows.

Let $\boldsymbol{Y}_0$ be a vector of the binary response at new, unobserved location s_{0i}, $i = 1,\ldots,n_0$. Following the maximum likelihood approach, the distribution of Y_0 is given by

$$P\left(Y_0\middle|\hat{\beta},\widehat{U},\hat{\sigma}^2,\widehat{\emptyset}\right) = \int P\left(Y_0\middle|\hat{\beta},U_0\right)P\left(U_0|\widehat{U},\hat{\sigma}^2,\widehat{\emptyset}\right)dU_0 \qquad (6.15)$$

where $\hat{\beta},\hat{\sigma}^2$ and $\widehat{\emptyset}$ are the maximum likelihood estimates of the corresponding parameters. As part of the iterative estimation process, for penalized quasi-likelihood (PQL), $\widehat{U}$ can be derived (Breslow and Clayton, 1993). $P\left(Y_0\middle|\hat{\beta},U_0\right)$ is the Bernoulli-likelihood at new locations and $P(U_0|\widehat{U},\hat{\sigma}^2,\widehat{\emptyset})$ is the distribution of the spatial random effects U_0 at new sites, given $\widehat{U}$ at observed sites and is assumed to follow the normal distribution that is

$$P\left(Y_0\middle|\widehat{U},\hat{\sigma}^2,\widehat{\emptyset}\right) = N(\varSigma_{01}\varSigma_{11}^{-1}\widehat{U},\varSigma_{00} - \varSigma_{01}\varSigma_{11}^{-1}\varSigma_{10}) \qquad (6.16)$$

with $\varSigma_{11} = E(UU_t)$, $\varSigma_{00} = E(U_oU_0^t)$ and $\varSigma_{01} = \varSigma_{01}^t = E(U_oU_0^t)$. The mean of the Gaussian distribution in (6.16) is the classical kriging estimator (Schabenberger and Gotway, 2005).

The Bayesian predictive distribution of $\boldsymbol{Y}_0$ is given by

$$P(Y_0 \mid Y\,) = \int P(\boldsymbol{Y}_0|\beta, U_0)P(U_0|U, \sigma^2, \emptyset) \times P(\beta, U, \sigma^2, \emptyset|Y)d\beta dU_0 dU d\sigma^2 d\emptyset \qquad (6.17)$$

where $P(\beta, U, \sigma^2, \emptyset|Y)$ is the posterior distribution of the parameters obtained by the Gibbs sampler or the sampling importance re-sampling (SIR) approach. Simulation-based Bayesian spatial prediction is performed by consecutive draws of samples from the posterior distribution, the distribution of the spatial random effects at new locations and the Bernoulli-distributed predicted outcome. The maximum likelihood predictor can be viewed or interpreted as the Bayesian predictor, with parameters fixed at their maximum-likelihood estimates. In contrast to Bayesian kriging, classical kriging does not account for uncertainty in estimation of β and the covariance parameters.

The data was analyzed by fitting a generalized linear mixed model (GLMM) using SAS 9.2 PROC GLIMMIX.

6.5 Data analysis using spatial statistics approach

Using the identified thirteen main effects and six two-way and three-way interaction effects (Ayele et al., 2012) several covariance structures including SP(EXP) (Exponential), SP(EXPA) (Anisotropic Exponential), SP(EXPGA) ((2D Exponential, Geometrically Anisotropic), SP(GAU) (Gaussian), SP(GAUGA) ((2D Gaussian), (Geometrically Anisotropic), SP(LIN) (Linear), SP(LINL) (Linear Log), SP(MATERN) (Matérn), SP(MATHSW) (Matérn (Handcock-Stein-yene maWallis)), SP(POW) (Power), SP(POWA) (Anisotropic Power), SP(SPH) (Spherical) and SP(SPHGA)((2D Spherical, Geometrically Anisotropic) were fitted but SP(GAU) (Gaussian) was found to be the best spatial covariance structure for the model (Kincaid, 2012).

The plots presented in Figure 6.2 are a spatial scatter plot of the observed data. The scatter plot suggests distribution which is not indicative of a uniformly spread of the RDT measurements throughout the prediction area. No direct inference can be made about the existence of a surface trend in the data. However, the apparent stratification of RDT values might indicate a non-random trend. The Spatial Autocorrelation is an inferential statistic tool, which is important to test for randomness. This means that the results of the analysis are always interpreted within the context of its null hypothesis of a random occurrence of events. For the randomness test Moran's and Geary's C tests can be used (Cliff and Ord, 1975, Cliff and Ord, 1981, Geary, 1954, Moran, 1950, Sokal and Oden, 1978).

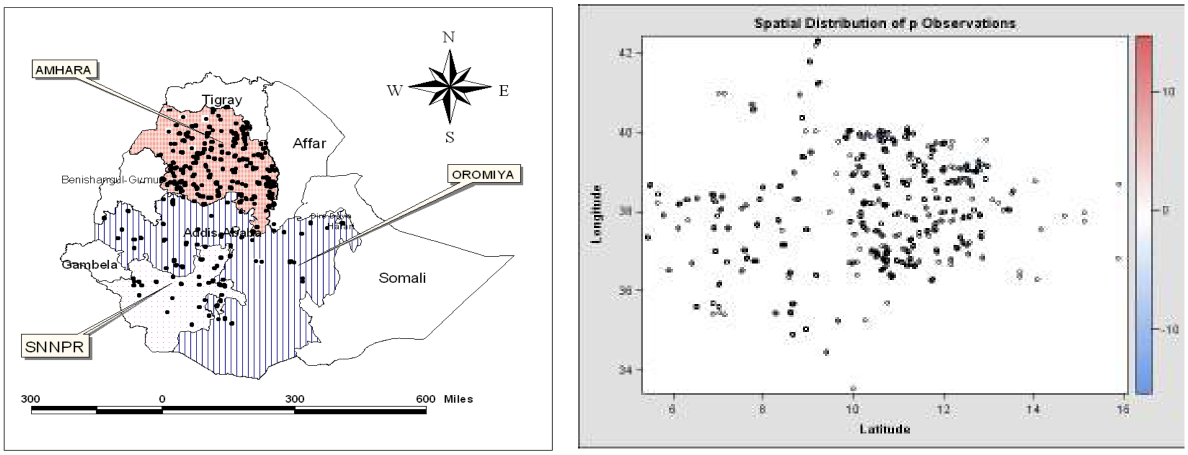


Figure 6. 2: Scatter plot for the malaria prevalence

For these tests, the null hypothesis states that the spatial distribution of feature values is the result of random spatial processes. The result from *Moran's* (Z value = -40.4 and p – value <.0001) and *Geary's c* (Z value = -11.2 and P-value <.0001) tests indicate that the spatial distribution of feature values is not the result of random spatial processes. The Z values are negative for both *Moran's* and *Geary's C* tests. This indicates that the spatial distribution of high values and low values in the dataset is more spatially dispersed than would be expected if underlying spatial processes were random. A dispersed spatial pattern often reflects some type of competitive process, i.e., a feature with a high value repels other features with high values; similarly, a feature with a low value repels other features with low values. The observed spatial pattern of feature values could not very well be one of many possible versions of complete spatial randomness.

Figure 6.3 represents different semivariogram estimators using classical and robust estimators. The classical estimator was suggested by (Matheron, 1963). The classical estimator can be calculated by

$$\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}(Z(s_i) - Z(s_j))^2,$$

where (s_i) is the anscombe residual.

$$N(h) = \{(s_i - s_j) : \|s_i - s_j\| = h \pm \in\} \text{ and } |N(h)|$$

is its cardinality. But, the classical estimator is sensitive to outliers. For this reason a robust estimator was proposed by (Cressie and Hawkins, 1980). Among the different types of isotropic covariograms given above, Gaussian type was selected. Thus as discussed earlier, the best spatial covariance structure from all possible types was found to be the SP(GAU) (Gaussian) covariance structure. Therefore, the Gaussian type of the variogram was used to perform variogram analysis. The figure (Figure 6.3) shows first a slow, then a rapid rise from the origin. Therefore, the shape of the graph suggests a Gaussian type form which is given by

$$\gamma(t) = c_0 + c_1\left[1 - \exp\left(-\frac{t^2}{R^2}\right)\right].$$

In general, from Figure 6.3, it is possible to distinguish three main features. The first one is the Y-axis well above zero, indicating the possible presence of a nugget effect. Moreover, the shapes of the semivariogram up through distances in the low 40s have roughly the shape of a spherical covariance model. Besides these, the semivariogram values are extremely high for the largest distances.

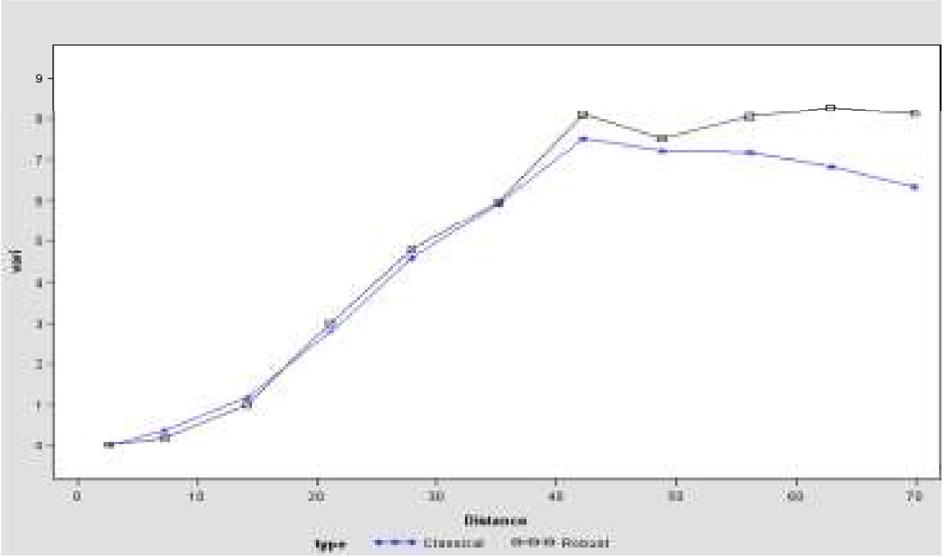


Figure 6. 3: Classical and robust semivariogram for malaria prevalence

Table 6.1 present the significant effects for the model which incorporate spatial variability using SP (GAU) (Gaussian) covariance structure. Among all significant effects namely family size, altitude, toilet facilities, availability of radio, number of rooms per person, main material of the room's wall, use of indoor residual spray, use of mosquito nets and number of nets per person, were not involved in the interaction effects. The significant two-way and three-way interaction effects found to be time to collect water and main material of the room's floor; age and gender; gender and availability of electricity; gender and main material of the room's floor; age, gender and main source of drinking water; and age, gender and availability of electricity. Based on these results for a unit increase in family size, the odds of positive rapid diagnosis test increases by 2.34% (OR = 1.0234, P-value < 0.0001). Furthermore, for a unit increase in altitude, the odds of positive rapid diagnosis test decreases by 1.4% (OR = 0.996, P - value <0.0001) (Table 6.2).

Table 6. 1: Type 3 analysis of effects for the GLMM with spatial correlation

Effect	Num DF	F Value	Pr > F
Age	1	0.59	0.9876
Gender	1	0.5906	0.9882
Family size	1	19.85	<.0001
Region	2	0.74	0.476
Altitude	1	14.25	<.0001
Main source of drinking water	2	29.25	<.0001
Toilet facility	3	37.15	<.0001
Time to collect water	2	16.29	<.0001
Availability of electricity	1	0.01	0.9185
Availability of radio	1	14.36	<.0001
Availability of television	1	2.67	0.1023
Total number of rooms	1	20.88	<.0001
Main material of the room's wall	2	49.01	<.0001
Main material of the room's roof	2	45.3	<.0001
Main material of the room's floor	2	27.36	<.0001
Use indoor residual spray	1	585.68	<.0001
Use of mosquito nets	1	22.14	<.0001
Total number of nets	1	22.36	<.0001
Main source of drinking water and main material of the room's roof	2	28.36	<.0001
Time to collect water and main material of the room's floor	2	27.36	<.0001
Age and gender	1	.0.691	0.9897
Gender and main source of water	2	12.43	<.0001
Gender and main Material of the room's floor	1	10.85	0.001
Gender and availability of electricity	2	0.08	0.9189
Age, gender and main source of drinking water	4	23.88	<.0001
Age, gender and Availability of electricity	2	24.11	<.0001
Age, gender and main material of the room's floor	2	0.65	0.5202

Table 6. 2: Socio-economic, demographic and geographic of effects on malaria RDT test for main effects

Effect	Estimate	OR	SE	P -value
Intercept	-0.2460	0.7819	5.8100	0.9995
Age	0.0209	1.0212	0.0503	0.6772
Gender (ref. male)				
Female	-2.5463	0.0784	3.0804	0.4084
Family size	0.02311	1.0234	0.0527	<.0001
Region (ref. SNNP)				
Amhara	-0.6896	0.5018	0.4502	0.1256
Oromiya	-0.837	0.4330	0.5796	0.1487
Altitude	-0.0037	0.9963	0.0001	<.0001
Main source of drinking water (ref. protected water)				
Tap water	-0.5557	0.5737	0.722	<.0001
Unprotected water	0.6372	1.8912	0.6871	0.005
Time to collect water (ref. > 90 minutes)				
< 30 minutes	-0.7829	0.4571	0.252	0.0019
between 30 to 40 minutes	-0.603	0.5472	1.2666	0.6341
between 40 - 90 minutes	-4.0189	0.0180	2.8957	0.1652
Toilet facility (Ref. No facility)				
Pit latrine	-0.4403	0.6438	0.6433	<.0001
Toilet with flush	-0.9177	0.3994	0.6413	<.0001
Availability of electricity (ref. no)				
Yes	-3.1219	0.0441	1.0961	0.0044
Availability of television (ref. no)				
Yes	0.6991	2.0119	0.2121	0.001
Availability of radio (ref. no)				
Yes	-0.6991	0.4970	0.2121	0.001
Number of rooms/person	-0.4631	0.6293	0.0688	<.0001
Main material of room's wall (ref. cement block)				
Mud block/wood	-4.1691	0.0155	1.2646	0.038
Corrugated metal	-3.1196	0.0442	1.2576	0.004
Main material of room's roof (ref. corrugate)				
Thatch	1.5031	4.4956	1.6732	0.005
Stick and mud	0.454	1.5746	0.6726	0.0058
Main material of room's floor (ref. earth/Local dung plaster)				
Wood	-1.1407	0.3196	0.803	0.004
Cement	-0.9273	0.3956	0.114	0.028
Use of indoor residual spray (Ref. Yes)				
No	1.237	3.4453	0.1734	<.0001
Use of mosquito nets (ref. no)				
Yes	-0.8741	0.4172	0.1541	<.0001
Number of months room sprayed	-0.7626	0.4665	0.1274	<.0001
Number of nets/person	-0.9349	0.3926	0.0977	<.0001

With reference to individuals with no toilet facilities, the odds of a positive malaria rapid diagnosis test is lower for those individuals using a flushing toilet to those who have septic tanks (OR = 0.399, P - value <0.0001) or pit latrine slabs (OR = 0.644, P - value <0.0001) compared to individuals who have no toilet facilities. Moreover, for a unit increase in the number of total rooms, the odds of malaria diagnosis test for an individual decreased by 37.07% (OR = 0.629, P - value <0.0001). Similarly, with a unit increase in the number of nets in the house, the odds of rapid diagnosis test of malaria for individuals decreased by 60.7% (OR = 0.393, P - value <0.0001). Furthermore, for a unit increase in the number of rooms in the household sprayed with indoor residual spray, the odds of a positive malaria diagnosis test decreased by 53.3% (OR = 0.467, P - value <0.0001).

Table 6. 3: Socio-economic, demographic and geographic of effects on malaria RDT test for interaction effects

Effect	Estimate	OR	SE	P-value
Gender and main source of drinking water (ref. Male & protected water)				
Female and Tap water	-2.747	0.064	0.861	0.001
Female and Unprotected water	1.224	3.402	1.064	0.250
Gender and material of room's floor (ref. Male and earth/Local dung plaster)				
Female and Cement	-0.839	0.432	0.571	<.0001
Female and Wood	0.762	2.143	0.387	<.0001
Age, gender and main source of drinking water (ref. Male & protected water)				
Female and Tap water	-0.045	0.956	0.000	<.0001
Female and Unprotected water	0.042	1.043	0.000	<.0001
Age, gender and availability of electricity (ref. Male & yes)				
Female and No	0.066	1.068	0.000	<.0001

Interaction effects

Figures 6.4 and 6.5 show the distribution of malaria rapid diagnosis test against age, main source of drinking water for both males and females respectively. As age increased, positive malaria diagnosis was less likely for males than females who were using protected, unprotected and tap water for drinking. Furthermore, as age of respondents increased, malaria rapid diagnosis test was less likely to be positive for individuals who use tap water for drinking for males and for females. More specifically, positive malaria diagnosis rate increases with age for females whereas it decreases as age increases for males (Figures 6.4 and 6.5). The figures further show that the gap in the rapid diagnosis test between respondents with unprotected, protected and tap water widens with increasing age.

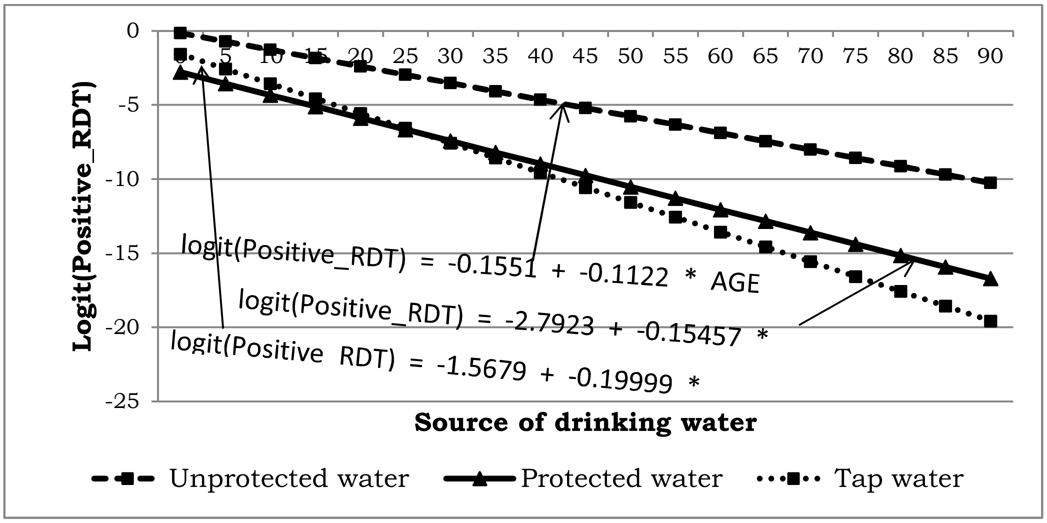


Figure 6. 4: Log odds associated with rapid diagnosis test and age for male respondents with source of drinking water

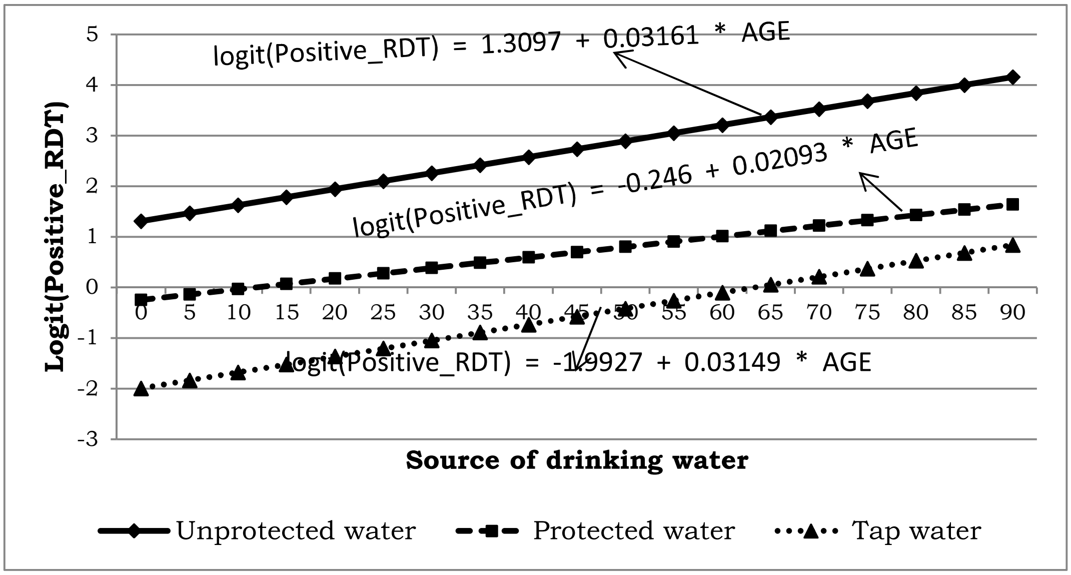


Figure 6. 5: Log odds associated with rapid diagnosis test and age for female respondents with source of drinking water

The relationship between age, gender and availability of electricity is presented in Figure 6.6. As the figure indicates, positive malaria rapid diagnosis test decreases as age increases for both male and female respondents, whether or not they have access to electricity, except for females who responded to having electricity. However, the rate of decrease was not the same for males and females after controlling for other covariates in the model.

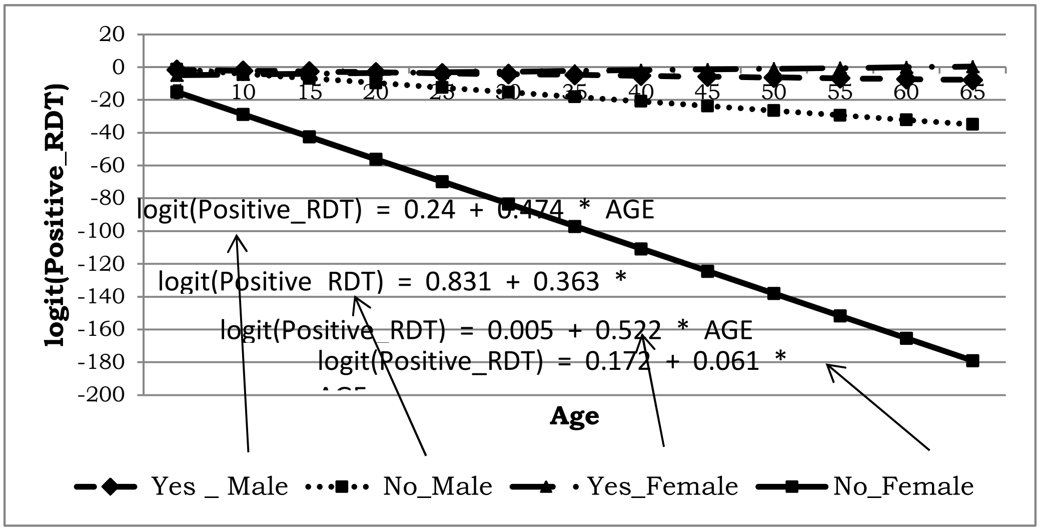


Figure 6. 6: Log odds associated with rapid diagnosis test with age for female and female respondents with availability of electricity

Interaction effects between the main source of water and the main material used for the room's roof is presented in Figure 6.7. From the figure, it is clearly seen that positive rapid diagnosis of malaria was significantly higher for households with a stick and mud roof followed by thatch and lastly a corrugated iron roof. This occurred with respondents who reported using tap water as well as protected and unprotected water for drinking (Figure 6.7)). Furthermore, there was a significant difference in rapid diagnosis test between tap, protected and unprotected sources of drinking water for those who reported having thatch and stick and mud roofs. It is also shown that for corrugated iron roofs, the positive rapid diagnosis test was significantly lower for respondents who reported using tap water for drinking than for those who used protected and unprotected water for drinking.

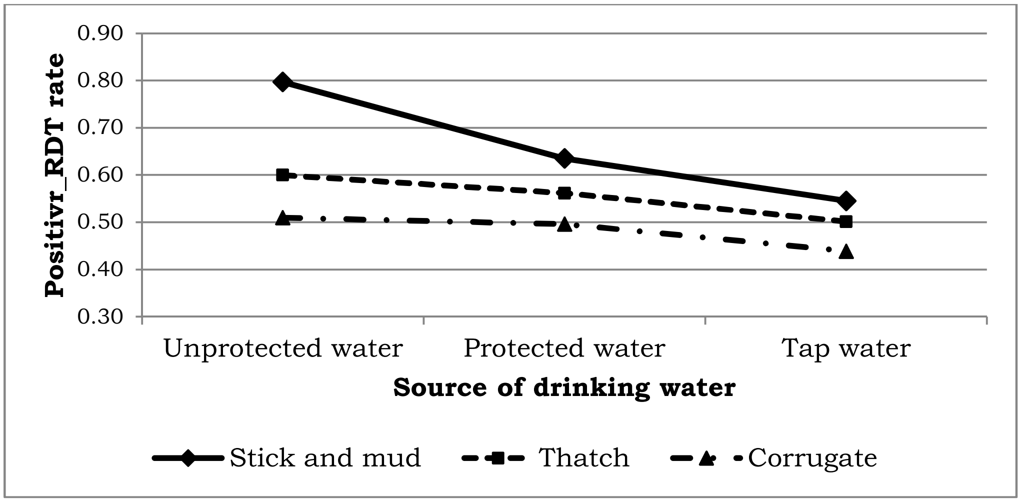


Figure 6. 7: Log odds associated with rapid diagnosis test and source of drinking water with material of the room's roof

The other significant two-way interaction effect was between the time taken to collect water and the main flooring material (Table 6.1). This result is presented graphically in Figure 6.8. A positive rapid diagnosis test was significantly higher in those rooms with earth and local dung plaster floors than for those with cement and wooden floors, for respondents who took < 30 minutes and >90 minutes to collect water. But, for respondents who took less than 30 minutes to collect water but had a cement floor, the positive rapid diagnosis was low. Furthermore, with respondents who took between 30 to 40 minutes to collect water, there was a lower positive rapid diagnosis test for those with earth and local dung plaster floors compared to wooden floors.

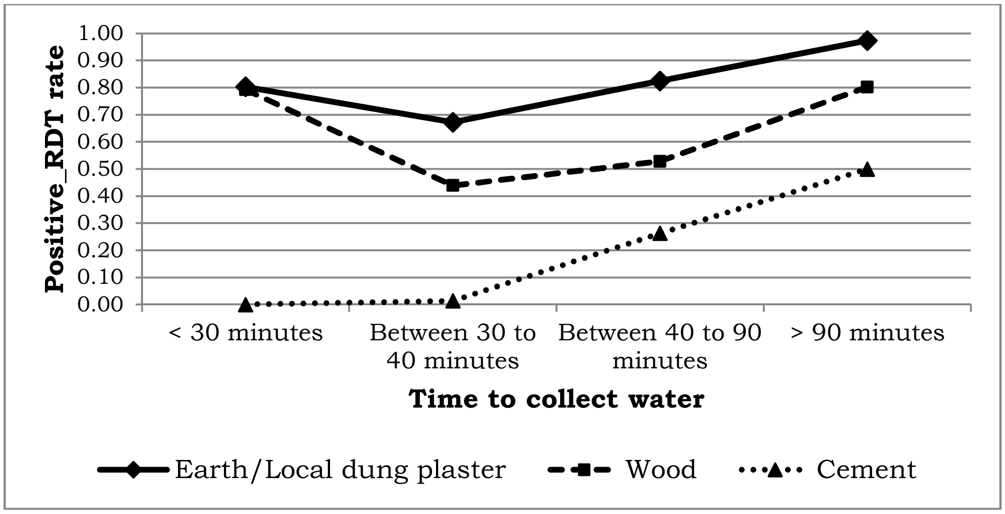


Figure 6. 8: Log odds associated with rapid diagnosis test and time to collect water with material of room's floor

The relationship between the main source of drinking water and gender is presented in Figure 6.9. As the figure indicates, a positive rapid diagnosis test was significantly higher for female respondents than for male respondents who reported using unprotected water. There was however, no significant difference in a positive rapid diagnosis test between females and males who reported using protected and tap water for drinking.

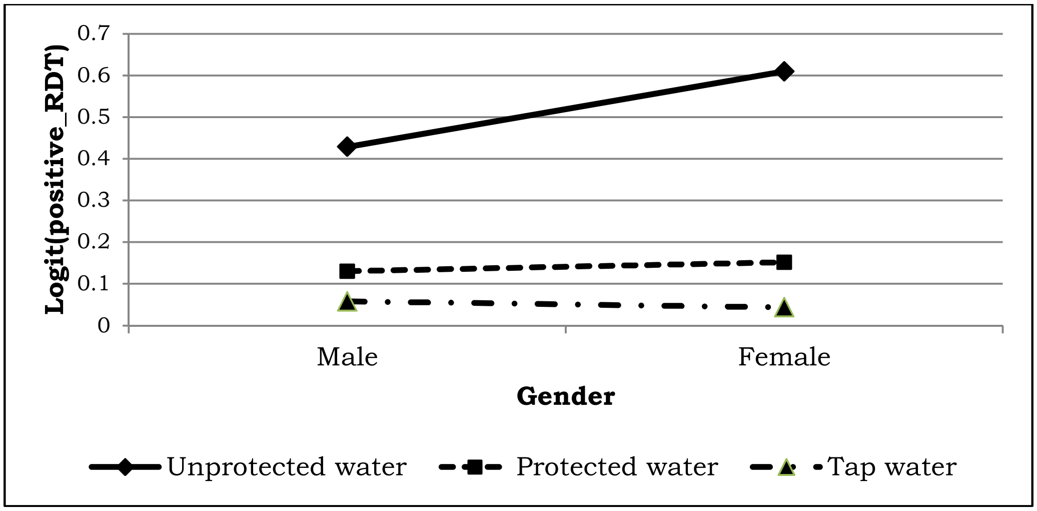


Figure 6. 9: Log odds associated with rapid diagnosis test and main source of drinking water with gender

Besides the fixed effects, Table 6.4 gives estimated spatial covariance parameters. An estimate of the variation between *kebeles* is $\boldsymbol{\sigma}^2 = \mathbf{1.0506}$. The estimate of the range $\boldsymbol{\rho}$ was estimated using SP(GAU) spatial structure and its estimate is 1.3805. The estimate of the sill $\boldsymbol{\sigma}^2$ is reported as "Residual." The estimate of the sill $\boldsymbol{\sigma}^2$ is 1.0506. Therefore, for the Gaussian model, the variance parameters, which is estimated by 4.5446, is called the partial sill. The sill is the sum of the partial sill and the nugget. In general, based on the result, it is observed that there is variability between *kebele*s.

Table 6. 4: Random effects estimates

Effect	Estimate	SE	Pr > Z
Variance	4.5446	0.5866	<.0001
SP(GAU)	1.3805	0.2165	0.0178
Residual	1.0506	0.0307	<.0001

The spatial model which is described above was used to produce a map of predicted prevalence of positive diagnosis of malaria RDT incidence rates for Amhara, Oromiya and SNNP regions of Ethiopia. When there is spatial data, the basic concern is the potential for spatial correlation in the observations. These spatial correlations could lead to incorrect estimates (estimates with underestimated standard errors). Spatial clustering of disease is almost to be expected since human populations generally live in spatial clusters rather than in a random distribution of space. An infectious disease that is highly associated with socio-economic, demographic and geographic factors is likely to be spatially clustered. This spatial clustering can occur even if the population distribution is not clustered. The model derived in this study explains some of the spatial patterns of the prevalence of malaria. The predicted prevalence of malaria is given in Figures 6.10 and 6.11. The prediction map (Figures 6.10 and 6.11) shows that the socio-economic, geographic and demographic factors are closely associated with the risk of malaria, mostly in the SNNP region followed by the Amhara and Oromiya regions. As can be seen from the map, the risk of transmission of malaria is of a moderately high in intensity in almost all parts of the SNNP region. But, for the Oromiya region, the majority of households experience a lesser prevalence of malaria. Furthermore, from the map it can be seen that there is a high predicted value for the prevalence of malaria around the borders. This could be caused by cross-border migration of infected persons and the proximity of uncontrolled areas across the border which may further add to the intensity of transmission in border areas.

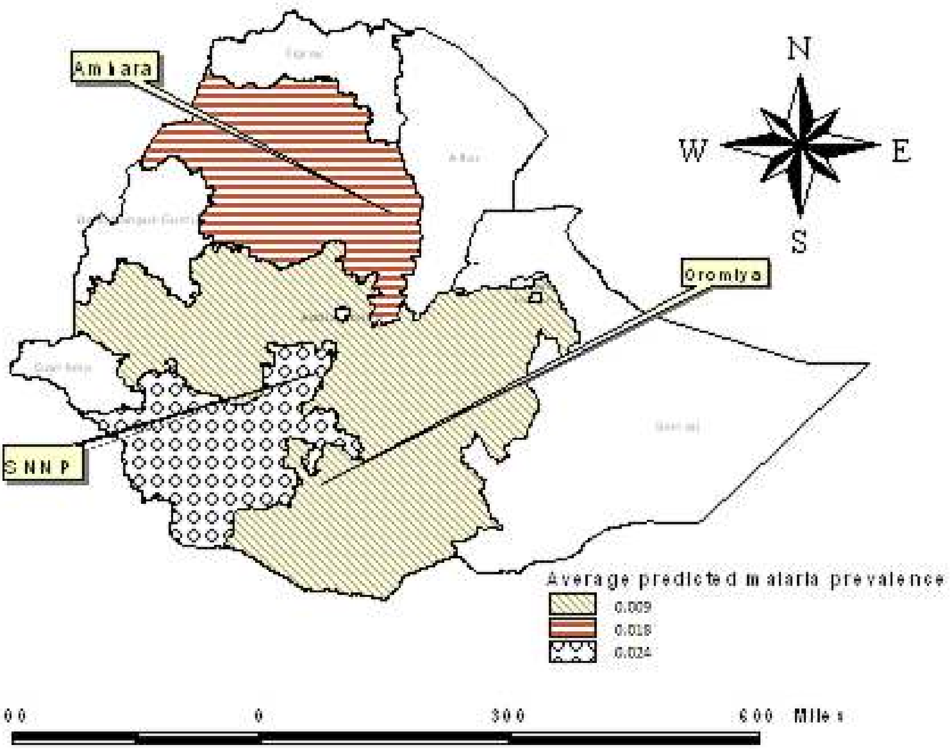


Figure 6. 10: Predicted average spatial effects from the malaria prevalence model

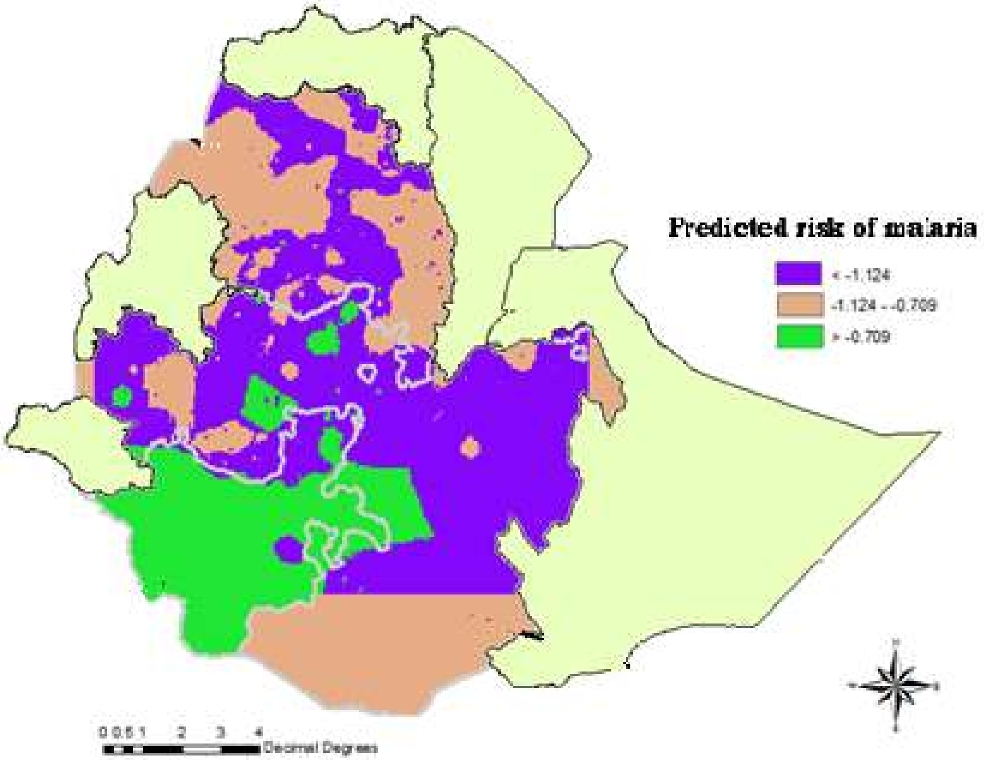


Figure 6. 11: Predicted spatial effects from the malaria prevalence model

6.6 Summary and discussion

Looking at the global distribution of malaria in the world suggested that the concentration of the disease is in the world's poorest continents and countries. Accurate information on the distribution of malaria in epidemic-prone areas on the ground permits interventions to be targeted towards the transmission and high risk locations and households. Such targeting greatly increases the effectiveness of control measures but the inadvertent exclusion of these locations causes potentially effective control measures to fail. The computerized mapping and management of location data assists the targeting of interventions against malaria at the focal and household levels, leading to improved efficacy and cost-effectiveness of control.

As the distribution of malaria infection suggests, it is important to understand the relationship between malaria and poverty. This relationship is important to enable the design of coherent and effective policies and tools to tackle the problem (Hay et al., 2004, Mendis et al., 2009). As is already known from the previous chapters, poverty is related to socio-economic factors. Therefore, it is important to identify those factors which are also related to the risk of malaria. Based on these facts, the findings from the current study show that the following socio-economic factors are related to the risk of malaria: main source of drinking water, time taken to collect water, toilet facilities, availability of radio, total number of rooms per person, main material of the room’s walls, main material of the room's roof, main material of the room's floor, use indoor residual spray, use of mosquito nets, total number of persons per net. Besides socio-economic factors, there are demographic and geographic factors which also have an effect on the risk of malaria. These include gender, age and family size. In addition to the main effects there were interactional effects between the socio-economic, demographic and geographic factors which also influenced the risk of malaria. Most notable of these were the interaction between main source of drinking water and main material of the room's roof, time taken to collect water and main material of the room's floor, age and gender, gender and availability of electricity, gender and main material of the room's floor, age, gender and main source of drinking water; and age, gender and availability of electricity.

Spatially correlated data cannot be regarded as independent observations. Therefore, ignoring the spatial variability might lead to an inaccurate estimation of parameters. Accordingly, we considered the spatial correlation structure and the significance of the variables were checked and predictions of the malaria risk levels for the sampled areas were produced. A useful way of providing up to date information is in the use of GIS-based management systems. This method helps to address effective malaria vector control and management. Therefore, the spatial distribution of malaria incidence was one of the points which were important for such GIS studies.

Spatial clustering of malaria is almost predictable as human populations generally live in spatial clusters rather than in random distributions of space. Disease which is highly correlated to socio-economic variables is likely to be spatially clustered. Therefore, the model explains some of the spatial patterns of malaria risk for Amhara, Oromiya and SNNP regions of Ethiopia. Moran's and Geary's C tests were used to test for randomness (Cliff and Ord, 1975, Cliff and Ord, 1981, Geary, 1954, Moran, 1950, Sokal and Oden, 1978). The interest was to test if the spatial distribution of feature values is the result of random spatial processes. However, the test favors that the spatial distribution of feature values is not the result of random spatial processes. Moreover, the spatial distribution of high values and low values in the dataset is more spatially dispersed than would be expected. A dispersed spatial pattern often reflects some type of competitive process, i.e. a feature with a high value repels other features with high values; similarly, a feature with a low value repels other features with low values.

The results of this study provide evidence on the spatial distribution of socio-economic, demographic and geographic risk factors in the occurrence of malaria. This forms the basis for this research. Therefore, the utilization of socio-economic, demographic and geographic data on malaria rapid diagnosis test, including the information on the spatial variability, clarifies the effects of these factors. From the study it was observed that residents living in the SNNP region were found to be more at risk of malaria than those living in Amhara and Oromiya regions. Similarly, houses which were treated with indoor residual spraying were less likely to be affected by malaria. However, a major challenge in the control of malarial infection was found to be in the type of toilet facilities available in the household. From the results, it was observed that individuals living in households which had no toilet facilities were more likely to be positive for malaria diagnosis tests. Furthermore, positive malaria diagnosis rates decreased with age and the risk of malaria increased per unit increase in family size. Generally, malaria parasite prevalence differed between age and gender, with the highest prevalence occurring in children and females.

From the findings of this study, it can be suggested that having toilet facilities, access to clean drinking water and the use of electricity offers a greater chance of knowing whether or not an individual in the household is at risk of malaria or not. In addition to this, using mosquito nets and spraying anti-mosquito treatment on the walls of the house were also found to be a way of reducing the risk of malaria. Similarly, having a cement floor and corrugated iron roof was found to be one means of reducing the risk of malaria. Based on the findings, different types of housing materials have an influence on the risk of malarial transmission with those houses constructed of poor quality materials having an increased risk. Moreover, the presence of particular structural features, such as bricks, that may limit contact with the mosquito vector, also helps to reduce infection. The risk of malaria therefore, is higher for households in a lower socio-economic bracket than for others who may enjoy a higher status and who are able to afford to take measures to reduce the risk of transmission. Therefore, with the correct use of mosquito nets, indoor residual spraying and other preventative measures, like having more rooms in a house, the incidence of malaria could be decreased. In addition to this, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria. To provide clean drinking water, proper hygiene and maintaining the good condition of a house is essential in controlling the transmission of malaria. With other control measures, including creating awareness about the use of mosquito nets, indoor residual spraying and malaria transmission, the number of malaria cases can be reduced. Furthermore, spatial statistics studies significantly contribute to the understanding of the distribution of malarial infections. The use of spatial statistics analysis is effective in monitoring and identifying high-rate malaria affected regions and helpful when implementing preventative measures. Finally, studies incorporating spatial variability are necessary for devising the most appropriate methodology for remedial action to reduce the risk of malaria.

It is important statistically to jointly model variables which ideally may be dependent rather than as independent. Therefore, in the next chapter, joint model between malaria RDT result, use of mosquito nets and use of indoor residual spraying in the last twelve months will be investigated.

Chapter 7
Modeling of the joint determinants of malaria Rapid Diagnosis Test result, use of mosquito nets and use of indoor residual spray

7.1 Introduction

In previous Chapters, the factors affecting malaria RDT result was explored. To assess malaria RDT result and the associated socio-economic, demographic and geographic factors, different statistical methods were used in the previous chapters (Ayele et al., 2012, Ayele et al., 2013a, Ayele et al., 2013b). But, in some studies, the interest could be with multiple outcomes. The different outcomes may be similar or different types (Verbeke et al., 2012). Therefore, the association between a primary outcome and another related outcome can disclose a great deal of understanding about the mechanism of changes to reduce risk of malaria. For this study, related outcomes for malaria RDT result are use of mosquito nets and use of indoor residual spray in the last twelve months which has been related to each individual. The aim of this study is to further investigate the joint effect of these predictor variables on malaria RDT result, use of mosquito nets and use of indoor residual spray for the last twelve months. Furthermore, the desire is to assess whether the explanatory variables that were found to be significantly related to malaria RDT result were still significant even when use of mosquito nets and use of indoor residual spray in the last twelve months were accounted for in the model. In addition to this, the association between these outcomes is also of interest. Therefore, the current problem can be addressed within the frame work of joint modelling of binary outcomes. Joint model has advantages over separate fitting of models. These advantages include that the joint models better control over type I error rates in multiple test, efficiency in estimating parameters and answer multivariate questions.

There are difficulties in answering the question for assessing the relationship between some covariates and all outcomes simultaneously. For such case of multiple outcomes, two types of correlations must be taken into account, i.e., the objectives of a multivariate analysis of binary data should include (1) the description of the dependency of each binary response on some covariates and (2) the characterization of the degree of association between pairs of responses and the dependence of this association on covariates (Verbeke and Davidian, 2008, Verbeke et al., 2012). Joint models are extensively used for many studies. The literature related to joint modelling is vast (Guo and Carlin, 2004, Tsiatis and Davidian, 2004, Verbeke and Davidian, 2008, Verbeke et al., 2012). On the other hand, methods focusing on models that jointly analyse discrete and continuous outcomes have been explored (Aerts et al., 2002, Faes et al., 2004, Faes et al., 2008, Molenberghs and Verbeke, 2005). The difficulties of joint modelling arise in the lack of multivariate distributions for combining both types of outcomes. Therefore, because the specification of a joint distribution of the response is not straight forward, there are two approaches adapted to joint modelling. The first approach is based on a conditioning argument that allows joint distribution to be factored in a marginal component and a conditional component, i.e., avoiding direct specification of a joint distribution (Catalano and Ryan, 1992, Faes et al., 2004, Fava Del et al., 2011). But this method has disadvantage, i.e., do not directly lead to marginal inference. Furthermore, the correlation among the two outcomes cannot be directly estimated (Verbeke and Davidian, 2008, Verbeke et al., 2012). Formulating a joint model for both outcomes directly is the second approach. For the second approach, Plackett-Dale approach has been used. This method assumed Plackett latent variable to model bivariate outcomes (Molenberghs and Verbeke, 2005).

Therefore, the main objective of joint modelling is to provide a framework within the interest of systematic relationships among the multiple outcomes and between them and other factors. To obtain valid inferences, joint models must account for the correction among the outcomes and other effects of different factors (Fitzmaurice et al., 2008). The joint generalized linear mixed model assumes GLMM for each outcome. The univariate models are combined through specification of a joint multivariate distribution for all random effects. Therefore, joint model can be considered as a new GLMM. Furthermore, the mixed model can be used by specification of the marginal distribution, conditional on correlated random effect. The generalized linear mixed model forms a very general class of models in the exponential family. Furthermore, the aim of this chapter is to review the extension of GLMM approach for multivariate data by assuming separate random effects and then combining the outcomes by imposing a joint multivariate distribution on the random effect.

This chapter, therefore, is organized as follows. In Section 7.2, a multivariate generalized linear mixed model for two outcomes is presented. The formulation of a joint model is binary outcomes in Section 7.3. Section 7.4 presents the results of a joint model between malaria RDT result, use of mosquito nets and use of indoor residual spray in the last twelve months. Section 7.5 presents summary and discussion.

7.2 Joint model formulation for multivariate GLMM

The primary objective of the joint modelling is to provide a framework where questions of scientific interest pertaining to relationships among and between multiple outcomes and other factors. Therefore, generalized linear mixed model introduced in previous studies can be easily be adapted to situations where various outcomes of a different nature are observed (Molenberghs and Verbeke, 2005). Consider a conditional random effects model with bivariate responses. Let the two outcomes be y_{i1} and y_{i2} and denoted by $\mathbf{y}_i = (\mathbf{y}_{i1}', \mathbf{y}_{i2}')'$, where $\mathbf{y}_{i1} = (y_{i11}, y_{i12}, \ldots, y_{i1n_{i1}})'$ and $\mathbf{y}_{i2} = (y_{i21}, y_{i22}, \ldots, y_{i2n_{i2}})'$ on the first and second outcome. Here, $y_{i1j}, j = 1, \ldots, n_{i1}$ and $y_{i2j}, j = 1, \ldots, n_{i2}$ are conditionally independent given b_{i1} and b_{i2} with densities $f_1(.)$ and $f_2(.)$ in the exponential family for the first and second outcomes. Furthermore, y_{i1} and y_{i2} are conditionally independent given $b_i = (b_{i1}, b_{i2})'$ and the responses are independent. In addition to this, $g_1(.)$ and $g_2(.)$ be appropriate link functions for f_1 and f_2. Moreover, the conditional means of y_{i1j} and y_{i2j} denoted by μ_{i1j} and μ_{i2j} respectively. Suppose $\boldsymbol{\mu}_{i1} = (\mu_{i1j}, \ldots, \mu_{i1n_i})'$ and $\boldsymbol{\mu}_{i2} = (\mu_{i2j}, \ldots, \mu_{i2n_i})'$ (Gueorguieva and Agresti, 2011). Therefore, at the first stage the mixed model specification is assumed to be

$$\mu_{i1} = g_1^{-1}(X_{i1}\beta_1 + Z_{i1}b_{i1}) \tag{7.1}$$

$$\mu_{i2} = g_2^{-1}(X_{i2}\beta_2 + Z_{i2}b_{i2}) \tag{7.2}$$

where β_1 and β_2 are $(p_1 \times 1)$ and $p_2 \times 1)$ dimensional unknown parameter vectors, X_{i1} and X_{i2} are $(n_{i1} \times p_1)$ and $(n_{i2} \times p_2)$ dimensional design matrices for the fixed effects, Z_{i1} and Z_{i2} are $(n_{i1} \times q_1)$ and $(n_{i2} \times q_2)$ design matrices for the random effects and g_1 and g_2 are applied component wise to μ_{i1} and μ_{i2} (Gueorguieva, 2001). Secondly,

$$b_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim \text{i.i.d}\, \text{MVN}\,(0, \Sigma) = \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right), \tag{7.3}$$

where Σ_{12}, $\Sigma_{11}$and $\Sigma_{22}$are unknown positive definite matrices. For a given value of $\Sigma_{12} = 0$, the above model is equivalent to two separate GLMM's for two outcome variables. This leads to the assumption of complete independence for both outcomes. Advantages of joint model include the control of the type I error rates in multiple tests. This leads to possible gains in efficiency in the parameter estimates and the ability to answer intrinsically multivariate questions (Molenberghs and Verbeke, 2005).

The marginal means and the marginal variances of y_{i1} and y_{i2} for the model defined by (7.1), (7.2) and (7.3) are the same as those of the GLMM considering one variable at a time

$$E(y_{i1}) = E[E(y_{i1}|b_{i1})] = E[\mu_{i1}]$$

$$E(y_{i2}) = E[E(y_{i2}|b_{i2})] = E[\mu_{i2}]$$

and

$$\mathrm{var}(y_{i1}) = E[\emptyset_1 V(\mu_{i1})] + \mathrm{Var}[\mu_{i1}]$$

$$\mathrm{var}(y_{i2}) = E[\emptyset_2 V(\mu_{i2})] + \mathrm{Var}[\mu_{i2}]$$

where $\mathrm{Var}[\mu_{i1}]$ and $\mathrm{Var}[\mu_{i2}]$ denote the variance function corresponding to the exponential family distributions for the two response variables. Therefore, $\mathrm{Var}[\mu_{i1}] = \mathrm{Var}[E(y_{i1}|b_{i1})]$ and $\mathrm{Var}[\mu_{i2}] = \mathrm{Var}[E(y_{i2}|b_{i2})]$. The marginal covariance matrix between y_{i1} and y_{i2} is found to be equal to the covariance between μ_{i1} and μ_{i2}, that is $\mathrm{Cov}\,(y_{i1}, y_{i2}) = \mathrm{Cov}(\mu_{i1}, \mu_{i2})$. The property is a consequence of the key assumption of conditional independence between the two response variables. This property allows the method to extend model fitting methods from the univariate to the multivariate GLMM.

To solve the problem of two outcomes, there are two strategies. These strategies accommodate mixed endpoints of the two outcomes. The product of the marginal distribution of one of the response variable and the conditional distribution of the other one given the first variable can be used to express the joint distribution of the binary variables. But, there is no simple expression to find the association between both endpoints. Therefore, to overcome this problem, it is important to treat the surrogate as binary variable. Therefore, the bivariate normal model for y_{i1} and y_{i2} can be described in Probit-linear model and an alternative can be formulated based on the bivariate Plackett density which Plackett-Dale modem (Plackett, 1965).

To use a Probit-normal formulation, assume the following models (Molenberghs and Verbeke, 2005).

$$y_{i1} = \mu_1 + \beta_1 X_i + \in_{i1} \qquad (7.4)$$

$$y_{i2} = \mu_2 + \beta_2 X_i + \in_{i2} \qquad (7.5)$$

where μ_1 and μ_2 are intercepts, β_i's are fixed effects and $\in_{i1}$ and $\in_{i2}$ are correlated errors. Therefore,

$$\begin{pmatrix} \in_{i1} \\ \\ \in_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \dfrac{\rho^\sigma}{\sqrt{1-\rho^2}} \\ & \dfrac{1}{1-\rho^2} \end{pmatrix}\right)$$

The bivariate normal density models are represented by (7.4) and (7.5). It is clear that y_{i1} univariate normal with variance σ^2. Therefore, μ_1, β_1 and σ^2 can be estimated with response y_{i1} and covariate Z_i. Therefore, the conditional density of y_{i2} for X_i and y_{i1} is

$$y_{i2} \sim N\left[\left(\mu_2 - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mu_1\right) + \left(\beta_2 - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\beta_1\right)X_i + \left(\frac{\rho}{\sigma\sqrt{1-\rho^2}}y_{i1}; 1\right)\right]$$

with unit variance. Therefore, the corresponding probability is

$$P(y_{i2} = 1 | y_{i1}, X_i) = \Phi_1(\lambda_0 + \lambda_x X_i + \lambda_T y_{i1}) \qquad (7.6)$$

where

$$\lambda_0 = \mu_2 - \frac{\rho}{\sigma\sqrt{1-\rho^2}}, \qquad (7.7)$$

$$\lambda_x = \beta_2 - \frac{\rho}{\sigma\sqrt{1-\rho^2}}\beta_1, \qquad (7.8)$$

$$\lambda_T = \frac{\rho}{\sigma\sqrt{1-\rho^2}}, \qquad (7.9)$$

and Φ_1 is the standard normal cumulative density function. To find the λ parameters, model (7.6) can be used to y_{i1} with covariate X_i and y_{i1}. Furthermore, regression parameters from $y_{i1}(\mu_1, \beta_1$ and $\sigma^2)$ and probit regression (λ_0, λ_x and λ_T) and parameters from y_{i2} can be obtained using equations (7.7) – (7.9)

$$\mu_2 = \lambda_0 + \lambda_T \mu_1, \tag{7.10}$$

$$\beta_2 = \lambda_2 + \lambda_x \beta_1, \tag{7.11}$$

$$\rho^2 = \frac{\lambda_T^2 \sigma^2}{1 + \lambda_T^2 \sigma^2}. \tag{7.12}$$

Where, $\widehat{\sigma}^2 = 2\,\sigma^2/N$. The asymptotic covariance of $(\hat{\lambda}_0, \hat{\lambda}_x, \hat{\lambda}_T)$ yields the covariance matrix of the parameters. The derivation of the asymptotic covariance of (μ_2, β_2, ρ) can be obtained from the calculations of equations (7.10 – 7.12) with respect to the six orthogonal parameters with delta method.

Therefore,

$$\frac{\partial(\mu_2,\ \beta_2,\ \rho)}{\partial(\mu_1,\ \beta_1,\ \sigma^2, \lambda_0, \lambda_x, \lambda_T)} = \begin{pmatrix} \lambda_T & 0 & 0 & 1 & 0 & \mu_1 \\ 0 & \lambda_T & 0 & 0 & 1 & \beta_1 \\ 0 & 0 & h_T & 0 & 0 & h_2 \end{pmatrix}$$

where

$$h_1 = \frac{1}{2\rho} \ \frac{\lambda_T^2}{(1 + \lambda_T^2 \sigma^2)^2},$$

$$h_2 = \frac{1}{2\rho} \ \frac{2\lambda_T^2 \sigma^2}{(1 + \lambda_T^2 \sigma^2)^2}.$$

Furthermore, the joint estimation can be obtained by maximizing the likelihood based influence of (7.1) and (7.2) (Molenberghs et al., 2001). To formulate Plackett-Dale, it is important to assume the cumulative distribution of y_{i1} and y_{i2} given by $F_{y_{i1}}$ and $F_{y_{i2}}$ (Plackett, 1965). Therefore,

$$F_{y_{i1,}y_{i2}} = \begin{cases} 1 + \dfrac{\left(F_{y_{i1}} + F_{y_{i2}}\right)(\psi_i - 1) - c\left(F_{y_{i1}}, F_{y_{i2}}, \psi_i\right)}{2(\psi_i - 1)} & \text{if } \psi_i \neq 1 \\ F_{y_{i1}} F_{y_{i2}} & \text{if } \psi_i = 1 \end{cases}.$$

Bivariate Plackett "density" function $G_i(y_{i1}, y_{i2})$ for mixed outcomes can be derived. Let y_{i2} be denoted by π_i, then define $G_i(y_{i1}, 0)$ by $G_i(y_{i1}, y_{i2})$and $G_i(y_{i1}, 1)$. In addition, the result can be a sum to $f_{y_{i1}}(t)$. Therefore,

$$G_i(t, 0) = \frac{\partial F_{y_{i1}}, F_{y_{i2}}(t, 0)}{\partial t}.$$

Then,

$$G_i(t, 0) = \begin{Bmatrix} f_{y_{i1}}(t)\left(1 - \dfrac{1 + F_{y_{i1}}(t)(\psi_i - 1) - F_{y_{i2}}(t)(\psi_i + 1)}{c\left(F_{y_{i1}}, 1 - \pi_i, \psi_i\right)}\right) & \text{if } \psi_i \neq 1, \\ F_{y_{i1}}(t)(1 - \pi_i) & \text{if } \psi_i = 1, \end{Bmatrix}$$

and

$$G_i(t, 1) = f_{y_{i1}}(t) - G_i(t, 0).$$

Moreover, assume $y_{i1} \sim N(\mu_i, \sigma^2)$ with $\mu_i = \mu_1 + \beta_1 X_i$ and $\text{logit}(\pi_i) = \mu_2 + \beta_2 X_i$.

For

$$\theta_i = \begin{pmatrix} \mu_i \\ \sigma^2 \\ \pi_i \\ \psi \end{pmatrix} \text{ and } \eta_i = \begin{pmatrix} \mu_i \\ \ln(\sigma^2) \\ \text{logit}\,(\pi_i) \\ \ln(\psi) \end{pmatrix},$$

estimation of parameters $\nu = (\mu, \beta_1, \beta_2, \ln(\sigma^2), \ln(\psi))$ easily obtained by solving the estimating equation, $U(\nu) = 0$, using Newton-Raphson iteration scheme, where $U(\nu)$ is given by

$$\sum_{i=1}^{n} \left(\frac{\partial \eta_i}{\partial \nu}\right)' \left\{\left(\frac{\partial \eta_i}{\partial \theta_i}\right)'\right\}^{-1} \left(\frac{\partial}{\partial \theta_i} \ln G_i(y_{i1}, y_{i2})\right).$$

The joint model can also be discussed based on the generalized linear mixed model formulation. For this approach, the formulation can be done on the presence of both random effects and serial correlations. The expressions

$$Y_i = \mu_i + \epsilon_i$$

is the general formulation and

$$\mathrm{Y_i} = \frac{\mathrm{e}^{\mathrm{X_i\beta+Z_ib}}}{1+\mathrm{e}^{\mathrm{X_i\beta+Z_ib}}} + \in_{\mathrm{i}}$$

is specific random effects logistic regression. For a bivariate response vectors $\mathrm{y_i} = (\mathrm{y}'_{\mathrm{i1}}, \mathrm{y}'_{\mathrm{i2}})'$ where $\mathrm{y_{i1}} = (\mathrm{y_{i11}}, \ldots, \mathrm{y_{i1n_i}})'$ and $\mathrm{y_{i2}} = (\mathrm{y_{i21}}, \ldots, \mathrm{y_{i2n_i}})'$ are for the two outcomes respectively (Goldstein, 2011).

In general,

$$\mu_{\mathrm{i}} = \mu_{\mathrm{i}}(\eta_{\mathrm{i}}) = \mathrm{g}^{-1}(\mathrm{X_i}\beta + \mathrm{Z_ib_i}). \tag{7.13}$$

Assume $\mathrm{b_i} \sim \mathrm{N}\,(0, \Sigma)$ are the q-dimensional random effects. Furthermore, the link function g^{-1} are allowed to change with the nature of outcomes in Y_i. X_i and Z_i are $(2n_i \times p)$ and $(2n_i \times q)$ dimensional matrices of the covariate values and β ia s p-dimensional vector of unknown fixed regression coefficients. The variance of $\in_i$ depends on the mean-variance link of various outcomes. In addition to this, the variance contains a correlation matrix $\mathrm{R_i}(\alpha)$ and a dispersion parameter $\o_i$.

The variance-covariance matrix of Y_i can be obtained from a general first-order approximate expression, which is given by

$$V_i = Var(Y_i) \simeq \Delta_i Z_i G Z'_i \Delta'_i + \Sigma_i \tag{7.14}$$

with

$$\Delta_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)|_{b_i=0},$$

and

$$V_i \simeq \phi_i^{\frac{1}{2}} A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \phi_i^{\frac{1}{2}},$$

where A_i a diagonal matrix containing the variance from the generalized linear model specification of $y_{ij}\,(j = 1{,}2)$ for a given random effects $\mathrm{b_i} = 0$. ϕ_i is a diagonal matrix with the overdispersion parameter along the diagonal. $R_i(\alpha)$ is a correlation matrix. Furthermore, the over dispersion is normally distributed with σ^2 and the variance function 1 (Molenberghs and Verbeke, 2005). For a binary outcome with logit link

$$\mu_{ij}\,(b_i = 0)(1 - \mu_{ij}(b_i = 0))$$

can be derived from Taylor series expression of the mean component around $b_i = 0$. When an exponential family specification is used for all components, with a canonical link, $\Delta_i = A_i$, the resulting GLMM has the variance-covariance matrix of y_i, i.e.,

$$\mathrm{var}(y_i) = \Delta_i Z_i G Z_i' \Delta_i' + \phi_i^{\frac{1}{2}} \Delta_i^{\frac{1}{2}} R_i(\alpha) \Delta_i^{\frac{1}{2}} \phi_i^{\frac{1}{2}}$$

under conditional independence R_i vanishes and

$$\mathrm{var}(y_i) = \Delta_i Z_i G Z_i' \Delta_i' + \phi_i^{\frac{1}{2}} \Delta_i^{\frac{1}{2}} \phi_i^{\frac{1}{2}}.$$

A model with no random effects for the marginal generalized linear model (MGLM) has a form

$$\begin{pmatrix} y_{i1} \\ \\ y_{i2} \end{pmatrix} = \begin{pmatrix} \mu_1 + \lambda b_i + \alpha X_i \\ \frac{\exp(\mu_2 + b_i + \beta X_i)}{1 + \exp(\mu_2 + b_i + \beta X_i)} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \\ \epsilon_{i2} \end{pmatrix} \tag{7.15}$$

The scale parameter λ is included in the continuous of random-intercept model, given two outcomes are measured. Therefore,

$$Z_i = \begin{pmatrix} \lambda \\ 1 \end{pmatrix}, \quad \Delta_i = \begin{pmatrix} 1 & 0 \\ 0 & v_{i2} \end{pmatrix}, \quad \emptyset = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}$$

with $v_{i2} = \mu_{i2}(b_i = 0)\big(1 - \mu_{i2}(b_i = 0)\big)$.

Suppose ρ is the correlation between ϵ_{i1} and ϵ_{i2}. But, Z_i is not a design matrix, because it contains unknown parameters. Therefore, variance-covariance function (7.14) leads to

$$\mathrm{V_i} = \begin{pmatrix} \lambda^2 & \mathrm{v_{i2}}\lambda \\ \mathrm{v_{i2}} & \mathrm{v_{i2}^2} \end{pmatrix}\tau^2 + \begin{pmatrix} \sigma^2 & \rho\sigma\sqrt{\mathrm{v_{i2}}} \\ \rho\sigma\sqrt{\mathrm{v_{i2}}} & \mathrm{v_{i2}} \end{pmatrix}$$

$$= \begin{pmatrix} \lambda^2\tau^2 + \sigma^2 & \mathrm{v_{i2}}\lambda\tau^2 + \rho\sigma\sqrt{\mathrm{v_{i2}}} \\ \mathrm{v_{i2}}\lambda\tau^2 + \rho\sigma\sqrt{\mathrm{v_{i2}}} & \mathrm{v_{i2}^2}\tau^2 + \mathrm{v_{i2}} \end{pmatrix}. \qquad (7.16)$$

Therefore, the derived approximate marginal correlation function is given by

$$\rho(\beta) = \frac{\mathrm{v_{i2}}\lambda\tau^2 + \rho\sigma\sqrt{\mathrm{v_{i2}}}}{\sqrt{\lambda^2\tau^2 + \sigma^2}\sqrt{\mathrm{v_{i2}^2}\tau^2 + \mathrm{v_{i2}}}}, \qquad (7.17)$$

Expression (7.17) depends on the fixed effects through $\mathrm{v_{i2}}$. A model with no random effects, it can be given as

$$\begin{pmatrix} \mathrm{y_{i1}} \\ \\ \mathrm{y_{i2}} \end{pmatrix} = \begin{pmatrix} \mu_2 + \beta_2\mathrm{X_i} \\ \frac{\exp(\mu_1+\beta_1\mathrm{X_i})}{1+\exp(\mu_1+\beta_1\mathrm{X_i})} \end{pmatrix} + \begin{pmatrix} \epsilon_{\mathrm{i1}} \\ \\ \epsilon_{\mathrm{i2}} \end{pmatrix} \qquad (7.18)$$

and expression (7.16) reduced to ρ.

Under conditional independence, ρ in expression (7.16) satisfies $\rho \equiv 0$ and equation (7.17) can be reduced to

$$\rho(\beta) = \frac{\mathrm{v_{i2}}\lambda\tau^2}{\sqrt{\lambda^2\tau^2+\sigma^2}\sqrt{\mathrm{v_{i2}^2}\tau^2+\mathrm{v_{i2}}}}. \qquad (7.19)$$

Equation (7.19) is simpler than equation (7.17). But, equation (7.19) is a function of the fixed effects. For the case of binary endpoints (both outcomes), equation (7.17) is

$$\rho(\beta) = \frac{\mathrm{v_{i2}}\mathrm{v_{i2}}\tau^2 + \rho\sigma\sqrt{\mathrm{v_{i1}}\mathrm{v_{i2}}}}{\sqrt{\mathrm{v_{i1}^2}\tau^2 + \mathrm{v_{i1}}}\sqrt{\mathrm{v_{i2}^2}\tau^2 + \mathrm{v_{i2}}}}.$$

Similarly, for a constant correlation ρ with no random effects and no residual correlation, we have

$$\rho(\beta) = \frac{v_{i2} v_{i2} \tau^2}{\sqrt{v_{i1}^2 \tau^2 + v_{i1}} \sqrt{v_{i2}^2 \tau^2 + v_{i2}}}. \tag{7.20}$$

Equation (7.20) can be performed with general random effects design matrices Z_i and for more than two components.

Full joint distribution is not necessary for the general model formulation. A full joint model specification needs full bivariate model specification, conditional upon the random effects. Furthermore, the generalized linear mixed model formulation can be extended to the hierarchical cases. The hierarchical cases include repeated measures, meta-analysis, cluster data, correlated data, etc. Model $Y_i = \mu_i + \in_i$ is sufficient to generate marginal and random effects models. For shared parameters between models of different types, it is important to ensure the models to be meaningful. For correlations in the model with random effects, the correlation structure can be derived from $V_i = Var(Y_i) \simeq \Delta_i Z_i G Z_i' \Delta_i' + \Sigma_i$. In general, the parameters from joint models can be estimated using numerical approximation method. These methods include Gaussian quadrature and Laplace approximation. Estimation based on data using pseudo-likelihood where pseudo data created based on a linearization of the mean. Furthermore, the pseudo-likelihood approach can be used to estimate parameters in marginal models and random effects with or without correlations. But, quadrature or Laplace approximations can only estimate parameters in the conditional independence random effects models.

7.4 Evaluation of malaria RDT result using joint models approach

In the malaria study, the primary outcome has been malaria RDT result of respondents where the overall goal has been to explore socio-economic, demographic and geographic factors associated with this outcome. A related outcome to the RDT result is use of mosquito nets and use indoor residual spray for the last twelve months, which has been collected from each individual. The effect of predictor variables on malaria RDT result was explored on the previous chapters (Ayele et al., 2012, Ayele et al., 2013a, Ayele et al., 2013b). In this Chapter, the aim is to further investigate the joint effect of these predictor variables (socio-economic, demographic and geographic factors) on malaria RDT result, use of mosquito nets and use of anti- mosquito spray in the last twelve months. More specifically, it is important to assess whether the explanatory variables that were found to be significantly related with RDT result in the previous studies would still have a significant effect on malaria RDT result even when use of mosquito nets and use of anti- mosquito spray is accounted for. Also assessing the association between the two outcomes (malaria RDT result and use of mosquito nets) and (malaria RDT result and use of anti- mosquito spray) is of interest. The advantages of fitting a joint model over a separate model that would contain use of mosquito nets and use of ant-malaria spray in a linear predictor include possible gains in efficiency of the parameter estimates (Gueorguieva, 2001). The respondent's malaria RDT result status (positive/negative) has been modelled as a binary variable that follows Bernoulli distribution.

To evaluate the association between malaria RDT result, use of mosquito nets and use indoor residual spray in the last twelve months, the generalized multivariate mixed effects model was fitted. The three response variables could be taken to be completely independent at any point. In this model, the correlation between the three outcomes as well as the correlation coming from the structure of the data is specified through the random effects structure. This is done by assuming separate random intercepts for each outcome variable and then combining them by imposing a joint multivariate distribution on the random intercepts. The SAS procedure GLIMMIX (SAS 9.3) was used to fit the marginal model. This procedure allows us to jointly model outcomes with different distributions and/or different link functions. The estimates from GLIMMIX were used as initial estimates for NLIMIXED procedure.

Table 7. 1: Parameter estimates for a joint marginal model for malaria RDT result, use of mosquito nets and use of indoor residual spray for main effects

Effects	Malaria RDT result			Use of mosquito nets			Use of indoor residual spray		
	Est	SE	P-value	Est	SE	P-value	Est	SE	P-value
Intercept	0.68	0.67	0.94	-9.21	2.70	0.00	-9.21	2.70	0.0001
Age	-1.01	0.00	<.0001	-1.04	0.01	<.0001	-1.04	0.01	<.0001
Gender (ref. Male)									
Female	2.99	0.61	0.53	4.10	0.26	<.0001	-4.78	0.73	0.99
Family size	0.07	0.01	0.02	0.01	0.01	0.14	0.04	0.01	0.01
Region (ref. SNNP)									
Amhara	0.09	0.05	0.10	-0.01	0.07	0.91	0.07	0.12	0.58
Oromiya	0.09	0.06	0.99	0.09	0.08	0.30	0.16	0.14	0.27
Altitude	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	0.00	0.31
Main source of drinking water (Ref. protected water)									
Tap water	-0.36	0.23	<.0001	-2.52	0.30	<.0001	-0.56	0.16	0.0001
Unprotected water	2.46	0.28	<.0001	2.43	0.38	<.0001	0.55	0.13	<.0001
Time to collect water (ref. > 90 minutes)									
< 30 minutes	-1.21	0.01	0.0001	-1.23	0.08	0.00	-1.64	0.49	0.0001
between 30 to 40 minutes	0.45	0.28	0.11	-0.12	0.05	0.03	-2.43	0.50	<.0001
between 40 - 90 minutes	-0.12	0.09	<.0001	0.47	0.36	0.19	-0.85	0.58	0.15
Toilet facility (ref. No facility)									
Pit Latrine	-0.74	0.23	<.0001	-0.01	0.29	0.97	-0.11	0.18	0.03
Toilet with flush	-0.92	0.23	<.0001	-0.54	0.29	0.06	-1.90	0.77	0.02
Availability of electricity (ref. no)									
Yes	2..072	0.08	<.0001	2.30	0.30	<.0001	2.04	0.12	<.0001
Availability of television (ref. no)									
Yes	-0.43	0.16	<.0001	0.25	0.16	<.0001	0.03	0.06	0.64
Availability of radio (ref. no)									
Yes	-0.63	0.03	0.72	-0.03	0.05	0.54	-0.60	0.16	0.0001
Total number of rooms	-0.23	0.04	0.0001	-0.49	0.14	0.00	-0.18	0.05	0.0001
Main material of room's wall (ref. Cement Blocks)									
Corrugated Metal	-0.53	0.11	<.0001	-3.30	0.34	<.0001	-0.76	0.03	<.0001
Mud Blocks	0.27	0.26	<.0001	3.05	2.65	0.25	-11.5	0.05	<.0001
Main material of room's roof (ref. Corrugate)									
Thatch	0.43	0.052	<.0001	0.51	0.07	<.0001	0.16	0.05	<.0001
Sticks and mud	1.21	0.12	<.0001	0.61	0.18	0.0001	0.24	0.18	<.0001
Main material of room's roof (ref. earth/Local dung plaster)									
Cement	-0.26	1.29	0.25	-3.83	2.87	<.0001	-6.13	0.41	0.25
Wood	-0.45	1.02	<.0001	-3.67	2.67	<.0001	-5.92	0.33	<.0001

The conditional independence random effects model was fitted with SAS 9.3 PROC NLMIXED using the general log-likelihood option. The NLMIXED procedure using the general log-likelihood function allows one to impose a joint multivariate distribution on the random effects from separate models. All statistical tests were conducted at a 5% level of significance.

The linear predictors which were used for fitted models consists the same variables which were used in the previous studies (Ayele et al., 2012, Ayele et al., 2013a, Ayele et al., 2013b). The following socio-economic, demographic and geographic variables were considered as explanatory variables. The socio-economic variables are main source of drinking water, time to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms, main material of the room's wall, main material of the room's roof and main material of the room's floor. Geographic variables are region and altitude, and demographic variables are gender, age and family size. In addition to the main effects, some two-way and three-way interaction effects which were significant in the previous studies were included in the model. These two-way and three-way interaction effects are drinking water and roof material, time to collect water and floor material, time to collect water and main material of room's roof, age and gender, gender and main source of drinking water, gender and availability of electricity, gender and floor material, age, gender and main source of drinking water, age, gender and electricity, and age, gender and floor material.

For this study, malaria RDT result, use of mosquito nets and use of indoor residual spray the last twelve months are binary outcome variables. Therefore, malaria RDT result, use of mosquito nets and use of indoor residual spray in the last twelve months will jointly be modelled using generalized linear mixed models. For this model, it is assumed uncorrelated random intercepts with correlated residual errors. The results from the generalized linear mixed model analysis are given in Tables 7.1 and 7.2. The result from joint models for malaria RDT result, use of mosquito nets and use of indoor residual spray in the last twelve months confirm the results obtained from other models in the previous Chapters.

Table 7. 2: Parameter estimates and their corresponding standard errors of a joint marginal model for malaria RDT result, use of mosquito nets and use of indoor residual spray for interaction effects

Effects	**Malaria RDT result**			**Use of mosquito nets**		
	Est	**SE**	**P-value**	**Est**	**SE**	**P-value**
Age and gender (ref. male)						
Female	1.426	0.215	<.0001	0.988	0.006	<.0001
Gender and main source of drinking water (ref. Male & protected water)						
Female and Tap water	-2.107	0.114	<.0001	-2.390	0.447	<.0001
Female and Unprotected water	0.534	0.162	<.0001	-1.592	0.483	0.001
Gender and availability of electricity (ref. Male & yes)						
Female and No	-2.152	0.291	<.0001	-3.256	0.593	<.0001
Age, gender and main source of drinking water (ref. Male & protected water)						
Female and Tap water	-0.335	0.159	0.017	-0.024	0.008	0.004
Female and Unprotected water	2.480	0.263	0.014	-0.008	0.008	0.286
Age and gender and material of room's floor (ref. Male and earth/Local dung plaster)						
Female and Cement	-0.468	1.026	<.0001	1.076	0.023	<.0001
Female and Wood	0.353	0.039	<.0001	1.064	0.000	<.0001

Effects	**use of indoor residual spray**		
	Est	**SE**	**P-value**
Age and gender (ref. male)			
Female	0.988	0.006	<.0001
Gender and main source of drinking water (ref. Male & protected water)			
Female and Tap water	-2.39	0.447	<.0001
Female and Unprotected water	-1.592	0.483	0.001
Gender and availability of electricity (ref. Male & yes)			
Female and No	-3.256	0.593	<.0001
Age, gender and main source of drinking water (ref. Male & protected water)			
Female and Tap water	-0.024	0.008	0.004
Female and Unprotected water	-0.008	0.008	0.286
Age and gender and material of room's floor (ref. Male and earth/Local dung plaster)			
Female and Cement	1.076	0.023	<.0001
Female and Wood	1.064	0	<.0001

The main significant socio-economic, demographic and geographic factors which were found from the joint model of malaria RDT result, use of mosquito nets and use of indoor residual spray in the last twelve months are age, family size, altitude, main source of drinking water, time to collect water, toilet facility, availability of radio, television and radio, total number of rooms, main material of room's wall, main material of room's roof and main material of room's floor. The two-way significant effects were drinking water and roof material, age and gender, gender and main source of drinking water; and gender and availability of electricity. Age, gender and main source of drinking water; and age, gender and floor material were found to be significant three-way interaction effects (Tables 7.1 and 7.2).

Furthermore, among the main effects age, gender, main source of drinking water, main material of room's roof and availability of electricity were involved in the interaction effects (Table 7.2). The estimates of the significant effects are given in Tables 7.1 and 7.2. Based on the results for a unit increase in family size, the odds of positive rapid diagnosis test increases by 7.6% (OR = 1.076, P-value = 0.02). With reference to individuals with no toilet facilities, the odds of a positive malaria rapid diagnosis test is lower for those individuals using a flushing toilet to those who have septic tanks (OR = 0.397, P-value <0.0001) or pit latrine slabs (OR = 0.477, P - value <0.0001). Moreover, for a unit increase in the number of total rooms, the odds of malaria diagnosis test for an individual decreased by 20.1% (OR = 0.799, P-value = 0.0001). With reference to individuals with no access to radio, the odds of a positive malaria rapid diagnosis test is lower for those individuals who have access to radio (OR = 0.535, P - value <0.0001). Similarly, for those households who have electricity, the odd of malaria RDT result to be positive is increased (OR=7.937, P – value < 0.0001) compared to households who have no electricity. Moreover, for households who have access to television, the odds of positive rapid diagnosis test increases (OR = 0.651, P - value <0.0001).

Interaction effects

From Table 7.2, it can be seen that there are significant two-way and three-way interaction effects. The estimates of these significant effects are given in Table 7.2. As the result indicates one of the three-way interaction effects which was found to be significant is age, gender and main source of drinking water. The result is presented in Figure 7.1 and 7.2.

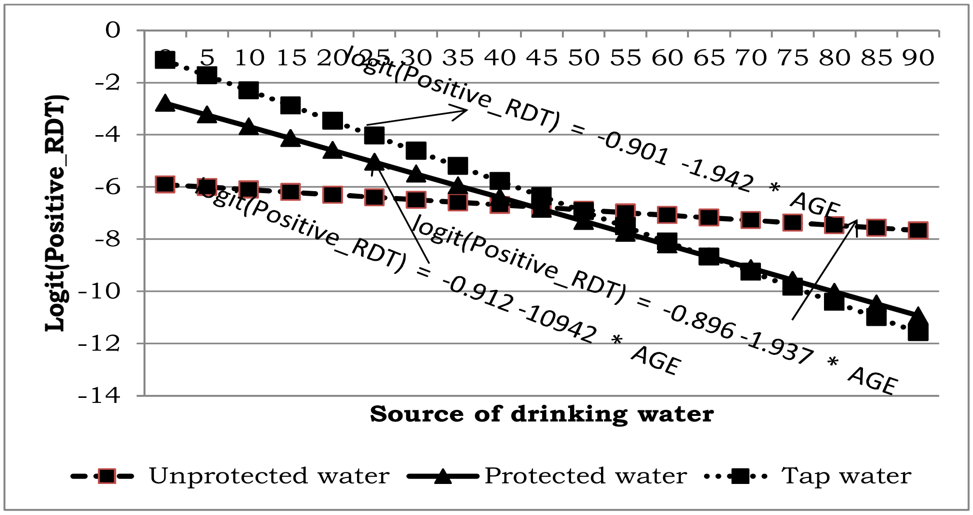


Figure 7. 1: Log odds associated with rapid diagnosis test and age for male respondents with source of drinking water

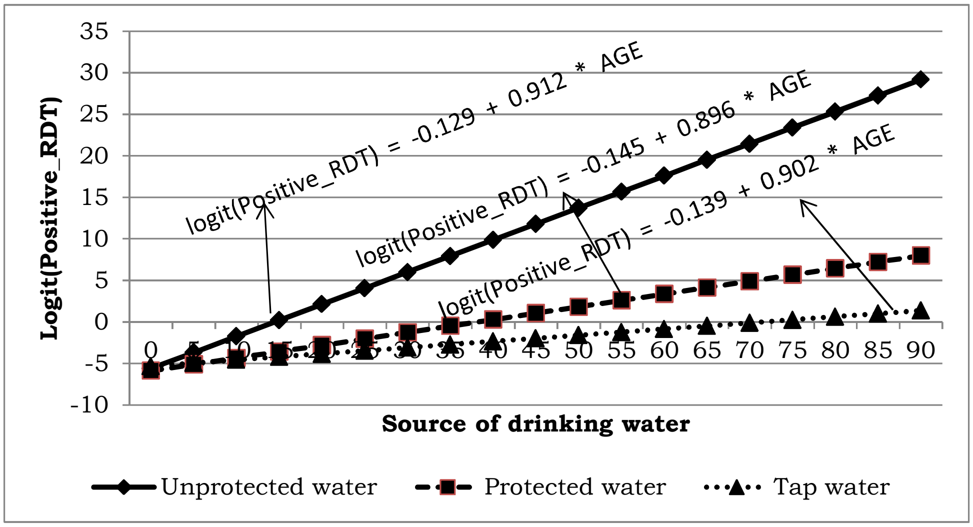


Figure 7. 2: Log odds associated with rapid diagnosis test and age for female respondents with source of drinking water

From the figures it can be seen that as age increased, positive malaria diagnosis was less likely for males than for females who were using protected, unprotected and tap water for drinking. Furthermore, as age of respondents increased, malaria RDT was less likely to be positive for individuals who used tap water for drinking for males and for females. More specifically, positive malaria diagnosis rates increased with age for females whereas it decreased for males as age increased (Figures 7.1 and 7.2). The Figures further show that the gap in the malaria RDT Test between respondents using unprotected, protected and tap water for drinking widens with increasing age for females.

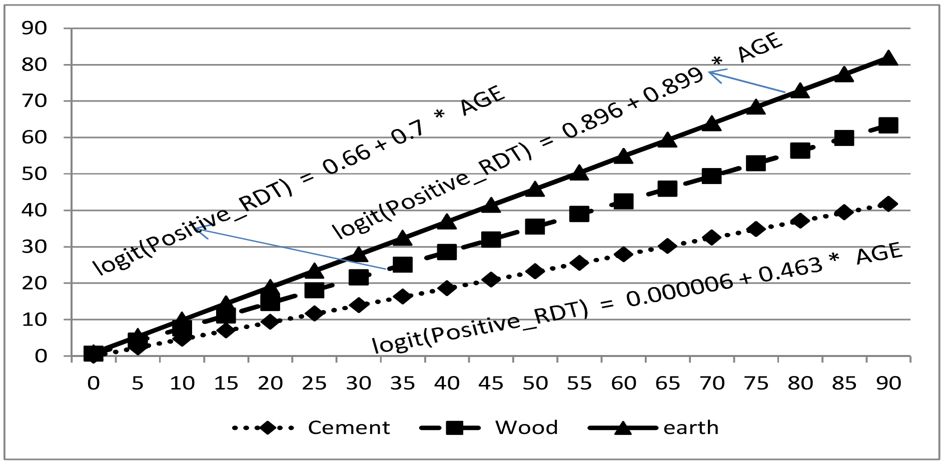


Figure 7. 3: Log odds associated with rapid diagnosis test and age for male respondents with material for room's floor

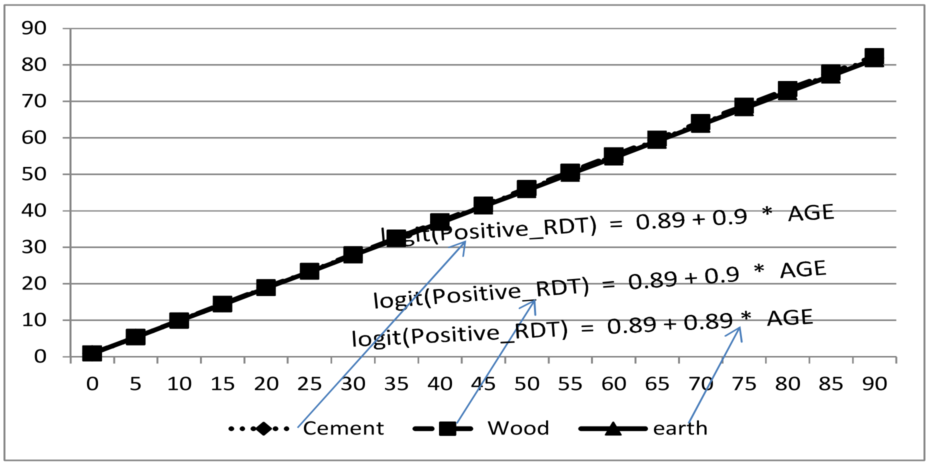


Figure 7. 4: Log odds associated with rapid diagnosis test and age for female respondents with material for room's floor

The other three-way significant interaction effect is between age, gender and material of room's floor (Table 7.2). The results are presented in Figures 7.3 and 7.4 show the interaction between age, gender and material of room's floor for male and female respectively. From the figures it can be seen that as age increased, positive malaria diagnosis was also increased for males for all kinds of material used for roof construction. As can be seen from the figures, individuals who has cement floor has less risk to be positive for malaria RDT result followed by wood and earth. Furthermore, as age of respondents increased, malaria RDT test was also increasing for females. Unlike males, for females the risk of malaria is the same for all type of house construction.

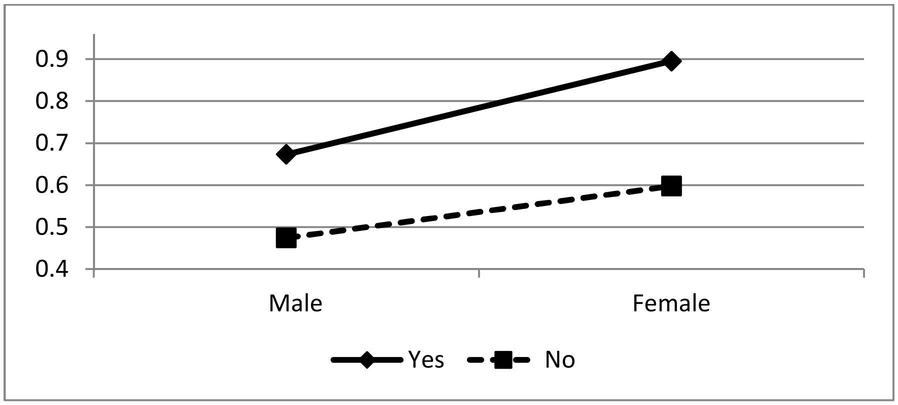


Figure 7. 5: Log odds associated with rapid diagnosis test and availability of electricity with gender

Figure 7.5 presents the interaction effect between availability of electricity and gender for individuals. Prevalence of malaria was significantly higher for female than for male respondents who were living in a house with electricity. Similarly, a female living in a house, which has no electricity, the positive malaria result was significantly higher than it was for males.

The random effects for malaria RDT result and use of mosquito nets are significantly negatively associated i.e., -0.468 (p-value <.0001) (Table 7.3). This indicates a negative correlation between malaria RDT result and use of mosquito nets. This means that increasing the use of mosquito nets tends to decrease the chance of being positive for malaria RDT result. Similarly, the random effect from the joint model of malaria RDT result and use indoor residual spray in the last twelve months (Table 7.3) are significant (-0.310, p-value <.0001). Based on the result, it can be seen that there is negative correlation between malaria RDT result and the use of indoor residual spray for the last twelve months. Therefore, an increase in the use of indoor residual spray leads to decrease for the chance of being positive for malaria RDT result. But, sometimes the conditional independence assumption might be too restrictive. Moreover, statistical tests to check the validity of the assumptions are not well-known in statistical literatures. Moreover, one way of solving conditional dependence is by including one response variable in the linear predictor variable for the other response. This approach was done by (Gueorguieva, 2001). But, models with malaria RDT result as the outcome and included use of mosquito nets and use indoor residual spray in the last twelve months as predictor variables were fitted in the previous studies (Ayele et al., 2012). Therefore, the results from all models fitted show that malaria RDT result is negatively associated with use of mosquito nets and use of indoor residual spray after controlling for the other socio-economic, demographic and geographic factors. Furthermore, if the use of mosquito nets and use of indoor residual spraying increased in the household, household members are less likely to be positive for malaria RDT result.

Table 7. 3: Variance components

Label	Est	SE	Pr > |t|
Var 1(RDT result)	0.632	0.042	<.0001
Var 2 (use of mosquito net)	0.694	0.211	<.0001
Var 3 (use of indoor residual spraying)	0.828	0.101	<.0001
Correlation between Var 1 & Var 2	-0.468	0.430	<.0001
Correlation between Var 1 & Var 3	-0.310	0.212	<.0001

7.5 Summary and discussion

Joint modelling provides efficient parameter estimates and the ability to answer multivariate research questions. This study makes a methodological contribution in the formulation and estimation of three discrete model systems by adopting a joint model methodology wherein flexible error dependency structures can be accommodated between discrete choice equations. To the knowledge of the researcher, this is the first instance in the malaria related literature of the development and application of joint model with an endogenous multinomial choice variable. Therefore, joint modelling provides efficient parameter estimates and the ability to answer multivariate research questions. The results from fitting a joint model of malaria RDT result, use of mosquito nets and use of indoor residual spray in the last twelve months indicate that malaria RDT result is negatively associated with use of mosquito nets and use of indoor residual spray for the last twelve months. That is, for households with less use of mosquito nets and use of indoor residual spray, individuals tend to be positive for malaria RDT result. Nevertheless the negative association between malaria RDT result and use of mosquito nets and use of indoor residual spray in the last twelve months further revealed that if the households have more nets in the house and use indoor residual spray in the last twelve months, the number of positive malaria RDT result might be less. The results reaffirm the significant determinants of socio-economic, demographic and geographic for malaria RDT result from previous studies, i.e., after accounting for the use of mosquito nets and use of indoor residual spray, age, family size, main source of drinking water, time to collect water, toilet facility, total room, main material of room's wall, main material of room's roof and main material of room's. The two-way significant effects were drinking water and roof material, age and gender, gender and main source of drinking water; and gender and availability of electricity. Age, gender and main source of drinking water; and age, gender and floor material were found to be significant three-way interaction effects. Therefore, the finding of this study reveals that for households with toilet facilities, clean drinking water and more living space, the chances of testing positive for malaria decreased. Moreover, using malaria nets and spraying the house walls were found to be effective control measures. In the next chapter, to allow for more flexible trajectory of the observed data, semiparametric approach was used to model the effect of age, family size, altitude, number of nets per person and number of rooms per person non-parametrically.

Chapter 8

Semiparametric models for malaria Rapid Diagnosis Test result

8.1 Introduction

In the previous chapters, malaria rapid diagnosis test data was reviewed and fitted using different statistical methods. These methods are: multiple correspondence analysis, the generalized linear model (Survey logistic), generalized linear mixed models (GLMMs), spatial statistics method and joint models (Ayele et al., 2012, Ayele et al., 2013a, Ayele et al., 2013b). These methods were used to identify the association between malaria RDT result and socio-economic, demographic and geographic factors. These models provide a powerful tool for modelling the relationship between a response variable and covariates. These parametric mean models are simple to use. Because of many sophisticated applications, many computationally intensive data analytic modelling techniques have been invented. These invented methods are useful to exploit possible hidden structures and to reduce modelling biases of the parametric methods. Therefore, because of the restrictions to use parametric models, there is strong demand in recent years on developing nonparametric regression methods. Using this method, flexible functional forms can be estimated from the data to capture possibly complicated relationships between outcomes and covariates. These data analytic approaches are also referred as nonparametric techniques (Lin and Carroll, 2000). Therefore, the basic principle of the nonparametric approaches is to determine the most suitable form of the functions for the available data structure.

The literature on nonparametric methods and their applications is discussed in various literatures (Devroye and Gyorfi 1985, Silverman, 1986, Eubank, 1988, Muller, 1988, Gyorfi et al., 1989 , Hastie and Tibshirani, 1990, Wahba, 1990, Scott, 1992, Green and Silverman, 1994, Wand and Jones, 1995, Fan and Gijbels, 1996, Simono 1996, Bowman and Azzalini, 1997, Hart, 1997 , Ramsay and Silverman, 1997, Ogden, 1997., Efromovich, 1999, Vidakovic, 1999).

Intensive efforts have been devoted to nonparametric function estimation. Over the past years, many new nonparametric models have been introduced. During the past years, to solve nonparametric problems massive arrays of new techniques have been invented. Many new phenomena have been unveiled and deep insights have been gained. The nonparametric modelling has progressed steadily and dynamically. The use of nonparametric techniques is important to reduce possible modelling biases of parametric models. Parametric models are simple and convenient linear models to facilitate computational convenience before 1980s'. But, parametric models are not derived from physical laws and cannot be expected to fit all data well. The purpose of nonparametric techniques is to fit a much larger class of models to reduce modelling biases. These models allow data to search for the appropriate nonlinear forms of the model which best describe the available data. They also provide useful tools for parametric nonlinear modelling and for model diagnostics.

For nonparametric methods, there are many regression and smoothing methods. The methods include kernel smoothing, spline fitting or smoothing, L-Smoothing, R-smoothing, M-smoothing, and Locally WEighted Scatterplot Smoothing (LOWESS) techniques. The techniques are mathematically related to each other. However, each techniques have different properties which are advantageous in different situations (Härdle, 1989, Wu and Zhang, 2006).

Many researchers have looked for possible remedies to solve nonparametric problems. A lot of effort has been allocated to developing methods which reduce the complexity of high dimensional regression problems. This developed methods help to reduce dimensionality as well as allowance for partly parametric modelling. But, the parametric and nonparametric methods, one follows the other. The resulting models can be considered as semiparametric models (Hardle, 1994, Härdle et al., 2004, Hastie and Tibshirani, 1990, Ruppert et al., 2003).

In many applications, the functional form of the relationship may only be partly specified, because the relationships between the response and some confounding covariates may have unknown functional form. This motivates to study the semiparametric generalized additive model (GAM). The proposed GAM generalizes the highly popular generalized additive model (Hastie and Tibshirani, 1986, Hastie and Tibshirani, 1990, Wood, 2006) by adding a parametric nonlinear component to the additive predictor on the link scale. This type of model structure has wide applications in scientific studies where some parametric nonlinear regression relationship is of main interest. Using pametric methods might have confounding effects for some covariates whose relationship to the response is of unknown functional form. In such cases, the parameters could be best estimated nonparametrically.

Therefore, the aim of this chapter is to review GAM and Generalized Additive Mixed Models (GAMM) and then fit them to malaria RDT result data. Specifically, interest is to model the effect of socio-economic, demographic and geographic factors on malaria rapid diagnosis test status non-parametrically. Application of GAMMS, a brief overview of nonparametric regression methods using generalized additive models (GAMs) for independent data is provided. This chapter is organized as follows. An overview of generalized additive models for independent data is presented in Sections 8.2. Section 8.3 reviews the generalized additive mixed models (GAMMs) data. The GAMM model is fitted to malaria RDT data in Section 8.4. Summary and discussion of the chapter is given in section 8.5.

8.2 Generalized Additive Models (GAMs)

Before introducing generalized additive models, it is important to introduce the additive model (AM). The additive model is a nonparametric regression method suggested by (Friedman and Stuetzle, 1981). An AM uses a one dimensional smoother to build a restricted class of nonparametric regression models. Therefore, AM is less affected by the dimensionality of smoother. However, the AM is more flexible and interpretable than a standard linear model. But, there are some problems with the additive model. These problems are model selection, overfitting, and multicollinearity. The AM, which was suggested by (Friedman and Stuetzle, 1981) and (Hastie and Tibshirani, 1990), have been widely used in multivariate nonparametric modelling. An AM, is a generalization of the linear regression model, and is defined by

$$y_i = \mu + \sum\nolimits_{j=1}^{p} f_j(\boldsymbol{x}_{ij}) + \varepsilon_i, \tag{8.1}$$

where y_i is the response variable, μ an intercept term, x_{ij} is the j^{th} component of x_i, f_j is an unkown one-dimensional smooth component function, $(f_1; \, ... \, ; f_p)$ are one for each covariate and ϵ_i is a random variable with mean 0 and finite variance σ^2 $(N(0, \sigma^2))$ (Hastie and Tibshirani, 1990). The optimization problem of additive models in the population setting is to minimize

$$L(f) = \frac{1}{2} E[(\boldsymbol{Y} - \sum_{j=1}^{p} f_j(\boldsymbol{x}_{ij})^2], \tag{8.2}$$

over $\{\boldsymbol{f} : f_j \in H_j\}$. The minimizers of (8.2) can be shown to satisfy

$$f_j = E[(\boldsymbol{Y} - \sum_{k \neq j} f_k) | X_j] := P_j(Y - \sum_{k \neq j} f_k), \tag{8.3}$$

where $P_j = E[.|x_j]$ is the projection operator onto H_j. Replacing P_j by a linear smoother with smoother matrix S_j in (8.3) immediately leads to a sample version of the above iteration procedure for fitting the additive model:

$$\hat{f}_j \leftarrow S_j\left(y - \sum_{k \neq j} \hat{f}_k\right), j = 1,\ldots,p. \tag{8.4}$$

Therefore, this simple algorithm is known as backshifting and is essentially a coordinate descent algorithm (Wood, 2006).

To fit the model, the smooth functions have to be represented. Smoothing of a dataset $\{(X_i, Y_i)\}_{i=1}^{n}$ involves the approximation of the mean response curve f in the regression relationship

$$Y_i = f(\boldsymbol{X}_i) + \in_i\,, \quad i = 1,\ldots,n, \tag{8.5}$$

where, f is a univariate function and $\in_i$ are $i.i.d.$ $N(0, \sigma^2)$. This model is fitted by maximizing the penalized log liklihood with respect to f, i.e.,

$$\max_f\{-\frac{1}{2}\big(\boldsymbol{Y} - f(x)\big)'(\big(\boldsymbol{Y} - f(x)\big) - \frac{1}{2}\lambda J(f) \tag{8.6}$$

where, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $f(x) = (f(x_1), f(x_2), \ldots, f(x_n))'$, and $J(f)$ is the wiggliness penality. Here, the regression curve and certain derivatives of it or functions of derivatives such as extrema point is the functional of interest. In different research approches, the data collection could have been performed in several ways. But, for most studies of a regression relationship, there is just a single response variable Y and predictor variables (Wood, 2006).

Consider representing a function of one variable, $f\ (x)$. Let $\{b_j\ (x): i = j \ldots m\}$ be a set of functions that are chosen to have convenient properties, and to have no unknown parameters. Here, $f\ (x)$ can be represented as:

$$f(x) = \sum_{j=1}^{m} \beta_j b_j(x) \tag{8.7}$$

where the β_j are m unknown coefficients. So $f(x)$ is made up of a linear combination of the basis functions $b_j(x)$, and estimating f is now equivalent to finding the $\beta's$.

Furthermore, for penalized regression spline, there are several examples of basis functions that can be considered. These examples include, cubic spline basis, thin plate regression splines and tensor product bases. The details for these examples is given in (Hastie and Tibshirani, 1990, Wood, 2006).

To model with basis functions, it is possible to control the wiggliness of the fitted model. This can be done by controlling the number of basis functions used. However, this can cause difficulties. These difficulties are:

- If the number of basis functions is large enough to be able to closely approximate the unknown underlying true function, then the model will overfit the data that contain any noise.
- If the number of basis functions is chosen to be low enough to avoid this overfitting, it will be too restrictive to closely approximate the underlying truth.

Using a relatively large number of basis functions, we can avoid over fitting by imposing a penalty during model fitting that is designed to ensure that the fitted model is smooth. This process is known as smoothing (Wood, 2006).

For the one basis function model the governing equation is given by,

$$E(y_i) = f(x_i)$$

where f is a smooth function. This smooth function can be estimated by minimizing

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda\int[f''(x)^2]dx \qquad (8.8)$$

where λ is a smoothing parameter that controls the trade-off between closely matching the data and having a smooth model. Choosing a basis for f requires a design matrix $\boldsymbol{X}$ and a penalty matrix $\boldsymbol{H}$ to be calculated. The fitting problem can be written as

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\beta^T\boldsymbol{H}\beta \qquad (8.9)$$

Therefore, this function can be re-written as:

$$(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta'\boldsymbol{H}\beta = \beta'[\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\lambda H}]\beta + 2\beta'X'y + y'y.$$

This function can be minimized by differentiating with respect to β and setting the resulting system of equations to zero. Therefore, the penalized least square estimator of β for a given λ is given by

$$\hat{\beta} = (\boldsymbol{X^TX} + \boldsymbol{\lambda H})^{-1}\boldsymbol{X'y} \qquad (8.10)$$

By estimating the smooth parameter λ, the degree of smoothness for the model can be obtained. The method of *ordinary cross validation* (OCV) and *generalized cross validation* (GCV) are used to estimate λ.

As it is mensioned earlier, (Hastie and Tibshirani, 1986) proposed Generalized Additive Models (GAM). These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. The GAM is an extension of the Generalized linear model replacing the linear form with the additive form. To determine the appropriate smooth function f, the steps in GLM are replaced by nonparametric addaptive regression steps. Therefore, the GAM using the notation of (Wood, 2006) can be presented as:

$$g(\mu_i) = \boldsymbol{X}_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \ldots \qquad (8.11)$$

where $\mu_i \equiv E(Y_i)$ and Y_i has a distribution that follows the exponential family distribution, $\boldsymbol{X}_i^*$ is the design matrix, θ is the corresponding parameter vector, and $f_j(.)$ are smooth functions of covariates. Model (8.11) is simply an additive model if g is the identity link and the response is normally distributed (Faraway, 2006).

Estimation of parameters for GAM depends on the choice of smoothing bases. Scatterplot smoothing functions, commonly referred to as smoothers, are central to GAM. A smoother is a tool used for summarizing the trend of a response measurement as a function of independent variables (Hastie and Tibshirani, 1990), i.e.,

$$y_i = f(x_i) + \in_i.$$

For choosing smooth function in the model, it can be seen that the basis is a way of defining the space of functions for which f is an element. Choosing a basis amounts to choosing a basis function b_j such that the regression splines $f(x_j)$ can be represented as:

$$f(x) = \sum_{j=1}^{m} \beta_j b_j(x)$$

where x may be a vector quantity and β_j are coeffcient of the smooth, which are estimated as part of model fitting. After selecting the bases, (8.11) reduces to a GLM problem. Each smooth function in the model can be written in terms of a model matrix $\breve{\boldsymbol{X}}_j$. Let f_j be a vector, so that $f = f(x_i)$ and $\breve{\beta}_j = [\beta_{j1}, \ldots, \beta_{jm}]'$, yields to $f_j = \breve{\boldsymbol{X}}_j \breve{\beta}_j$ where $\breve{\boldsymbol{X}}_{j,ik} = b_{jk}(x_{ji})$. Therefore, model (8.11) is not identifiable unless each smooth function is subjected to a centering constraint. For the smooth terms which are re-parameterized in terms of $m - 1$ new parameters, β_j, such that $\breve{\beta}_j = \boldsymbol{Z}\beta_j$ with $\boldsymbol{Z}$ being a matrix such that $m - 1$

columns are orthogonal and the matrix also satisfies $\mathbf{1}'\breve{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{Z} = \mathbf{0}$, a new model matrix for the j^{th} term, $\boldsymbol{X}_j = \breve{\boldsymbol{X}}_j\boldsymbol{Z}$, is obtained such that $f_j = \boldsymbol{X}_j\beta_j$ satisfies the centering constraint. For a given centered model matrices for the smooth function, (8.11) can be written as $g(\mu_i) = \boldsymbol{X}_i\beta$, where $\boldsymbol{X} = [\boldsymbol{X}^*: \boldsymbol{X}_{\mathbf{1}}: \boldsymbol{X}_{\mathbf{2}}: \ldots]$ and $\beta' = [\theta, \beta_1', \beta_2', \ldots]$. The GAM is usually estimated by penalized likelihood maximization, where penalties are designed to suppress overly wiggly estimates of f_j terms. In fact, this is the idea behind the penalized regression approach of GAM estimation. This is because, for m which is large enough, there is a reasonable chance of accurately representing the unknown f_j's, and β is estimated by ordinary likelihood maximization. But, there is a good chance of over-fitting (Wood, 2006).

Interpolating the points $\{x_i, y_i: i = 1, \ldots, n\}$ with $x_i < x_{i+1}$ for the natural cubic spline, g(x) is defined as a function composed of sections of cubic polynomial. For each interval $[x_i;\ x_n]$ joined together so that the function is continuous in value. The first and second derivatives of the function, i.e., $g(x_i) = y_i$ and $g\ (x_1) = g\ (x_n) = 0$ is also continious. The points at which the sections are joined are referred to as the knots of the spline. This function is not only the smoothest interpolator through any data set, but also provides interpolation that is optimal in various respects. The properties of spline indicates that splines are deemed as capable of closely approximating any smooth function. Therefore, splines are considered intuitively appropriate in representing smooth terms in the models (Hastie and Tibshirani, 1990, Wood, 2006).

The cubic smoothing splines arise as a solution to the smoothing objective, which is expressed as a minimization of

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx$$

where λ controls the trade-off between closely fitting the data and producing a smooth fuction. Here, computation becomes expensive in cases of many covariates because they have many free parameters as there are data to be smoothed. To retain the good properties of splines and computational efficiency, using penalized regression splines is a compromise solution. Cubic regression splines are a subset of penalized regression smoothers. There are many ways of defining a cubic regression spline basis. This method is appropriate to have the spline parameterized at its values at the knots. There are other spline frameworks. These are: thin plate regression splines, thin plate regression splines with shrinkage, cubic regression splines (CRS) with shrinkage and P-splines. However, CRS have the advantage that they are computationally cheap when compared to other splines (Hastie and Tibshirani, 1990, Wood, 2006).

Penalized likelihood for the model can be witten because it is possible to capture each smooth function in the model. Penalties, which measure quadratic forms in the function coefficients, are considered. For the j^{th} function, $\breve{\beta}_j'$, $\bar{S}_j$, $\bar{\beta}_j$, can be evaluated as a penalty matrix of known coefficients. By re-parameterization through centering and re-writing the penalty in terms of the full coefficient vector β, it can be expressed as $\beta' S_j \beta$ where S_j is $\overline{\boldsymbol{S}}_j$ padded with zeros such that $\beta' S_j \beta = \beta' \bar{S}_j \beta$ where $\bar{S}_j = \boldsymbol{Z}' \overline{\boldsymbol{S}}_j \boldsymbol{Z}$. The penalized likelihood is therefore defined as

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_j \lambda_j \beta' S_j \beta$$

where λ_j are smoothing parameters, which control the trade-off between model fit and smoothness. Given λ_j, l_p can be maximized with respect to β. Though, λ_j have to be estimated as well. Assuming that λ_j are known and defining $S = \sum\_j \lambda_j S_j$, then

$$l_p(\beta) = l(\beta) - \frac{1}{2}\beta' S_j \beta$$

can be maximized with respect to β_j using

$$\frac{\partial l_p}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} - [S\beta]_j$$
$$= \frac{1}{\emptyset}\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} - [S\beta]_j = 0$$

where $[.]_{\mathrm{j}}$ denotes the j^{th} raw vector.

Using penalized maximum likelihood estimation for a given λ_j, $\hat{\beta}$ can be estimated by iterating the two steps to convergence. These steps are:

1. Given the current $\mu^{[k]}$ calculate the pseudo-data $z^{[k]}$ and weights $w_i^{[k]}$ where,

   $$w_i^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2} \text{ and } z_i = g'\left(\mu_i^{[k]}\right)\left(y_i - \mu_i^{[k]}\right) + X_i\hat{\beta}^{[k]},$$

   g is the model link function, $z^{[k]}$ is a vector of pseudo-data and $w_i^{[k]}$ is a diagonal matrix with diagonal elements $w_i^{[k]}$.

2. Minimize

   $$\left\|\sqrt{W^{[k]}}(z^{[k]} - X\beta)\right\|^2 + \beta' S\beta$$

   with respect to β to find $\hat{\beta}^{|k+1|}$. Evaluate the linear predictor $\eta^{[k+1]} = X\beta^{[k+1]}$ and fitted values $\mu_i^{[k+1]} = g^{-1}\left(\eta_i^{[k+1]}\right)$. Increment k until convergence.

The influence matrix of a GAM fit is $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S})^{-1}\boldsymbol{X}'\boldsymbol{W}$, the influence matrix of the penalized working least square problem of final step of the P-IRLS.

The degree of freedom of GAM, defined as $tr(\boldsymbol{A})$, where $\boldsymbol{A}$ is the influence matrix, indicate the fexibility of the fitted model. Large values of smoothing parameters would result in a model with very few degree of freedom. But, it is very inflexible. The application of penalities reduce the model degree of freedom. The effective degrees of freedom of the model can be divided to each smooth function in the model separately. The effective degrees of freedom for the model parameters in the general weighted case are given by the diagonal of $\boldsymbol{F} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$, where $S = \sum_j \lambda_j \boldsymbol{S}_j$. Furthermore, $\boldsymbol{F}$ is the matrix that maps the un-penalized estimates to the penalized ones and F_i measures the effective degree of freedom of the i^{th} penalized parameters. To estimate the residual variance, σ^2 for additive model, the procedures used for estimation in linear regression are applicable so that

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\|^2}{n - tr(\boldsymbol{A})}$$

while the despersion parameter in the case of GAMs is estimated by the pearson estimator. The model coefficient, β, given smoothing parameters λ can be estimated by penalized likelihood maximization. There are two approaches suggested by (Wood, 2006) to estimate parameters. When the scale parameter, $\emptyset$, is known, then the Mallow's C_p or Un-Biased Risk Estimation (UBRE) can be used for estimation. For an unknown scale parameter, estimation can be done using genalized cross validation (GCV) (Craven and Wahba, 1979, Mallows, 1973). The ordinary cross validation (OCV) criterion is based on minimizing the average mean squared error in predicting a new observation y using the fitted model. To fit the model, using the model to predict $E(y_i)$, y_i is omitted. Repeated procedure to the data gives the estimate for OCV in additive model. This estimate is given by

$$v_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{[-t]})^2$$

where $\hat{\mu}_i^{[-t]}$ denotes the prediction of $E(y_i)$ obtained by leaving out y_i. Estimation of v_0 does not have to proceed by fitting the model n times; it can be estimated as

$$v_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}$$

which simply requires fitting the original model once. Furthermore, the OCV is computationally expensive when there are several smoothing parameters and has slightly disturbing lack of invariance. Therefore, to overcome this lack of invariance, generalized cross validation score can be used. For AM, it can be given by

$$v_g = \frac{n\|.y - \mu\|^2}{[n - tr(A)]^2}.$$

This estimate provides valid prediction error for estimates. OCV has also valid prediction error, but has the invariant property. Generalization of OCV to the GAM can be done by writing the GAM fitting objective in terms of model deviance. This leads to the GCV approach. This can be given as follows:

$$D(\beta) + \sum_{j=1}^{m}\lambda_j\beta' S_j\beta.$$

The GVC score application can be defined as

$$v_g = \frac{nD(\widehat{\beta})}{[n-tr(A)]^2}.$$

Performance iteration and outer iteration are the two numerical strategies for estimation of the smooth parameters (Wood, 2006).

8.3 Generalized Additive Mixed Models (GAMMs)

For data which consists correlated measurnment or other variabilities, the variabilities introduce a new source of randomness and creates an extension to GAM. Similar to generalized linear mixed models (GLMM) which are extensions of GLM, generalized additive mixed models (GAMM) are extensions of GAM and allows the parametric fixed effects to be modelled nonparametically using additive smooth functions. Therefore, GAMM's include random effects (Breslow and Clayton, 1993, Hastie and Tibshirani, 1990, Lin and Zhang, 1999). Generalized additive mixed model (GAMM) has the following structure (Wood, 2006).

$$y_i = \boldsymbol{x}_i\beta + f_1(x_{i1}) + f_2(x_{i2}, x_{i3}) + \ldots + f_p\left(x_{ip}\right) + \boldsymbol{Z}_i\boldsymbol{b} + \in_{\boldsymbol{i}}, \qquad (8.12)$$

where, y_i, $i = 1,\ldots,n$ is outcome variable, p covariates $X_i = (1, x_{i1},\ldots,x_{ip})'$ associated with fixed effects and $q \times 1$ vector of covariates Z_i associated with random effects. Therefore, given a $q \times 1$ vector of $\boldsymbol{b}$ of random effects, the observations y_i are assumed to be conditionally independent with means $E(y_i|\boldsymbol{b}) = \mu_i$ and variances $var\,(y_i|\boldsymbol{b}) = \emptyset v(\mu_i)$, where $v(.)$ is a specified variance function and $\emptyset$ is a scale parameter. Moreover, $g(.)$ is a monotonic differential link function, $f_i(.)$ is a centred twice-differentiable smooth function, the random effects are assumed to be distributed as $N\{0, \boldsymbol{G}(\gamma)\}$ and γ is a $c \times 1$ vector of variance components. To model correlations between observations, the adaptive nonparametrics are used (Ruppert et al., 2003).

For a given variance component θ, the log-quasi-likelihood function of $\{\beta, f_i, \theta\}$, a part from a constant

$$\exp[l\{y; \beta_0, f_1(.),\ldots,f_p(.),\theta] \propto |\boldsymbol{G}|^{-\frac{1}{2}} \int \exp\left\{-\frac{1}{2\emptyset}\sum_{i=1}^{n} d_i(y;\mu_i) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b}\right\} d\boldsymbol{b} \qquad (8.13)$$

where

$$y_i = (y_1, \dots, y_n)' \text{ and } d_i(y;\mu) \propto -2\int_{y_i}^{\mu_i} \frac{y_i - u}{v(u)du}$$

defines the conditional deviance function of $\{\beta, f_i, \theta\}$ given $\boldsymbol{b}$.

The estimation of smooth parameters, λ, and inference on variance component θ is required for GAMM statistical inference on the nonparametric functions $f_j(.)$. It has to be noted that smoothing spline estimators and linear mixed models have close connections (Lin and Zhang, 1999, Wang, 1998, Verbyla et al., 1999). As explained in (Green and Silverman, 1994), for a given value of λ and θ, the natural cubic smoothing spline estimators of $f_i(.)$ maximize the penalized log quasi-likelihood

$$l\{y;\beta_0, f_i(.), \theta\} - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \int_{S_j}^{l_j} f_j''(x)^2 dx = l\{y_i;\beta_0, f_i(.), \theta\} - \frac{1}{2}\sum_{j=1}^{p}\lambda_j f_j' S_{jf_j} \qquad (8.14)$$

where (S_j, t_j) defines the range of the *j^th^* covariate and $\lambda = (\lambda_1, \dots, \lambda_p)'$ is a vector of smoothing parameters. The trade-off between goodness of fit and the smoothness of the estimated functions is controlled by λ. Furthermore, $f_j(.)$ is an $r_j \times 1$ unknown vector of the values of $f_j(.)$ eveluated at the r_j ordered values of the $x_{ij} = (i = 1, \dots, n)$ and S_j is the smoothing matrix.

Using the matrix notation the GAMM model, which is given in (8.12), can be written as

$$g(\mu_i) = \mathbf{1}\beta_0 + \boldsymbol{N_1}f_1 + \dots + \boldsymbol{N}_p f_p + \boldsymbol{Zb} \qquad (8.15)$$

where $g(\mu_i) = \{g(\mu_1), \dots, g(\mu_n)\}'$, and $\boldsymbol{Z} = (Z_1, \dots, Z_n)'$ (Lin and Zhang, 1999).

The numerical integration is required to evaluate the expression given in (8.13). To calculate full natural cubic smoothing spline estimators of f_j by directly maximizing (8.14) is sometimes difficult. Therefore, to avoid this problem an alternative approximation is proposed by (Lin and Zhang, 1999). This proposed method is a double penalized quasi-likelihood (DPQL). Therefore, the nonparametric functions (f_j) estimation can be obtained by using double quasi-likelihood. Here, f_j (centered parameter vector) can be re-parametrized in terms of β_j and $\boldsymbol{a_j}((r_j-2)\times 1)$ through a one-to-one transformation as

$$f_j = \boldsymbol{X}_j^*\beta_j + \boldsymbol{B}_j a_{\boldsymbol{j}}, \tag{8.16}$$

where $\boldsymbol{X}_j^*$ is $r_j \times 1$ vector containing the r_j centered district values of the x_{ij} $(i = 1,\dots,n)$, and $\boldsymbol{B}_j = L_j(L_j'L_j)^{-1}$ and $\boldsymbol{L}_j$ is an $\boldsymbol{r_j} \times (r_j-2)$ fullrank matrix satisfying $\boldsymbol{S}_j = \boldsymbol{L_j L_j'}$ and $\boldsymbol{L_j' x_j^*} = \boldsymbol{0}$. Therefore, the double penalized quasi-likelihood with respect to (β_0, f_i) and $\boldsymbol{b}$ becomes

$$-\frac{1}{2\emptyset}\sum_{i=1}^{n} d_i(y;\mu_i) - \frac{1}{2}\boldsymbol{b'G^{-1}b} - \frac{1}{2}\boldsymbol{a'}\Lambda^{-1}\boldsymbol{a}, \tag{8.17}$$

where $f_j'\boldsymbol{S}_j f_j = \boldsymbol{a}_j'\boldsymbol{a}_j$, $\boldsymbol{a} = (a_1',\dots,a_p')'$ and $\Lambda = diag(\tau_1\boldsymbol{I},\dots,\tau_p\boldsymbol{I})$ with $\tau_j = 1/\lambda_j$. Note that small values of $\tau = (\tau_1,\dots,\tau_p)'$ corresponds to oversmoothing (Breslow and Clayton, 1993).

Using (Breslow and Clayton, 1993) penalized-likelihood approach, by plugging (8.16) into (8.15), equation (8.17) suggests that θ and τ, the DPQL estimators $\hat{f}_j$ can be obtained by fitting the following linear mixed model

$$g(\mu_i) = \boldsymbol{X}\beta + \boldsymbol{Ba} + \boldsymbol{Zb}, \tag{8.18}$$

where, $X = (1, N_1X_1,\dots,N_pX_p)$, $B = (1, N_1B_1,\dots,B_pX_p)$, $\beta = (\beta_0,\dots,\beta_p)'$ is a $(p+1)\times 1$ vector of regression coefficients and $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent random effects with distribution $a \sim N(0,\Lambda)$ and $b \sim N(0,G)$.

$$\hat{f}_j = X_j^*\hat{\beta}_j + B_j\hat{a}_j$$

gives the DPQL estimator $\hat{f}_j$. This estimator is a linear combination of the penalized quasi-likelihood estimators of the fixed effects $\hat{\beta}_j$ and the random effects $\hat{a}_j$ in equation (8.10) (Breslow and Clayton, 1993).

Using Fisher scoring algorithm, maximization of (8.17) with respect to $(\beta, \boldsymbol{a}, \boldsymbol{b})$ can be solved as

$$\begin{bmatrix} \boldsymbol{X'WX} & \boldsymbol{X'WB} & \boldsymbol{X'WZ} \\ \boldsymbol{B'WX} & \boldsymbol{B'WB} + \wedge^{-1} & \boldsymbol{B'WZ} \\ \boldsymbol{Z'WX} & \boldsymbol{Z'WB} & \boldsymbol{Z'WZ} + \boldsymbol{G^{-1}} \end{bmatrix} \begin{bmatrix} \beta \\ \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'WX} \\ \boldsymbol{B'WX} \\ \boldsymbol{Z'WX} \end{bmatrix} \tag{8.19}$$

where y is the working vector defined as

$$y = \beta_0 1 + \sum_{j=1}^{p} N_j f_j + Zb + \Delta(y - \mu),$$

$\Delta = diag(g'(\mu_i))$, $\boldsymbol{W} = diag\ [\{\emptyset v(\mu_i) g'(\mu_i)^2\}^{-1}]$.

The expression (8.19) shows that it corresponds to the normal equation of the best linear unbiased predictors (BLUPs) of β and (a, b) under the linear mixed model

$$Y = \boldsymbol{X}\beta + \boldsymbol{Ba} + \boldsymbol{Zb} + \in \tag{8.20}$$

where a and b are independent random effects with $a \sim N(0, \wedge)$, $b \sim N(0, \boldsymbol{G})$ and $\in\ \sim N(0, \boldsymbol{W}^{-1})$. Iteratively fitting equation (8.20) to the working vector Y, the DPQL estimators $\hat{f}_j$ and the random effect estimators $\hat{b}$ can be easily obtained using the BLUPs.

The covariance matrix of $\hat{f}_j$ can be obtained by calculating β and a using

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}B} \\ \boldsymbol{B'R^{-1}X} & \boldsymbol{B'R^{-1}B} + \wedge^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}Y} \\ \boldsymbol{B'R^{-1}Y} \end{bmatrix} \tag{8.21}$$

where $\boldsymbol{R} = \boldsymbol{W}^{-1} + \boldsymbol{ZGZ}'$. Let the left hand side of equation (8.21) be denoted by $\boldsymbol{H}$ and $H_0 = (X,B)'R^{-1}(X,B)$, the approximate covariance matrix of $\hat{\beta}$ and $\hat{a}$ is

$$cov(\hat{\beta},\hat{a}) = \boldsymbol{H}^{-1}\boldsymbol{H}_0\boldsymbol{H}^{-1}$$

The approximate covariance matrix of $\hat{f}_j$ is $(X_j,B_j)cov(\hat{\beta},\hat{\boldsymbol{a}})(X_j,B_j)'$, where $cov(\hat{\beta},\hat{\boldsymbol{a}})$ can be easily obtained from the corresponding blocks of $\boldsymbol{H}^{-1}\boldsymbol{H}_0\boldsymbol{H}^{-1}$. Note that $f_j(.)$ are fixed smooth functions in calculating the covariance of the $\hat{f}_j$. For the nonparametric function f_i, the smoothing parameters λ and the variance components θ are unknown. But, the estimates can be obtained from the data under the classical nonparametric regression model

$$y = f(\boldsymbol{X}) + \in \tag{8.22}$$

where $\in$ are independent random errors following $N(0,\sigma^2)$. Estimation of the smoothing parameter λ by maximizing a marginal likelihood was proposed by (Wahba, 1985, Kohn et al., 1991). Assuming $f(x)$ has a prior as $f_i = \boldsymbol{X}_j^*\beta_j + \boldsymbol{B}_j\boldsymbol{a}_j$ with $\boldsymbol{a} \sim N(0,\tau\boldsymbol{I})$ leads to the construction of $\tau = 1/\lambda$. A flat prior for β and integrating out a and β as follows

$$\exp\{l_M(y;\tau,\sigma^2)\} \propto \tau^{-1/2}\int\left\{l(y;\beta,\mathbf{a},\sigma^2) - \frac{1}{2r}\,\mathbf{a}'\mathbf{a}\right\}d\mathbf{a}d\beta \tag{8.23}$$

where $l(y;\beta,a,\sigma^2)$ is the log likelihood (normal) of f under (8.22). The maximum marginal likelihood estimator of τ is called the generalized maximum likelihood (GML) estimator. The marginal likelihood which is specified in (8.23) of τ is the REML likelihood under the linear mixed model

$$y = 1\beta_0 + \boldsymbol{X}\beta_1 + \boldsymbol{Ba} + \in,$$

where $a \sim N(0, \tau I)$, $\epsilon \sim N(0, \sigma^2 I)$ and τ is regarded as a variance component. Therefore, the maximum marginal likelihood estimator of τ is an REML estimator.

The smoothing parameter λ and the variance component θ using REML with normally distributed outcomes and a nonparametric mean function can be written as

$$y = f\ (\boldsymbol{X}) + \boldsymbol{Zb} + \in \qquad (8.24)$$

where $f(X)$ denotes the value of the nonparametric function $f(.)$ evaluated at the design point of $\boldsymbol{X}(n \times 1)$, $\boldsymbol{b} \sim N\{o,\ \boldsymbol{G}(\theta)\}$ and $N\ \{\boldsymbol{0}, \boldsymbol{V}(\theta)\}$. For $f(.)$ which is estimated using a natural cubic smoothing spline, equation (8.24) can be as a linear mixed model using equation $f_i = \boldsymbol{X}_j^* \beta_j + \boldsymbol{B}_j \boldsymbol{a}_j$ (Zhang et al., 1998)

$$y = 1\beta_0 + \boldsymbol{X}\beta_1 + \boldsymbol{B}a + \boldsymbol{Z}b + \in,$$

where $\boldsymbol{a} \sim N(0, \tau I)$ and the distribution of $\boldsymbol{b}$ and $\in$ are the same as those in (8.24). Here, τ is treated as an extra variance component in addition to θ. Furthermore, τ and θ can be estimated jointly using REML. The REML likelihood corresponds to the marginal likelihood of (τ, θ) constructed by assuming f takes the form $f_i = \boldsymbol{X}_j^* \beta_j + \boldsymbol{B}_j \boldsymbol{a}_j$ with $\boldsymbol{a} \sim N(0, \tau \boldsymbol{I})$ and a flat prior for β, and integrating out a and β as follows (Harville, 1974)

$$\exp\{l_M(\boldsymbol{y}; \tau, \theta)\} \propto\ |G|^{-1/2}\tau^{-1/2} \int \exp\left\{l(\boldsymbol{y}; \beta, \boldsymbol{a}, \boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b} - \frac{1}{2}\boldsymbol{a}'\boldsymbol{a}\right\} d\boldsymbol{b} d\boldsymbol{a} d\beta \qquad (8.25)$$

where $l(y; \beta, a, b) = l(y; f, b)$ is the conditional loglikelihood of f given the random effects b. The marginal likelihood given in (8.25) has a closed form expression. Based on the simulation result which was done by (Zhang et al., 1998), REML performs very well in estimating both the nonparametric function $f(.)$ and the variance component θ.

8.4 Fitting malaria RDT result using GAMM

In preceding chapters, the malaria RDT result was fitted to predictor variables using parametric models and assumed a linear age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed. But, the objective of this Chapter is to model the effect of age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed nonparametrically while the other covariates remain parametric using GAMM. Recall that the final GAMM model consists of the following socio-economic, demographic and geographic factors. These factors are gender, age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms per person, main material of the room's walls, main material of the room's roof, main material of the room's floors, incidence of indoor residual spray in the past twelve months, use of mosquito nets and total number of nets per person. Malaria test (RDT result), age and sex were collected at individual level. Main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms per person, main material of the room's walls, main material of the room's roof, main material of the room's floor, use of indoor residual spray in the past twelve months, use of mosquito nets and total number of nets per person were collected at household level. Therefore, malaria RDT result with semiparametric logistic regression model was fitting with all these variables including possible interaction effects. Unlike the previous models, age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed in the last twelve months were fitted nonparametically. Therefore, the final model is given as follows.

$$\begin{aligned} & g(\mu_{ij}) \\ & = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Region}_i + \beta_3 \text{drinking\_water}_i + \beta_4 \text{time\_to\_get\_water}_i + \beta_5 \text{toilet\_facility}_i \\ & + \beta_6 \text{elect}_i + \beta_7 \text{tv}_i + \beta_8 \text{radio}_i + \beta_9 \text{room\_wallroom\_roof}_i + \beta_{10} \text{room\_wall}_i + \beta_{11} \text{anti\_malaria}_i \\ & + \beta_{12} \text{net\_use}_i + \beta_{13} \text{Gender} * \text{drinking\_water}_i + \beta_{14} \text{Gender} * \text{elect}_i + \beta_{15} \text{Gender} * \text{room\_wall}_i \\ & + f_1(\text{age}_i) + f_2(\text{altitude}_i) + f_3(\text{famsize}_i) + f_4(\text{total\_room}_i) + f_5(\text{total\_nets}_i) \\ & + f_6(\text{months\_sprayed}_i) \\ & + b_{0i} \end{aligned} \tag{8.26}$$

where $g(.)$ is the logit link function, $\beta_j{}'$s and $\beta_{ij}{}'$s are parametric regression coefficients, f_j are centred smooth functions and the random effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{G}(\theta))$. Therefore, the estimation procedures discussed for fitting GAMMs in the previous section can be used to fit model (8.26). For the analysis, R package (*mgcv*) was used. There are many smoothing spline options in R package. Among the number of options, to fit model (8.26), several different penalized regression smoothers were used. Because of the size of the model and the size of the dataset, the model failed to converge for more interaction effects. Model (8.26) contains reduced parameters by removing the three-way parametric interactions.

Thin plate shrinkage smoothers was used to fit model (8.26). The use of shrinkage smoothers have different advantage, i.e., these methods helps to avoid the knot placement. Furthermore, these methods can be constructed to smooth of any number of predictor variables. Construction of shrinkage smoothers depends on the smooth terms which can be penalized away and this makes no contribution to the model (Wood, 2006).

Table 8.1 presents the significant effects for the parametric coefficients of the model. The result shows that gender, region, main source of drinking water, time to collect water, toilet facility, availability of electricity, availability of radio, main materials for the construction of room's wall, main materials for the construction of room roofs, main materials for the construction of room floors, use of indoor residual spray and use of mosquito nets were found to have significant effects on malaria rapid diagnosis test result. Among all significant effects gender, main source of drinking water, availability of electricity, main materials for the construction of room’s wall and main materials for the construction of room’s roof were involved in the intraction effects. These interaction effects are gender and main source of drinking water, gender and availability of electricity, gender and main material of room's wall and main source of drinking water and main material of the room's roof (Table 8.1 and Table 8.2).

The results from GAMM analysis showed that the odds of positive RDT for households who lives in Amhara region were 0.969 ($e^{-0.031}$) times less likely to be positive for malaria rapid diagnosis test than for those who live in SNNP region. Similarly, the odds of positive RDT for respondents who live in Oromiya region were found to be 0.807 ($e^{-0.215}$) times less likely to be positive for malaria rapid diagnosis test compared to SNNP region. Also, the odds of positive RDT for respondents who travelled greater than 40 minutes found to be 0.361 ($e^{-1.019}$) times less likely to be positive for malaria RDT test than those who travelled greater than 90 minutes followed by for respondents travelled between 30–40 minutes (0.293 ($e^{-1.226}$)) and less than 30 minutes (0.291 ($e^{-1.233}$)). Similary, the odds of positive RDT for respondents who were using toilet with flush were found to be 0.5 ($e^{-0.694}$) times less likely be positive for malaria RDT result compared to households who have no toilet facility followed by pit latrine toilet (0.656 ($e^{-0.421}$)). On the other hand, households who have no access to radio were 2.158 ($e^{0.769}$) times more likely to be positive for malaria RDT test result than those who have access to radio. Also, respondents who lives in house with cement floor where found to be 0.052 ($e^{-2.957}$) times less likely to be positive for malaria RDT result compared to houses with earth/local dung floors followed by houses with wood floor (0.198 ($e^{-1.621}$)).

Table 8. 1: The parameter estimates of the GAMM model of the main parametric coefficients

Effects	Estimate	OR	SE	t Value	Pr > |t|
Intercept	0.260	1.297	0.849	-2.490	<.0001
Gender (ref. Male)					
Female	-1.720	0.179	0.272	-1.080	<.0001
Region (ref . SNNP)					
Amhara	-0.031	0.969	0.082	-0.380	0.7041
Oromiya	-0.215	0.807	0.094	-2.280	0.0225
Main source of drinking water (ref. protected water)					
Tap water	-0.107	0.899	0.079	-1.360	0.1744
unprotected water	0.585	1.795	0.104	5.640	<.0001
Time to collect water (ref. greater than 90 minutes)					
Less than 30 minutes	-1.233	0.291	0.159	-7.760	<.0001
Between 30 - 40 minutes	-1.226	0.293	0.162	-7.570	<.0001
Between 40 - 90 minutes	-1.019	0.361	0.336	-3.040	0.0024
Toilet facility (Ref. No facility)					
Pit latrine	-0.421	0.656	0.365	2.760	0.0057
Toilet with flush	-0.694	0.500	0.362	4.990	<.0001
Availability of electricity (ref. no)					
Yes	0.111	1.117	0.129	16.390	<.0001
Availability of television (ref. no)					
Yes	0.049	1.050	0.057	0.870	0.383
Availability of radio (ref. yes)					
No	0.769	2.158	7.950	92.420	<.0001
Main material of room's wall (ref. cement block)					
Corrugated metal	-1.100	0.333	5.516	131.28	<.0001
Mud block/stick/wood	-0.851	0.427	15.872	412.02	<.0001
Main material of room's roof (ref. corrugate)					
Thatch	1.192	3.294	0.073	16.380	<.0001
Stick and mud	0.855	2.351	0.232	-3.680	0.0002
Main material of room's floor (ref. earth/Local dung plaster)					
Wood	-1.621	0.198	16.451	850.89	<.0001
Cement	-2.957	0.052	15.875	411.83	<.0001
Use of indoor residual spray (ref. yes)					
No	1.235	3.438	0.103	-31.490	<.0001
Use of mosquito nets (ref. no)					
Yes	-0.682	0.506	0.128	20.880	<.0001

Interaction effects

In addition to the main parametric effects, the fitted GAMM model contains four two-way interaction effects. These effects are gender and main source of drinking water, gender and availability of electricity, gender and main material of room's wall and main source of drinking water and main material of the room's roof (Table 8.2).

Table 8. 2: The parameter estimates of the GAMM model of the interaction parametric coefficients

Effects	Estimate	SE	t Value	Pr > |t|
Gender and main source of drinking water (ref. Male & protected water)				
Female and Tap water	1.757	0.198	8.870	<.0001
Female and Unprotected water	-1.605	0.183	-8.770	<.0001
Gender and availability of electricity (ref. Male & yes)				
Female and No	1.851	9.964	61.510	<.0001
Gender and main material of room's wall (ref. Male & earth/Local dung plaster)				
Female and wood	-0.517	119.547	-65.110	<.0001
Female and cement	-4.634	117.851	-0.140	0.888
Main source of drinking water and main material of the room's roof (ref. Protected water & thatch)				
Tap water and Mud block/stick/ wood	-3.732	0.138	-6.030	<.0001
Tap water and Corrugated metal	-3.852	258.258	-2.750	0.006
Unprotected water and Mud block/stick/wood	-4.003	0.143	-4.840	<.0001
Unprotected water and corrugated metal	-1.324	15.990	298.51	<.0001

Interaction effects between the main source of water and the main material used for the room's roof is presented in Figure 8.1. From the figure, it is clearly seen that positive rapid diagnosis of malaria was significantly higher for households with a stick and mud roof followed by thatch and lastly a corrugated iron roof. This occurred with respondents who reported to use tap water as well as protected and unprotected water for drinking (Figure 8.1).

Furthermore, there was a significant difference in rapid diagnosis test between tap, protected and unprotected sources of drinking water for those who reported having thatch and stick and mud roofs. It is also shown that for corrugated iron roofs, the positive rapid diagnosis test was significantly lower for respondents who reported using tap water for drinking than for those who used protected and unprotected water for drinking.

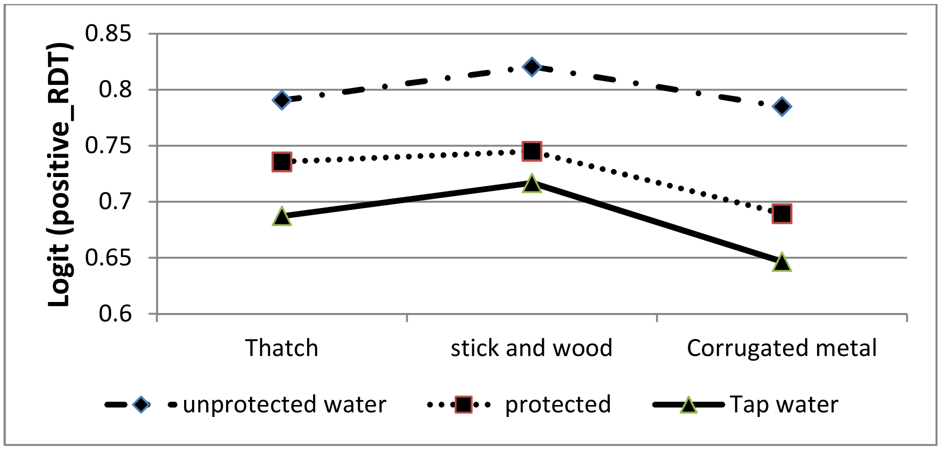


Figure 8. 1: Log odds associated with rapid diagnosis test and source of drinking water with material of the room's roof

The other significant two-way interaction effect was between gender and main source of drinking water (Table 8.2). This result is presented graphically in Figure 8.2. The probability of a positive rapid diagnosis test was significantly higher in those female household members who used unprotected water for drinking than for those respondents who used protected and tap water for females. Generally, for male households who use protected and unprotected water are less likely to be positive for malaria RDT result compared to female household members. But, for female household members who use tap water malaria RDT result found to be less compared to male household members.

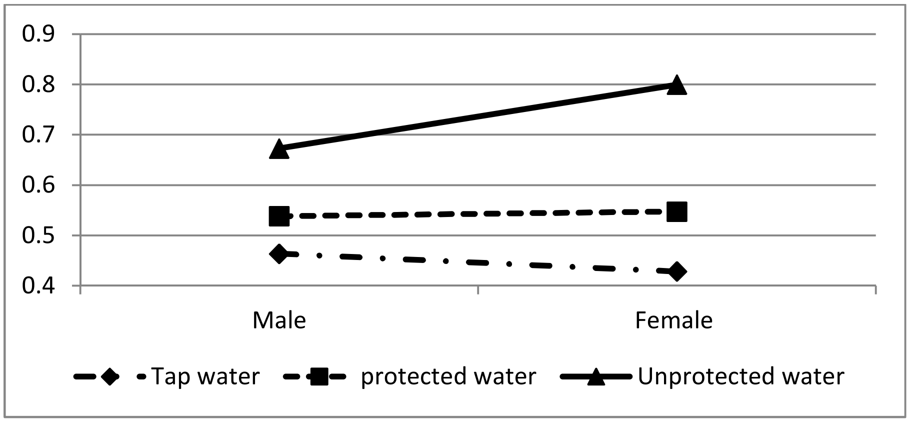


Figure 8. 2: Log odds associated with rapid diagnosis test and main source of drinking water with gender

Figure 8.3 presents the interaction effect between availability of electricity and gender for individuals. Prevalence of malaria was significantly higher for female than for male respondents who were living in a house with electricity. Similarly, a female living in a house, the positive malaria result was significantly higher than it was for males which have no electricity.

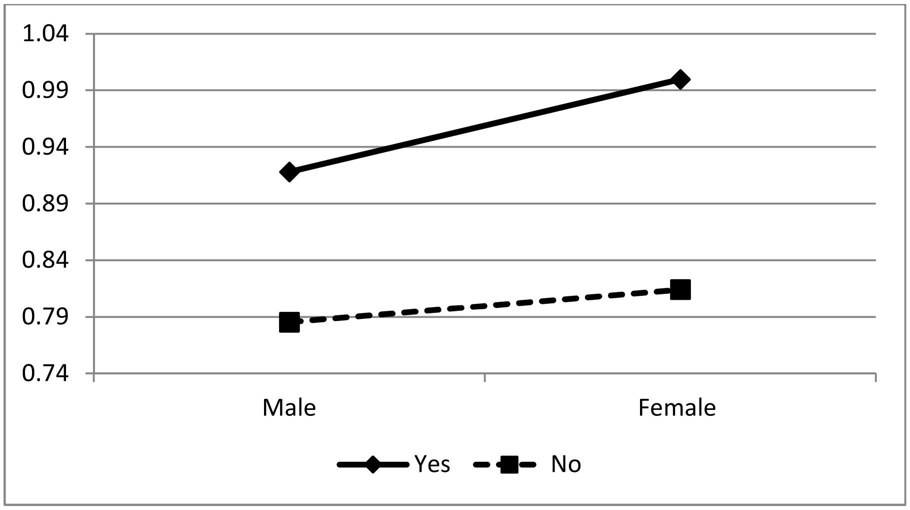


Figure 8. 3: Log odds associated with rapid diagnosis test and availability of electricity with gender

The interaction effect between gender and main material of floor is presented in Figure 8.4. The Figure shows that the odds of positive RDT for households with earth/local/dung floor are significantly higher than for those households with wood and cement floors for both males and females. Moreover, for female members of the household, the odds of malaria RDT was higher for those households who reported having earth/local dung floor.

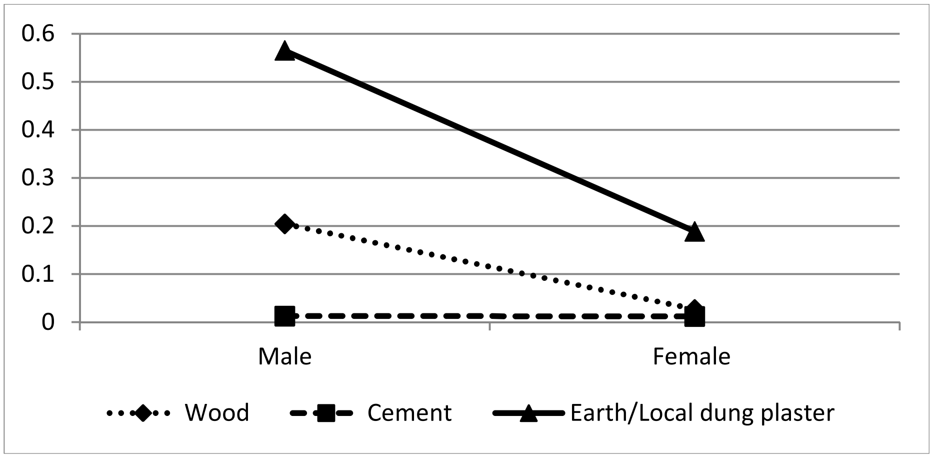


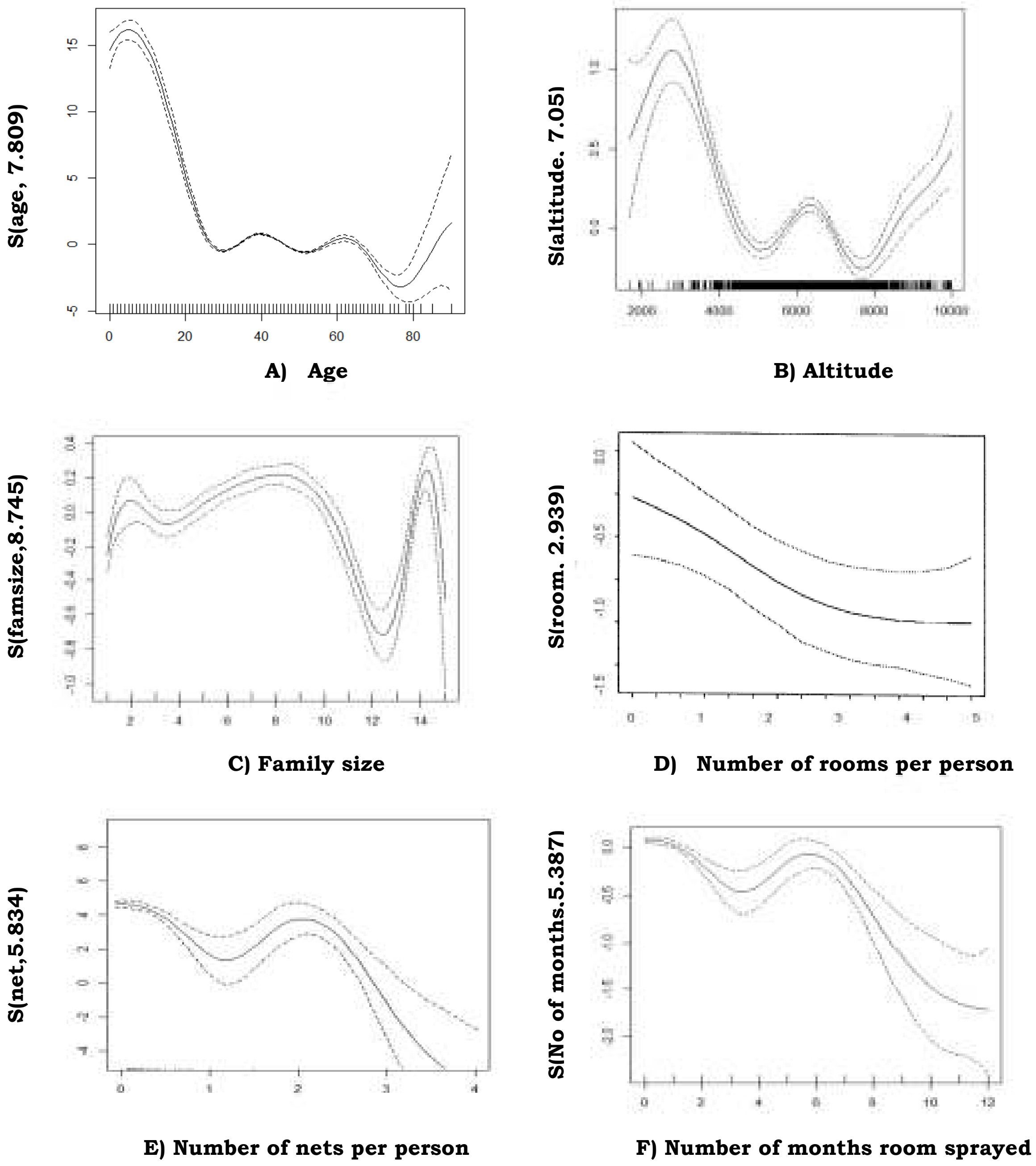
Figure 8. 4: Log odds associated with rapid diagnosis test and main material of floor with gender

In addition to parametric effects, there were effects which were handled non-parametrically to the model. Therefore, age, altitude, family size, total number of rooms per person, total number of nets per person and number of months the room sprayed have been fitted as a smooth. The result in Table 8.3 shows that age, altitude, family size, total number of rooms per person, total number of nets per persons and number of months the room sprayed had a significant effect on malaria RDT result. The smooth term for these effects has been presented in Figure 8.5. The figure suggests that age, altitude, family size, total number of rooms per person, total number of nets per person and number of months the room sprayed effects departs dramatically from linearity.

Table 8. 3: Approximate significance of the smooth terms

Source	Edf*	F-value	P-value
S(age)	7.809	461.1	<.0001
S(altitude)	7.050	39.25	<.0001
S(family size)	8.745	25.07	<.0001
S(total number of room)	2.939	24.56	<.0001
S(total number of nets)	5.834	15.62	<.0001
S(no month room sprayed)	5.387	16.01	<.0001

* Estimated degree of freedom



A) Age

B) Altitude

C) Family size

D) Number of rooms per person

E) Number of nets per person

F) Number of months room sprayed

Figure 8. 5: Smoothing components for malaria RDT with A) age, B) Altitude, C) Family size, D) Total number of rooms per person, E) Total number of nets per person and F) number of months room sprayed

Figure 8.5 gives the estimated smoothing components for malaria RDT result with A) age, B) altitude, C) family size, D) total number of rooms per person, E) total number of nets per person and F) number of months room sprayed. In each panel, the smooth line is the estimated trend from a generalized additive mixed model for the model with spherical Gaussian covariance structure. Figure 8.5a shows the estimated smooth function of age ($\hat{f}(age)$) and its 95% confidence interval. The y-axis represents the effect of the age term, where s is a smoother term and the number in parentheses is the estimated degrees of freedom (edf). Furthermore, the figure suggests that the malaria RDT result is higher at early age, i.e., increased during the first five years of life and then steadily decreased afterwards. The test statistic was 461.1 with 7.809 degrees of freedom, providing strong evidence (p-value < 0.0001) against the assumption that age is linearly associated with malaria RDT result (Table 8.3). Figure 8.5b shows the estimated smooth function for altitude. Larger edfs value in the figure (7.05) corresponds to increasingly nonlinear relationships. Moreover, the malaria RDT result is higher for the first 3000 meters then starts to decrease.

In addition to this, family size had significant effect on malaria RDT test result (Table 8.3). The estimated smooth function for family size is presented in Figure 8.5c. The result in figure shows that edf is 8.745, which shows increasing nonlinear relationship. Moreover, the F–value is 25.07 with p-value <.0001 suggested that family size is not linearly associated with malaria RDT test result. The other significant results were found to be total number of rooms, total number of nets number of months the room sprayed with anti-mosquito. The estimated degrees of freedom are 2.939, 5.834 and 5.387 respectively. These figures suggested nonlinear relationship with malaria RDT result.

8.5 Summary and discussion

The result in this study using GAMM model with nonparametric age, altitude, total number of rooms, total number of nets, family size and number of months room sprayed presented in the above section. The result from this study supports the results from the previous models fitted. In addition to this, the results gave more insight regarding the distribution of age, altitude, total number of rooms, total number of nets, family size and number of months room sprayed. The results from the nonparametric part of the model confirm that malaria RDT test result is high for children. Moreover, persons with more mosquito nets and more number of rooms have greater chance to reduce the risk of malaria. Furthermore, with the correct use of mosquito nets, indoor residual spray and other preventative measures, like having more rooms in a house, the incidence of malaria could be decreased. In addition to this, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria. To provide clean drinking water, proper hygiene and maintaining the good condition of a house is essential in controlling the transmission of malaria. With other control measures, including creating awareness about the use of mosquito nets, indoor residual spraying and malaria transmission, the number of malaria cases can be reduced.

Chapter 9

Using Rasch modeling to re-evaluate malaria Rapid Diagnosis Test analyses

9.1 Introduction

In the previous chapters, malaria rapid diagnosis test data was reviewed and fitted using different parametric and semiparametric statistical methods. These methods are: multiple correspondence analysis, the generalized linear model (Survey logistic), generalized linear mixed models (GLMMs), spatial statistics method, joint models and semiparametric modes. These methods were used to identify the association between malaria RDT result and socio-economic, demographic and geographic factors. These models provide a powerful tool for modelling the relationship between a response variable and covariates. Using these models, it was possible to identify socio-economic, demographic and geographic factors (variables) which have effect on malaria RDT result. The purpose of the current chapter is to confirm the results from the previous models using the Rasch model. The use of Rasch model seeks to answer questions like which items are biased and its source, which items define the trait to be measured, and which individuals are properly identified by the items that define the trait. Furthermore, the objective is to test how well the observed data fit the expectation of malaria RDT result model. Moreover, this model helps to identify if a person's measure on any trait is a simple function of their ability and the items difficulty.

Item response theory (IRT) is paradigm for the design, analysis and scoring of tests, questionnaires and similar instruments, measuring abilities, attitudes or other variables. Item response models (IRM) are a class of probabilistic models that explains the response of a person to a set of items. IRT concerns models and methods where the responses to variables are assumed to depend on nonmeasurable respondent characteristics and on item characteristics. The responses to the items (generally binary or polytomous ordinal variables) and the latent trait are linked nonlinearly. As a link function, the logistic function is often used. IRT models consider a unidimensional latent trait (Van der Linden and Hambleton, 1997). The responses to items are influenced by a unidimensional variable characterizing the individuals. To perform IRT models, general statistical software packages, like Stata, R, or SAS, allow estimating parameters of IRT models in the scope of generalized linear mixed models. In addition to these software, RUMM software also can be used. The literature on the item response theory is presented in many research works and books (De Boeck and Wilson., 2004, Hardouin, 2007, Matschinger, 2006, Rizopoulos, 2006, Skrondal and Rabe-Hesketh, 2004, Van der Linden and Hambleton, 1997, Weesie, 2000).

The Rasch model is a mathematical formula that specifies the form of the relationship between individuals and the items that operationalize one trait. The Rasch model assumes that item responses are governed by individual's position on the underlying trait and item difficulty. As implied by the theory’s name, item responses are modelled rather than sum total responses. The model makes no allowance for deliberate or unconscious deception, guessing, or any other variable that might impinge on the responses provided. Therefore, the Rasch model is the best known model using IRT for binary variables because it has useful property. Some of the properties of Rasch model includes the following (Fitz-Gibbon, 2000).

- The abilities of individuals and difficulties of items are along the same scale so that abilities and difficulties can be compared.
- The Rasch model produces item difficulty levels independent of examinee samples and individual abilities independent of the particular test administered.

- Sufficient statistics exist, i.e., all the information about the ability of individual on a given dimension is contained in the number of correct responses
- The model is a theoretical model and is relatively simpler than other logistic models. Therefore, it is less expensive and easier to apply in solving practical measurement problems.

Furthermore, the score in the model is a sufficient statistic on the latent trait and can be computed easily by summing the responses to all the items. Therefore, all the individuals with the same score have the same estimation of the latent trait.

The aim of this chapter is to review Rasch model and then fit them to malaria RDT result data. Application of the Rasch model, a brief overview of the model for Item response theory is provided. This chapter is organized as follows. An overview of Rasch models is presented in Sections 9.2 and 9.3. The Rasch model is fitted to malaria RDT data in Section 9.4. Summary and discussion of the chapter is given in section 9.5.

9.2 Rasch models

Item response theory (IRT)

Notations

In discussing dichotomous items with a positive response coded 1 and a negative response coded 0, the following notations will be used.

- N is the number of individuals;
- J is the number of items;

- X_{nj} is the random variable representing the response of the n^{th} individual ($n = 1, \ldots, N$) to the j^{th} item ($j = 1, \ldots, J$), and x_{nj} is the realization of this variable;
- $S_n = \sum_{j=1}^{J} X_{nj}$ is the random variable, containing the score of the n^{th} individual, and its realization $s_n = \sum_{j=1}^{J} x_{nj}$;
- N_s is the number of individuals with a score equal to s;
- θ_n is the value of the latent trait for the nth individual ($n = 1, \ldots, N$); and
- $y = (y_j)$, $j = 1, \ldots, Y$ is a vector of size Y composed of the elements y_j.

To use IRT, certain assumptions have to be considered. These assumptions are:

- One of the assumption is unidimensionality: the responses to the items depend on only one latent trait, θ, to characterize the individuals;
- Monotonicity: the probability $Pr(X_{nj} = 1|\theta)$ is a monotone nondecreasing function in θ;
- Local independence: the variables $X_{nj}$and X_{nk} with $j, k = 1, \ldots, J$, and $j \neq k$ are independent conditionally to θ.

Model estimation

In the Rasch model, a set of fixed effects $\theta_n, n = 1, \ldots, N$ or a set of random variables can be considered as latent traits. But, using the Rasch model, the estimations of the parameters obtained by maximum likelihood are not consistent. Therefore, the use of conditional maximum likelihood (CML) method, gives possible solution as a better way to obtain consistent estimate of the parameters. But, for random effects, the parameters can be estimated by the marginal maximum likelihood (MML) method (Ghosh, 1995, Andersen, 1970, Molenaar, 1983).

The item response functions (IRFs) was defined by considering the latent trait as a set of fixed effects (Molenaar, 1983). Therefore, the Rasch model can be specified as

$$\Pr\left(X_{nj} = x_{nj} \middle| \theta_n, \delta_j\right) = \frac{exp\left\{x_{nj}(\theta_n - \delta_j)\right\}}{1 + \exp\left(\theta_n - \delta_j\right)}, j = 1, \dots, J \qquad (9.1)$$

where the δ_j parameters represents the difficulty of the j^{th} item (difficulty parameter); the probability $\Pr\left(X_{nj} = x_{nj} \middle| \theta_n, \delta_j\right)$ decreases, for a given value θ_n, as the value of this parameter increases (Fisher and Molenaar, 1995).

The Rasch model is composed of N parameters $\theta_n (n = 1, \dots, N)$ and of J parameters δ_j $(j = 1, \dots, J)$. The likelihood of the n^{th} individual is given by the following equation

$$L_n(\delta, \theta_n | X_n) = \left(X_{nj} = x_{nj} \middle| \theta_n, \delta_j\right)$$

with $X_n = \left(x_{nj}\right) j = 1, \dots, J$ and $\delta = \left(\delta_j\right) j = 1, \dots, J$. This equation is appropriate under local independence assumption. The conditional maximum likelihood (CML) consists of estimating the difficulty parameters conditionally to the score S_n. The equality

$$\Pr(X_n = x_n / \theta_n, \delta, S_n = s_n) = \frac{\exp(-\sum_{j=1}^{J} x_{nj}\delta_j)}{\gamma_{S_n}(\delta)} = \Pr(X_n = x_n | \delta, S_n = s_n) \qquad (9.2)$$

is independent of the parameters θ_n $(n = 1, \dots, N)$. The denominator also known as the gamma function $\gamma_{S_n}(\delta)$ is defined by

$$\gamma_{S_n}(\delta) = \sum_{y \epsilon \Omega / \sum_{j=1}^{J} y_j = 0} \exp\left(-\sum_{j=1}^{J} y_j \delta_j\right)$$

with Ω, is the set of possible vectors y=(y_j) $j = 1, \dots, J$, with values of 0 and 1 (Ayala, 2009, Van der Linden and Hambleton, 1997).

Maximizing the conditional likelihood (9.2) gives

$$L_c(\delta/\boldsymbol{x},\boldsymbol{s}) = \prod_{n=1}^{N} \Pr(X_n = x_n|\delta, S_n = s_n)\ . \qquad (9.3)$$

For this estimate, an identifiably constraint is necessary. But, the difficulty parameter can be fixed to 0 or the sum $\sum_{j=1}^{J} \hat{\delta}_j$ is fixed to 0. The null score ($s_n = 0$) or a perfect score ($s_n = J$) does not provide any information. As a result, it cannot be used to estimate the difficulty parameters (Molenaar, 1983).

All individuals with score S_n have the same estimation θ_n for only $J + 1$ different parameters θ_n which can be estimated. For all $n, n' = 1, \dots, N, s_n = s_{n'} \Rightarrow \theta_n = \theta_{n'}$. The value of θ_n parameters with $s_n = s(s = 0, \dots, J)$ is denoted by θ_s.

As presented in (Hoijtink and A. Boomsma, 1995), the estimation of θ_s parameters by maximizing the likelihood conditionally to CML estimations of the difficulty parameters are biased and cannot be estimated when $s = 0$ or $s = J$ (Molenaar, 1983, Ayala, 2009, Hardouin, 2007). Furthermore, the weighted likelihood estimators of the θ_s parameters are unbiased. Therefore, the equation can be obtained by maximizing the quantities

$$\hat{\theta}_s = \max_{\theta} \frac{\exp(s^{\theta})}{\prod_{j=1}^{J} 1 + \exp\left(\theta - \hat{\delta}_j\right)} \sqrt{I(\theta)}, s = 0, \dots, J$$

with $I(\theta)$, the information function, defined by

$$I(\theta) = \sum_{j=1}^{J} \frac{\exp\left(\theta - \hat{\delta}_j\right)}{\left\{1 + \exp\left(\theta - \hat{\delta}_j\right)\right\}^2} \qquad (8.4)$$

The distribution of the latent trait θ is assumed as a Gaussian distribution with parameters (μ, σ^2) denoted $(\theta/\mu, \sigma^2)$ for model with random effects (Hardouin, 2007). The IRF of the j^{th} item under the Rasch model is written as

$$\Pr(X_{nj} = x_{nj}) = \frac{exp\{x_{nj}(\theta - \delta_j)\}}{1 + \exp(\theta - \delta_j)}.$$

The marginal likelihood is

$$L_M(\delta, \mu, \sigma^2/X) = \prod_{n=1}^{N} \int_{-\infty}^{+\infty} \prod_{j=1}^{J} \Pr(X_{nj} = x_{nj}/\theta; \delta_j)\, G(\theta/\mu, \sigma^2) d\theta. \qquad (8.5)$$

To obtain consistent estimator of the parameters $\delta_j\ (j = 1, \ldots, J), \mu$ and σ^2 equation 8.5 have to be maximized. For such purpose, an indentifiability constraint is used, i.e. $\mu = 0$. Using all individuals in the estimation process, random effect estimations of all the θ_s parameters can be obtained where $(s = 0, \ldots, J)$. The estimations of the θ_n parameters are obtained by approximating the posterior mean of the latent trail for each individual as

$$\hat{\theta}_n = \frac{\int_{-\infty}^{+\infty} \theta G(\theta/\hat{\mu}, \hat{\sigma}^2) \prod_{j=1}^{J} Pr(X_{nj} = x_{nj}/\hat{\delta}_j, \theta) d\theta}{\int_{-\infty}^{+\infty} G(\theta/\hat{\mu}, \hat{\sigma}^2) \prod_{j=1}^{J} Pr(X_{nj} = x_{nj}/\hat{\delta}_j, \theta) d\theta}.$$

In the Rasch model, the individuals who have the same score s have equal posterior mean of the latent trait. Therefore, the value is equal to $\hat{\theta}_s$. Moreover, the posterior means are also referred to as empirical Bayes prediction (Rabe-Hesketh et al., 2004, Skrondal and Rabe-Hesketh, 2004).

9.3 Tests for Rasch models

For tests using Rasch model, there are different methods. These methods include Andersen Likelihood-ratio Z test, Splitting test, First-order test, U-test, outfit and infit indices.

In the Rasch model with fixed effects, the Andersen Z test allows testing the assumptions that the estimations of the difficulty parameters are the same, whatever the level of the latent trait (Andersen, 1970). To perform the test, the sample has to be divided into G groups, as a function of the score s_n, and the difficulty parameters are estimated in each of these groups.

Let $ll_C(\hat{\delta})$ be the conditional log-likelihood obtained in the sample and $ll_c^{(g)}(\hat{\delta}^{(g)})$ the conditional log-likelihood obtained in the g^{th} group, $g = 1,\dots,G$. The statistic

$$Z = -2\{ll_c(\hat{\delta})\} + 2\sum_{g=1}^{G} ll_c^{(g)}(\hat{\delta}^{(g)})$$

follows, under the null assumption, a χ^2 distribution with $(J-1)(G-1)$ degrees of freedom.

To make a fair comparison, it is important to rely on the test data. If some of the items used as a criteria measure, then the test can be constructed on the other items. For only one item, the technique is called splitter-item technique (Molenaar, 1983). The splitting test consists of splitting the sample as a function of the responses to one given item. For the two groups, the equality of estimations is realized using the Andersen test. A graphical representation of the estimations of the parameters allows detecting the splitter items that give different estimations of the difficulty parameters of the remaining items. Special analysis is needed for the items that have difficulty parameters greater in the group of positive responses than in the group of negative responses.

Using Rasch model, the first-order tests allow testing the fit of the data to the model. The first order tests are sensitive to the nonrespect of the monotonicity assumption.

Let N_{gj} be the number of individuals in the g^{th} group, $g = 1,\dots,G$, these values have positive response to the j^{th} item. Furthermore, the expectation of the number under the Rasch model is represented by $\widehat{N}_{gj}$.

Suppose $d_{gj} = (N_{gj} - \widehat{N}_{gj})$ and $d_g = (d_{gj})$ $j = 1, \ldots, J$. The first-order statistic under the contribution of the g^{th} group is given by

$$T_g = d_g' V_g^{-1} d_g. \tag{8.6}$$

Therefore, from the expression (8.6) V_g is matrix of weights. In the literature, there exist several first-order statistics. This values depend on the nature of the latent trait, on the estimations of $\widehat{N}_{gj}$, and on the used matrix V_g (Fisher and Molenaar, 1995).

First-order tests for the Rasch model with fixed effects

The Wright–Panchapakesan test is based on the estimations

$$\widehat{N}_{gj} = \sum_{s \in I_g} N_s \frac{\exp(\hat{\theta}_s - \hat{\delta}_j)}{1 + \exp(\hat{\theta}_s - \hat{\delta}_j)}$$

where I_g is the set of scores associated with group g. The matrix V_g is a diagonal matrix where the diagonal elements are

$$\epsilon_{gjj} = \sum_{s \in I_g} N_s \hat{\pi} W P_{sj} (1 - \hat{\pi} W P_{sj}), j = 1, \ldots, J.$$

Therefore, the Wright–Panchapakesan statistic Y is given by $Y = \sum_{g=1}^{G} T_g$. This Statistic follows χ^2 distribution with $(G-1)(J-1)$ degrees of freedom under the null assumption (Wright and Panchapakesan, 1969). In the construction of the statistic there were some logical errors. These logical errors were pointed out by (Van den Wollenberg, 1982) and discouraged its use, especially for small J.

In the R_{1c} test, N_{gj} is estimated by

$$\widehat{N}_{gj} = \sum_{s\epsilon I_g} N_s \frac{\exp\left(-\hat{\delta}_j\right)\gamma_{s-1}\left(\hat{\delta}^{(-j)}\right)}{\gamma_s^{(\hat{\delta})}}$$

where $\hat{\delta}^{(-j)} = \left(\hat{\delta}_k\right)\ k = 1, \ldots, j-1, j+1, \ldots, J.$

Furthermore, the V_g matrix is composed of $\mathrm{e}_{gjj} = \widehat{\mathrm{N}}_{gj}$ for the J diagonal elements ($j = 1, \ldots, J$) and

$$\mathrm{e}_{gjk} = \sum_{s\in I_g} \mathrm{N}_s \frac{\exp\left(-\hat{\delta}_j\right)\exp\left(-\hat{\delta}_k\right)\gamma_{s-2}\left(\hat{\delta}^{(-j,k)}\right)}{\gamma_s^{(\hat{\delta})}}. \qquad (8.7)$$

Expression (8.7) is working for the off diagonal elements ($\mathrm{j} = 1, \ldots, \mathrm{J}$, $\mathrm{k} = 1, \ldots, \mathrm{J}$, $\mathrm{j} \neq \mathrm{k}$ with $\hat{\delta}^{(-j,k)} = \left(\hat{\delta}_l\right)_{l=1,\ldots,J,l\neq j,l\neq k}$. By definition, $\forall \mathrm{j}, \mathrm{k} = 1, \ldots, \mathrm{J}\ \ \mathrm{e}_{ljk} = 0$. Under the null assumption, $\mathrm{R}_{1c} = \sum_{g=1}^{G} \mathrm{T}_g$ follows, a χ^2 distribution with $(\mathrm{G}-1)(\mathrm{J}-1)$ degrees of freedom. R_{1c} is approximated by Q_1 statistic (Glas and Verhelst, 1995). The Q_1 statistics is $\mathrm{Q}_1 = \frac{J-1}{J}\sum_{g=1}^{G} T_g$ and follows χ^2 distribution with $(\mathrm{G}-1)(\mathrm{J}-1)$ degrees of freedom under the null assumption (Van den Wollenberg, 1982).

The R_{1c} statistic is replaced by R_{1m} if we use the Rasch model with a random effect (Glas and Verhelst, 1995). This statistic is calculated using

$$\widehat{\mathrm{N}}_{gj} = \mathrm{N}\sum_{s\in I_g} \exp\left(-\hat{\delta}_j\right)\gamma_{s-1}\left(\hat{\delta}^{(-j)}\right) \int_{-\infty}^{+\infty} \frac{\exp(s^{\infty})}{\prod_{j=1}^{J}\left\{1+\exp\left(\theta - \hat{\delta}_j\right)\right\}} \mathrm{G}(\theta/\hat{\mu}, \hat{\sigma}^2)\mathrm{d}\theta$$

and the V_g matrix is composed of $e_{gjj} = \widehat{N}_{gj}$ for J diagonal elements ($j = 1, \ldots, J$) and

$$e_{gjk} = N\sum_{s\in I_g} \exp\left(-\hat{\delta}_j\right)\exp\left(-\hat{\delta}_k\right)\gamma_{s-2}\left(\hat{\delta}^{(-j,k)}\right) \int_{-\infty}^{+\infty} \frac{\exp(s^{\theta})}{\prod_{j=1}^{J}\left(1+\exp\left(\theta - \hat{\delta}_j\right)\right)} \mathrm{G}(\theta/\hat{\mu}, \hat{\sigma}^2)\mathrm{d}\theta$$

for the off-diagonal elements ($j = 1,\ldots,J, k = 1,\ldots,J, j \neq k$). For s = 1, let the off-diagonal elements equal 0.

The individuals with $s_n = J$ can be used in the MML method. Let

$$c_0 = \frac{\left(N_0 - \widehat{N}_0\right)^2}{\widehat{N}_0}$$

and

$$c_J = \frac{\left(N_J - \widehat{N}_J\right)^2}{\widehat{N}_J}$$

with

$$\widehat{N}_0 = N \int_{-\infty}^{+\infty} \frac{1}{\prod_{j=1}^{J}\left\{1 + exp\left(\theta - \hat{\delta}_j\right)\right\}} G(\theta/\hat{\mu}, \hat{\sigma}^2) d\theta$$

$$\widehat{N}_J = N \int_{-\infty}^{+\infty} \frac{exp\left(J\theta - \sum_{j=1}^{J} \hat{\delta}_j\right)}{\prod_{j=1}^{J}\left\{1 + exp\left(\theta - \hat{\delta}_j\right)\right\}} G(\theta/\hat{\mu}, \hat{\sigma}^2) d\theta$$

The R_{1m} statistics is

$$R_{1m} = c_0 + \sum_{g=1}^{G} T_g + c_J$$

and followsa χ^2 distribution with $G(J-1) - 1$ degrees of freedom under the null assumption.

The contribution of each item to the first-order statistic can be estimated by using the vector

$$\sum_{g=1}^{G} W^{-1/2} d_g$$

where $W^{-1/2}$represents the Cholesky decomposition of the positive-definite matrix $W^{-1}(W^{-1/2'} W^{-1/2}) = W^{-1}$. The j^{th} element of this vector represents the contribution of the j^{th} item to the first-order statistic, and follows, under the null assumption, a χ^2 distribution with $G-1$ degrees of freedom (Van den Wollenberg, 1982).

Using Rasch model, the equality of the mean slopes of the item characteristic curves can be tested using U test. ICCs are graphical representations of the IRF. The CML estimations of the difficulty parameters were used to develop the u test. For this estimation, the sample is divided in three subsamples as a function of the values of the score of the individuals.

From these divided samples, the first subsample is composed of all the individuals with a score inferior or equal to a thresholdc_1. On the other hand, the third subsample of all the individuals with a score superior or equal to a threshold c_2. Lastly, the second subsample of the remaining individuals, c_1 and c_2 are computed as follows

$$\sum_{\mathrm{s}=1} \mathrm{N}_{\mathrm{s}} \geq 25\%\mathrm{N} \text{ and } \sum_{\mathrm{s}=\mathrm{c}_2} \mathrm{N}_{\mathrm{s}} \geq 25\%\mathrm{N}$$

The statistic U_{j}, $\mathrm{j} = 1, \dots, \mathrm{J}$, is equal to

$$\mathrm{U}_{\mathrm{j}} = \frac{\mathrm{z}_1 - \mathrm{z}_2}{\sqrt{\mathrm{c}_1 + \mathrm{J} - \mathrm{c}_2}}$$

with

$$z_1 = \sum_{s=1}^{c_1} \frac{\pi_{sj} - \hat{\pi}_{sj}}{\sqrt{N_s \hat{\pi}_{sj}(1 - \hat{\pi}_{sj})}} \text{ and } z_2 = \sum_{s=c_2}^{c_1} \frac{\pi_{sj} - \hat{\pi}_{sj}}{\sqrt{N_s \hat{\pi}_{sj}(1 - \hat{\pi}_{sj})}}$$

where π_{sj} is the observed proportion of positive responses to the j^{th} item. Moreover, for the individuals with a score $s_n = s$, $\hat{\pi}_{sj}$ is an estimation of this quantity under the Rasch model ($\hat{\pi}wP_{sj}$ or $\hat{\pi}wRC_{sj}$). Furthermore, U_j statistic follows the assumption of equality of the slope of the item j to the mean of the slopes of the other items of the model. This statistic follows standardized normal distribution (Molenaar, 1983, Glas and Verhelst, 1995).

The other method of the test is OUTFIT and INFIT indices. The OUTFIT and INFIT indices are commonly used like indices of fit of the items and of the individuals.

The residuals used for the two indices are

$$r_{nj} = x_{nj} - \hat{\pi}_{s_{nj}}$$

The OUTFIT index for the j^{th} item is

$$OUTFIT_j = \frac{1}{N}\sum_{n=1}^{N}\frac{r_{nj}^2}{\hat{\pi}_{nj}\left(1 - \hat{\pi}_{nj}\right)}$$

The INFIT index for the j^{th} item is

$$INFIT_j = \frac{\sum_{n=1}^{N} r_{nj}^2}{\sum_{n=1}^{N}\left(1 - \hat{\pi}_{nj}\right)\hat{\pi}_{nj}}$$

Using $E\left(OUTFIT_j\right) = E\left(INFIT_j\right) = 1$, the OUTFIT and INFIT indices can be standardized. Therefore,

$$V\left(OUTFIT_j\right) = \frac{1}{N^2}\sum_{n=1}^{N}\frac{C_{ni}}{W_{ni}^2}$$

$$V\left(INFIT_j\right) = \frac{\sum_{n=1}^{N}(C_{ni} - W_{ni}^2)}{\sum_{n=1}^{N} W_{ni}^2}$$

where W_{ni} is the variance of x_{ni} and C_{ni} is the 4th order moment of x_{ni}. Since $OUTFIT_j$ and $INFIT_j$ are sum of squares, using the transformations

$$OUTFIT_j^* = \frac{3\left(\sqrt[3]{OUTFIT_J} - 1\right)}{\sqrt{V(OUTFIT_j)}} - \frac{\sqrt{V(OUTFIT_j)}}{3}$$

and

$$INFIT_j^* = \frac{3\left(\sqrt[3]{INFIT_J} - 1\right)}{\sqrt{V(INFIT_j)}} - \frac{\sqrt{V\left(INFIT_j\right)}}{3}.$$

It is possible to obtain indices whose distributions are close to a standardized Gaussian distribution. The outliers can be detected using these two indices (Molenaar, 1983, Linacre and Wright, 1994).

9.4 Application of Rasch models

In this Chapter, the malaria data was fitted to the Rasch model using the RUMM2030 software. The objective is to test how well the observed malaria RDT result data fit the expectations of the model. To check the accuracy of the model, the overall fit statistics can be considered. These methods are related to item–person interaction statistics (Fisher and Molenaar, 1995). Using these methods, it can be transformed to approximate a *z* score. The Z scorerepresents standardized normal distribution. Furthermore, if the items (socio-economic, demographic and geographic variables) and persons (RDT, indoor residual spraying and use of mosquito nets) fit the model, it is expected to see a mean of approximately zero and a standard deviation of one. The other method is an item–trait interaction statistic. This statistics is reported as a chi-square and reflects the property of invariance across the trait. Therefore, if the chi-square is significant, then it means the hierarchical ordering of the items varies across the trait. This means that the value compromises the required property of invariance.

Besides these overall summary fit statistics, individual person and item-fit statistics are presented, both as residuals and as a chi-squared statistic. Therefore, residuals between $\pm$ 2.5 are deemed to indicate adequate fit to the model. In addition to this, misfit to the model can also be viewed graphically where observed model fit is groups of responders across class intervals. The graph can be plotted against the expected model curve (item characteristic curve, ICC). Items with good fit will show each of the group plots lying on the curve. But, plots which are steeper than the curve would be considered to be over-discriminating and those flatter than the curve considered being under-discriminating. The summed chi-square within each group provides the overall chi-square for the item. The summary of overall chi-square for items is summed given as the item trait interaction statistic. In the analysis, Bonferroni corrections are applied to adjust the chi-squared p-value (Tennant et al., 2004). This is done to account for multiple testing.

Furthermore, examination of person fit is important for item fit. If there are few respondents who deviate from model expectation, this may cause significant misfit at the item level. In case of validation of a scale, the misfit runs the risk of discarding the scale. But, the scale would be more appropriate to find out why a few respondents may be responding in a way different to others. Indication of how well-targeted the items are in the sample can be obtained from the comparison of the mean location score obtained for the persons with that of the value of zero set for the items. For a well-targeted measure the mean location would also be around the value of zero. The positive mean value indicate that the sample as a whole was located at a higher level than the average of the scale. On the other hand, a negative value would suggest the opposite.

From the analysis, an estimate of the internal consistency reliability of the scale can be obtained. This is obtained based on the Person Separation Index (PSI) where the estimates on the logit scale for each person are used to calculate reliability (Molenaar, 1983). To see the improvement of scale construction, the sources of deviation from model expectation can be examined. Good fitting model can be obtained for each of the items if respondents with high levels of the attribute being measured would endorse high scoring responses. But, individuals with low levels of the attribute would consistently endorse low scoring responses. In Rasch analysis, thresholds can be used to indicate ordered set of response thresholds for each of the items. The term threshold refers to the point between two response categories where either response is equally probable.

To investigate responses to an item, the category probability curves can be inspected. For a well-fitting item, it is expected across the whole range of the trait to be measured. In addition to this, each response option would systematically take turns showing the highest probability of endorsement. Disordered thresholds indicate the most common source of item-misfit, i.e., the failure of respondents to use the response categories in a manner consistent with the level of the trait being measured. Disordered thresholds occur for respondents with difficulty consistently discriminating between response options. The problem can occur for too many response options and when the labelling of options is potentially confusing. To overcome this problem, collapsing of categories where disordered thresholds occur improves overall fit to the model.

Differential item functioning (DIF) is the other issue that can affect model fit in the form of item bias. This occurs when different groups within the sample respond in a different manner to an individual item. This can occur despite equal levels of the underlying characteristic being measured. From the analysis, two types of DIF may be identified. DIF's also shows a consistent systematic difference in their responses to an item. This is referred as uniform DIF. When there is non-uniformity in the differences between the groups then this is referred to as non-uniform DIF. When non-uniformity is detected, the problem can be remedied by splitting the file by group and separately calibrating the item for each group. But, there is little that can be done to correct the problem. Therefore, it is often necessary to remove the item from the scale.

In RUMM, the statistical and graphical methods can be used to detect the presence of DIF. Analysis of variance is conducted for each item comparing scores across each level of the person factor and across different levels of trait. Uniform DIF is indicated by a significant main effect for the person factor, and the presence of non-uniform DIF is indicated by a significant interaction effect.

A principal component analysis (PCA) of residuals can be used to detect the sign of multidimensionality when there are issues of threshold disordering and DIF. If there is no meaningful pattern of residuals, the result suggests the assumption of local independence. This leads to unidimensionality of the scales. Moreover, the subsets of items can be determined by allowing the factor loading of the first residual. The use of paired t-test helps to see if the person estimate derived from the subsets significantly differs from that derived from all items. Furthermore, violation of the assumption of local independence can be detected if the person estimate is found to differ between the subset and the full scale (De Boeck and Wilson., 2004, Fitz-Gibbon, 2000, Wright and Panchapakesan, 1969).

For the RUMM analysis, baseline household cluster malaria survey which was conducted by The Carter Center in 2007 was used. For the study, malaria RDT result, indoor residual spray and use of mosquito nets were used as person items. The other variables which were considered as items are main source of drinking water, time to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms, main material of the room's wall, main material of the room's roof, main material of the room's floor, total number of nets, region, altitude, age and family size. For the analysis, altitude, age and gender were categorized to be appropriate for the RUMM2030 analysis because RUMM 2030 is appropriate for categorical variables.

The residual mean value for items in the anxiety subscale is .0205 with a standard deviation (SD) of 1.0187. To be a good fit, SD would be expected to be much closer to 1. Since the value is close to 1, the fit is adequate to the model. This result is supported by a non-significant chi-squared interaction of 96.994 with p-value = 0.3491. Therefore, the scale fits the Rasch model. The value of the Person-Separation-Index for the original set of sixteen items with the response categories was 0.832. This result indicates that the scale worked well to separate the persons. The Power of Test-of-Fit is a visual representation of the Person-Separation-Index. It is indicative of the power of the construction to discriminate amongst the respondents. Based on the values, 0.7 is the minimum accepted level of Person-Separation-Index. This value indicates that it is possible to differentiate statistically between two groups of respondents. Furthermore, a value of 0.9 means that we can statistically differentiate between four or more groups. The Person-Separation-Index is also an indicator of how much we can rely on the Fit Statistics. If the Person-Separation-Index is low, then the Fit Statistics that have been obtained may not be reliable as there will be a substantial amount of error surrounding them. If the Person-Separation-Index is high, then the Fit Statistics that have been generated can be deemed to be more reliable. Based on this, because our Person-Separation-Index 0.832, it can be concluded that the fit statistics is reliable.

Figure 9.1 shows the person-item threshold distribution for the original set of items. To find person-item threshold distribution, person and item locations are logarithmically transformed and plotted on the same continuum. For the plot common unit of measurements were termed as logit. The ordinal data was converted as equal-interval data. Furthermore, Figure 9.1 illustrates how person and item locations can be plotted on the same continuum along the x axis. The upper part of the graph represents groups of respondents who have tested for malaria infection and their ability to respond the questions. The lower part of the graph represents the item locations and their distribution. Both respondent’s ability level and item difficulty level are being shown on the same linear scale. Some items are located in the same place in terms of difficulty and this common location is represented as one block on top of another. A lot of item thresholds are clustered around the central locations. The plot endpoints are known as the floor and ceiling of the scale. The respondents that are located outside of the range measured by the scale were not included in the analysis but excluded as extreme scores.

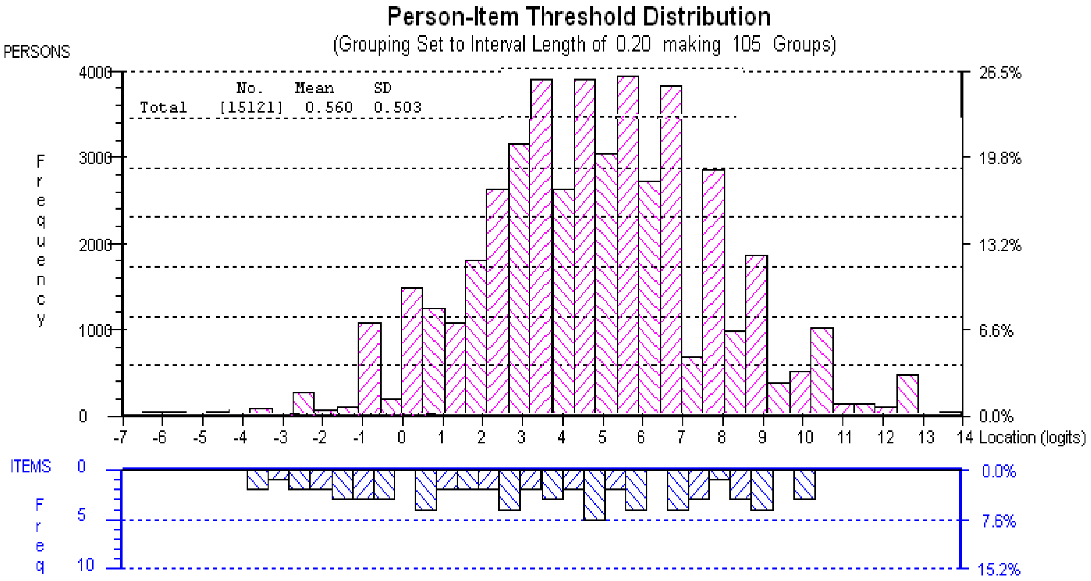


Figure 9. 1: Person-item threshold distribution (16 items)

It can be seen that little information is being derived from those respondents with maximum score (at the top end of the scale). Maximum information for any given item is derived when the respondents have the same logit ability as the item's logit difficulty. Besides the person-item threshold distribution, another useful function of this display screen is the option to look at the location, or 'ability', differences between person factor (RDT result, use of indoor residual spray and use of mosquito nets) groups. Moreover, the statistical relationship between person factors (RDT result, use of indoor residual spray and use of mosquito nets) can be assessed. Whether there is a statistical difference between the person factors groups can be seen using the ANOVA results of the location differences between person factor subgroups. The ANOVA value is given in Table 9.1. The result from the ANOVA analysis reveals that there is statistical difference in ability between malaria RDT result of positive and negative subgroups (p=0.00156). Similarly, there is statistical difference between use of indoor residual spraying and not using (p=0.00327) and between respondents who are using and not using mosquito nets (p=0.006027).

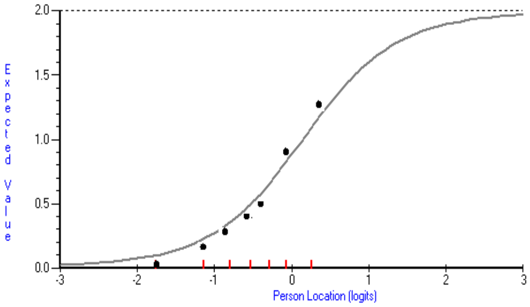
Table 9. 1: ANOVA table for Malaria RDT result, indoor residual spraying and use of mosquito nets

Source	Sum of squares	DF	Mean sum of squares	F-Stat	Prob
Malaria RDT result					
Between	6.28	1	6.28	14.81	0.00156
Within	6408.78	15119	0.42		
Total	6415.06	15120			
Indoor residual spray					
Between	87.75	1	87.75	209.68	0.00327
Within	6327.31	15119	0.42		
Total	6415.06	15120			
Use of mosquito nets					
Between	68.47	1	68.47	163.02	0.006027
Within	6346.59	15119	0.42		
Total	6415.06	15120			

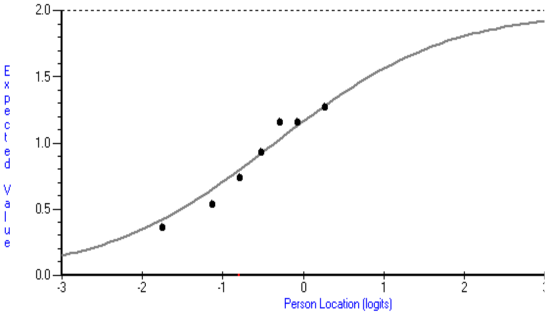
DF = degree of freedom

Targeting and reliability is important that the measures used are appropriately targeted to assess the analysis. The other inspection method is the graphical inspection of the Item Characteristic Curves (ICC). For each item, the ICC was made to examine the fit between expected and observed values. Using the ICC graphical method the average response of persons within each class interval (CI) is represented graphically by a dot and expected values are represented by the solid curve.

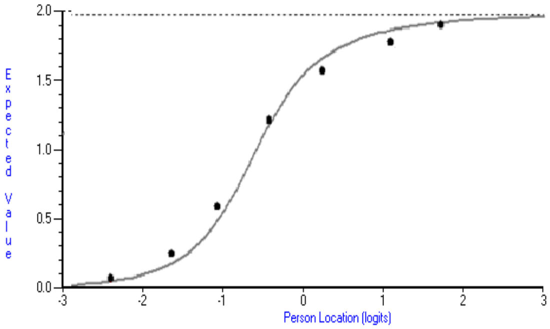
Item characteristic curve (ICC) plot for the sixteen items and three person items are divided into several groupings, or class intervals, of approximately equal size to create contingency tables of expected and observed values. To assess the probability of the degree of divergence between observed and expected values, the chi-square can be derived. Divergence between observed and expected values can occur by chance. Therefore, the number of intervals is determined by the size of the calibration sample. From the plot, the curved line represents the expected scores for the item, and the dots represent the observed scores for the class intervals at the different ability levels. The side of the expected score is represented. The plot can be helpful to observe the behaviour of the variables by the class interval fit (Black Dots) compared with the expected model. The ICC plot for the sixteen items is presented in Figures 9.2 and 9.3. From the two figures, it can be seen that region, age group, availability of electricity, total number of rooms and total number of nets have classic fit pattern. On the other hand, material of roof, wall, availability of television, gender, family size and altitude have marginal under-discrimination pattern, i.e, the response from the lowest group are above what is expected by the model and those for the highest group, are below model expectation. Unlike for the two cases, source of drinking water, distance to fetch water, toilet facility and material for floor have marginal over-discrimination pattern. Thus, the response from the highest group are above what is expected by the model and those for the lowest group, are below model expectation.



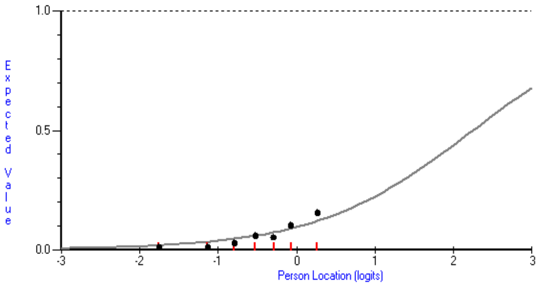
a) Region



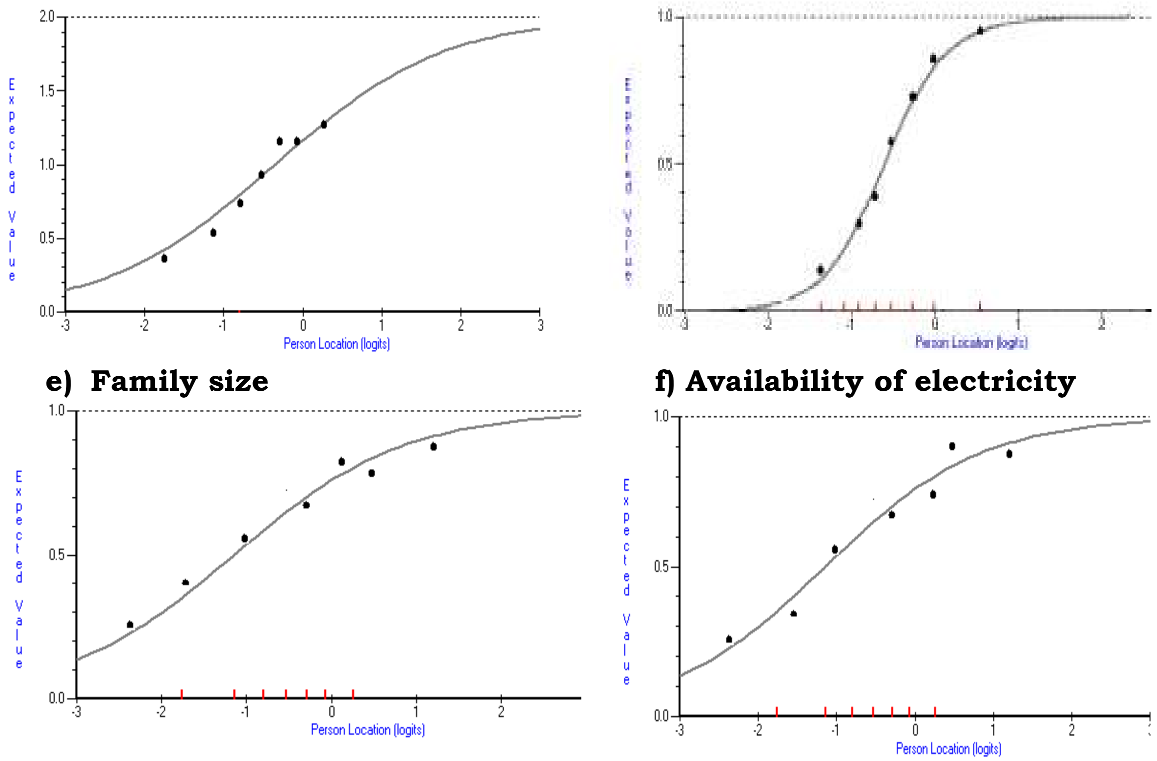
b) Altitude



c) Age group



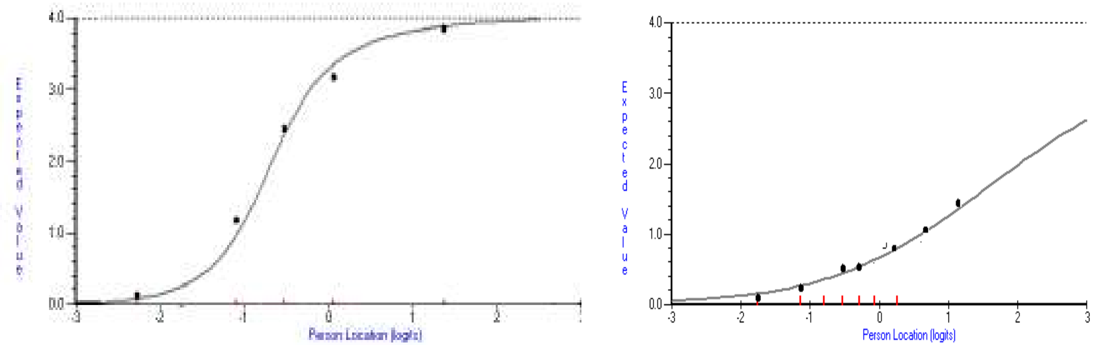
d) gender



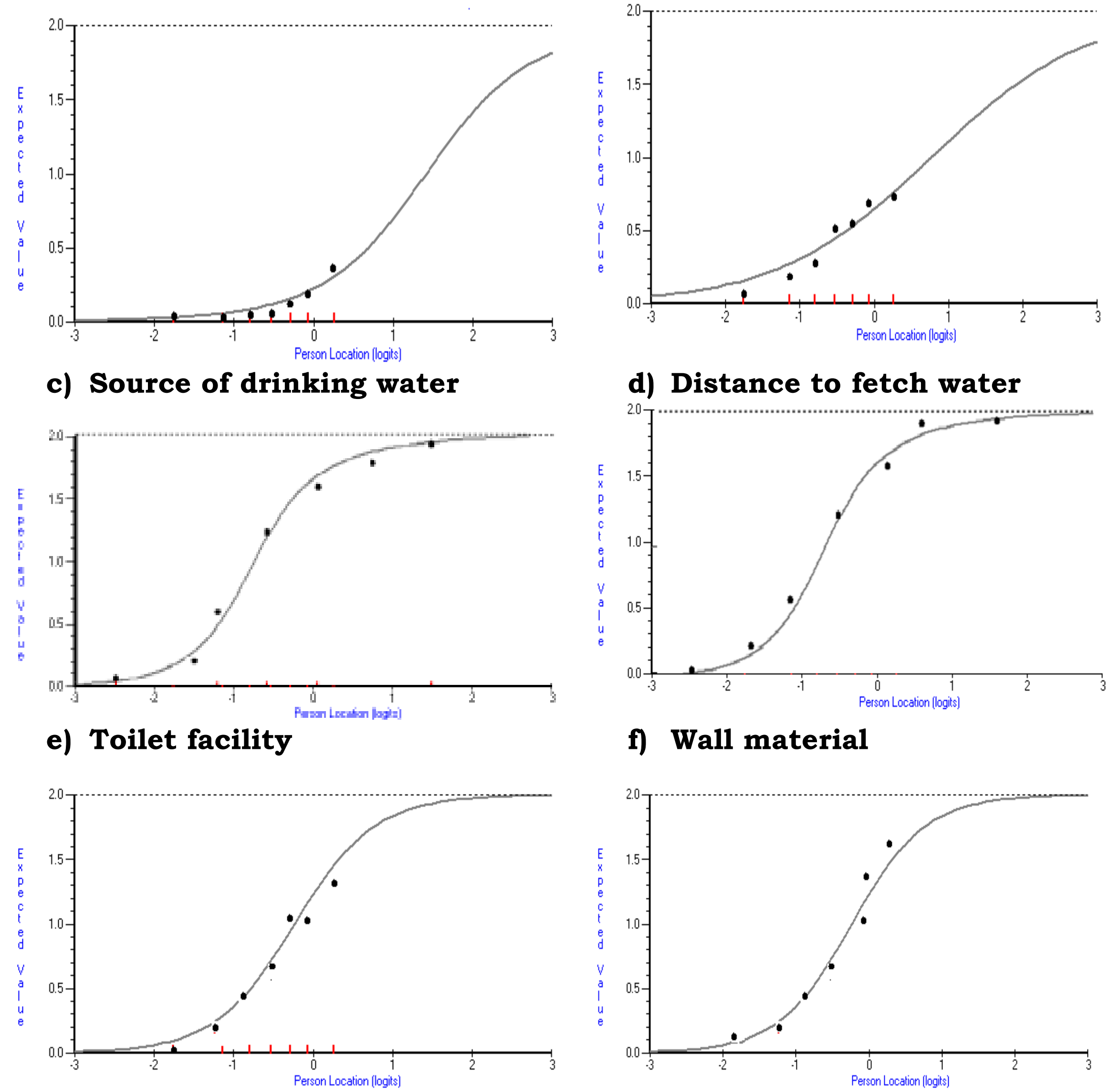
e) Family size **f) Availability of electricity**

g) Availability of radio **h) Availability of television**

Figure 9. 2: ICC of an item for region, altitude, age group, gender, family size, availability of electricity, radio and television



a) Total number of rooms **b) Total number of nets**



c) **Source of drinking water**

d) **Distance to fetch water**

e) **Toilet facility**

f) **Wall material**

g) **Material of roof**

h) **Availability of television**

Figure 9. 3: ICC of an item for total number of rooms, number of nets, source of drinking water, distance to fetch water, toilet facility, material for wall, roof and floor

Another source of misfit in the data could be due to the Differential Item Functioning of certain items. Therefore, DIF can be used to diagnosis the model. For DIF analysis, there are two groups. We consider the two groups of equal status. In the use of DIF the perspective is that there is a standard or main group and that there is a subgroup, sometimes referred to as a focal group, which might have items which are biased. When using DIF analysis, the sample sizes of the two groups should be as close as possible. This is because if the sample sizes are different, and there is DIF, then the estimates will be weighted by the estimates that would be present for group with the larger sample size.

The use of analysis of variance for residuals provides the facility to identify two kinds of DIF: first, uniform DIF and non-uniform DIF. The two-way ANOVA structure involves the class intervals as one of the factors, and the groups as the other factor. Then it is possible to study the main effect of the class intervals, the main effect of the groups and the interaction between the two. The main effect across class intervals is a general test of fit of the responses to the ICC, irrespective of any classification by groups. Items can show fit to the model using this criterion, while showing DIF.

Non-uniform DIF occurs where the observed means of responses in the class intervals of two groups are different systematically. In ANOVA, there is an interaction between the class intervals and the groups. If there is no non-uniform DIF, then uniform DIF can be interpreted directly. Uniform DIF occurs where the observations of responses in the class intervals of two groups are different systematically and are parallel. This means that for the best estimate of locations of persons on the continuum one group tends to have a higher mean than the other group.

Groups can have different means, but some items have DIF. This means that DIF detects an interaction between some items and the rest of the items, not an absolute effect. Suppose an item has DIF. Then suppose a whole set of items that has this characteristic are put together, and they all, individually show DIF in the same direction. Then these items put together would show no DIF, but the mean of one group would be greater than the other.

Table 9. 2: DIF Summary with Bonferroni corrected for malaria RDT result

Item	Class Interval		RDT		Interaction	
	F	P-value	F	P-value	F	P-value
Region	15.182	0.082	4.678	0.290	2.884	0.491
Availability of electricity	41.121	0.074	3.749	0.315	2.918	0.516
Availability of radio	2.534	0.239	1.967	0.089	2.952	0.541
Availability of television	4.951	0.111	4.703	0.043	2.986	0.567
Total number of rooms	3.826	0.214	3.968	0.050	3.054	0.617
Number of nets	1.660	0.240	3.996	0.096	3.088	0.642
Gender	4.724	0.265	4.023	0.143	3.122	0.667
Source of drinking water	4.387	0.290	4.050	0.189	3.157	0.692
Distance to get water	2.686	0.315	4.077	0.235	3.191	0.717
Toilet facility	5.329	0.340	4.104	0.282	3.225	0.743
Wall material	4.746	0.365	4.132	0.328	3.259	0.768
Roof material	1.220	0.390	4.159	0.374	3.293	0.793
Floor material	3.700	0.416	4.186	0.421	3.327	0.818
Family size	5.294	0.441	4.213	0.467	3.361	0.843
Age group	2.685	0.466	4.240	0.513	3.395	0.868

The initial summary of DIF for malaria RDT result, use of indoor residual spraying and use of mosquito nets show misfit across the continuum as evidenced by the class interval, malaria RDT result, indoor residual spraying and use of mosquito nets fit statistics show misfit. These items are item 5 due to malaria RDT, items 5, 6, 7 and 11 (total number of rooms, total number of nets, sex and wall material) due to indoor residual spraying and items 1, 5 (region and total number of rooms) and due to use of mosquito nets.

To resolve this problem, there are suggestions for correction of the significance level in the literature, and a common one is the Bonferroni correction. This is very simple to carry out; the chosen probability value of significance is simply divided by the number of tests of fit. There is some controversy with this correction. In RUMM, both the numbers with correction, and the numbers without correction are provided to give the user discretion in making decisions. It also permits them to report both. Tables

9.2, 9.3 and 9.4 show the ANOVA of residuals after misfitted items has been resolved. Therefore, no item shows any misfit.

Table 9. 3: DIF Summary with Bonferroni corrected for indoor residual spraying

Item	Class Interval		Use of indoor residual spray		Interaction	
	F	P-value	F	P-value	F	P-value
Region	4.672	0.516	3.056	0.214	3.056	0.111
Availability of electricity	8.837	0.541	2.942	0.240	2.942	0.240
Availability of radio	2.593	0.567	2.827	0.290	2.827	0.089
Availability of television	3.582	0.592	2.712	0.315	2.712	0.050
Total number of rooms	1.159	0.265	2.597	0.089	2.597	0.099
Source of drinking water	8.998	0.340	2.253	0.050	1.159	0.290
Distance to get water	4.034	0.365	2.138	0.096	1.084	0.239
Toilet facility	3.472	0.390	2.023	0.099	5.419	0.080
Roof material	4.744	0.441	1.793	0.087	4.034	0.062
Floor material	1.260	0.466	1.678	0.080	3.472	0.037
Family size	3.350	0.491	1.564	0.074	5.362	0.093
Age group	2.772	0.516	1.449	0.068	4.744	0.068
Altitude	3.182	0.541	1.334	0.062	5.003	0.315

Diagnosis and detection of violations of independence can be reflected in the fit of data to the model. Over-discriminating items often indicate response dependence and under-discriminating items. This situation indicates multidimensionality. Response dependence increases the similarity of the responses of persons across items. Therefore, responses are more Guttman-like than they should be under no dependence. Multidimensionality acts as an extra source of variation in the data, and the responses are less Guttman-like than they would be under no dependence. Violations of local independence can be assessed by examining patterns among the standardized item residuals.

High correlations between standardized item residuals indicate a violation of the assumption of independence. A principal component analysis (PCA) of the item residuals provides further information about dependence. After extracting the 'Rasch factor' there should be no further pattern among the residuals. If a PCA indicates a meaningful pattern the scale or test is not unidimensional.

Table 9. 4: DIF Summary with Bonferroni corrected for indoor residual spraying

Item	Class Interval		Use of mosquito nets		Interaction	
	F	**P-value**	**F**	**P-value**	**F**	**P-value**
Electricity	0.671	0.697	2.296	0.130	5.475	0.751
Radio	0.884	0.519	0.933	0.335	5.633	0.189
Television	0.468	0.858	31.943	0.000	0.513	0.825
Number of nets	1.039	0.403	0.597	0.440	0.163	0.992
Source of drinking water	1.044	0.400	1.124	0.290	0.892	0.512
Distance to get water	1.281	0.258	5.049	0.025	0.989	0.439
Toilet facility	1.617	0.128	0.819	0.366	0.487	0.844
Wall material	1.420	0.215	8.606	0.267	0.905	0.581
Roof material	1.472	0.180	8.409	0.282	0.899	0.590
Floor material	1.525	0.145	8.212	0.298	0.893	0.599
Family size	1.578	0.110	8.014	0.313	0.887	0.608
Age group	1.631	0.075	7.817	0.328	0.881	0.617
Altitude	1.683	0.283	7.619	0.343	0.875	0.626

Table 9.5 shows the results of a PCA on a data set. Items are sorted according to their loadings on principal component one (PC1). The table shows that there is meaningful pattern. Therefore, the scale or test is not unidimentional.

Table 9.6 shows the summary of the PCA. The Eigenvalue of 2.42 for the first component is considerably larger than the Eigenvalues for the other components. The first principal component explained 15.14% of the total variance among residuals. This all suggests multidimensionality with items 1 to 16 tapping into a second factor, after the main factor had been extracted.

Table 9. 5: Results of a PCA, items sorted according to their loadings on principal component (PC)1

Item	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Region	-0.09	-0.05	0.02	-0.01	0.03	0.05	-0.16	-0.09
Electricity	-0.06	0.03	0.01	-0.19	-0.04	0.12	-0.09	-0.10
Radio	-0.05	-0.08	-0.01	-0.13	-0.04	-0.96	-0.02	-0.09
Television	-0.04	-0.06	0.01	0.88	-0.05	0.17	-0.07	-0.16
Total Number of Rooms	-0.03	-0.17	0.02	-0.13	0.05	0.10	0.01	0.01
Number of nets	-0.03	0.00	0.19	0.02	0.00	0.07	-0.01	-0.04
Sex	-0.03	-0.01	-0.02	-0.07	-0.03	0.10	-0.04	-0.09
Source of drinking water	-0.01	-0.03	0.01	-0.05	-0.01	0.02	0.98	-0.02
Distance to get water	-0.01	-0.06	0.03	-0.04	0.99	0.04	-0.01	-0.04
Toilet facility	-0.01	0.01	-0.98	-0.01	-0.03	-0.01	-0.01	-0.01
Wall material	0.00	-0.02	0.00	-0.07	-0.11	0.01	0.06	-0.02
Roof material	0.00	0.96	-0.01	-0.05	-0.06	0.08	-0.03	-0.11
Floor material	0.01	-0.11	0.01	-0.12	-0.05	0.09	-0.03	0.97
Family size	0.03	-0.16	-0.03	-0.19	-0.03	0.04	0.00	-0.02
Age group	0.20	0.01	0.01	-0.06	0.00	0.04	0.12	0.01
Altitude	0.98	0.00	0.01	-0.03	-0.01	0.04	-0.01	0.01
	PC9	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**	**PC15**	**PC16**
Region	0.03	0.06	-0.14	-0.04	-0.94	-0.02	-0.18	-0.05
Electricity	0.06	0.20	-0.05	-0.93	-0.04	-0.05	-0.05	-0.09
Radio	-0.10	0.09	-0.01	0.11	0.05	-0.08	-0.03	-0.04
Television	-0.09	0.16	-0.10	0.23	0.01	0.03	-0.07	-0.24
Total Number of Rooms	-0.11	-0.94	0.02	0.19	0.06	-0.02	0.02	0.01
Number of nets	-0.01	0.01	-0.01	0.04	0.02	0.97	-0.04	-0.06
Sex	0.97	0.10	-0.05	-0.05	-0.03	-0.01	-0.09	-0.12
Source of drinking water	-0.03	-0.01	0.06	0.07	0.14	-0.01	0.10	0.00
Distance to get water	-0.02	-0.04	-0.11	0.03	-0.03	0.00	0.00	-0.03
Toilet facility	0.02	0.02	0.00	0.01	0.02	-0.19	-0.01	0.02
Wall material	-0.04	-0.02	0.97	0.04	0.12	-0.01	0.10	0.04
Roof material	-0.01	0.16	-0.02	-0.03	0.05	0.00	0.01	-0.15
Floor material	-0.09	-0.01	-0.02	0.09	0.08	-0.04	0.01	-0.02
Family size	-0.13	-0.01	0.04	0.09	0.05	-0.07	0.07	0.95
Age group	-0.10	-0.02	0.11	0.05	0.19	-0.04	0.94	0.07
Altitude	-0.03	0.03	0.00	0.05	0.09	-0.03	0.18	0.02

Table 9. 6: Summary of the PCA

Code	PC	Eigen	Per cent	CPer cent	StdErr
I0001	Region	2.422	15.14%	15.14%	0.332
I0002	Electricity	1.642	10.26%	25.40%	0.221
I0003	Radio	1.539	9.62%	35.02%	0.204
I0004	Television	1.288	8.05%	43.06%	0.169
I0005	Total Number of Rooms	1.204	7.53%	50.59%	0.158
I0006	Number of nets	1.105	6.91%	57.50%	0.143
I0007	Sex	1.05	6.57%	64.06%	0.137
I0008	Source of drinking water	0.946	5.91%	69.97%	0.121
I0009	Distance to get water	0.879	5.49%	75.46%	0.108
I0010	Toilet facility	0.817	5.10%	80.57%	0.107
I0011	Wall material	0.719	4.50%	85.07%	0.098
I0012	Roof material	0.677	4.23%	89.29%	0.093
I0013	Floor material	0.622	3.89%	93.18%	0.086
I0014	Family size	0.566	3.54%	96.72%	0.08
I0015	Age group	0.483	3.02%	99.74%	0.078
I0016	Altitude	0.041	0.26%	100.00%	0.054

Response dependence occurs when a person's response to an item depends on the person's response to a previous item. Table 9.7 shows the correlations between the standardized item residuals for a data set in which a dichotomous item depend on another dichotomous item. The correlation between item 2 and 8 is 0.41 and is considerably larger than the correlations of other items, which are mostly negative. The table is presented as follows.

Table 9. 7: Correlations between standardized item residuals

Item	**I0001**	**I0002**	**I0003**	**I0004**	**I0005**	**I0006**	**I0007**	**I0008**
Region	1							
Electricity	-0.123	1						
Radio	-0.14	0.186	1					
Television	-0.092	0.386	0.08	1				
Total Number of Rooms	0.134	0.013	-0.272	0.013	1			
Number of nets	-0.343	0.066	0.1	-0.032	-0.315	1		
Sex	-0.052	-0.012	-0.041	-0.014	-0.112	0.107	1	
Source of drinking water	-0.133	0.41	-0.117	-0.201	-0.054	0.116	0.081	1
Distance to get water	-0.151	0.053	0.039	-0.033	0.192	-0.092	0.069	-0.148
Toilet facility	-0.186	0.206	-0.239	-0.063	-0.04	-0.195	0.035	0.108
Wall material	-0.006	-0.008	0.061	-0.006	-0.028	-0.038	0.055	-0.036
Roof material	-0.285	-0.141	-0.356	-0.064	-0.008	-0.185	0.043	0.064
Floor material	-0.065	-0.076	-0.126	-0.057	-0.018	0.01	0.013	-0.01
Family size	-0.161	-0.046	-0.007	-0.07	-0.109	-0.114	0.063	-0.064
Age group	-0.087	0.239	0.117	-0.025	0.02	-0.065	0.214	-0.28
Altitude	-0.151	0.248	0.036	-0.009	0.04	-0.053	0.024	-0.312
	I0009	**I0010**	**I0011**	**I0012**	**I0013**	**I0014**	**I0015**	**I0016**
Distance to get water	1							
Toilet facility	-0.153	1						
Wall material	-0.01	0.032	1					
Roof material	-0.205	-0.066	0.021	1				
Floor material	-0.061	0	0.369	0.099	1			
Family size	-0.125	-0.163	-0.033	-0.198	0.133	1		
Age group	-0.006	-0.099	-0.001	-0.151	0.021	0.014	1	
Altitude	-0.016	-0.079	-0.011	-0.094	0.008	-0.006	0.147	1

9.5 Summary and discussion

The purpose of the chapter was to introduce the Rasch model and to show an application of the model in malaria research. Using Rasch model, according to standard statistical tests, it is possible to use the model to diagnosis the empirical ordering of the categories. The initial descriptive analysis of the frequency distributions indicated that the sixteen items (socio-economic, dempgraphic and geographic factors) scale with each response categories mistargeted the current sample. This conclusion was confirmed and the analysis elaborated taking advantage of the Rasch model that places independently estimated item and person parameters.

This was the first study to undertake an examination of the socio-economic, demographic and geographic factors on the malaria RDT result, use of indoor residual spraying and use of mosquito nets using Rasch analysis and to assess item bias. The Rasch analysis support for the measurement properties, internal consistency reliability, targeting, and unidimensionality of the different levels of malaria RDT result, use of indoor residual spraying and use of mosquito nets. During the analysis, it was necessary to remove some item from each of the scales to achieve fit to the Rasch model. Using differential item functioning analysis, it was found for malaria RDT result, use of indoor residual spraying and use of mosquito nets the items responding good. Further examination of fit of data from the malaria RDT result, use of indoor residual spraying and use of mosquito nets to the Rasch measurement model in larger and appropriately targeted samples is recommended to confirm the findings of the current study. The categorisation of the items was examined using the Rasch model for the ordering of the item thresholds. From the analysis, few items showed disordered thresholds indicating some problems with the categorization of items.

In conclusion, application of the Rasch model in this study has supported the viability of total sixteen (socio-economic, demographic and geographic) items for measuring malaria RDT result, use of indoor residual spraying and use of mosquito nets. Therefore, from the analysis it can be seen that the scale shows high reliability. But, there were little disordering of thresholds and no evidence of differential item functioning. Therefore, the result supports the analysis carried out in previous chapters.

Chapter 10

Discussion and conclusion

The focus of this study was to model and analyze malaria rapid diagnostic test outcome data in Ethiopia using different statistical methods. Malaria is related to poor socio-economic factors and normally referred to as a disease of the poor or is a disease normally associated with poverty (Hay et al., 2004). Malaria disproportionately affects those who cannot afford treatment or have limited access to health care. Families and communities are then trapped in a downward spiral of poverty (Worrall et al., 2002). It is known that socio-economic factors are related to poverty. Therefore, to introduce the most advanced level of care for people with malaria infections in the health care system, it is important to scale up the malaria treatment programmes. This process requires continuous monitoring and counseling of patients in order to optimize medication benefits. Based on this fact, it is important to understand the linkages between malaria and poverty. Identifying the factors that increase the risk of malaria can be used to guide government policy to create and implement more effective policies to tackle the problem.

The development of in-depth advanced statistical methods for analysis of malaria data with discrete outcomes is important area of research. In this study, we have been concerned with statistical methods for binary data, which is used in applied statistics. Data analysis for binary response face the challenge of choosing the appropriate method of analysis to address the research questions. In addition, the other challenge relates to the estimation procedure. To solve the problem, there will be more than one estimation procedure methods to choose from. Choosing the appropriate estimation procedure is important to obtain the appropriate inferences. Therefore, these methodologies have been demonstrated with in-depth analyses of a practical data set with a binary outcome. The data relates malaria rapid diagnosis test which was collected from December 2006 to January 2007 in Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia.

In this study, malaria rapid diagnosis test (RDT) result was the response variable and the independent predictor variables consisted of baseline socio-economic, demographic and geographic variables. The socio-economic variables were as follows: main source of drinking water; time taken to collect water; toilet facilities; availability of electricity, access to radio and television; total number of rooms; main construction material of the rooms' walls, main construction material of the room's roof and main construction material of the room's floor; incidence in the past twelve months of indoor residual spraying; use of mosquito nets and total number of nets. Geographic variables were region and altitude, while demographic variables were gender, age and family size. Of these variables, age and sex were collected at the individual level, while altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms, main construction material of walls, roof and floor, incidence of indoor residual spraying and use of mosquito nets were all collected at the household level.

Because of the importance of malaria rapid diagnosis test outcome, the study began by identifying factors affecting malaria RDT status of respondents. Multiple Correspondence Analysis (MCA) was used to explore associations between sets of categorical variables. MCA is a method for breaking down the value of the goodness-of-fit statistic into components due to the rows and columns of the contingency table. Moreover, the MCA approach involves defining a set of points, with associated masses, in a multidimensional space structured by Euclidean distance. The technique allows the analysis of the relationships between the variables and different levels of one variable. Furthermore, the results of the analysis can be seen analytically and visually. This method of display gives detailed information of the relationship between variables and their associations.

The result from multiple correspondence analysis shows that there is association between malaria RDT result and the different socio-economic, demographic and geographic variables. Moreover, there was an indication that some socio-economic, demographic and geographic factors have joint effects. Therefore, the interaction effects between socio-economic, demographic and geographic variables were included for advanced statistical analysis techniques. For identifying socio-economic, demographic and geographic predictors of malaria RDT result, generalized linear models were employed. These include several broad model families that include survey logistic, GLMM, GLMM with spatial covariance structure, joint models and semiparametric additive models. These models can be viewed as direct extensions of generalized linear models for independent observations to the context of correlated data. Between these models, there are differences in the way the dependency in the data is addressed. To test how well the observed data fit the expectation of the model, Rasch model was used. All the models were used to assess the determinants of malaria RDT result and each one of the methods has its own strengths and weaknesses.

The nature of data used for this study can be described as one from a complex survey. The first attempt of this study was to use survey logistics methodology. Survey logistic is a method which is able to handle complex survey information. The findings using this method show that some socio-economic, demographic and geographic factors are related to malaria risk. It was observed that houses that were treated with indoor residual spraying were less likely to be affected by malaria. One of the major challenges in the control of malarial infection was the use of toilet facilities. From the results, it was observed that households with no toilet facilities were more likely to have occupants who are positive with malaria diagnosis test. Furthermore, positive malaria diagnosis rate decreased with age. For household size, the risk of malaria increased per unit increase in family size. Generally, malaria parasite prevalence differed between age and gender with the highest prevalence occurring in children and females.

Although factors associated with malaria RDT at the intial stages of the analysis give important information, identifying factors that are associated with malaria RDT with other variabilities is important. The survey logistic model is survey based, which only allows for a fixed clustering variable whereas the *Kebeles* are chosen at random which could result in some variability between the sampling units. Therefore, the survey logistic method does not incorporate variability between sampling units (*kebeles*). For this reason, generalized linear mixed models (GLMM) were used to explore socio-economic, geographic and demographic factors affecting malaria rapid diagnosis test result. GLMM explore the idea of statistical models that incorporate random factors into generalized linear models. These models add random effects or correlations among observations to a model, where observations arise from a distribution in the exponential family. Furthermore, the use of GLMMs can allow random effects to be properly specified and computed and errors can also be correlated. In addition to this, GLMMs can allow the error terms to exhibit non constant variability while also allowing investigation into more than one source of variation. This ultimately leads to greater flexibility in modelling the dependent variable.

The same socio-economic, demographic and geographic variables were used for analysis using generalized linear mixed model. The study indicates that socio-economic, demographic and geographic factors are correlated with the transmission of malaria. Compared to the survey logistic method, the generalized linear mixed model explains the model better. This is supported by the fact that the standard errors for the estimation of parameters is small compared to the survey logistic method. Furthermore, the number of significant effects was found to be more compared to the survey logistic methods.

It was also of interest to the researcher to know whether the data display any spatial autocorrelation, i.e., to check whether regions or areas that are near in space have malaria prevalence or incidence that is similar with the surveys that are far apart. This is important because spatially correlated data cannot be regarded as independent observations. If the analysis does not take account of the correlation structure of the data, the estimates obtained from the analysis may be inaccurate because of the underestimated standard errors. Therefore, spatial statistics analysis was used to identify important socio-economic, demographic and geographic variables associated with the malaria RDT result and to produce prevalence maps of the area illustrating the variation in malaria risk. The same variables used in GLMM were used with spatial correlation structure. From all possible spatial covariance structures, SP(GAU) (Gaussian) was found to be the best spatial covariance structure for the data.

Therefore, the results of the study provide evidence on the spatial distribution of socio-economic, demographic and geographic risk factors in the occurrence of malaria. The utilization of socio-economic, demographic and geographic data on malaria rapid diagnosis test, including the information on the spatial variability, clarifies the effects of these factors. From the study it was observed that residents living in the SNNP region were found to be more at risk of malaria than those living in Amhara and Oromiya regions.

In addition to the different models used so far fo analysis, a joint modeling approach was used to further investigate the joint effect of the predictor variables on malaria RDT result, use of mosquito nets and use of indoor residual spraying in the last twelve months, i.e., the customary two variables joint modelling approch was extended to three variables joint effect. The study assessed whether the explanatory variables that were found to be significantly related with malaria RDT result in random effect model would still have a significant effect on long-term malaria transmission even when use of mosquito nets and use of indoor residual spraying were accounted for. Also assessing the association among the three outcomes (malaria RDT result use of mosquito nets and use of indoor residual spraying) was of interest.

Joint models have different advantages. Using joint model is useful to control over type I error rates in multiple tests. This way, there will be possible gains in efficiency in the parameter estimates and the ability to answer multivariate questions (Gueorguieva, 2001). Joint models are useful to solve the problem of correlations. When using these models, two types of correlations must be taken into account. These correlations are correlations between different variables and correlations between the same variable. To evaluate the association between malaria RDT result, use of mosquito nets and use of indoor residual spraying in the last twelve months, conditional random-intercepts models were fitted. The variables malaria RDT result, use of mosquito nets and use of indoor residual spraying were specified as binary variables. In this model, the correlation among the three outcomes as well as the correlation coming from the structure of the data is specified through the random effects structure. This is done by assuming separate random intercepts for each outcome variable and then combining them by imposing a joint multivariate distribution on the random intercepts. The linear predictors which were used for joint models consist the same variables which were used in the previous models. The result from joint models for malaria RDT result, use of mosquito nets and use of indoor residual spraying in the last twelve months confirm the result from the previous models.

Different parametric statistical models were employed for the analysis of malaria RDT result data. But, these models may not have been flexible enough to capture the main features of the data structure. A semiparametric approach was adopted to identify the non-parametric relationships. To resercher’s knowledge, this method is the first method to be used in malaria research. To identify the non-parametric relationships, generalized additive mixed models were used. To GAMM, the effect of age, family size, number of rooms per person, number of nets per person, altitude and number of months the room sprayed were fitted non-parametrically. The result from GAMM approach supports the results from the previous models fitted. In addition to this, the results gave more insight into the distribution of age, altitude, total number of rooms per person, total number of nets per person, family size and number of months room sprayed. The results from the non-parametric part of the model confirm that malaria RDT result is high for children. They reveal that persons with more mosquito nets and more number of rooms have greater chance to reduce the risk of malaria. Furthermore, with the correct use of mosquito nets, indoor residual spraying and other preventative measures, like having more rooms in a house, is a major contributing factor or determinant for the incidence of malaria to be decreased.

Finally, the Rasch model was used to show an application of the model in malaria research (Rasch, 1960, Rasch, 1980). Rasch model can be used to diagnosis the empirical ordering of the categories of the socio-economic, demograpic and geographic variables in the model. This method was the first study to undertake an examination of the socio-economic, demographic and geographic properties on the malaria RDT result, use of indoor residual spraying and use of mosquito nets to assess item bias. The Rasch analysis supports the measurement properties, internal consistency reliability, targeting, and unidimensionality of the different levels of malaria RDT result, use of indoor residual spraying and use of mosquito nets. During the analysis, it was necessary to remove some item from each of the scales to achieve better fit to the Rasch model.

Using differential item functioning analysis, it was found that malaria RDT result, use of indoor residual spraying and use of mosquito nets, the items answering reseasonably. Further examination model fit to the data to malaria RDT result, use of indoor residual spraying and use of mosquito nets to the Rasch measurement model in larger and appropriately targeted samples is recommended to confirm the findings of the current study. The categorisation of the items was examined using the Rasch model for the ordering of the item thresholds. From the analysis, few items showed disordered thresholds indicating some problems with the categorization of items.

Application of the Rasch model in this study has supported the viability of total sixteen items (socio-economic, deographic and geographic variables) for measuring malaria RDT result, use of indoor residual spraying and use of mosquito nets. From the analysis it can be seen that the scale shows high reliability. But, there was little disordering of thresholds and no evidence of differential item functioning. Differential item functioning (DIF) referred to measurement bias. DIF analysis provides an indication of unexpected behavior of items on a test. An item does not display DIF if respondents from different groups have a different probability to give a certain response. Therefore, from the analysis, there was no evidence of differential item functioning.

The government of Ethiopia has adopted various strategies to control malaria. These include early diagnosis, prompt treatment, selective vector control, epidemic prevention and control. In addition to this, the government has supported strategies such as human resource development, monitoring and evaluation. One of the government's key goals in the control of malaria is to achieve the complete elimination of malaria within those geographical areas with historically low malaria transmission and achieve near zero malaria transmission in the remaining malarious areas of the country. For this reason, evidence based strategies to prevent malaria is an attractive strategy for the country (FMH, 2006b).

The results of this this study showed that malaria is associated with socio-economic, demographic and geographic factors, mainly influenced by poverty levels. Malaria is generally regarded as a disease of the poor. The poor socio-economic condition is a major contributing factor or determinat for malaria burden. Hence, wealthier households who can afford toilet facilities, a greater number of rooms in the house, clean drinking water, and well built houses were found to be less affected by malaria. It was also found that women and children are more vulnerable to malaria. Lack of bed nets contributes to this vulnerability. As the results indicate having more bed nets is one means of reducing malaria and the evidence suggests that households who are unable to afford sufficient mosquito nets, due to large families and low incomes, are more affected by malaria. Women and children are also exposed to mosquito bites while they are travelling long distances to fetch water. As expected wealthier households were found to be less vulnerable to malaria than the poor households, thus the living conditions of the communities could be one way of achieving the malaria control goals set by the health professionals.

In conclusion, different family of models were reviewed and applied to malaria RDT result data. The finding of the analyses performed using different statistical models demonstrated that these models are useful in the study of binary responses. Furthermore, the analysis and the result of the thesis highlighted the direction and development of malaria RDT result data analyses. In practice, there are complicated binary data which comes from complex survey designs. The structure and complexity of the data pose major challenges. Theoretically, it is very interesting to find statistical methods which incorporates survey design informations. For this issue, several authors suggested some corrections to chi-square statistic and F – values (Rao and Scott, 1981, Thomas, 1989, Solomon and Stephens, 1977). This issue provides great opportunities and the advancement of important research areas. On the other hand, developing a comparison of different models is one of the challenges. Therefore, one of the future directions of this thesis is to compare the different families of methods and diagnosis of these methods using simulations method.

References

ADHANOM, T., DERESSA, W., WITTEN, H. K., GETACHEW, A. & SEBOXA, T. 2006. Malaria. In The Eipdemiology and Ecology of Health and Disease in Ethiopia 1st edition. *Edited by: Berhane Y, Hailemariam D, Kloos H and Shama PLC. Addis Ababa, Ethiopia.*

AERTS, M., GEYS, H., MOLENBERGHS, G. & RYAN, L. 2002. *Topics in Modelling of Clustered Data,* London, Chapman and Hall.

AGRESTI, A. 1984. *Analysis of ordinal categorical data* New York John Wiley & Sons.

AGRESTI, A. 1990. *Categorical Data Analysis,* New York, John Wiley & Sons.

AGRESTI, A. 2002. *Categorical Data Analysis,* New Jersey, John Wiley & Sons.

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19**,** 716-723.

ANDERSEN, E. B. 1970. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society,* Series B 32**,** 283-301.

ARCHERA, K. J., LEMESHOW, S. & HOSMER, D. 2007. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis* 51**,** 4450-4464.

ASKELL-WILLIAMS, H. & LAWSON, M. J. 2004. A Correspondence Analysis of Child-Care Students' and Medical Students' Knowledge about Teaching and Learning. *International Education Journal,* 5 (2)**,** 176-206.

AYALA, R. J. D. 2009. *The theory and practice of item response theory,* New York, The Guilford Press.

AYELE, D. G., ZEWOTIR, T. & MWAMBI, H. 2012. Prevalence and risk factors of malaria in Ethiopia. *Malaria Journal,* 11:195 doi:10.1186/1475-2875-11-195.

AYELE, D. G., ZEWOTIR, T. & MWAMBI, H. 2013a. The risk factor indicators of malaria in Ethiopia *International Journal of Medicine and Medical Sciences,* 5**,** doi:10.5897/IJMMS2013.0956 335-347

AYELE, D. G., ZEWOTIR, T. & MWAMBI, H. 2013b. Spatial distribution of malaria problem in three regions of Ethiopia. *Malaria Journal* 12:207**,** doi:10.1186/1475-2875-12-207.

BANGUERO, H. 1984. Socio-economic factors associated with malaria in Colombia. *Social Science & Medicine* 19**,** 1099-1104.

BARNDORFF-NIELSEN, E. & COX, D. R. 1989. *Asymptotic Techniques for Use in Statistics,* London, Chapman and Hall.

BEAN, J. A. 1975. Distribution of properties of variance estimatorsfor complex multistage probability samples. *Vaital and health Statistics,* Series 2, No 65. Washington, DC: National Center for Health Statistics, Public Health Service.

BENDIXEN, M. 1996. *A practical guide to the use of correspondence analysis in marketing research. Marketing Research On-Line, 2003, 16-38.* *http://mro.massey.ac.nz/ca.html* [Online]. [Accessed Oct 03 2011].

BENZÉCRI, J.-P. 1973. *Analyse des Donnkes, Tome 2: Analyse de Correspondences,* Paris, Dunod.

BERRIDGE, D. M. & CROUCHLEY, R. 2011. *Multivariate Generalized Linear Mixed Models Using R* Lancaster, UK, CRC Press.

BIEWEN, M. & JENKINS, S. P. 2006. Variance Estimation for Generalized Entropy and Atkinson Inequality Indices: the Complex Survey Data Case. *OXFORD BULLETIN OF ECONOMICS AND STATISTICS,* 68, 3**,** Blackwell Publishing Ltd

BIVAND, R. S., PEBESMA, E. J. & GOMEZ-RUBIO, V. 2008. *Applied Spatial Data Analysis with R,* New York, Springer.

BOWMAN, A. W. & AZZALINI, A. 1997. *Applied Smoothing Techniques for Data Analysis,* Oxford, Oxford University Press.

BRESLOW, N. E. & CLAYTON, D. G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association,* 88**,** 9-25.

CANTY, A. J. & DEVISON, A. C. 1999. Resampling-based variance estimation for labour force survey. *The Statistician,* 48(3)**,** 379 - 391.

CATALANO, P. J. & RYAN, L. M. 1992. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association,* 87**,** 651-658.

CHECCHI, F., COX, J., BALKAN, S., TAMRAT, A., PRIOTTO, G., ALBERTI, K. P., ZUROVAC, D. & GUTHMANN, J.-P. 2006. Malaria Epidemics and Interventions, Kenya, Burundi, Southern Sudan, and Ethiopia, 1999–2004. *Emerging Infectious Diseases* *www.cdc.gov/eid*, 12

CHEN, Z. & MANTEL, H. 2009. Analysis of binary data from a complex survey with misclassification in an ordinal covariate. *Proceeding of the Survey Methods Section. Statistical Society of Canada Annual Meeting.*

CHILES, J. P. & DELFINER, P. 1999. *Geostatistics. Modelling Spatial Uncertainty,* Chichester, Wiley.

CLAUSEN, S.-E. 1998. *Applied Correspondence Analysis. An Introduction,* Thousand Oaks, CA, Sage.

CLAY, D. E. & SHANAHAN, J. F. 2011. *GIS Applications in Agriculture: Nutrient Management for Energy Efficiency,* Boca Raton, CRS press.

CLIFF, A. & ORD, J. 1975. The choice of a test for spatial autocorrelation. *In J. C. Davies and M. J. McCullagh (eds) Display and Analysis of Spatial Data,* John Wiley and Sons, London 54-77.

CLIFF, A. D. & ORD, J. K. 1981. *Spatial processes - models and applications* London, Pion.

COMMENGES, D. & JACQMIN-GADDA, H. 1997. Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society,* Series B, 59**,** 157-171.

COMMENGES, D., LETENNEUR, H., JACQMIN-GADDA, H., MOREAU, T. & DARTIGUES, J. F. 1994. Test of homogeneity of binary data with explanatory variables. *Biometrika,* 50**,** 613-620.

CRAIG, M. H., SNOW, R. W. & SUEUR, D. L. 1999. A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitol Today,* 15 (3)**,** 105-11.

CRAVEN, P. & WAHBA, G. 1979. Smoothing noisy data with spline functions. *Numerische Mathematik,* 31 (4)**,** 377-403.

CRESSIE, N. 1985. Fitting variogram models by weighted least squares. *Journal of Mathematical Geology,* 17**,** 563 - 586.

CRESSIE, N. 1993. *Statistics For Spatial Data,* New York, John Wiley & Sons.

CRESSIE, N. & HAWKINS, D. H. 1980. Robust estimation of the variogram. *Mathematical Geology,* 12 (2)**,** 115-125.

CROUX, C. & ROUSSEEUW, P. J. 1992. A Class of High-Breakdown Scale Estimators Based on Subranges. *Communications in Statistics, Theory and Methods,* 21**,** 1935- 195 1.

CSA 2000. Central Statistics Agency of Ethiopia and ORC Macro: Ethiopia demographic and health survey 2000. Addis Ababa and Calverton, MD: Central Statistics Agency and ORC Macro.

CSA 2006. Central Statistics Agency of Ethiopia and ORC Macro: Ethiopia demographic and health survey 2005. Addis Ababa and Calverton, MD: Central Statistics Agency and ORC Macro. [http://www. measuredhs.com/pubs/pdf/FR179/FR179.pdf].

CSA 2012. Central Statistics Agency of Ethiopia and ORC Macro: Ethiopia demographic and health survey 2011. Addis Ababa and Calverton, MD: Central Statistics Agency and ORC Macro.

DE BOECK, P. & WILSON., M. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach,* New York, Springer.

DEMPSTER, A. P., LARID, N. M. & RUBIN, D. B. 1997. Maximum likelihood from incomplete data via the Em algorithm. *Journal of the Royal Statistical Society.,* B, 59**,** 1-38.

DENG, L. Y. & WU, C. F. J. 1987. Estimation of variance of the regression estimator. *Journal of the American Statistical Association,* 82**,** 568 - 576.

DER, G. & EVERITT, S. 2002. *A Handbook of Statistical Analysis: Using SAS,* New York, Chapman and Hall/CRC.

DERESSA, W., ALI, A. & ENQUSELLASSIE, F. 2003. Self-treatment of malaria in rural communities, Butajira, southern Ethiopia. *Bulletin World Health Organization,* 81**,** 261-268.

DEVROYE, L. P. & GYORFI , L. 1985. *Nonparametric Density Estimation: The L_1 View,* New York, Wiley.

DIGGLE, P. J., TAWN, J. A. & R.A., M. 1998. Model-based geostatistics. *Applied Statistics,* 47**,** 299-350.

DIPPO, C. S. & WOLTER, K. M. 1984. A comparison of variance estimators usinf Taylor series approximation. *Proceeding of the Section on Survey Research Methods, American Statistical Association***,** 113-121.

DRAY, S., CHESSEL, D. & THIOULOUSE, J. 2003. Co-Inertia Analysis and the Linking of Ecological Data Tables. *Ecology* 84**,** 3078-3089.

EFROMOVICH, S. 1999. *Nonparametric Curve Estimation: Methods, Theory and Applications,* New York., Springer-Verlag.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Mathematical Statistics,* 7**,** 1-26.

EFRON, B. 1981. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics,* 9**,** 139-172.

EFRON, B. 1982. The jackknife, the bootstrap and other resampling plans. *Philadelphia: SIAM monograph,* 38.

EMERSON, P. M., NGONDI, J., BIRU, E., GRAVES, P. M., YESHEWAMEBRAT, EJIGSEMAHU, GEBRE, T., ENDESHAW, T., GENET, A., MOSHER, A. W., ZERIHUN, M., MESSELE, A. & JR, F. O. R. 2008. Integrating an NTD with One of "The Big Three": Combined Malaria and Trachoma Survey in Amhara Region of Ethiopia. *PloS Neglected tropical diseases,* 2.

EUBANK, R. L. 1988. *Spline Smoothing and Nonparametric Regression,* New York, Marcel Dekker

FAES, C., GEYS, H., AERTS, M., MOLENBERGHS & G. AND CATALANO, P. 2004. Modelling combined continuous and ordinary outcomes in a clustered setting. *Journal of Agricultural Biological and Environment Statistics,* 9**,** 515-530.

FAES, C., GEYS, H. & CATALONO, P. 2008. *Joint models for continuous and discrete longitudinal data. In: Longitudinal Data Analysis: A handbook of modern statistical methods. Fitzmaurice, G., Davidian, M. Verbeke, G and Molenberghs, G. (Eds),* Boca Raton, Chapman & Hall/CRC.

FAN, J. & GIJBELS, I. 1996. *Local Polynomial Modelling and Its Applications,* London, Chapman and Hall.

FARAWAY, J. J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models,* Boca Raton, Chapman and Hall/CRC.

FAVA DEL, E., SHKEDY, Z., HENS, N., AERTS, M., SULIGOI, B., CAMONI, L., VALLEJO, F., WIESSING, L. & KRETZSCHMAR, M. 2011. Joint Modeling of HCVand HIV Co-Infection among Injecting Drug Users in Italy and Spain Using Individual Cross-Sectional Data. *Statistical Communications in Infectious Diseases,* 3**,** DOI: 10.2202/1948-4690.1009.

FEDERAL MINISTRY OF HEALTH (FMH) 1999. Malaria and Other Vector-borne Diseases Control Unit. Addis Ababa, Ethiopia.

FINNEY, D. 1952. *Probit Analysis,* Cambridge, Cambridge University Press.

FISHER, G. H. & MOLENAAR, I. W. 1995. *Rash Models: Foundations, Recent Developments, and Applications,* New York, Springer-Verlag.

FITZ-GIBBON, C. T. 2000. Value Added for those in Despair: Research Methods Matter. *The Vemon-Wall Lecture for the annual meeting of the Education Section of the British Psychological Society.*

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. & MOLENBERGHS, G. 2008. *Longitudinal Data Analysis,* Boca Raton, Chapman & Hall/CRC Press.

FITZMAURICE, G. M., LAIRD, N. M. & WARE, J. H. 2004. *Applied Longitudinal Analysis,* New York, John Wiley & Sons.

FMH 2004. Federal Ministry of Health: Guideline for malaria epidemic prevention and control in Ethiopia. 2nd edition. Addis Ababa, Ethiopia, Federal democratic Republic of Ethiopia, Ministry of Health.

FMH 2006a. Federal Democratic Republic of Ethiopia Ministry of Health: National five year strategic plan for malaria prevention and control in Ethiopia:2006 - 2010 Addis Ababa. 1-52.

FMH 2006b. National five-year strategic plan for malaria prevention and control in Ethiopia 2006 – 2010.

FMH 2008. Ethiopia National Malaria Indicator Survey 2007. Addis Ababa, Ethiopia.

FOX, J. 2008. *Applied Regression Analysis and Generalized Linear Models,* California, Sage Publications.

FRANKEL, M. R. 1971. Inference for Survey Samples: An Emprical Investigation. *Ann Arbor, MI: Institute for Social Research.*

FRIEDMAN, J. H. & STUETZLE, W. 1981. Projection Pursuit Regression. *Journal of the American Statistical Association* 76**,** 817-823.

GAETAN, C. & GUYON, X. 2010. *Spatial Statistics and Modeling,* New York Springer.

GEARY, R. 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5 115-45

GELFAND, A. & CARLIN, B. 1993. Maximum-likelihood estimation for constrained- or missing data models. *Canadian Journal of Statistics,* 21**,** 303-311.

GENTON, M. G. 2000. The correlation structure of Matheron's classical variogram estimator under elliptically contoured distributions. *Mathematical Geology,* 32 127-137.

GENTON, M. G. 2001. Robustness problems in the analysis of spatial data. *In spatial statistics: Methodological Aspects and Applications,* Marc Moore (Editor) 21-37.

GENTON, M. G., HE, L. & LIU, X. 2001 Moments of skew-normal random vectors and their quadratic forms. *Statistics and Probability Letters,* 51**,** 319-325.

GEYER, C. J. & THOMPSON, E. A. 1992. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society,* B, 54**,** 657-699.

GHEBREYESUS, T., HAILE, M., WITTEN, K., GETACHEW, A., YOHANNES, M. & LINDSAY, S. 2000. Household risk factors for malaria among children in the Ethiopian highlands. *Trans R Soc Trop Med Hyg,* 94**,** 17-21.

GHOSH, M. 1995. Inconsistent maximum likelihood estimators for the Rasch model. *Statistics and Probability Letters* 23**,** 165-170

GIFI, A. 1990. *Nonlinear Multivariate Analysis,* New York, John Wiley & Sons.

GLAS, C. A. W. & VERHELST, N. D. 1995. Testing the Rasch model. In Rasch Models, Foundations, Recent Developments and Applications. *ed. G. H. Fisher and I. W. Molenaar,* **,** 69-95. New York: Springer.

GOLDSTEIN, H. 2011. *Multilevel Statistical Models,* United Kingdom, John wiley and sons, Ltd.

GOOD, I. J. 1967. Analysis of Log-Likelihood Ratios,'ANO' (A contribution to the discussion of the paper on least squares by F. J. Anscombe). *Journal of the Royal Statistical Society,* B, 29**,** 39-42.

GOOVAERTS, P. 1997. *Geostatistics for Natural Resources Evaluation,* New York, Oxford University Press.

GOOVAERTS, P., JACQUEZ, G. M. & GREILING, D. 2005. Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. *Geographical Analysis,* 37**,** 152-182.

GRAVES, P. M., RICHARDS, F. O., NGONDI, J., EMERSON, P. M., SHARGIE, E. B., ENDESHAW, T., CECCATOD, P., EJIGSEMAHU, Y., MOSHERA, A. W., HAILEMARIAME, A., ZERIHUN, M., TEFERI, T., AYELE, B., MESELE, A., YOHANNES, G., TILAHUN, A. & GEBRE, T. 2009. Individual, household and environmental risk factors for malaria infection in Amhara, Oromia and SNNP regions of Ethiopia. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 10.

GREEN, P. J. 1990. On use of EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society,* B, 52**,** 443-452.

GREEN, P. J. & SILVERMAN, B. W. 1994. *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach,* London, Chapman and Hall.

GREENACRE, M. 2000. Correspondence analysis of square asymmertric matrices. *Applied Statistics,* 49**,** 297-310.

GREENACRE, M. J. 1984. *Theory and Applications of Correspondence Analysis*, Academic Press.

GREENACRE, M. J. & BLASIUS, J. 2006. *Multiple correspondence analysis and related methods* Boca Raton, Chapman & Hall/CRC.

GUEORGUIEVA, R. 2001. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling,* 1: 177, DOI: 10.1177/1471082X0100100302.

GUEORGUIEVA, R. V. & AGRESTI, A. 2011. A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses. *Journal of the American Statistical Association,* 96:455, 1102-1112.

GUITONNEAU, G. G. & ROUX, M. 1977. Sur la Taxinomie du Genre Erodium. *Cahiers de l'Analyse des Donnees,* 2 (1), 97-113.

GUO, X. & CARLIN, B. P. 2004. Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician,* 58, 16-24.

GYORFI , L., HARDLE, W., SARDA, P. & VIEU, P. 1989 Nonparametric Curve Estimation from Time Series. *Lecture Notes in Statistics, 60. Springer-Verlag, Berlin.*

HAIR, J. F., ANDERSON, R. E., TATHAM, R. L. & BLACK, W. C. 1995. *Multivariate Data Analysis,* Upper Saddle River, NJ, Prentice Hall.

HANDCOCK, M. S. & STEIN, M. L. 1993. A Bayesian analysis of kriging. *Technometrics* 35, 403-410.

HANDCOCK, M. S. & WALLIS, J. R. 1994. An approach to statistical spatio-temporal modeling of meteorological fields. *Journal of American Statistical Association,* 89, 368-378.

HARDLE, W. 1994. *Applied nonparametric regression* Berlin, Cambridge University Press.

HÄRDLE, W. 1989. *Applied Nonparametric Regression,* New York, Cambridge University Press.

HÄRDLE, W., MÜLLER, M., SPERLICH, S. & WERWATZ, A. 2004. *Nonparametric and Semiparametric Models: An Introduction* Berlin, springer.

HARDOUIN, J.-B. 2007. Rasch analysis: Estimation and tests with raschtest. *The Stata Journal,* 7**,** 22-44.

HART, J. D. 1997 *Nonparametric Smoothing and Lack-of-Fit Tests,* New York, Springer.

HARVILLE, D. A. 1974. Bayesian inference for variance components using only error contrasts. *Biometrika,* 61**,** 383 - 385.

HASTIE, T. J. & TIBSHIRANI, R. 1986. Generalized additive models. *Statistical Science* 1**,** 297-318.

HASTIE, T. J. & TIBSHIRANI, R. 1990. *Generalized Additive Models,* London, Chapman and Hall.

HAY, S. I., GUERRA, C. A., TATEM, A. J., NOOR, A. M. & SNOW, R. W. 2004. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Disease* 4**,** 327-336.

HAYASHI, C. 1954. Multidimensional quantication{with applications to analysis of social phenomena. *Annals of the Institute of Statistical Mathematics,* 5(2) 121-143.

HENGL, T. 2007. A Practical Guide to Geostatistical Mapping of Environmental Variables. Italy: European Commission, Joint Research Centre, Institute for Environment and Sustainability.

HILL, M. O. 1974. Correspondence Analysis: A Neglected Multivariate Method. *Journal of the Royal Statistical Society,* 23 (3)**,** 340-354.

HOIJTINK, H. & A. BOOMSMA, A. 1995. On person parameter estimation in the dichoto-mous Rasch model. In Rasch Models, Foundations, Recent Developments and Applications *ed. G. H. Fisher and I. W. Molenaar*, 53-68. New York: Springer.

HOLMES, D. G. & SKINNER, C. J. 2000. Variance estimation for labour force survey estimates of level and change. *GSS Methodology Series,* No 21.

HOSMOER, D. W. & LEMESHOW, S. 1980. Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics, Theory and Methods,* Part A 9**,** 1043-1069.

JACQMIN-GADDA, H. & COMMENGES, D. 1995. Test of homogeneity for generalized linear models. *Journal of American Statistical Association,* 90**,** 1237-1246.

JIANG, J. 2001. A non-standard X^2-test with application to generalized linear model diagnostics. *Statistics and Probability Letters,* 53**,** 101-109.

JIANG, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications,* New York, USA, Springer.

JIMA, D., GETACHEW, A., BILAK, H., STEKETEE, R. W., PAUL M EMERSON4, P. M. G., GEBRE, T., REITHINGER, R. & HWANG, J. 2010. Malaria indicator survey 2007, Ethiopia: coverage and use of major malaria prevention and control interventions. *Malaria Journal,* 9:58.

JOHNSON, R. A. & WICHERN, D. W. 2007. *Applied multivariate statistical analysis,* New Jersey, Pearson/Prentice Hall.

JURNEL, A. G. & HUIJBREGTS, C. J. 1978. *Mining Geostatistics,* London, Academic press.

KAKKILAYA, B. S. 2003. Rapid Diagnosis of Malaria. *K.S. Hegde Medical Academy,* Deralakatte, India.

KINCAID, C. 2012. *Guidelines for Selecting the Covariance Structure in Mixed Model Analysis:* *http://www2.sas.com/proceedings/sugi30/198-30.pdf* [Online]. Portage: SAS and all other SAS Institute Inc. [Accessed May 2012].

KITANIDIS, P. K. 1983. Statistical Estimation of Polynomial Generalized Covariance Functions and Hydrologic Applications. *Water Resources Research,* 19 (4)

KOHN, R., AUSLEY, C. F. & THARM, D. 1991. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of American Statistical Association,* 86.

KORAM, K., BENNETT, S., ADIAMAH, J. & GREENWOOD, B. 1995. Socio-economic risk factors for malaria in a peri-urban area of the Gambia. *Transactions of the royal society of tropical medicine and hygiene,* 89 146-150.

KREWSKI, D. & RAO, J. N. K. 1981. Inference from stratified samples: Properties of the linearisation, jackknife, and balanced repeated replication methods. *Annals of Statistics,* 9**,** 1010-1019.

KUK, A. Y. C. & CHENG, Y. W. 1997. The Monte Carlo Newton-Raphson Algorithm *Journal of Statistical Computing Simulation,* 59**,** 233-250.

KUTNER, M. H., NACHTSHEIM, C. J., NETER, J. & LI, W. 2005. *Applied Linear Statistical Models,* New York, McGraw-Hill Irwin.

LABARRE, P., GERLACH, J., WILMOTH, J., BEDDOE, A., SINGLETON, J. & WEIGL, B. 2010. Non-Instrumented Nucleic Acid Amplification (NINA): Instrument-Free Molecular Malaria Diagnostics for Low-Resource

Settings. *32nd Annual International Conference of the IEEE EMBS* Buenos Aires, Argentina.

LAIRD, N. M. A. W., J.H. 1982. Random-effects models for longitudinal data. *Biometrics,* 38**,** 963-974.

LE, N. D. & ZIDEK, J. V. 1992. Interpolation with uncertain spatial covariances: A Bayesian alternative to Kriging. *Journal of Multivariate Analysis,* 43 351-374.

LEE, E. S. & FORTHOFER, R. N. 2006. *Analyzing Complex Survey Data,* California, SAGE PUBLICATIONS.

LEE, P. M. 2004. *Basyesian Statistics: an introduction,* London, Hodder.

LEHTONEN, R. & PAHKINEN, E. 2004. *Practical methods for design and analysis of complex surveys,* England, John wiley & sons, Ltd.

LESAFFRE, E. & SPIESSENS, B. 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics***,** 325-335.

LIN, X. 1997. Variance component testing in generalized linear models with random effects. *Biometrika,* 84**,** 309-326.

LIN, X. 2007. Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models. *Lifetime Data Analysis,* 13**,** 533-544.

LIN, X. & BRESLOW, N. E. 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of American Statistical Association,* 91**,** 1007-1016.

LIN, X. & CARROLL, R. J. 2000. Nonparametric function estimation for clustered data when the predictor variable is measured with/without error. *Journal of the American Statistical Association* 95**,** 520-534.

LIN, X. & ZHANG, D. 1999. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B,* 61, Part 2**,** 381 - 400.

LINACRE, J. M. & WRIGHT, B. D. 1994. (Dichotomous mean-square) chi-square fit statistics. *Rasch Measurement Transactions,* 8**,** 360.

LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D. & SCHABENBERGER, O. 2006. *SAS for Mixed Models. 2nd Edition,* California, SAS Institute Inc.

MADSEN, H. & THYREGOD, P. 2010. *Introduction to General and Generalized linear models,* Baca Raton, CRC Press : Imprint of the Taylor & Francis Group.

MALLOWS, C. L. 1973. Some comments on Cp. *Technometrics,* 15(4)**,** 661-675.

MARDIA, K. V. A. M., R. J. 1984. Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. *Biometrika,* 71 135-146.

MATÉRN, B. 1960. Spatial variation. *Meddelanden från statens skogsforskningsinstitut,* 49.

MATHERON, G. 1962. Traite de geostatistique appliquee, Tome I. *Memoires du Bureau de Recherches Geologiques et Minieres.,* 14, Editions Technip, Paris, 333 p.

MATHERON, G. 1963. Principles of Geostatistics. *Economic Geology,* 58**,** 1246-1266.

MATSCHINGER, H. 2006. *Estimating IRT models with gllamm. 4th German Users Group meeting. http://econpapers.repec.org/paper/bocdsug06/03.htm* [Online].

MCCULLAGH, C. E. 2008. *Generalized, Linear and Mixed models,* Hoboken, N.J., John wiley and sons, inc.

MCCULLAGH, C. E. & SEARLE, S. R. 2001. *Generalized, Linear, and Mixed Models,* New York, John wiley and sons, inc.

MCCULLAGH, P. & NELDER, J. 1983. *Generalized Linear Models,* New York, Chapman & Hall.

MCCULLAGH, P. & NELDER, J. 1989. *Generalized Linear Models,* New York, Chapman & Hall.

MCCULLOCH, C. E. 1994. Maximum likelihood variance components etimation for binary data. *Journal of American Statistical Association,* 89**,** 330-335.

MCCULLOCH, C. E. 1997. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association,* 92, No. 437 162-170.

MENDIS, K., RIETVELD, A., WARSAME, M., BOSMAN, A., GREENWOOD, B. & WERNSDORFER, W. H. 2009. From malaria control to eradication: The WHO perspective. *Tropical Medicine and International Health,* 14 1-7.

MENS, P. F., SCHOONE, G. J., KAGER, P. A. & SCHALLIG, H. D. 2006. Detection and identification of human Plasmodium species with real-time quantitative nucleic acid sequence-based amplification. *Malaria Journal,* 5:80.

MOLENAAR, I. W. 1983. Some improved diagnostics for failure of the Rasch model. *Psy-chometrika,* 48**,** 49-72.

MOLENBERGHS, G., GEYS, H. & BUYSE, M. 2001. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine,* 20**,** 401-414.

MOLENBERGHS, G. & VERBEKE, G. 2005. *Models for Discrete Longitudinal Data,* New York, Springer-Verlag.

MOODY, A. 2002. Rapid Diagnostic Tests for Malaria Parasites. *Clinical Microbiology Reviews***,** 66–78.

MORAN, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37**,** 17-23.

MULLER, H. G. 1988. Nonparametric Regression Analysis of Longitudinal Data. Lecture Notes in Statistics.

MULRY, M. N. & WOLTER, K. M. 1981. The effect of Fisher's Z-transformation on confidence intervals for the correlation coefficient. *Proceeding of the Section on Survey Research Methods, American Statistical Association***,** 113 - 121.

MURRAY, C. K. & BENNETT, J. W. 2009. Rapid Diagnosis of Malaria. *Hindawi Publishing Corporation Interdisciplinary Perspectives on Infectious Diseases,* Article ID 415953**,** 7 pages.

NATARAJAN, S., ET AL., 2008. Variance estimation in complex survey sampling for generalized linear models. *Applied Statistics,* 57, Part 1**,** 75-87.

NATHANIEL, P. 2003. Limiting the spread of communicable diseases caused by human population movement. *Journal of Rural and Remote Environmental Health,* 2(1)**,** 23-32.

NELDER, J. & WEDDERBURN, R. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society,* Series A (General) 135 (3)**,** 370–384.

NISHISATO, S. 1980. *Analysis of categorical data: Dual Scaling and its applications,* Toronto, University of Toronto Press.

NISHISATO, S. 1994. *Elements of dual scaling: An introduction to practical data analysis* Hillsdale, NJ, Erlbaum.

OGDEN, T. 1997. *Essential wavelets for statistical applications and data analysis,* Birkhuser Boston.

PARK, I. 2008. Primary sampling unit (PSU) masking and variance estimation in complex surveys. *Survey Methodology,* 34, No. 2**,** 183-194.

PATTERSON, H. D. & THOMPSON, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3)**,** 545.

PENDERGAST, J. F., GANGE, S. J., NEWTON, M. A., LINDSTROM, M. J., PALTA, M. & FISHER, M. R. 1996. A Survey of Methods for Analyzing Clustered Binary Response Data. *International Statistical Review / Revue Internationale de Statistique,* 64**,** 89-118.

PETERSON, I., BORRELL, L. N., EL-SADR, W. & TEKLEHAIMANOT, A. 2009. Individual and Household Level Factors Associated with Malaria Incidence in a Highland Region of Ethiopia: A Multilevel Analysis. *American Journal of Tropical Medicine and Hygiene,* 80(1)**,** pp. 103-111.

PFEFFERMANN, D. 1993. The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review,* 61(2)**,** 317-337.

PINHEIRO, J. C. & CHAO, E. C. 2006. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. . *Journal of Computational and Graphical Statistics,* 15**,** 58-81.

PLACKETT, R. L. 1965. A class of bivariate distribution. *Journal of American Statistical Association,* 60**,** 516-522.

QUENOUILLE, M. H. 1949. Problems in plane sampling. *Annals of Mathematical Statistics,* 20**,** 355-375.

QUENOUILLE, M. H. 1956. Notes on bias in estimation. *Biometrika,* 43**,** 353 - 360.

RABE-HESKETH, S., A. , SKRONDAL, A. & PICKLES, A. 2004. *GLLAMM Manual. University of California-Berkeley, Division of Biostatistics, Working Paper Series. Paper No. 160. http://www.bepress.com/ ucbbiostat/paper160/* [Online].

RABE-HESKETH, S. & SKRONDAL, A. 2006. Multilevel modelling of complex survey data. *Royal Statistical Society,* 169, Part 4**,** 805-827.

RAMSAY, J. O. & SILVERMAN, B. W. 1997. *The Analysis of Functional Data,* Berlin, Springer-Verlag.

RAO, C. R. 1979. Separation theorem for singular values of matrices and their application in multivariate analysis. *Journal of Multivariate Analysis,* 9**,** 362-377.

RAO, J. N. K. 1997. Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics,* 25**,** 1-21.

RAO, J. N. K. & SCOTT, A. J. 1981. The Analysis of categorical data from complex surveys: chi-square tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association,* 76(374)**,** 221-230.

RAO, J. N. K. & SHAO, A. J. 1992. Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika,* 79**,** 811-822.

RAO, J. N. K. & SHAO, J. 1996. On balanced half-sample variance estimation in stratified sampling. *Journal of the American Statistical Association,* 91**,** 343-348.

RAO, J. N. K. & WU, C. F. J. 1985. Inference from stratified samples: Second order analysis of three methods for non-linear statistics. . *Journal of the American Statistical Association,* 80**,** 620-630.

RAO, J. N. K. & WU, C. F. J. 1988. Resampling Inference with complex survey data. *Journal of the American Statistical Association,* 83 (401)**,** 231 - 241.

RASCH, G. 1960. *Probabilistic models for some intelligence and attainment tests,* Copenhagen, Danish Institute for Educational Research, Chicago, The University of Chicago Press.

RASCH, G. 1980. *Probabilistic models for some intelligence and attainment tests,* Copenhagen, Danish Institute for Educational Research, Chicago, The University of Chicago Press.

RBMM 2005. Roll Back Malaria Monitoring and Evaluation Reference Group: Malaria indicator survey: basic documentation for survey design and implementation. Geneva.

REYBURN, H., MBAKILWA, H., MWANGI, R., MWERINDE, O., OLOMI, R., DRAKELEY, C. & WHITTY, C. J. M. 2007. Rapid diagnostic tests compared with malaria microscopy for guiding outpatient treatment of febrile illness in Tanzania: randomised trial. *BMJ, doi:10.1136/bmj.39073.496829.*

RIZOPOULOS, D. 2006. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software* 17**,** 1-25.

ROUSSEEUW, P. J. & CROUX, C. 1993. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association,* 88, No. 424**,** 1273 - 1283.

RUPPERT, D., WAND, M. P. & CARROLL, R. J. 2003. *Semiparametric Regression,* Cambridge Cambridge University Press.

SACKS, J., WELCH, W. J., T.J., M. & WYNN, H. P. 1989. Design and analysis of computer experiments. *Statistical Science,* 4 (4)**,** 400-423.

SAMPSON, P. D. & GUTTORP, P. 1992. Nonparametric estimation of non-stationary spatial covariance structure. *Journal of American Statistical Association,* 87**,** 108-119.

SÄRNDAL, C. E., SWENSSON, B. & ., J. W. 1992. *Model assisted survey sampling,* New York, Springer-Verlang. Inc.

SAS. 9.2. *The GLIMMIX Procedure: http://www.ats.ucla.edu/stat/sas/glimmix.pdf* [Online]. [Accessed May 2011].

SCHABENBERGER, O. & GOTWAY, C. A. 2005. *Statistical Methods for Spatial Data Analysis,* New York, Chapman and Hall/CRC.

SCHABENBERGER, O. & PIERCE, F. J. 2002. *Contemporary Statistical Models:for the Plant and Soil Science,* New York, CRC Press.

SCHAEFER, E., POTTER, F., WILLIAMS, S., DIAZ-TENA, N., RESCHOVSKY, J. D. & MOORE, G. 2003. Comparison of Selected Statistical Software Packages for Variance Estimation in the CTS Surveys. 600 Maryland Avenue, SW, Suite 550, Washington, DC 20024: Technical Publication No. 40.

SCHALL, R. 1991. Estimation in generalized linear models with random effects. *Biometrika,* 78**,** 719-727.

SCHELLENBERG, J. A., NEWELL, J. & SNOW, R. 1998. An analysis of the geographical distribution of severe malaria in children in Kilifi District, Kenya. *International Journal of Epidemiology,* 27(2)**,** 323-9.

SCHIMEK, M. 1997. Non- and Semiparametric Alternatives to Generalized Linear Models. *Computational Statistics,* 12**,** 173-191.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Annals of Statistics,* 6**,** 461-464.

SCOTT, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization,* New York, Wiley.

SHAO, J. & TU, D. 1995. *The Jackknofe and Bootstrap,* New York, Springer-Verland.

SHARGIE, E. B., GEBRE, T., NGONDI, J., GRAVES, P. M., MOSHER, A. W., EMERSON, P. M., EJIGSEMAHU, Y., ENDESHAW, T., OLANA, D., WELDEMESKEL, A., TEFERRA, A., TADESSE, Z., TILAHUN, A., YOHANNES, G. & JR, F. O. R. 2008. Malaria prevalence and mosquito net coverage in Oromia and SNNPR regions of Ethiopia. *BMC Public Health,* 8:321.

SHARGIE, E. B., NGONDI, J., GRAVES, P. M., GETACHEW, A., HWANG, J., GEBRE, T., MOSHER, A. W., CECCATO, P., ENDESHAW, T., JIMA, D., TADESSE, Z., TENAW, E., REITHINGER, R., EMERSON, P. M., JR, F. O. R. & GHEBREYESUS, T. A. 2010. Rapid increase in ownership and use of long-lasting insecticidal nets and decrease in prevalence of malaria in three regional states of Ethiopia, 2006-2007. *Journal of Tropical Medicine***,** doi: 10.1155/2010/750978.

SILVERMAN, B. W. 1986. *Density Estimation for Statistics and Data Analysis,* London, Chapman and Hall.

SIMONO , J. S. 1996. *Smoothing Methods in Statistics,* New York, Springer.

SINTASATH, D. M., GHEBREMESKEL, T. & LYNCH, M. 2005. Malaria Prevalence and Associated Risk Factors in Eritrea. *The American Journal of Tropical Medicine and Hygiene,* 72(6)**,** 682-687.

SKINNER, C. J., HOLT, D. & SMITH, T. M. F. 1989. *Analysis of Complex Survey,* New York, John Wiley & Sons.

SKRONDAL, A. & RABE-HESKETH, S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models,* Boca Raton, FL: Chapman & Hall/CRC. .

SMITH, D. E., WALKER, B., COOPER, D., ROSENBURG, E. & KALDOR, J. 2004. Is antiretroviral treatment of primary HIV infection clinically justified on the basis of current evidence? *AIDS,* 18**,** 709-718.

SMITH, P., FULL, S., UNDERWOOD, C., CHAMBER, R. & HILMES, D. 1998. Simulation study of alternative variance estimation methods. *Model Quality report in Business Statistics, Vol II: Comparison of variance Estimation Software and Methods.*

SNOW, R. W., PESHU, N. & FORSTER, D. 1998. Environmental and entomological risk factors for the development of clinical malaria among children on the Kenyan coast. *Transactions of the royal society of tropical medicine and hygiene,* 92**,** 381 - 385.

SOKAL, R. R. & ODEN, N. L. 1978. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society,* 10 199-228.

SOLOMON, H. & STEPHENS, M. A. 1977. Distribution of a sum of weighted chi-squrae variables. *Journal of the American Statistical Association,* 72**,** 881-885.

STEIN, M. L. 1987. Minimum norm quadratic estimation of spatial variograms. *Journal of American Statistical Association,* 82.

TEKOLA, E., TESHOME, G., JEREMIAH, N., PATRICIA, M. G., ESTIFANOS, B. S., EJIGSEMAHU, Y., AYELE, B., YOHANNES, G., TEFERI, T., MESSELE, A., ZERIHUN, M., GENET, A., ARYC, M. W., EMERSON, P. M. & RICHARDS, F. O. 2008. Evaluation of light microscopy and rapid diagnostic test for the detection of malaria under operational field conditions: a household survey in Ethiopia. *Malaria Journal,* 7:118**,** 1475-2875.

TENENHAUS, M. & YOUNG, F. W. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika,* 50**,** 91-119.

TENNANT, A., PENTA, M., TESIO, L., GRIMBY, G., THONNARD, J.-L. & SLADE, A. 2004. Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: The Pro-ESOR project. *Medical Care,* 42(Suppl 1)**,** 37-48.

THE CARTER CENTER (TCC) 2007. Prevalence and risk factors for malaria and trachoma in Ethiopia. The Carter Center.

THOMAS, D. R. 1989. Simultaneous confidenceintervals for propoetions under cluster sampling. *Survey Methodology,* 15**,** 187-201.

TIAN, D., SOROOSHIAN, S. & MYERS, D. E. 1993. Correspondence Analysis with MATLAB. *Computers & Geosciences* 19 (7)**,** 1007-1022.

TIERNEY, L. & KADANE, J. B. 1986. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association,* 81**,** 82-86.

TSIATIS, A. A. & DAVIDIAN, M. 2004. Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica,* 14**,** 809-834.

TULU, N. A. 1993. Malaria. In The Ecology of Health and Disease in Ethiopia 2nd edition. Edited by: Kloos H and Zein AZ. Boulder, USA, Westview Press Inc. 341-352.

VAN DEN WOLLENBERG, A. L. 1982. Two new test statistics for the Rasch model. *Psy-chometrika* 47**,** 123-140.

VAN DER HEIJDEN, P. G. M. & DE LEEUW, J. 1985. Correspondence analysis used complementary to loglinear analysis. *Psychometrika,* 50 (4) 429-447.

VAN DER LINDEN, W. J. & HAMBLETON, R. K. 1997. *Handbook of Modern Item Response Theory* New York, Springer.

VERBEKE, G. & DAVIDIAN, M. 2008. *Joint models for longitudinal data: Introduction and overview. In: Longitudinal Data Analysis: A handbook of modern statistical methods. Fitzmaurice, G., Davidian, M. Verbeke, G and Molenberghs, G. (Eds),* Boca Raton, Chapman & Hall/CRC.

VERBEKE, G., FIEUWS, S., MOLENBERGHS, G. & DAVIDIAN, M. 2012. The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research,* DOI: 10.1177/0962280212445834.

VERBEKE, G. & MOLENBERGHS, G. 2000. *Linear Mixed Models for Longitudinal Data,* New York, Springer-Verlag.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. & WELHAM, S. J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics,* 48**,** 269-311.

VIDAKOVIC, B. 1999. *Statistical Modeling by Wavelets,* New York, Wiley

VITTINGHOFF, E., GLIDDEN, D. V., SHIBOSKI, S. C. & MCCULLOCH, C. E. 2005. *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models. ,* New York:, Springer.

WAHBA, G. 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics,* 13**,** 1378 - 1402.

WAHBA, G. 1990. Spline Models for Observational Data. *SIAM, Philadelphia.*

WALLER, L. A. & GOTWAY, C. A. 2004. *Applied Spatial Statistics for Public Health Data,* New Jersey, John Wiley and sons, Inc.

WAND, M. P. & JONES, M. C. 1995. *Kernel Smoothing,* London, Chapman and Hall.

WANG, Y. 1998. Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society,* B 60**,** 159-174.

WEDDERBURN, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss - Newton method. 61**,** 439-447.

WEESIE, J. 2000. Estimating Rasch models with Stata. Rasch Measurement Transactions 13**,** 724.

WEI, G. C. G. & TANNER, M. A. 1990. A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association,* 85**,** 699-704.

WHO 2006a. The role of laboratory diagnosis to support malaria disease management.

WHO. 2006b. *World Health Organization: Health action in crises: Horn of Africa Health Review. [http://www.who.int/hac/crises/international/hoafrica/en/index.html]* [Online]. [Accessed February 2011].

WHO 2006c. World Health Organization: Systems for the early detection of malaria epidemics in Africa: an analysis of current practices and future priorities, country experience. Geneva, Switzerland, World Health Organization.

WHO 2009. Malaria Rapid Diagnostic Test Performance. Geneva, Switzerland.

WHO. 2011. *World Health Organization: Malaria http://www.who.int/ mediacentre/factsheets/fs094/en/* [Online]. [Accessed Nov 15 2011].

WOLFINGER, R. D. & O'CONNELL, M. 1993 Generalized Linear Mixed Models: a Pseudo-likelihood Approach. *Journal of Statistical Computation and Simulation,* 48 233-243.

WOLFINGER, R. W. 1993. Laplace's approximation for nonlinear mixed models. *Biometrika,* 80**,** 791-795.

WOLTER, K. M. 1985. *Introduction to variance estimation,* Tokyo, Springer-Verlag

WONGSRICHANALAI, C., BARCUS, M. J., MUTH, S., SUTAMIHARDJA, A. & WERNSDORFER, W. H. 2007. A Review of Malaria Diagnostic Tools: Microscopy and Rapid Diagnostic Test (RDT). *The American Journal of Tropical Medicine and Hygiene,* 77 (Suppl 6)**,** 119-127.

WOOD, S. N. 2006. *Generalized Additive Models: An Introduction with R,* Florida, Chapman & Hall/CRC.

WORRALL, E., BASU, S. & HANSON, K. 2002. The relationship between socio - economic status and malaria: a review of the literature. *Background paper for Ensuring that malaria control interventions reach the poor,* London 5th - 6th September

WRIGHT, B. D. & PANCHAPAKESAN, N. 1969. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 29**,** 23-48.

WU, H. & ZHANG, J.-T. 2006. *Nonparametric Regression Methods for Longitudinal Data Analysis,* New Jersey, John Wiley & Sons, Inc.

WU, L. 2010. *Mixed Effects Models for Complex Data,* Vancouver, Canada, CRC press.

YUNG, W. & RAO, J. N. K. 2000. Jackknife variance estimation under imputation for estimators using post-stratification information. *Journal of the American Statistical Association,* 95**,** 903-915.

ZHANG, D., LIN, X., RAZ, J. & SOWERS, M. 1998. Semi-parametric stochastic mixed models for longitudinal data. *Journal of American Statistical Association,* 93**,** 710-719.

ZHOU, G., MINAKAWA, N., GITHEKO, A. & YAN, G. 2004. Association between climate variability and malaria epidemics in the East African highlands. *Proceedings of the National Academy of Sciences* 101**,** 2375-2380.

ZIMMERMAN, D. L. & ZIMMERMAN, M. B. 1991. A comparison of spatial semivariogram estimators and corresponding kriging predictors. *Technometrics,* 33**,** 77 - 91.

ZURR, A. F., IENO, E. N. & SMITH, G. M. 2007. *Analysing Ecological Data,* New York, Springer.

ZUUR, A. F., IENO, E. N., WALKER, N. & SAVELIEV, A. A. 2009. *Mixed Effects Models and Extensions in Ecology with R* New York, Springer.

Appendices: Published papers

Ayele *et al. Malaria Journal* 2012, **11**:195
http://www.malariajournal.com/content/11/1/195

MALARIA JOURNAL

RESEARCH Open Access

Prevalence and risk factors of malaria in Ethiopia

Dawit G Ayele*, Temesgen T Zewotir and Henry G Mwambi

Abstract

Background: More than 75% of the total area of Ethiopia is malarious, making malaria the leading public health problem in Ethiopia. The aim of this study was to investigate the prevalence rate and the associated socio-economic, geographic and demographic factors of malaria based on the rapid diagnosis test (RDT) survey results.

Methods: From December 2006 to January 2007, a baseline malaria indicator survey in Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia was conducted by The Carter Center. This study uses this data. The method of generalized linear model was used to analyse the data and the response variable was the presence or absence of malaria using the rapid diagnosis test (RDT).

Results: The analyses show that the RDT result was significantly associated with age and gender. Other significant covariates confounding variables are source of water, trip to obtain water, toilet facility, total number of rooms, material used for walls, and material used for roofing. The prevalence of malaria for households with clean water found to be less. Malaria rapid diagnosis found to be higher for thatch and stick/mud roof and earth/local dung plaster floor. Moreover, spraying anti-malaria to the house was found to be one means of reducing the risk of malaria. Furthermore, the housing condition, source of water and its distance, gender, and ages in the households were identified in order to have two-way interaction effects.

Conclusion: Individuals with poor socio-economic conditions are positively associated with malaria infection. Improving the housing condition of the household is one of the means of reducing the risk of malaria. Children and female household members are the most vulnerable to the risk of malaria. Such information is essential to design improved strategic intervention for the reduction of malaria epidemic in Ethiopia.

Keywords: Generalized linear model, Odds ratio, Rapid diagnosis test, Risk factors, Survey design

Background

Malaria is a life-threatening caused by *Plasmodium* parasite infection. Malaria is the most deadly, and it predominates in Africa [1]. The problem of malaria is very severe in Ethiopia where it has been the major cause of illness and death for many years [1,2]. According to records from the Ethiopian Federal Ministry of Health, 75% of the country is malarious with about 68% of the total population living in areas at risk of malaria [1,2]. That is, more than 50 million people are at risk from malaria [3], and four to five million people are affected by malaria annually [4,5]. The transmission of malaria in Ethiopia depends on altitude and rainfall with a lag time varying from a few weeks before the beginning of the rainy season to more than a month after the end of the rainy season [6,7]. Epidemics of malaria are relatively frequent [8,9] involving highland or highland fringe areas of Ethiopia, mainly areas 1,000-2,000 m above sea level [1,7,10]. Malaria transmission peaks bi-annually from September to December and April to May, coinciding with the major harvesting seasons. This has serious consequences for Ethiopia's subsistence economy and for the nation in general. Major epidemics occur every five to eight years with focal epidemics as the commonest form. Early diagnosis and prompt treatment is one of the key strategies in controlling malaria. For areas where laboratory facilities are not available, clinical diagnosis is widely used [11,12]. To diagnose malaria, microscopy remains the standard method, but it is not accessible or affordable in most peripheral health facilities. The recent introduction of rapid diagnostic tests (RDT) for malaria is a significant step forward in case detection, management and reduction of unnecessary treatment. RDT could be used in malaria diagnosis during population-

* Correspondence: ejigmul@yahoo.com
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209 South Africa

based surveys and to provide immediate treatment based on the results.

Rapid diagnostic tests (RDTs) for malaria offer the potential to extend accurate malaria diagnosis to areas when microscopy services are not available, such as in remote locations or after regular laboratory hours. Rapid malaria diagnostic tests have been developed in the lateral flow format [13]. These tests use finger-stick blood, take only 10 to 15 minutes to complete, and do not require a laboratory. Non-clinical staff can easily learn to perform the test and interpret the results [14]. The objective of this paper is to identify the socio-economic, geographic and demographic risk factors of malaria using the rapid diagnosis test (RDT).

Methods

Study design

A baseline household cluster malaria survey was conducted by The Carter Center from December 2006 to January 2007. A questionnaire was developed as a modification of the Malaria Indicator Survey (MIS) Household Questionnaire. The questionnaire had two parts; the household interview and malaria parasite form. For this survey, the sampling frame was the rural populations of Amhara, Oromiya and SNNP regions, which is *kebele* (the smallest administrative unit in Ethiopia). Firstly, 224 *kebeles* of 25 household each were selected. From each *kebele*, out of the 25 households 12 even-numbered households were selected for malaria tests. All members of the household were tested for malaria by using RDT. In the survey, each room in the house was listed separately. During the study period, 5,708 households which were located in 224 clusters, covered in the survey. From the total of 5,708 households, Amhara, Oromiya and SNNP regions cover 4,101 (71.85%), 809 (14.17%) and 798 (13.98%) households respectively [15].

For the baseline household cluster malaria survey which was conducted by The Carter Center, a multi-stage cluster random sampling was used. By assuming the lowest measurement of prevalence malaria indicator, the sample size was estimated. Based on the assumption that prevalence of malaria to be the lowest indicator to be measured, the prevalence in the population was taken to be 8%. In Amhara region, each zone was regarded as a separate domain, while in Oromiya and SNNPR, the community-directed treatment with ivermectin (CDTI) areas combined were one domain. All ten Amhara zones were surveyed as separate domains, with 16 clusters in each zone (total 160 clusters). Bahir Dar town and two *woredas* with less than 10% of the population living in malarious areas were excluded. In Oromiya and SNNPR, sampling was done directly at the *kebele* level. From the total number of individuals who participated in the survey, 7,745 in Amhara, 1,996 in Oromiya and 1,860 in SNNP from all age groups were tested using RDT [15]. Further studies on the sampling procedure for the survey were studied by different researchers [16,17].

Malaria parasite testing was performed on consenting residents. A blood sample was collected by taking finger-prick blood from participants for malaria RDT. The test is capable of detecting both *Plasmodium falciparum* and other *Plasmodium* species. Participants with positive rapid tests were immediately offered treatment according to national guidelines.

Using the baseline household cluster malaria survey which was conducted by The carter Center in Amhara, Oromiya and SNNP regions, a number of research papers have been published. Individual, household and environmental risk factors of malaria in Amhara, Oromiya and SNNP regions of Ethiopia was studied by Graves *et al.* in 2008 [18]. To assess malaria infections in relation to socio-economic, demographic and environmental factors, they used univariate analysis. From the result it can be seen that overall prevalence of malaria was found to be low. The detailed report for this survey is presented by The Carter Center [15]. The other research paper which was conducted using this population-based survey is evaluation of light microscopy and rapid diagnosis test. This was done by Endeshaw *et al.* in 2008 [19]. The finding of this study suggested that blood slide microscopy found to be the best option for population-based prevalence survey of malaria *parasitaemia*. Similarly, Sharge *et al.* studied net coverage in Oromiya and SNNP regions of Ethiopia and ownership and use of long lasting insecticidal nets in 2008 and 2010 [17,20]. The result from these studies implies that malaria continues to be a significant public health problem in the surveyed regions of Ethiopia. The use of mosquito nets resulted in the decline of the prevalence of malaria in Amhara, Oromiya and SNNP regions of Ethiopia. These studies focused only to univariate analysis, but advanced statistical analysis is very important to identify the socio-economic, demographic and geographic factors which have influence to the risk of malaria. Multivariate statistical methods used for this study. Therefore, in this study the variables of interest are as follows.

Response variable

The outcome of interest is malaria RDT result. RDTs assist in the diagnosis of malaria by detecting evidence of malaria parasites in human blood and are an alternative to diagnosis based on clinical grounds or microscopy, particularly where good quality microscopy services cannot be readily provided. Thus, the response variable is binary, indicating whether or not a person was positive for malaria.

Independent variables

The independent covariates comprised the baseline socio-economic, demographic, and geographic variables that included gender, age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms, main material of the room's wall, main material of the room's roof, main material of the room's floor, incidence of anti-malarial spraying in the past 12 months, use of mosquito nets and total number of nets. Malaria test (RDT result), age and sex were collected at individual level. Altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms, main material of the room's walls, main material of the room's roof, main material of the room's floor, use of anti-malarial spray in the past 12 months, use of mosquito nets and total number of nets were all collected at household level.

The statistical model

Data was analysed by fitting a generalized linear model (GLM). The GLM generalizes linear regression by relating the response variable to predictor variables via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

The class of GLM includes many well-known statistical models such as: multiple regression for normal responses; logistic and probit regression for binary responses; binomial counts, or proportions; Poisson and negative binomial regression; log-linear categorical data analysis models; gamma regression for variance models; and exponential and gamma models for survival time models.

The literature on GLM and their extensions is vast [21-24]. Generalized linear models have been extended in many ways, such as accommodating random and mixed effects, accommodating correlated data, relaxing distributional assumptions, allowing semi-parametric linear predictors [25,26].

The logistic regression model is classified under GLM. This model is used to model binary data. The logistic regression model used to analyse data from complex sampling designs is referred to as survey logistic regression models. Survey logistic regression models have the same theory as ordinary logistic regression models. The difference between ordinary and survey logistic is that survey logistic accounts for the complexity of survey designs. But, for data from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical.

For ordinary logistic regression, a method of maximum likelihood estimation is used to estimate parameters of the model. But, estimation of the standard errors of the parameter estimates is very complicated for data that comes from complex designs. The complexities in variance estimation arise partly from the complicated sample design and the weighting procedure imposed. Therefore, the incorporation of sampling information is important for the proper assessment of the variance of a statistic [27-29]. Since weighting and specific sample designs are particularly implemented for increasing the efficiency of a statistic, their incorporation in the variance estimation methodology is of major importance [30]. Thus, the bias induced under this simplifying approach depends on the particular sampling design and should be investigated circumstantially. Therefore, there are several methods to obtain the covariance matrix [31]. These methods include the Taylor expansion approximation procedure, jack-knife estimator, bootstrap estimator, balanced repeated replication method and random groups method [32,33].

Results

The data analysis for this study was done using SAS version 9.2. The deviance was used to compare alternative models during model selection. Change in the deviance was used to measure the extent to which the fit of the model improves when additional variables were included. To avoid confounding effects, the model was fitted in two steps. The model was fitted to each predictor variables one at a time. In stage two the significant predictors were retained in a multivariate logistic regression model. In addition to the main effects, possible combinations of up to three-way interaction terms were added and assessed to further avoid and mitigate the problem of confounding.

The objective of the analysis is to identify the individual characteristics that could be associated with the malaria rapid diagnosis test outcome. On the other hand, this study focused on identifying the household characteristics which could be associated with the increase/decrease of the number of malaria infected household members. These household characteristics which were included in the model are main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, number of persons per room, main material of the room's wall, main material of the room's roof, main material of the room's floor, use of anti-malaria spray in the past 12 months, use of mosquito nets, number of nets per person, family size, region and altitude of region. The individual characteristics are gender and age.

To make statistically valid inferences, the analysis of the data must account for the design of the study. The SAS procedure which performs logistic regression for categorical responses in sample survey data was used [34].

The maximal model with significant effects is given in Tables 1 and 2. These models have the smallest deviance (−2logL) amongst all the nested models with the three-way interaction effects. Based on the final model, six interactions reduced the deviance (−2logL). Therefore, the final model includes all the main effects and the six interaction effects.

Toilet facilities, availability of television, number of rooms per person, main material for walls, number of months the room was sprayed, number of mosquito nets per person, age and family size were found to be significant main effects. In addition to the main effects, five significant two-way interaction terms and one three-way interaction terms was obtained. The two-way interaction terms were: the interaction between main source of drinking water and main material of the room's roof; use of anti-malarial spray and use of mosquito nets; time taken to collect water and floor material; gender and main source of drinking water; gender and main material of the room's floor; and gender and use of anti-malarial spray. Three-way interaction between gender, main source of drinking water and availability of electricity was also significant. Age, family size, toilet facilities, availability of television, number of persons per room, wall material and number of months anti-malarial spray was used were the significant main effects, which were not involved in significant interaction terms (Table 2). Accordingly, the effect of these variables can be directly interpreted using the odds ratio (OR).

Table 1 Estimates and odds ratios of socio-economic, demographic and geographic factors on RDT

	Estimate	OR	95% CI		P -value
			Lower	Upper	
Intercept	−3.030	0.048	0.016	0.125	0.001
Age	−0.031	0.970	0.319	2.505	0.0001
Sex (ref. male)					
Female	−1.820	0.162	0.053	0.418	<.0001
Family size	0.049	1.057	1.014	1.124	<.0001
Region (ref. SNNP)					
Amhara	−0.099	0.906	0.178	0.183	0.521
Oromiya	−0.184	0.832	0.238	8.581	0.183
Toilet facility (Ref. No facility)					
Pit latrine	−0.3213	0.725	2.575	2.147	<.0001
Toilet with flush	−0.5935	0.552	2.632	4.909	<.0001
Main source of drinking water (ref. protected water)					
Tap water	−0.038	0.963	0.316	0.373	<.0001
Unprotected water	0.717	2.048	0.673	5.289	0.007
Availability of television (ref. no)					
Yes	0.304	1.356	0.446	3.500	0.024
Number of rooms/person	−0.473	0.623	0.205	1.610	0.044
Main material of room's wall (ref. cement block)					
Mud block/stick/wood	−2.326	0.098	0.032	0.252	0.048
Corrugated metal	−0.620	0.538	0.471	0.826	0.001
Main material of room's roof (ref. corrugate)					
Thatch	1.325	3.761	1.236	9.712	<.0001
Stick and mud	−1.960	0.141	0.046	0.364	<.0001
Main material of room's floor (ref. earth/Local dung plaster)					
Wood	−1.701	0.183	0.149	0.443	<.0001
Cement	−3.927	0.014	1.014	4.876	0.018
Anti-malarial spraying					
No	1.857	6.405	2.105	16.539	0.046
Use of mosquito nets (ref. no)					
Yes	−0.095	0.910	0.299	2.349	<.0001
Number of nets/person	−0.782	0.457	0.150	1.181	<.0001

Tables 1 and 2 present estimates of socio-economic, demographic and geographic factors on RDT. Based on the result for a unit increase in age, implies a reduction of the odds of a positive malaria test by 3.0% (OR = 0.970, p - value = 0.0001). Furthermore, for a unit increase in family size, the number of persons infected by malaria in the household increased by 5.1% (OR = 1.057, p - value < .0001). Furthermore, compared to households which had no toilet facilities, those with a pit latrine were at lower risk of malaria diagnosis (OR = 0.725, p-value = <.0001) as well as households with flush toilets (OR = 0.552, p - value = <.0001). Households who were using mosquito nets were found to be at a lower risk of malaria compared to the households who were not using mosquito nets (OR = 0.91, p - value = <.0001). Furthermore, for a unit increase in the number of nets, the odds of positive malaria diagnosis test decreases by 54% (OR = 0.46, p - value = <0.0001) for the household.

Interaction effects

The relationship between gender, main source of drinking water and availability of electricity is presented in Figure 1 to indicate the risk of positive malaria RDT is higher for unprotected water use by female respondents. However, for both males and females, positive RDT is low for households using tap water and electricity.

With reference to households that have tap water for drinking and corrugated iron-roofed houses, the risk of positive malaria RDT was significantly lower than for households living in stick and mud-roofed houses and drinking unprotected water (OR = 8.09624, p-value < 0.0001). As Figure 2 indicates, higher positive malaria diagnosis test was found for households that reportedly used unprotected water for drinking.

Table 2 Estimates and odds ratios of socio-economic, demographic and geographic factors on RDT for interaction effects

	Estimate	OR	95% CI		P -value
			Lower	Upper	
Main source of drinking water and main material of the room's roof (ref. Protected water & cement block)					
Tap water and Mud block/stick/wood	−3.339	0.035	0.007	0.177	<.0001
Tap water and Corrugated metal	−3.377	0.034	0.007	0.184	<.0001
Unprotected water and Mud block/stick/wood	−4.008	0.018	0.003	0.130	<.0001
Unprotected water and Cement block	−1.857	0.156	0.022	1.119	<.0001
Time to collect water and material of room's floor (ref. Less than 30 minutes and earth/local dung plaster)					
Greater than 90 minutes and Cement	−0.423	0.655	0.066	1.478	<.0001
Greater than 90 minutes and Wood	−0.721	0.486	0.160	1.478	0.0013
Between 30–40 minutes and Cement	−1.901	0.149	0.049	1.478	<.0001
Between 30–40 minutes and Wood	1.554	4.729	0.821	9.220	<.0001
Between 40–90 minutes and Cement	−0.739	0.933	0.129	1.258	0.0011
Between 40–90 minutes and Wood	0.554	3.769	1.835	7.232	<.0001
Gender and main source of drinking water and main material of the room's roof (ref. Male & protected water)					
Female and Tap water	−0.069	0.933	0.624	1.397	0.0972
Female and Unprotected water	1.327	3.769	1.948	7.293	<.0001
Gender and material of room's floor (ref. Male and earth/Local dung plaster)					
Female and Cement	−0.372	0.689	0.158	1.254	<.0001
Female and Wood	−4.893	0.008	0.003	0.017	<.0001
Anti-malarial spraying and use of mosquito nets (ref. Yes & no)					
No and Yes	0.104	1.110	0.898	1.372	0.0319
Gender, main source of drinking water and electricity (ref. Male, protected water & yes)					
Female, tap water and no	0.550	1.734	1.137	2.643	0.0172
Female, unprotected water and no	−1.319	0.267	0.132	0.542	0.0049

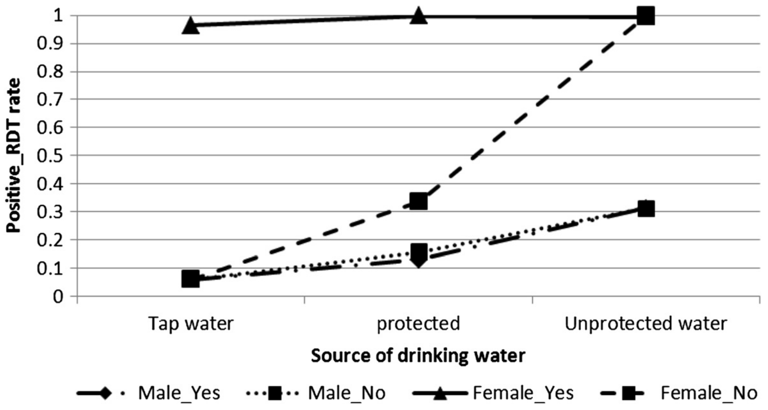


Figure 1 **Log odds associated with rapid diagnosis test and gender, source of drinking water with availability of electricity.**

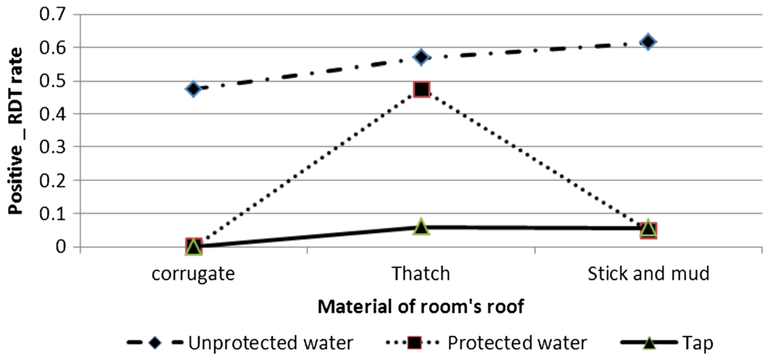


Figure 2 Log odds associated with rapid diagnosis test and material of room's roof with main source of drinking water.

The OR values for the interaction between gender and main material of the room's floor is given in Figure 3. Based on the result, positive malaria diagnosis test was significantly higher for females than for males who reported that the material of the room’s floor was earth/ local dung (OR = 1.358, p - value < .0001) as well as those who reported that the material of the room’s floor was wood (OR = 2.415, p - value < 0.0001). There was however, higher positive malaria diagnosis test found for both males and females who reported that the material of the room’s floor was wood.

Positive RDT was significantly higher for respondents living in a room with a wooden or earth/local dung floor than for those living in a room with a cement floor for respondents who took 40–90 minutes to collect water. But, for respondents who took less than 40 minutes to collect water, positive RDT was low (refer Figure 4).

Prevalence of malaria was significantly higher for male than for female respondents who were living in a house treated with anti-malarial spray (refer Figure 5). For both males and females who were living in a house that had not been sprayed, the risk of positive malaria was significantly higher. On the other hand, for males living in a house that had not been treated with anti-malarial spray, the risk of malaria infection for males is more than that of females.

The use of mosquito nets and applying anti-malarial spray to the walls of the house altered the risk of malaria. The risk of malaria was low for individuals who lived in houses that had been sprayed and used malaria nets. It is shown in Figure 6 that the estimated risk of malaria was higher for individuals with no mosquito nets.

Discussion

The government of Ethiopia has developed strategies related to human resource development, monitoring, and evaluation to control malaria and reduce the hardships it causes. However, the key goals and targets set by the government are aimed at making those areas with historically low malaria transmission, malaria free and a near zero malaria transmission in the remaining malarious areas of the country [35]. Some studies conducted so far have suggested that malaria should be regarded as a disease of the poor or a disease of poverty [36]. This claim can be substantiated by noting the global distribution of malaria where the concentration of the disease is in poorest continents and countries. Being a primary

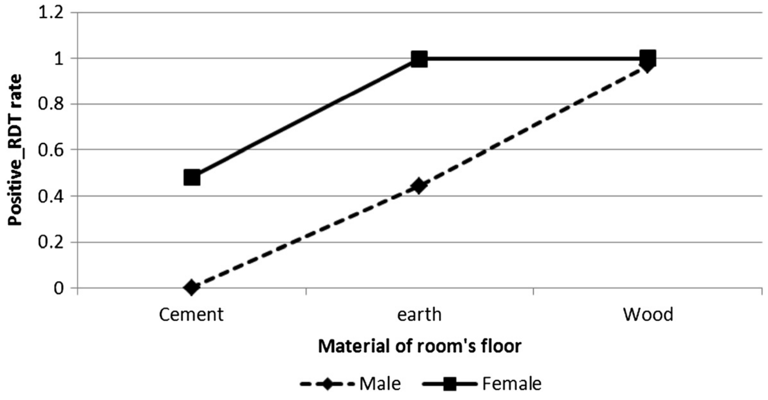


Figure 3 Log odds associated with rapid diagnosis test and gender with material of room's floor.

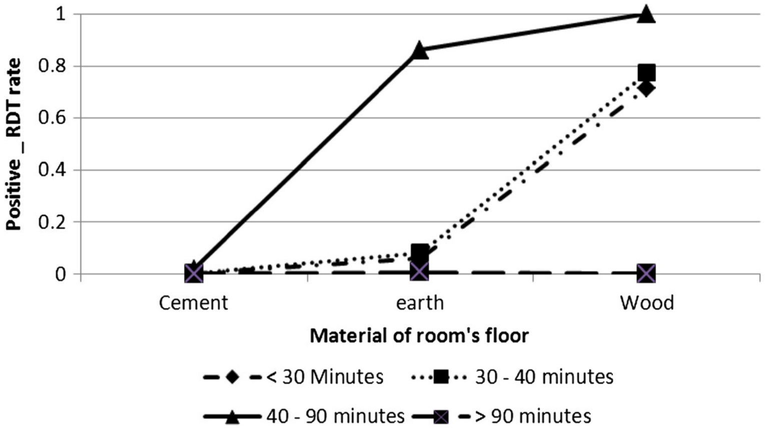


Figure 4 Log odds associated with rapid diagnosis test and material of room's floor with time to collect water.

cause of poverty, some studies suggest that a better understanding of the relationships between malaria and poverty is needed to enable the design of coherent and effective policies and tools to tackle the problem. Since poverty is related to socio-economic factors, it is important to identify those factors that are also related to the risk of malaria [37,38].

The present study was conducted based on the 2006 baseline malaria indicator survey in Amhara, Oromiya and Southern Nation Nationalities and People (SNNP) regions of Ethiopia. This survey was a population-based household cluster survey. There were 224 clusters and each cluster consists of 25 households. For this survey, the sampling frame was the rural population of Amhara, Oromiya and SNNP regions. Therefore, the data used for this study was from complex survey. For the statistical analysis, the study used generalized linear model. For this study, gender, age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms, main material of the room's wall, main material of the room's roof, main material of the room's floor, incidence of anti-malarial spraying in the past 12 months, use of mosquito nets and total number of nets with up to three-way interaction effects were used for the analysis.

Based on these facts, the findings of this study show that the following socio-economic factors are related to malaria risk: construction material of walls, roof and floor of house; main source of drinking water; time taken to collect water; toilet facilities and availability of electricity. Besides socio-economic factors, there are demographic and geographic factors that also had an effect on the risk of malaria. These include gender, age, family size and the region where the respondents lived. In addition to the main effects, there were interactional effects between the socio-economic, demographic and geographic factors that also influenced the risk of malaria. Most notable of these were the interaction between the main source of drinking water and the main construction material of the room's roof; the time taken to

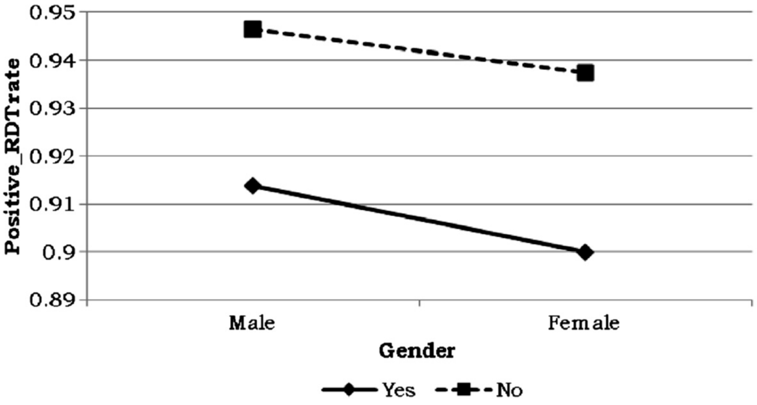


Figure 5 Log odds associated with rapid diagnosis test and anti-malaria spray with gender.

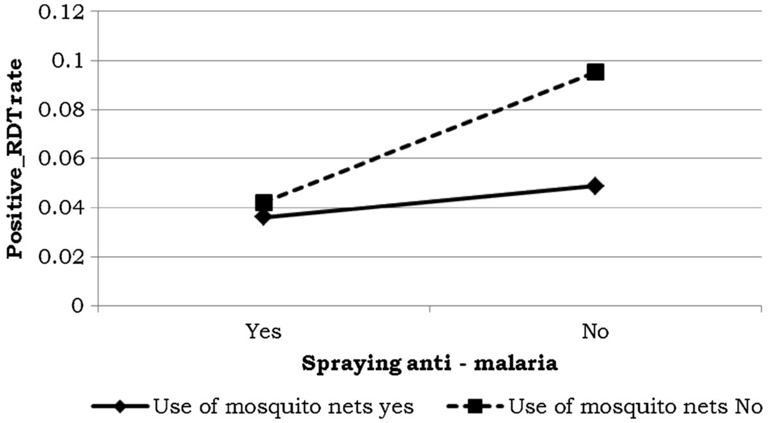


Figure 6 Log odds associated with rapid diagnosis test and use of anti-malaria with use of mosquito nets.

collect water and the main construction material of the room's floor; gender and the main source of drinking water; gender and the availability of electricity; gender and the main construction material of the room's floor and finally, interaction between gender, main source of drinking water and the availability of electricity.

From the study, it was observed that residents living in the Amhara region were found to be more at risk of malaria than those living in the SNNP and the Oromiya regions. Similarly, houses that were treated with anti-malarial spray were less likely to be affected by malaria. One of the major challenges in the control of malarial infection was found to be the use of toilet facilities. From the results, it was observed that households with no toilet facilities were more likely to be positive for malaria diagnosis test. Furthermore, positive malaria diagnosis rate decreased with age. But, for households, the risk of malaria increased per unit increase in family size. Generally, malaria parasite prevalence differed between age and gender with the highest prevalence occurring in children and females. The findings of the association between socio-economic factors and malaria prevalence are similar to some of the results from previous studies [39-41]. In addition to this in 1998 and 2000, study was conducted by Ghebreyesus *et al.* and Snow *et al.* [42,43] in Ethiopia and Kenya, respectively. The objectives of the studies were to assess different types of materials used in the construction of walls, roofs and floors of a house. They used generalized linear models, Poisson and logistic models, for their study. Based on their findings, they observed association between any roof, wall and floor material and risk of malaria. Therefore, the finding of this study is similar to the previous results.

This study suggest that having toilet facilities, access to clean drinking water and the use of electricity offers a greater chance of not being positive for malaria diagnosis. Using mosquito nets and spraying anti-malarial treatment on the walls of the house were also found to be a way of reducing the risk of malaria. In addition to this, having a cement floor and corrugated iron roof was found to be one means of reducing the risk of malaria. Based on the study findings, different types of housing have an influence on the risk of malarial transmission with those houses constructed of poor quality materials having an increased risk. Moreover, the presence of particular structural features, such as bricks, that may limit contact with the mosquito vector, also reduces infection. Therefore, the risk of malaria is higher for households in a lower socio-economic bracket than for those that enjoy a higher status and who are able to afford to take measures to reduce the risk of transmission.

This study suggests that with the correct use of mosquito nets, anti-malarial spraying and other preventative measures, coupled with factors such as the number of rooms in a house, the incidence of disease is decreased. However, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria.

Ethical clearance

The ethical protocol received approval from the Emory University Institutional Review Board (IRB 1816) and Amhara, Oromiya and SNNPR regional health bureaux. Informed consent was sought in accordance with the tenets of the declaration of Helsinki.

Abbreviations

FMH: Federal ministry of health of ethiopia; GLM: Generalized linear model; OR: Odds ratio; RDT: Rapid diagnosis test; SNNP: Southern nation nationalities and people; WHO: World health organization.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DGA acquired the data, performed the analysis and drafted the manuscript. TTZ and HGM designed the research. All authors discussed the results and implications and commented on the manuscript at all stages. All authors contributed extensively to the work presented in this paper. All authors read and approved the final manuscript.

Acknowledgements

We thank, with deep appreciation, The Carter Center of Ethiopia, for providing and giving permission to use the data for this study.

Received: 6 March 2012 Accepted: 30 May 2012
Published: 12 June 2012

References

1. Adhanom TDW, Witten HK, Getachew A, Seboxa T: **Malaria**. In *The Epidemiology and Ecology of Health and Disease in Ethiopia 1st edition*. Edited by Berhane Y, Hailemariam D, Kloos H. Ababa Addis, Ethiopia: Shama PLC; 2006:556–576.
2. Federal Ministry of Health (FMH): *Malaria and Other Vector-borne Diseases Control Unit*. Addis Ababa, Ethiopia: Federal Ministry of Health of Ethiopia; 1999.
3. Lesaffre E, Spiessens B: **On the effect of the number of quadrature points in a logistic random-effects model: an example.** *Applied Statistics* 2001, **50**:325–335.
4. Federal Ministry of Health (FMH): *Guideline for malaria epidemic prevention and control in Ethiopia*. 2nd edition. Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2004.
5. World Health Organization: *Health action in crises: Horn of Africa Health Review*. http://www.who.int/hac/crises/international/hoafrica/en/index.htm.
6. Deressa W, Ali A, Enqusellassie F: **Self-treatment of malaria in rural communities, Butajira, southern Ethiopia.** *Bull World Health Organ* 2003, **81**:261–268.
7. Tulu NA: **Malaria**. In *The Ecology of Health and Disease in Ethiopia*. 2nd edition. Edited by Kloos H, Zein AZ. Boulder, USA: Westview Press Inc; 1993:341–352.
8. WHO: *Systems for the early detection of malaria epidemics in Africa: an analysis of current practices and future priorities, country experience*. Geneva, Switzerland: World Health Organization; 2006.
9. Zhou G, Minakawa N, Githeko A, Yan G: **Association between climate variability and malaria epidemics in the East African highlands.** *Proc Natl Acad Sci U S A* 2004, **101**:2375–2380.
10. Federal Ministry of health (FMH): *National five-year strategic plan for malaria prevention and control in Ethiopia 2006 – 2010*. Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2006.
11. Federal Ministry of Health (FMH): *Malaria: Diagnosis and Treatment Guidelines for Health Workers in Ethiopia*. Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2004.
12. WHO: *New Perspectives: Malaria Diagnosis. Report of a Joint WHO/USAID: Informal Consultation held on 25–27 October 1999*. Geneva, Switzerland: World Health Organization; 2000:4–48. 1999.
13. WHO: *Malaria rapid diagnostic test performance*. Geneva, Switzerland: World Health Organization; 2009.
14. Wongsrichanalai C, Barcus MJ, Muth S, Sutamihardja A, Wernsdorfer WH: **A Review of Malaria Diagnostic Tools: Microscopy and Rapid Diagnostic Test (RDT).** *AmJTrop Med Hyg* 2007, **77**(Suppl 6):119–127.
15. TCC: *Prevalence and risk factors for malaria and trachoma in Ethiopia*. Addis Ababa, Ethiopia: The Carter Center; 2007.
16. Emerson PM, Ngondi J, Biru E, Graves PM, Ejigsemahu Y, Gebre T, Endeshaw T, Genet A, Mosher AW, Zerihun M, Messele A, Richards FO: **Integrating an NTD with one of "The Big Three": Combined malaria and trachoma survey in Amhara region of Ethiopia.** *PLoS Negl Trop Dis* 2008, **2**:e197.
17. Shargie EB, Gebre T, Ngondi J, Graves PM, Mosher AW, Emerson PM, Ejigsemahu Y, Endeshaw T, Olana D, WeldeMeskel A, Teferra A, Tadesse Z, Tilahun A, Yohannes G, Richards FO Jr: **Malaria prevalence and mosquito net coverage in Oromia and SNNPR regions of Ethiopia.** *BMC Publ Health* 2008, **8**:321.
18. Graves PM, Richards FO, Ngondi J, Emerson PM, Shargie EB, Endeshaw T, Ceccatod P, Ejigsemahu Y, Aryc W, Moshera, Hailemariame A, Zerihun M, Teferi T, Ayele B, Mesele A, Yohannes G, Tilahun A, Gebre T: **Individual, household and environmental risk factors for malaria infection in Amhara, Oromia and SNNP regions of Ethiopia.** *Trans R Soc Trop Med Hyg* 2009, **103**:1211–1220.
19. Tekola E, Teshome G, Jeremiah N, Patricia MG, Estifanos BS, Ejigsemahu Y, Ayele B, Yohannes G, Teferi T, Messele A, Zerihun M, Genet A, Mosher AW, Emerson PM, Richards FO: **Evaluation of light microscopy and rapid diagnostic test for the detection of malaria under operational field conditions: a household survey in Ethiopia.** *Malar J* 2008, **7**:118.
20. Shargie EB, Ngondi J, Graves PM, Getachew A, Hwang J, Gebre T, Mosher AW, Ceccato P, Endeshaw T, Jima D, Tadesse Z, Tenaw E, Reithinger R, Emerson PM, Richards FO, Ghebreyesus TA: **Rapid increase in ownership and use of long-lasting insecticidal nets and decrease in prevalence of malaria in three regional states of Ethiopia, 2006–2007.** *J Trop Med 2010* 2010, pii 750978.
21. Schabenberger O: *Gotway CA: Statistical Metods for Spatial Data Analysis*. New York: Chapman and Hall/CRC; 2005.
22. Goovaerts P: *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press; 1997.
23. Goovaerts P, Jacquez GM, Greiling D: **Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms.** *Geogr Anal* 2005, **37**:152–182.
24. Chiles JP, Delfiner P: *Geostatistics*. Modelling Spatial Uncertainty. Chichester: Wiley; 1999.
25. Schimek M: **Non- and semiparametric alternatives to generalized linear models.** *Comput Stat* 1997, **12**:173–191.
26. Smith D, Walker B, Cooper D, Rosenburg E, Kaldor J: **Is antiretroviral treatment of primary HIV infection clinically justified on the basis of current evidence?** *AIDS* 2004, **18**:709–718.
27. Bivand RS, Pebesma EJ, Gomez-Rubio V: *Applied Spatial Data Analysis with R*. New York: Springer; 2008.
28. Hengl T: *A Practical Guide to Geostatistical Mapping of Environmental Variables*. Italy: European Commission, Joint Research Centre, Institute for Environment and Sustainability; 2007.
29. Matheron G: **Principles of Geostatistics.** *Econ Geol* 1963, **58**:1246–1266.
30. Schaefer E, Potter F, Williams S, Diaz-Tena N, Reschovsky JD, Moore G: *Comparison of Selected Statistical Software Packages for Variance Estimation in the CTS Surveys*. 600 Maryland Avenue, SW, Suite 550, Washington, DC 20024: Technical Publication; 2003.
31. Cressie N, Hawkins DH: **Robust estimation of the variogram.** *Math Geol* 1980, **12**:115–125.
32. Wolter KM: *Introduction to variance estimation*. Tokyo: Springer; 1985.
33. Lee ES, Forthofer RN: *Analyzing Complex Survey Data*. California: Sage Publications; 2006.
34. SAS 9.2: *SAS/STAT® 9.2 User's Guide The SURVEYLOGISTIC Procedure (Book Excerpt)*. Cary, NC, USA: In SAS Institute Inc; 2008.
35. FMH: *Ethiopia National Malaria Indicator Survey 2007*. Addis Ababa, Ethiopia: Federal Ministry of Health of Ethiopia; 2008.
36. Abegunde D, Stanciole A: *An estimation of the economic impact of chronic noncommunicable diseases in selected countries*. Department of Chronic Diseases and Health Promotion (CHP): World Health Organization; 2006.
37. Mendis K, Rietveld A, Warsame M, Bosman A, Greenwood B, Wernsdorfer WH: **From malaria control to eradication: The WHO perspective.** *Trop Med Int Health* 2009, **14**:1–7.
38. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW: **The global distribution and population at risk of malaria: past, present, and future.** *Lancet Infect Dis* 2004, **4**:327–336.
39. Banguero H: **Socio-economic factors associated with malaria in Colombia.** *Soc Sci Med* 1984, **19**:1099–1104.
40. Koram K, Bennett S, Adiamah J, Greenwood B: **Socio-economic risk factors for malaria in a peri-urban area of the Gambia.** *Trans R Soc Trop Med Hyg* 1995, **89**:146–150.
41. Sintasath DM, Ghebremeskel T, Lynch M: **Malaria prevalence and associated risk factors in Eritrea.** *AmJTrop Med Hyg* 2005, **72**:682–687.
42. Ghebreyesus T, Haile M, Witten K, Getachew A, Yohannes M, Lindsay S: **Household risk factors for malaria among children in the Ethiopian highlands.** *Trans R Soc Trop Med Hyg* 2000, **94**:17–21.
43. Snow RW, Peshu N, Forster D: **Environmental and entomological risk factors for the development of clinical malaria among children on the Kenyan coast.** *Trans R Soc Trop Med Hyg* 1998, **92**:381–385.

doi:10.1186/1475-2875-11-195
Cite this article as: Ayele *et al.*: Prevalence and risk factors of malaria in Ethiopia. *Malaria Journal* 2012 **11**:195.

academicJournals

Vol. 5(7), pp. 335-347, July 2013
DOI: 10.5897/IJMMS2013.0956
ISSN 2006-9723 ©2013 Academic Journals
http://www.academicjournals.org/IJMMS

International Journal of Medicine and Medical Sciences

Full Length Research Paper

The risk factor indicators of malaria in Ethiopia

Dawit Getnet Ayele, Temesgen T. Zewotir and Henry G. Mwambi

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa.

Accepted 18 June, 2013

This study evaluates the effects of socio-economic, demographic and geographic indicators on the malaria rapid diagnosis test (RDT), using the baseline malaria indicator survey of 2007. This survey covered the Amhara, Oromiya and Southern Nations, Nationalities, and People's Region (SNNPR) of Ethiopia. A total of 224 clusters of, on average, 25 households each were selected. In total, 28,994 individuals participated in the survey. A generalized linear mixed model was used to analyze the data where the response variable was the presence or absence of malaria using the RDT. The results showed that for households with toilet facilities, clean drinking water and more living space, the chances of testing positive for malaria RDT decreased. Moreover, using malaria nets and spraying the house walls with anti-mosquito were found to be effective control measures.

Key words: Cluster sampling, interaction effect, mixed model, odds ratio, rapid diagnostic test.

INTRODUCTION

While malaria has long been a cause of human suffering and mortality in Sub-Saharan Africa (Eisele et al., 2010), in Ethiopia the problem is particularly severe. Here, it is the major cause of illness and death (Schabenberger and Gotway, 2005), with 75% of the total area being malarious (Cressie, 1991), and approximately 68% of the Ethiopian population living in these affected areas. Annually, about 4 to 5 million Ethiopians are affected by malaria (Federal Ministry of Health (FMH), 2004a; World Health Organization (WHO), 2006a). Malaria transmission in Ethiopia is seasonal, depending mostly on altitude and rainfall, with a lag time varying from a few weeks before the beginning of the rainy season to more than a month after the end of the rainy season (Deressa et al., 2003; Tulu, 1993).

Malaria epidemics in Ethiopia are relatively frequent (WHO, 2006b; Zhou et al., 2004), involving highland or highland fringe areas, mainly 1,000 to 2,000 meters above sea level (Adhanom, 2006; FMH, 2006; Tulu, 1993). Malaria transmission peaks bi-annually from September to December and April to May, coinciding with the major harvesting seasons (FMH, 2004a). This seasonality has serious consequences for the subsistence economy of Ethiopia's countryside and for the nation in general. Early diagnosis and prompt treatment is one of the key strategies in controlling malaria. For areas where laboratory facilities are not available, clinical diagnosis is widely used (FMH, 2004b; WHO, 1999). To diagnose malaria, microscopy remains the standard method. However, it is not accessible and affordable in most peripheral health facilities. The recent introduction of rapid diagnosis test (RDT) for malaria has become a significant step forward in case detection, management and reduction of unnecessary treatment in Ethiopia (Tekola et al., 2008).

In order to estimate the prevalence of malaria parasites in Ethiopia, a population based survey was conducted in 2006/2007. Rapid diagnostic tests as well as the conventionally accepted diagnostic tests using standard microscopy of peripheral blood slides were used for this survey. Both tests use *finger-stick* or *venous* blood. The level of disagreement in this survey between the results of microscopy and RDT was studied by Tekola et al. (2008) and found to be insignificant.

*Corresponding author. E-mail: ejigmul@yahoo.com, 208517203@ukzn.ac.za. Tel: +27739720957. Fax: +27332605648.

The objective of this study is to identify the socio-economic, demographic and geographic risk factors associated with the prevalence of malaria obtained from the rapid diagnosis tests.

METHODS AND MATERIALS

Study design

The Carter Center (TCC) conducted a baseline household cluster malaria survey in Ethiopia in 2007. The questionnaire was developed as a modification of the malaria indicator survey (MIS) household questionnaire. The questionnaire had two parts; the household interview and malaria parasite form.

For the baseline household cluster malaria survey which was conducted by TCC, a multi-stage cluster random sampling was used. By assuming the lowest measurement of prevalence malaria indicator, the sample size was estimated. Therefore, for TCC baseline household cluster malaria survey in Amhara, Oromiya and the Southern Nations, Nationalities and People's (SNNP) regions of Ethiopia which was conducted in 2007, the design was a population-based household cluster survey. Based on these clusters, zone-level estimates of indicators were obtained for Amhara region, and sub-regional estimates were obtained for Oromiya and SNNPR. Furthermore, the sampling design was involved to select households within each first-stage cluster, or Kebele (smallest administrative unit in Ethiopia). From the 224 selected Kebeles, 25 households were chosen, from which even-numbered households were selected for the malaria (RDT). All individuals in these 12 households (even-numbered households) were eligible for individual interviews. Furthermore, each room in the house was listed separately. By using the mosquito nets as a guide, it was possible to determine the number of persons sleeping in each room. This information was useful in determining the number of sleeping rooms both within and outside the house. In addition to the number of rooms and number of nets, the persons sleeping under each net were listed. Further studies on the sampling procedure for the survey were studied by different researchers (Emerson et al., 2008; Shargie et al., 2008).

Malaria parasite testing was performed on consenting residents. The blood sample for malaria RDT was collected by taking finger-prick blood samples from participants. The RDT used was ParaScreen which is capable of detecting both *Plasmodium falciparum* and other *Plasmodium* species. Participants with positive rapid tests were immediately offered treatment according to national guidelines.

Variable of interest

Response variables

The outcome of interest is the RDT result. RDTs assist in the diagnosis of malaria by detecting evidence of malaria parasites in human blood and are an alternative to diagnosis based on clinical grounds or microscopy, particularly where good quality microscopy services cannot be readily provided. Thus, the response variable was binary, indicating that either a person was positive or not positive.

Independent variables

The independent predictor variables consisted of baseline socio-economic, demographic and geographic variables, which were collected from each household. The socio-economic variables were the following: main source of drinking water; time taken to collect water; toilet facilities, availability of electricity, access to radio and television, total number of rooms, main construction material of the rooms' walls, main construction material of the room's roof and main construction material of the room's floor, incidence in the past 12 months of anti-mosquito spraying, use of mosquito nets and total number of nets. Geographic variables were region and altitude, and demographic variables were gender, age and family size. Of these variables, age and sex were collected at the individual level, while altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms, main construction material of walls, roof and floor, incidence of anti-mosquito spraying and use of mosquito nets were all collected at the household level.

The statistical model

A generalized linear mixed model (GLMM) was used to analyze the data. Classical linear models can be generalized using the generalized linear models (GLMs) to the exponential family of sampling distributions. These models have an immense impact on both theoretical and practical aspects in statistics. The term 'mixed' in the GLMMs means that the random effects and the fixed effects are mixed together to get a modified model. This can overcome the over-dispersion in the data and at the same time, accommodate the population heterogeneity. Therefore, the addition of random effects allows accommodating correlation in the context of a broad class of models for non-normally distributed data. These models become more applicable in practical situations. The logistic regression model, which includes the mixed effects, is a common choice for analysis of multilevel dichotomous data. In the GLMM, this model utilises the logit link, namely:

$$g(\mu_{ijk}) = logit\,(\mu_{ijk}) = log\left[\frac{\mu_{ijk}}{1-\mu_{ijk}}\right] = \eta_{ijk},$$

The conditional expectation $\mu_{ij} = E(Y_{ijk}|v_i,x_i)$ equals $p(Y_{ij} = 1|v_i,x_{ij})$, i.e., the conditional probability of a response given the random effects. Here, Y_{ijk} corresponds to the i^{th} respondent in the j^{th} household within k^{th} probabilistic sampling unit (PSU). Therefore, this model can also be written as:

$$P(Y_{ijk} = 1|v_i, x_{ijk}, z_{ijk}) = g^{-1}(\eta_{ijk})$$

Where, the inverse link function $g^{-1}(\eta_{ij})$ is the logistic cumulative distribution function (cdf), namely:

$$g^{-1}(\eta_{ijk}) = [1 + exp(-\eta_{ijk})]^{-1}$$

Table 1. Type 3 analysis of effects for the GLMM.

Effect	Num DF	F value	P > F
Age	1	10.16	0.0014
Gender	1	0.12	0.7257
Family size	1	75.32	<0.0001
Region	2	0.02	0.9761
Altitude	1	215.47	<0.0001
Main source of drinking water	2	6.59	0.0014
Time to collect water	3	7.46	<0.0001
Toilet facilities	2	5.2	0.0055
Availability of electricity	1	17.61	<0.0001
Availability radio	1	2.82	0.0732
Availability television	1	4.5	0.034
Number of rooms/person	1	38.49	<0.0001
Main material of the room's wall	2	12.94	<0.0001
Main material of the room's roof	2	2.07	0.1262
Main material of the room's floor	2	13.37	<0.0001
Spraying of anti- mosquito	1	986.9	<0.0001
Number of months room sprayed	1	944.72	<0.0001
Use of mosquito nets	1	11.62	0.0027
Number of nets/person	1	13.48	0.0002
Age and gender	1	0.027	0.9784
Main source of drinking water and main material of the room's roof	4	4.57	0.0004
Gender and use of mosquito nets	1	11.59	<0.0001
Time to collect water and main material of the room's floor	4	14.57	0.0024
Gender & main source of drinking water	1	33.46	<0.0001
Gender and main material of the room's floor	2	5.67	0.0035
Gender and spraying anti-mosquito spray	1	849.57	<0.0001
Use of mosquito nets and number of nets per person	1	849.57	<0.0001
Age, gender and source of drinking water	4	8.42	<0.0001
Age, gender and availability of electricity	2	7.8	0.0004

Num DF = Number difference.

The logistic distribution simplifies parameter estimation because the probability density function (pdf) is related to the cdf (Agresti, 2002).

The survey logistics model is an alternative statistical methodology (Natarajan et al., 2008) used to identify factors affecting the malaria risk. Studies conducted by Ayele et al. (2012), using survey logistic method, concluded that malaria epidemic in Amahara, Oromia and SNNP regions of Ethiopia is associated with the socio-economic, demographic and geographic factors (Ayele et al., 2012). But this model is survey based, whereas the Kebeles are chosen at random which could result in some variability between the sampling units. Such a study of the identification of the socio-economic, demographic and geographic risk factors is helpful to identify households who are in a critical need of intervention. Generalized linear mixed models (GLMM) explore the idea of statistical models that incorporate random factors into generalized linear models. GLMMs add random effects or correlations among observations to a model, where observations arise from a distribution in the exponential family. The generalized linear mixed model has many advantages. The use of GLMMs can allow random effects to be properly specified and computed, and errors can also be correlated. In addition to this, GLMMs can allow the error terms to exhibit non constant variability while also allowing investigation into more than one source of variation. This ultimately leads to greater flexibility in modelling the dependent variable.

RESULTS

Model selection was achieved by first including into the model all predictor variables and then evaluating whether or not any interaction terms needed to be incorporated. This was determined by fitting to the model, one at a time, each of the interaction terms formed from the predictor variables, and retaining in the model only those interaction terms which were significant. This process continued until the final maximal model was obtained. The final chosen model for the malaria rapid diagnosis test contained all main effects as well as six two-way interaction terms, and two three-way interaction terms. The final model is presented in Table 1.

Age, family size, altitude, main source of drinking water, time taken to collect water, availability of toilet facilities, availability of television, number of rooms per person, main construction material of the rooms' walls, roof and floors, incidence in the past 12 months of anti-mosquito spraying, number of months the room sprayed and total number of nets per person were found to be significant main effects. From these main effects, the following were involved in the interaction effects: main source of drinking water; time to collect water; availability of electricity; main construction material of the rooms' walls, roof and floor; incidence of anti-mosquito spraying; and the use of mosquito nets. There are two three-way and eight two-way significant interaction terms. The three-way interaction term is between age, gender and main source of drinking water and between age, gender and availability of electricity. The two-way interaction terms are between source of water and roof material; between number of nets per person and use of mosquito nets; between gender and availability of electricity; between gender and floor material; between time to collect water and construction material of room's floor; between gender and application of anti-mosquito spray; and between gender and number of months the room was sprayed. The interpretation of the results is presented as follows.

Tables 2 and Table 3 presents odds ratio estimates associated with age, gender, family size, region, altitude, toilet facilities, main source of drinking water, time to collect water, availability of electricity, radio and television, number of rooms per person, main construction material of room's roof, wall and floor, application of anti-mosquito spray, number of months the room sprayed, use of mosquito nets and number of nets per person. Our result reveals that malaria risk is high for young household members (OR = 0.992, P-value < 0.0002). Based on the results, for a unit increase in family size, the odds of positive RDT for individuals increases by 3.76% (OR = 1.0376, P-value < 0.0001). Furthermore, for a unit increase in altitude, the odds of positive RDT decreases by 2.2% (OR = 0.978, P - value <0.0001). With reference to individuals with no toilet facility, malaria RDT was seen to be positive for more individuals with toilet with flush (OR = 0.894, P-value = 0.0141) followed by pit latrines (OR = 0.878, P-value = 0.005). Moreover, for a unit increase in the number of total rooms, the odds of malaria diagnosis test for individuals decreased by 5.5% (OR = 0.945, P-value = 0.004).

Interaction effects

Figures 1 and 2 shows the distribution of malaria RDT against the main source of drinking water for both males and females, respectively. As age increased, positive malaria diagnosis was less likely for males than for females who were using protected, unprotected and tap water for drinking. Furthermore, as age of respondents increased, malaria RDT was less likely to be positive for individuals who used tap water for drinking (OR = 0.98, P - Value < 0.0001) for males and (OR = 1.077, P - Value < 0.0001) for females. More specifically, positive malaria diagnosis rates increased with age for females whereas it decreased for males as age increased (Figures 1 and 2). The figures further show that the gap in the RDT between respondents using unprotected, protected and tap water for drinking widens with increasing age.

The relationship between age, gender and availability of electricity is presented in Figure 3. As the figure indicates, positive malaria RDT decreases as age increases for both male and female respondents, whether or not they had access to electricity. However, the rate of decrease was not the same for males and females after controlling for other covariates in the model. The rate of increase for females who responded positively to having electricity was 9.14% higher than the other categories (OR = 1.0914, p-value < 0.001). Probabilities for this interaction are presented in Figure 3.

Interaction effects between main source of water and main construction material of the room's roof is presented in Figure 4. From the figure, it is clearly seen that with respondents who reported using tap water as well as protected and unprotected water for drinking, positive rapid diagnosis of malaria was significantly higher when the roof of the house was thatched, followed by those who occupied a stick and mud roof and finally respondents living in a house with a corrugated iron roof. The difference in RDT between the respondents' use of tap, protected and unprotected sources of drinking water and having a thatch or stick/mud roof was particularly significant. It has also shown that for a corrugated iron roof, positive RDT was significantly lower for respondents who reported using tap water for drinking than for those who were using protected and unprotected water. The other two-way interaction effect which is significant is between the time taken to collect water and main construction material of the room's floor (Table 1). This result is presented graphically in Figure 5. Positive RDT was significantly higher in a room with an earth or dung and plaster floor than in one with cement or wooden floors for respondents who took < 30 min and > 90 min to collect water. But for respondents who took less than 90 min to collect water and had a cement floor, positive rapid diagnosis is low. Furthermore, with respondents who took between 30 to 40 min to collect water, there was lower positive RDT for respondents with an earth or dung and plaster floor and a wooden floor.

The relationship between the main construction material of the room's floor and gender for a household is presented in Figure 6. As the figure indicates, positive RDT was significantly higher for males than females with respondents who reported having an earth or dung and plaster floor (OR = 4.911, P-value = 0.001) as well as for

Table 2. Estimates of odds ratio for main effects.

Effect	Estimate	OR	95% CI		P-value
			Lower	Upper	
Intercept	0.622	1.863	1.369	2.536	<0.0001
Age	-0.009	0.992	0.987	0.996	0.0002
Gender (Ref. Male)					
Female	-0.027	0.973	0.637	1.487	0.8995
Family size	0.037	1.038	1.018	8.118	<0.0001
Region (Ref. SNNP)					
Amhara	0.004	1.044	0.972	1.036	0.8271
Oromiya	0.002	1.072	0.963	1.043	0.9053
Altitude	-0.007	0.978	0.945	0.998	<0.0001
Main source of drinking water (Ref. protected water)					
Tap water	1.591	4.909	1.892	7.751	<.0001
Unprotected water	0.725	2.065	1.066	3.902	0.031
Time to collect water (Ref. less than 30 min)					
30 - 40 min	0.721	2.056	1.066	3.900	0.031
40 - 90 min	1.470	4.349	2.284	8.373	<0.0001
> 90 min	0.069	1.071	0.959	1.065	0.6932
Availability of toilet facility (Ref. No facility)					
Pit latrine	-0.130	0.878	0.694	0.940	0.005
Toilet with flush	-0.112	0.894	0.610	0.956	0.0141
Availability of electricity (ref. no)					
Yes	0.166	1.181	0.987	1.133	0.1098
Availability of radio (ref. yes)					
No	-0.022	0.978	0.980	1.009	0.4328
Availability of television (ref. yes)					
No	-0.104	0.901	0.845	0.960	0.0013
Number of rooms/person	-0.057	0.945	0.908	0.982	0.004
Main material of room's wall (Ref. cement block)					
Corrugated metal	-0.329	0.719	0.700	0.740	<0.0001
Mud block/stick/wood	-0.322	0.725	0.570	0.922	0.0086
Main material of room's roof (Ref. Corrugate)					
Thatch	0.006	1.006	0.995	1.018	0.0269
Stick and mud	0.045	1.046	1.016	1.077	0.0024
Main material of room's floor (Ref. /Local dung plaster)					
Cement-floor	-0.174	0.840	0.624	1.132	0.2532
Wood-floor	-0.136	0.872	0.657	1.158	0.3456
Use of anti-mosquito spray (ref. No)					
Yes	-0.396	0.673	0.656	0.690	<0.0001
Number of months the room sprayed	-0.053	0.949	0.945	0.953	<0.0001
Use of mosquito nets (ref. No)					
Yes	-0.009	0.991	0.999	1.019	0.0778
Number of nets/person	-0.034	0.966	0.949	0.984	0.0002

Table 3. Estimates and odd ratios for interaction effects.

Effect	Estimate	OR	95% CI		P-value
			Lower	Upper	
Main source of drinking water and main material of the room's roof (ref. Protected water and cement block)					
Tap water and mud block/stick/wood	-0.034	0.967	0.944	0.991	0.006
Tap water and corrugated metal	-0.264	0.768	0.626	0.829	0.019
Unprotected water and Mud block/stick/wood	-0.008	0.992	0.966	1.000	0.020
Unprotected water and Cement block	-0.032	0.968	0.906	1.035	0.549
Time to collect water and material of room's floor (ref. less than 30 min and earth/local dung plaster)					
Greater than 90 min and Cement	-0.039	0.962	0.857	1.079	0.5048
Greater than 90 min and Wood	-0.294	0.745	1.201	1.500	<0.0001
Between 30 - 40 min and Cement	-0.016	0.985	0.980	1.053	0.3901
Between 30 - 40 min and Wood	0.145	1.156	1.147	1.165	0.0048
Between 40 - 90 min and Cement	-0.172	0.842	1.226	1.151	<0.0002
Between 40 - 90 min and Wood	0.200	1.221	1.312	1.137	0.3901
Gender and main source of drinking water (ref. male and protected water)					
Female and tap water	0.0169	1.017	0.941	1.099	0.0488
Female and unprotected water	-0.0795	0.924	0.854	0.999	0.0467
Gender and material of room's floor (ref. male and earth/local dung plaster)					
Female and cement	-0.0175	0.983	0.619	0.998	0.0408
Female and wood	0.2741	1.315	0.859	2.014	0.0075
Gender and use of mosquito nets (ref. male and yes)					
Female and no	-0.034	0.967	0.964	0.969	<0.0001
Gender and use of anti-mosquito spray (ref. male and no)					
Female and yes	0.0018	1.002	0.985	1.030	0.0055
Number of nets per person and use of mosquito nets (ref. No)					
Yes	0.00491	1.005	1.000	1.010	0.0467
Age and gender (ref. Male)					
Age and female	0.0336	1.034	0.992	1.002	0.4011
Age, gender, main source of drinking water (ref. male and protected water)					
Female and tap water	-0.00098	0.999	0.998	1.000	0.0119
Female and unprotected water	0.00199	1.002	1.001	1.003	<0.0001
Age, gender and electricity (ref. Male and yes)					
Female and no	0.00335	1.003	0.995	1.105	0.0003

those who reported having a wooden floor in their house (OR = 2.039, P-value = 0.031). There was however, no significant difference in positive RDT between females and males who reported having a room with a cement floor. The interaction effect between gender and main source of drinking water is presented in Figure 7. The figure shows that the risk of malaria for households using unprotected water is significantly higher than for those households who reported having protected and tap water for both males and females. Moreover, for female members of the household, the risk of malaria was higher for those households who reported having unprotected

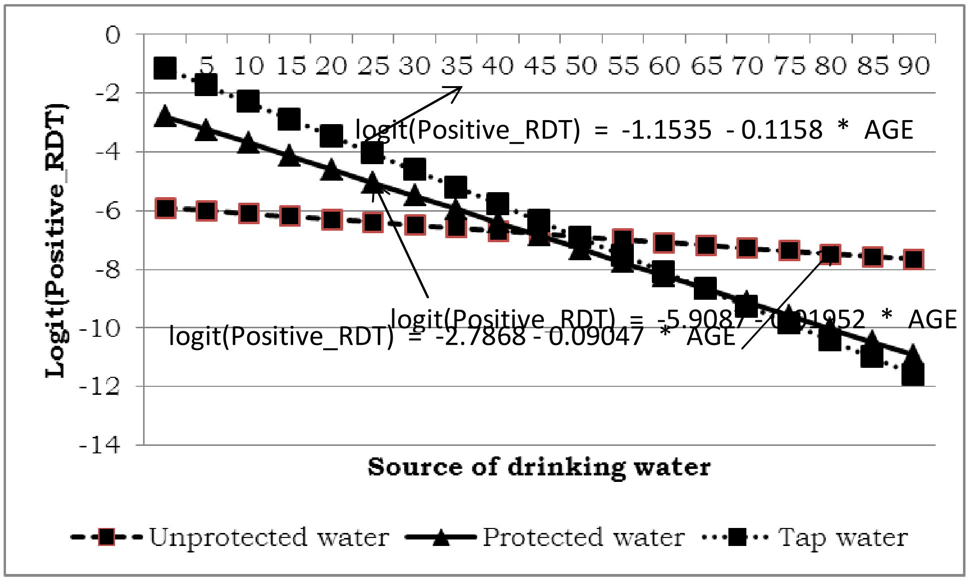


Figure 1. Log odds associated with rapid diagnosis test and age for male respondents with source of drinking water.

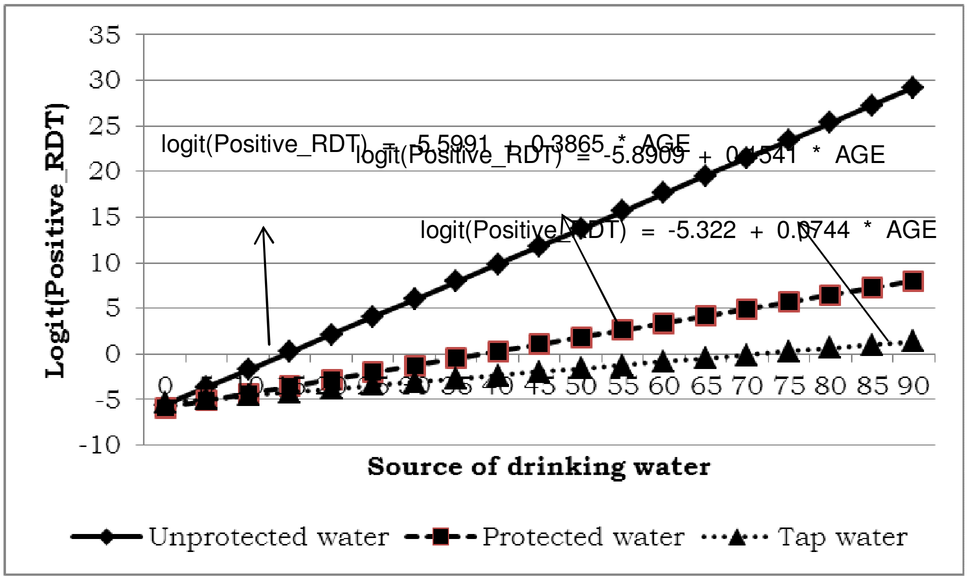


Figure 2. Log odds associated with rapid diagnosis test and age for female respondents with source of drinking water.

water.

Figure 8 presents the interaction effect between the use of anti-mosquito spray and gender for individuals. Prevalence of malaria was significantly higher for male than for female respondents who were living in a house treated with anti-mosquito spray. For males living in a house, which was not treated with anti-mosquito spray, the positive malaria result was significantly higher than it was for females. Similarly, the interaction effect between use of mosquito nets and gender is presented in Figure 9. As the figure indicates, the risk of malaria is higher for males than for females using mosquito nets when sleeping. As the number of mosquito nets increased, the risk of malaria was less likely for household members with and without nets. However, the risk of malaria was found to be much lower for individuals as the number of nets increased (Figure 10). This figure shows that for individuals with and without the use of mosquito nets, the

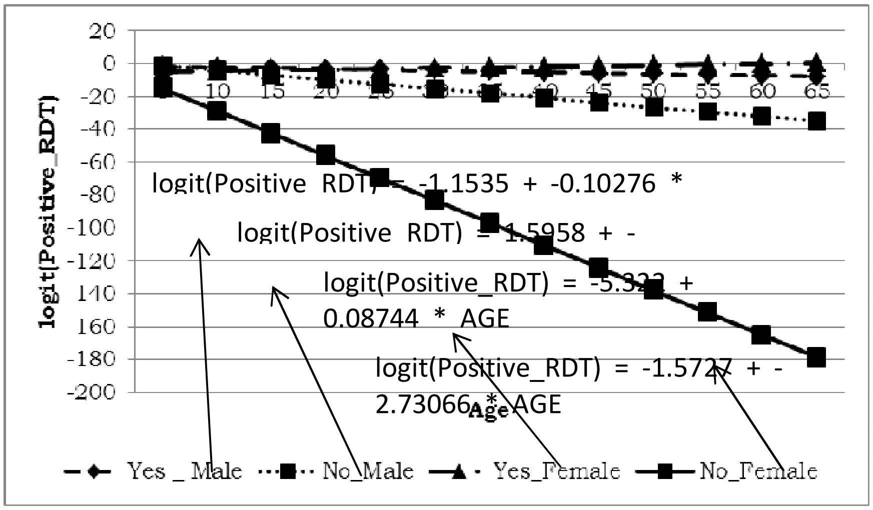


Figure 3. Log odds associated with rapid diagnosis test with age for male and female respondents with availability of electricity.

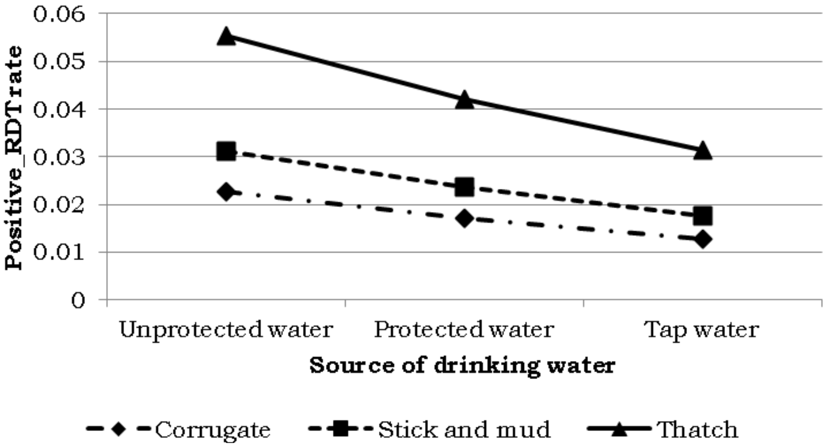


Figure 4. Log odds associated with rapid diagnosis test and source of drinking water with material of the room's roof

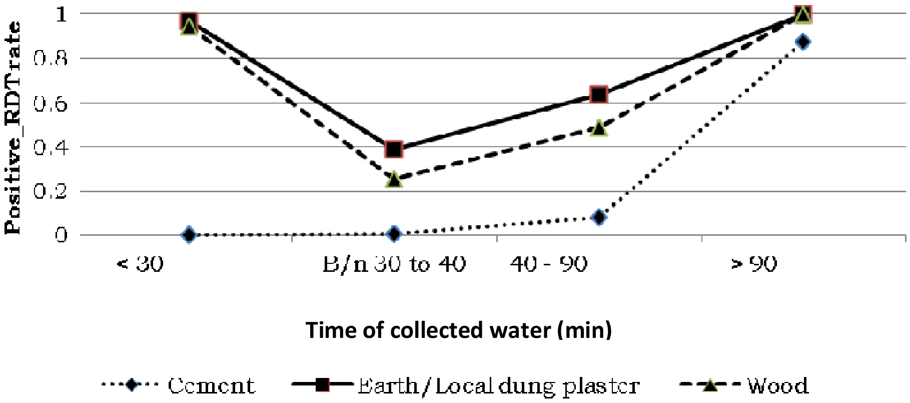


Figure 5. Log odds associated with rapid diagnosis test and time to collect water with material of the room's floor.

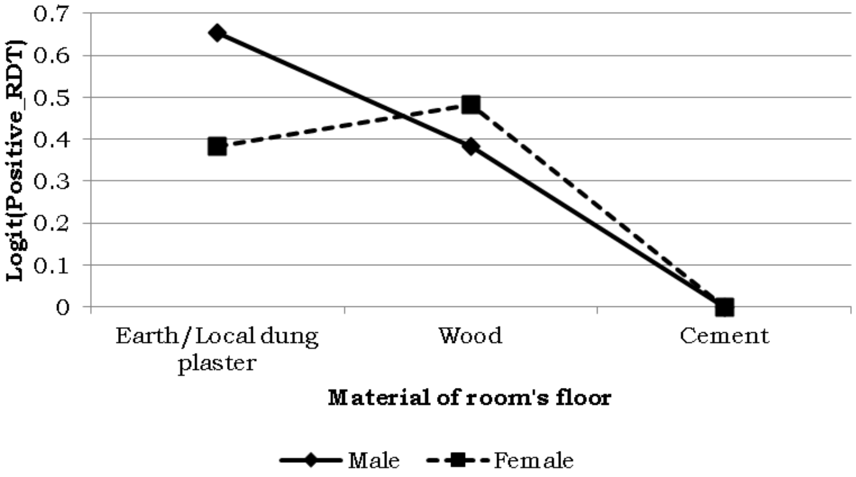


Figure 6. Log odds associated with rapid diagnosis test and material of room's floor with gender.

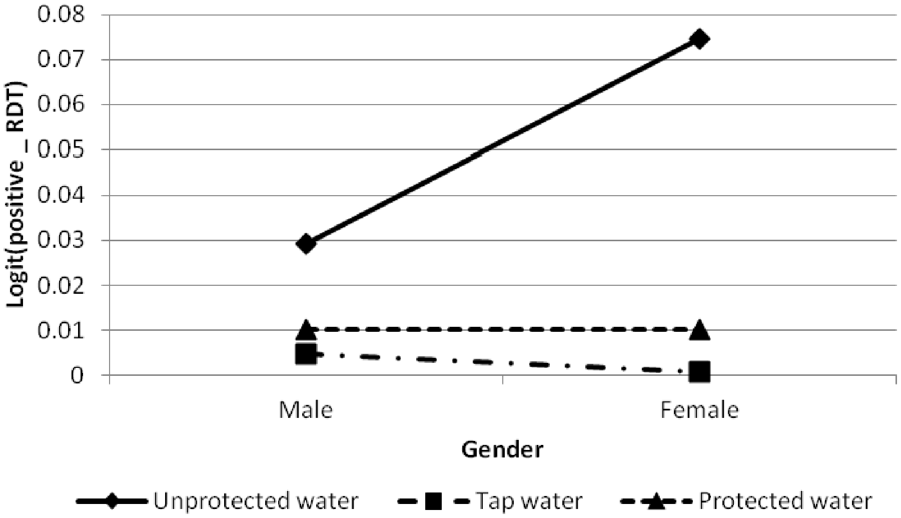


Figure 7. Log odds associated with rapid diagnosis test and main source of drinking water with gender.

risk of malaria decreased as the number of net ownerships in the household increased.

DISCUSSION

Malaria is normally referred to as a disease of poverty and related to poor socio-economic factors (Hay et al., 2004). Malaria disproportionately affects poor people who cannot afford treatment or have limited access to health care. Families and communities are then trapped in a downward spiral of poverty (Worrall et al., 2002). Since poverty is related to socio-economic factors, it is important to understand the linkages between malaria and poverty. Identifying the factors that increase the risk of malaria can be used to guide government policy-makers into creating and implementing more effective policies to tackle the disease.

SAS version 9.2 was used for the analysis of the data. Because of the nature of the methodology of the study and socio-economic, demographic and geographic variables are related. This might cause the confounding

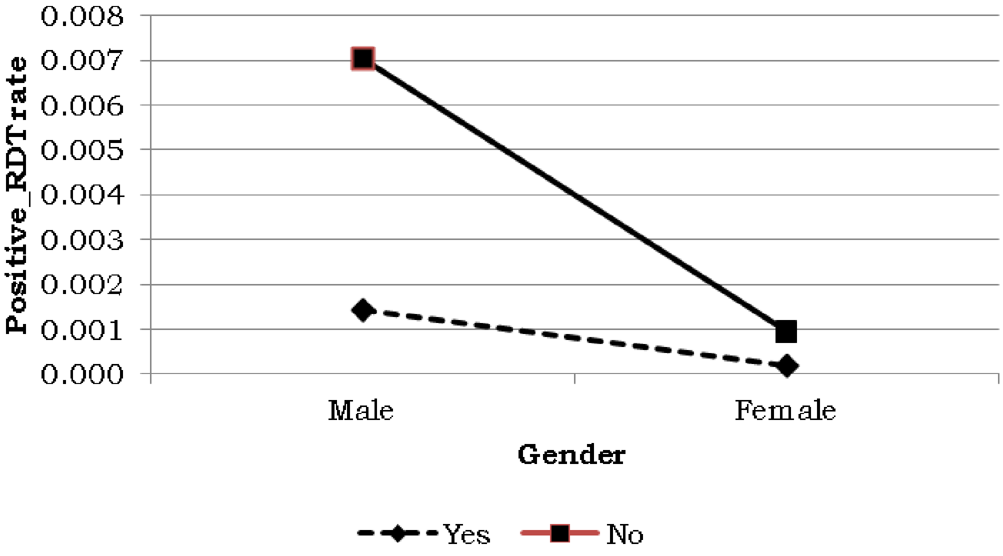


Figure 8. Log odds associated with rapid diagnosis test and anti-mosquito spraying of respondents with gender.

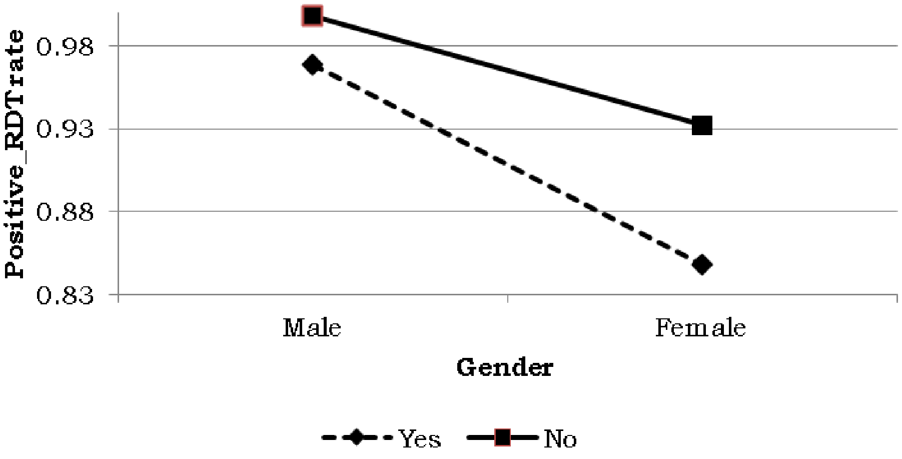


Figure 9. Log odds associated with rapid diagnosis test and use of mosquito nets with gender at individual level.

problem. Therefore, to avoid confounding effects, the model was fitted in two steps. The model was fitted to each predictor variables one at a time. In stage two, the significant predictors were retained in the model. In addition to the main effects, possible combinations of up to three-way interaction terms were added and assessed to further avoid and mitigate the problem of confounding.

Majority of studies conducted so far have suggested that malaria could be linked to poverty. The global distribution of malaria also supports this claim because malaria is concentrated to the poorest continents and countries. Therefore, our study supports the fact that malaria is related to poverty. The study indicates that socio-economic, demographic and geographic factors are responsible for the transmission of malaria. These factors are age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, availability of radio, total number of rooms, main construction material of the room's walls, main construction material of the room's floor, use of anti-mosquito spray, use of mosquito nets and total number of nets were the major factors associated with malaria RDT results. In addition to the main effects, three-way and two-way interaction effects were identified. The three-way interactions were between age, gender and main source of drinking water and age, gender and availability of electricity. The two-way interaction effects were between main source of drinking water and main construction

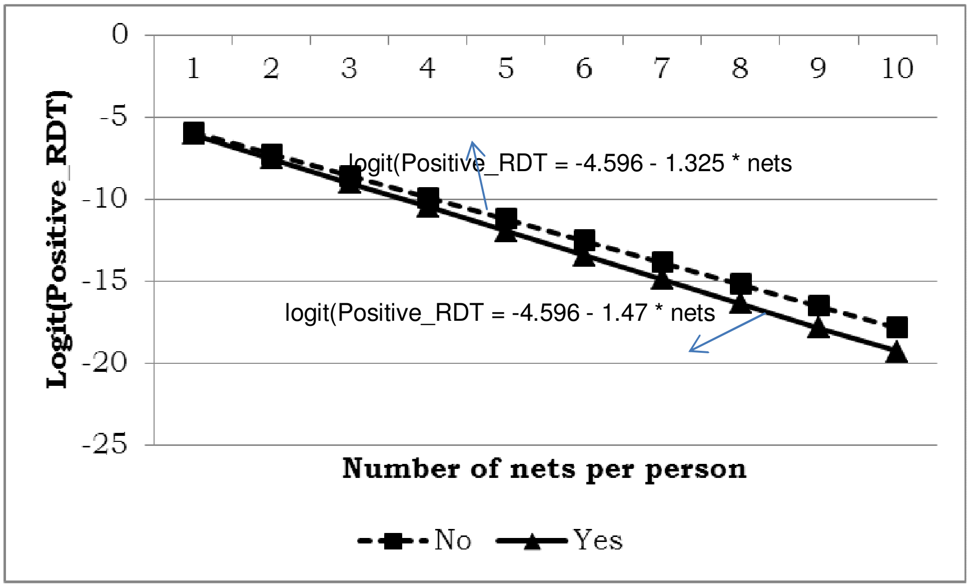


Figure 10. Log odds associated with rapid diagnosis test and use of mosquito nets with number of nets per person.

material of the room's roof, time taken to collect water and main construction material of the room's floor, age and gender, gender and main source of drinking water, gender and availability of electricity, and gender and main construction material of the room's floor.

In the present study, the effect of socio-economic factors shows that residents with no toilet facilities were found to be at more risk of malaria than those with toilet facilities. Additionally, malaria prevalence is low for households with a greater number of rooms in the house. On the other hand, having more mosquito nets over beds was found to be one way of reducing the risk of malaria. The prevalence of malaria for households with access to clean water was found to be less. Malaria rapid diagnosis was found to be higher for those respondents living in thatched houses, or ones with stick and mud roofs. Therefore, having a house with a corrugated iron roof was found to reduce the risk of malaria. Furthermore, the prevalence of malaria for households with earth and local dung and plaster floors was found to be higher. Moreover, the treatment of walls of houses with anti-mosquito spray was found to be one means of reducing the risk of malaria.

Based on demographic factors associated with malaria, our findings showed that females and children are at a greater risk. Furthermore, the malaria prevalence rate was found to be less for households with fewer people in the house. Malaria prevalence was similarly associated with geographic factors. The association between malaria and altitude showed that malaria prevalence is higher for households who are living at lower altitudes.

The result of this study supports the result from the majority of previous studies. These studies were conducted to understand the distribution of malaria. Moreover, these studies have suggested that malaria could be linked to poverty. Therefore, better understanding of the relationships between malaria and poverty is important to design effective policies (Hay et al., 2004; Mendis et al., 2009). Furthermore, the findings of this study have similar results to some of the results from previous studies (Banguero, 1984; Koram et al., 1995; Sintasath et al., 2005). In 1998 and 2000, study was conducted by (Ghebreyesus et al., 2000; Snow et al., 1998) in Ethiopia and Kenya, respectively. In this study, the assessment of different types of materials used in the construction of walls, roofs and floors of a house was done. Therefore, from the study, it was possible to observe association between any roof, wall and floor material and risk of malaria. Therefore, the finding of this study supports the result from the previous studies. Similarly, the use of mosquito nets was studied by different researchers. Therefore, the findings of these studies support the outcome of this study (Messina et al., 2011).

CONCLUSION

The government of Ethiopia has adopted various strategies to control malaria. These include early diagnosis, prompt treatment, selective vector control, epidemic prevention and control. In addition to this, the government has supporting strategies such as human resource development, monitoring and evaluation. One of the government’s key goals in the control of malaria is to achieve the complete elimination of malaria within those geographical areas with historically low malaria transmission and achieve near zero malaria transmission in the remaining malarious areas of the country. For this reason, evidence based strategies to prevent malaria is an attractive strategy for the country (Goovaerts, 1997). Therefore, the results from this study showed that malaria is associated with socio-economic, demographic and geographic factors, mainly influenced by poverty levels. Malaria is generally regarded as a disease of poverty. The more wealthy households who can afford to have toilet facilities, a greater number of rooms in the house, clean drinking water, and well built houses were found to be less affected by malaria. Furthermore, it was found that women and children are more vulnerable to malaria. Lack of bed nets contributes to this vulnerability. Moreover, as our results indicate having more bed nets is one means of reducing malaria and evidence suggests that households which are unable to afford sufficient mosquito nets, due to large families and low incomes, are more affected by malaria. Women and children are also exposed to mosquito bites while they are travelling long distances to fetch water. As the wealthier households were found to be less vulnerable to malaria than the poor households, improving the living conditions of the communities could be one way of achieving the malaria control goals set by the health professionals.

ACKNOWLEDGEMENT

We thank, with deep appreciation, Amhara RHB and The Carter Center, for providing the data.

REFERENCES

Adhanom TDW, Witten HK, Getachew A, Seboxa T (2006). Malaria. In The Eipdemiology and Ecology of Health and Disease in Ethiopia 1st edition. *Edited by: Berhane Y, Hailemariam D, Kloos H and Shama PLC. Addis Ababa, Ethiopia* 556-576.

Agresti A (2002). Categorical Data Analysis, 2nd edition/Ed. John Wiley & Sons, New Jersey.

Ayele DG, Zewotir T, Mwambi H (2012). Prevalence and risk factors of malaria in Ethiopia. *Malaria* J. 11:195 doi:10.1186/1475-2875-11-195.

Banguero H (1984). Socio-economic factors associated with malaria in Colombia. Soc. Sci. Med. 19, 1099-1104.

Cressie N (1991). Statistics For Spatial Data, Wiley, New York.

Deressa W, Ali A, Enqusellassie F (2003). Self-treatment of malaria in rural communities, Butajira, southern Ethiopia. Bull. World Health Organ. 81, 261-268.

Eisele TP, Larsen D, Steketee RW (2010). Protective efficacy of interventions for preventing malaria mortality in children in Plasmodium falciparum endemic areas. Int. J. Epidemiol. 39, i88-i101.

Emerson PM, Ngondi J, Biru E, Graves PM, Yeshewamebrat E, Gebre T, Endeshaw T, Genet A, Mosher AW, Zerihun M, Messele A, Jr FOR (2008). Integrating an NTD with One of "The Big Three": Combined Malaria and Trachoma Survey in Amhara Region of Ethiopia. *PloS* Neglected tropical diseases 2.

FMH (2004a). Federal Ministry of Health: Guideline for malaria epidemic prevention and control in Ethiopia. 2nd edition. Addis Ababa, Ethiopia, Federal democratic Republic of Ethiopia, Ministry of Health.

FMH (2004b). Federal Ministry of Health: Malaria: Diagnosis and Treatment Guidelines for Health Workers in Ethiopia. Addis Ababa, Ethiopia, Federal democratic Republic of Ethiopia, Ministry of Health.

FMH (2006). Federal Ministry of Health: National Five Year strategic plan for malaria prevention and control in Ethiopia, 2006-2010. Addis Ababa, Ethiopia, Federal democratic Republic of Ethiopia, Ministry of Health.

Ghebreyesus T, Haile M, Witten K, Getachew A, Yohannes M, Lindsay S (2000). Household risk factors for malaria among children in the Ethiopian highlands. Trans. R. Soc. Trop. Med. Hyg. 94, 17-21.

Goovaerts P (1997). Geostatistics for Natural Resources Evaluation Oxford University Press, New York.

Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW (2004). The global distribution and population at risk of malaria: past, present, and future. Lancet Infect. Dis. 4(6):327-336.

Koram K, Bennett S, Adiamah J, Greenwood B (1995). Socio-economic risk factors for malaria in a peri-urban area of the Gambia. Trans R. soc. trop. med. hyg. 89 146-150.

Mendis K, Rietveld A, Warsame M, Bosman A, Greenwood B, Wernsdorfer WH (2009). From malaria control to eradication: The WHO perspective. Trop. Med. Int. Health 14 No. 7, 1-7.

Messina J, Taylor S, Meshnick S, Linke A, Tshefu A, Atua B, Mwandagalirwa K, Emch M (2011). Population, behavioural and environmental drivers of malaria prevalence in the Democratic Republic of Congo. Malaria J. 10:161 doi:10.1186/1475-2875-10-161.

Natarajan S, Lipsitz SR, Fitzmaurice G, Moore CG, Gonin R (2008). Variance estimation in complex survey sampling for generalized linear models. Applied Statistics 57, Part 1, 75-87.

Schabenberger O, Gotway CA (2005). Statistical Metods for Spatial Data Analysis, Chapman and Hall/CRC, New York.

Shargie EB, Gebre T, Ngondi J, Graves PM, Mosher AW, Emerson PM, Ejigsemahu Y, Endeshaw T, Olana D, WeldeMeskel A, Teferra A, Tadesse Z, Tilahun A, Yohannes G, Jr FOR (2008). Malaria prevalence and mosquito net coverage in Oromia and SNNPR regions of Ethiopia. BMC Public Health 8:321.

Sintasath DM, Ghebremeskel T, Lynch M (2005). Malaria Prevalence and Associated Risk Factors in Eritrea. The Am. J. Trop. Med. Hyg. 72(6), 682-687.

Snow RW, Peshu N, Forster D (1998). Environmental and entomological risk factors for the development of clinical malaria among children on the Kenyan coast. Trans. R. soc. trop. med. hyg. 92, 381-385.

Tekola E, Teshome G, Jeremiah N, Patricia MG, Estifanos BS, Ejigsemahu Y, Ayele B, Yohannes G, Teferi T, Messele A, Zerihun M, Genet A, Aryc MW, Emerson PM, Richards FO (2008). Evaluation of light microscopy and rapid diagnostic test for the detection of malaria under operational field conditions: a household survey in Ethiopia. Malaria J. 7:118, 1475-2875.

Tulu NA (1993). Malaria. In The Ecology of Health and Disease in Ethiopia 2nd edition. Edited by: Kloos H and Zein AZ. Boulder, USA, Westview Press Inc. 341-352.

WHO (1999). World Health Organization: New Perspectives: Malaria Diagnosis, Report of a Joint WHO/USAID: Informal Consultation held on 25-27 October 1999. Geneva, Switzerland, World Health Organization; 2000:4-48.

WHO (2006a). World Health Organization: Health action in crises: Horn of Africa Health Review. [http://www.who.int/hac/crises/international/hoafrica/en/index.html]. Vol. 2011.

WHO (2006b).World Health Organization: Systems for the early detection of malaria epidemics in Africa: an analysis of current practices and future priorities, country experience. Geneva, Switzerland, World Health Organization.

Worrall E, Basu S, Hanson K (2002). The relationship between socio - economic status and malaria: a review of the literature. *Background paper for Ensuring that malaria control interventions reach the poor* London 5th - 6th September

Zhou G, Minakawa N, Githeko A, Yan G (2004). Association between climate variability and malaria epidemics in the East African highlands. Proceedings National Acad. Sci. 101, 2375-2380.

Ayele *et al. Malaria Journal* 2013, **12**:207
http://www.malariajournal.com/content/12/1/207

MALARIA JOURNAL

RESEARCH Open Access

Spatial distribution of malaria problem in three regions of Ethiopia

Dawit G Ayele*, Temesgen T Zewotir and Henry G Mwambi

Abstract

Background: The transmission of malaria is the leading public health problem in Ethiopia. From the total area of Ethiopia, more than 75% is malarious. The aim of this study was to identify socio-economic, geographic and demographic risk factors of malaria based on the rapid diagnosis test (RDT) survey results and produce the prevalence map of the area illustrating variation in malaria risk.

Methods: This study accounts for spatial correlation in assessing the effects of socio- economic, demographic and geographic factors on the prevalence of malaria in Ethiopia. A total of 224 clusters of about 25 households each were selected from the Amhara, Oromiya and Southern Nation Nationalities and People's (SNNP) regions of Ethiopia. A generalized linear mixed model with spatial covariance structure was used to analyse the data where the response variable was the presence or absence of malaria using the RDT.

Results: The results showed that households in the SNNP region were found to be at more risk than Amhara and Oromiya regions. Moreover, households which have toilet facilities clean drinking water, and a greater number of rooms and mosquito nets in the rooms, have less chance of having household members testing positive for RDT. Moreover, from this study, it can be suggested that incorporating spatial variability is necessary for understanding and devising the most appropriate strategies to reduce the risk of malaria.

Keywords: Mixed model, Rapid diagnostic test, Spatial statistics, Variogram, Kriging

Background

Malaria is a life-threatening disease affecting the world's most under-developed countries and regions where basic healthcare infrastructure is lacking [1] as well some developed countries. Malaria is a major cause of morbidity and mortality in Africa, especially in sub-Saharan African countries [1]. It is a leading cause of death amongst children in many African countries [2]. With 68% of the total population of Ethiopia living in areas at risk of malaria [3], it is a major public health problem and for many years the prime cause of illness and death [3,4]. From the total population of Ethiopia (77,127,000 in 2007), more than 50 million people are at risk from malaria [5]. In general, 4–5 million people are affected by malaria annually [6,7].

Epidemics of malaria are relatively frequent [8,9] involving highland or highland fringe areas of Ethiopia, mainly areas 1,000-2,000 meters above sea level [10-12]. Notably this altitude covers 48% of the regions of Amhara, Oromiya and Southern Nations Nationalities and People's regions of Ethiopia. Malaria epidemics have serious consequences for Ethiopia's subsistence economy as the malaria transmission peaks during the major harvesting seasons. To control the risk of malaria, early diagnosis and prompt treatment is one of the key strategies. To diagnose malaria, clinical diagnosis is the most widely used. But, laboratory facilities are not available in all areas of the country [13,14]. The standard method to diagnose malaria is microscopy. However, this form of diagnosis is not accessible or affordable in most peripheral health facilities. The recent introduction of rapid diagnostic tests (RDT) for malaria is a significant step forward in case detection, timely treatment and management, and reduction of unnecessary treatment. RDT could be used in malaria diagnosis during population-based surveys and to provide immediate treatment based on the results.

RDTs offer the potential to extend accurate malaria diagnosis to areas where microscopy services are not available such as in remote locations or after regular laboratory

* Correspondence: ejigmul@yahoo.com
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa

hours. Rapid malaria diagnostic tests have been developed in the lateral flow format [15]. These tests use finger-stick or venous blood, which takes only 10 to 15 minutes to complete, and do not require a laboratory. Non-clinical staff can easily learn to perform the test and interpret the results [16].

It is essential to identify the socio-economic and demographic risk factors associated with the prevalence of malaria using data obtained from the rapid diagnosis test. Such a study of the identification of the socio-economic and demographic risk factors is helpful in identifying households who have a critical need for intervention. In previous studies, Ayele, Zewotir and Mwambi (2012) have concluded that malaria problem in Amahara, Oromiya and SNNP regions of Ethiopia are associated with key socio-economic, demographic and geographic factors, in particular it was noted that poverty levels of households are highly associated with the risk of malaria. Nevertheless the spatial distribution of malaria was not considered or investigated [17]. Though identification of the household characteristics is essential for grass root level intervention, the government goals and targets are focused on achieving malaria eradication/reduction within specific geographical areas. Such studies are limited, and hence the conception of this study. Therefore, the objective of this study is to undertake a statistical analysis of malaria incidence. This will identify important socio-economic, demographic and geographic variables associated with the disease and ultimately a prevalence map of the area illustrating variations in malaria risk.

Methods

Study design

From December 2006 and January 2007, a baseline household cluster malaria survey was conducted by The Carter Center (TCC). The questionnaire was developed as a modification of the Malaria Indicator Survey (MIS) Household Questionnaire. The questionnaire had two parts; the household interview and malaria parasite form. For this survey, the sampling frame in each of the rural populations of Amhara, Oromiya and SNNP regions was a *Kebele* (the smallest administrative unit in Ethiopia). The study area with the selected households is presented in Figure 1. From the three regions, 5,708 households located in 224 clusters were included in the survey. Out of these households, Amhara, Oromiya and SNNP regions covered 4,101 (71.85%), 809 (14.17%) and 798 (13.98%) households respectively. Prior to conducting the survey, 224 *Kebeles* were selected. From each *Kebele*, 12 households (even numbered households) were selected for malaria tests. In the survey each room in the house was listed separately. In addition to the number of rooms and number of nets, the persons sleeping under each net were listed. The detailed sampling procedure is presented in [17-19].

Before testing for malarial parasites, consent was obtained from the participants. To collect the sample, finger-prick blood was collected from the participants for the malaria rapid diagnostic test. The test used is known as *ParaScreen* which is capable of detecting both *Plasmodium falciparum* and other *Plasmodium* species. Participants with positive rapid tests were immediately offered treatment according to national guidelines.

Variables of interest

Response variable

The outcome of interest is the malaria rapid diagnosis test (RDT) result. RDTs assist in the diagnosis of malaria by detecting evidence of malaria parasites in human blood and are an alternative to diagnosis based on clinical grounds or microscopy, particularly where good quality microscopy services cannot be readily provided. Thus, the response variable is binary, indicating whether or not a person is positive for malaria using the RDT.

Independent variables

The independent variables or covariates were the baseline socio-economic status, demographic and geographic variables including gender, age, family size, region, altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio and television, total number of rooms, main material of the room’s walls, main material of the room’s roof, main material of the room’s floors, incidence of anti-mosquito spraying in the past 12 months, use of mosquito nets and total number of nets. Malaria test (RDT result), age and sex were collected at individual level. Altitude, main source of drinking water, time taken to collect water, toilet facilities, availability of electricity, radio, television, total number of rooms, main material of the room’s walls, main material of the room’s roof, main material of the room’s floor, use of anti- mosquito spray in the past 12 months, use of mosquito nets and total number of nets were all collected at household level.

The statistical model

The distribution of malaria is nonrandom across a landscape in areas of higher or lower transmission intensity and malaria risk. The transmissions are separated by greater or lesser distances from each other. Based on geographical aggregation, there are two distinct levels. These are, the focal unit of malaria transmission, the area over which human malaria is actively transmitted originating from a specific aquatic breeding site and the household or other reasonably identified point of contact between a small group of humans and mosquito vectors. The baseline household cluster malaria survey which was conducted by The Carter Center from December 2006 to January 2007 includes the geographical locations

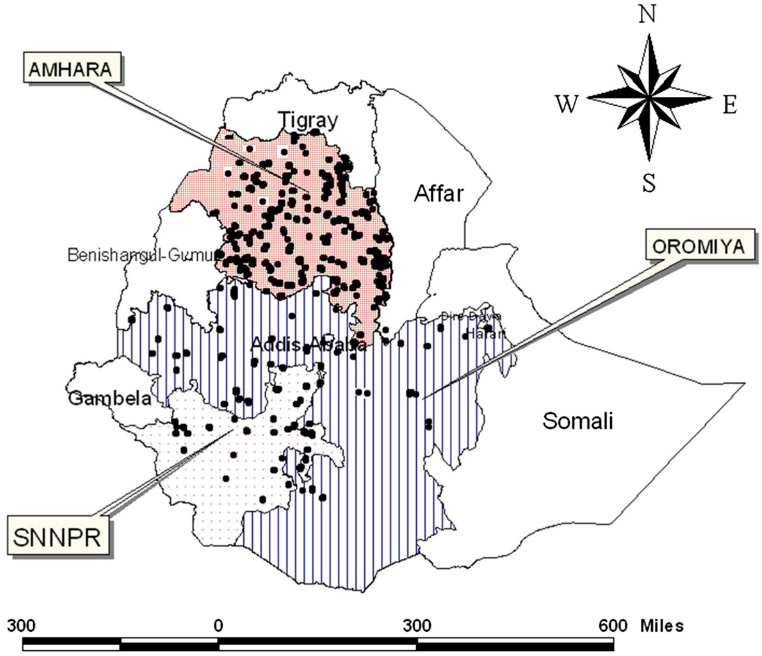


Figure 1 Map of Ethiopia showing the surveyed households.

of the reference for each household. Therefore, it is of interest to know whether the data display any spatial autocorrelation. Furthermore, it is important to check whether surveys that are near in space have malaria prevalence or incidence that is similar with the surveys that are far apart. This is important because spatially correlated data cannot be regarded as independent observations. If the analysis does not take account of the correlation structure of the data, the estimates obtained from the analysis may be inaccurate because of the underestimated standard errors. Therefore, the objective of this study is to undertake statistical analysis of malaria incidence to identify important socio-economic, demographic and geographic variables associated with the disease and to produce prevalence maps of the area illustrating the variation in malaria risk using spatial statistics analysis. Spatial statistics can be divided into three methods. These are: point pattern analysis, methods for lattice data and geostatistics [20,21]. *Point referenced data* is often called geocoded or geostatistical data. *Areal data* is often called lattice data. Some spatial data sets feature both point and areal-level data. *Point pattern data*: The response is often fixed (occurrence of the event), and only the locations where it occurs are thought of at random. Of these, the geostatistical approach is most relevant to epidemiological analysis which is conducted at the landscape scale and based on remote sensing [22-24].

A common approach to integrate spatially correlated data with the random effects and proceed with maximum likelihood based approaches for estimating the covariate and covariogram parameters, is based on the theory of generalized linear mixed models (GLMM). Using GLMM, numerical approximation can be implemented [20,25].

Non-Gaussian spatial problems may be formally analysed in the context of generalized linear mixed models (GLMM). Specification of the likelihood of the random variable $y(s)$ is required where s generally denotes the location the observation is made. As in classical generalized linear models (GLMs), there is a canonical parameter corresponding to the distribution, which is nominally a function of the location parameter via the link function $g(.)$ for the distribution. This function is assumed to be linear in the explanatory variables. In the classical formulation of GLMs containing only fixed effects, $\mathrm{g}(\mu) = \boldsymbol{X\beta}$, where $\mathbf{X}$ is the matrix of explanatory variables [26-30]. To incorporate a spatial process, we assume $\mathrm{y}(\mathrm{s}_i|\alpha)$ is conditionally independent for any location s_i with conditional mean $E[y(s_i)|\alpha] = \mu(s_i)$. The parameter α is used to define the distribution of s. Then, the spatially correlated random effect is incorporated into the linear predictor as:

$$g(\mu) = \boldsymbol{X}\beta + \boldsymbol{Z}\alpha \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrix. The error term accommodates over-dispersion relative to the mean-variance relationship implied by the distribution under consideration. The random effect at location $\mathrm{s_i}$, $\alpha \sim Gau(0, \Sigma_\alpha(\theta))$ and $\varepsilon \sim Gau\left(0, \sigma_\varepsilon^2 I\right)$, with spatial correlation is parameterized by θ in $\Sigma_\alpha(\theta)$ [20]. Note that s_i is just one location. $s = (s_1, ,s_k)$' denotes a vector of k locations with variance-covariance matrix .

Spatial dependence may be represented by a range of functions [31]. To describe spatial correlation of observations, there are three major functions used in geostatistics. These major functions are the correlogram, the covariance, and the semivariogram. Semivariogram is also more simply called the variogram. In geostatistics, the variogram is the key function and is used to fit a model for the spatial correlation in the data. The model which is obtained using the variogram is used in kriging estimation procedures, a method which was first used in minimizing [23]. Moreover, variogram models are also used to understand maximum distances of spatial autocorrelation which can further be used in construction of search parameters for different interpolation techniques. A variogram represents both structural and random aspects of the data under consideration. The structural part of the variogram model is represented by the range of a variogram. Furthermore, the variogram values increase with increases in the distance of separation until it reaches the maximum at a distance known as the "range". To develop the variogram, assume μ(s) is a constant, that is constant mean μ(s), and define

$$var\{Z(s_1)–Z(s_2)\} = 2y(s_1–s_2) \tag{2}$$

In statement (2), the variance of s_1 and s_2 is through their difference s_1-s_2, and the process which satisfies this property is called intrinsically stationary. The function $2y(.)$ is called the variogram and $y(.)$ the semivariogram.

The other concept here is isotropy. Suppose the process is intrinsically stationary with semivariogram $[y(h), h \in \mathbf{R}^d$. If $y(h) = Y_0(\|h\|)$ for some function Y_0, i.e. if the semivariogram depends on its vector argument h only through its length $\|h\|$, then the process is isotropic. Therefore, a process which is both intrinsically stationary and isotropic is also called homogeneous. Isotropic processes are convenient to deal with because there are a number of widely used parametric forms for y_0 (h). Using semivariance $y_0(t)$ for interval distance class t, lag distance interval t, c_0 (nugget variance) ≥ 0,c_1 (structural variance) ≥ c_0 and R is the range parameter R, some of the examples are:

1. Spherical:

$$y_0(t) = \begin{array}{cc} 0 & ift = 0, \\ \{c_0 + c_1 t\left\{\frac{3}{2}\frac{t}{R} – \frac{1}{2}\left(\frac{t}{R}\right)^3\right\} & ift < t \leq R, \\ c_0 + c_1 t & ift \geq R. \end{array}$$

   It is a convenient form because it increases from a positive value c_0 when t is small, levelling at the constant $c_0 + c_1$ at $t = R$. This is the so-called "nugget/range/ sill" form which is often considered a realistic and interpretable form for a semivariogram.

2. Exponential:

$$\gamma_0(t) = \begin{cases} 0 ift = 0, \\ c_0 + c_1\left(1–e^{-t/R}\right) ift > 1. \end{cases}$$

   This is simpler in functional form than the spherical case (and valid for all d) but without the finite range of the spherical form. The parameter ***R*** has a similar interpretation to the spherical model however, of fixing the scale of variability.

3. Gaussian:

$$\gamma_0(t) = \begin{cases} 0 ift = 0, \\ c_0 + c_1\left(1–e^{-t^2/R^2}\right) ift > 1. \end{cases}$$

4. Exponential-power form:

$$\gamma_0(t) = \begin{cases} 0 ift = 0, \\ c_0 + c_1\left(1–e^{-|t/R|^p}\right) ift > 1. \end{cases}$$

   Here $0 < p \leq 2$. This form generalizes both the exponential and Gaussian forms, and forms the basis for the families of spatial covariance functions introduced by Sacks *et al.* in 1989 [32]. However, in generalizing the results from one dimension to higher dimensions, these authors used a product form of covariance function in preference to constructions based on isotropic processes [33].

Spatial prediction

Modelling spatial data is not only useful for identifying significant covariates but for producing smooth maps of the outcome by predicting it at unsampled locations. Spatial prediction is usually referred to as kriging. Kriging is an optimal interpolation based on regression against observed values of surrounding data points, weighted according to spatial covariance values. Interpolation refers to an estimation of a variable at an unmeasured location from observed values at surrounding locations [34]. Kriging has some advantages. These advantages are that it

- helps to compensate for the effects of data clustering, assigning individual points within a cluster less weight than isolated data points,
- gives an estimate of estimation error (kriging variance), along with an estimate of the variable,
- ensures availability of estimation error which provides a basis for stochasticity,
- allows simulation of possible realization.

The spatial prediction which is called kriging can statistically be defined as follows.

Let $\boldsymbol{Y}_0$ be a vector of the binary response at a new, unobserved location s_{0i}, $i = l, ,n_0$. Following the maximum likelihood approach, the distribution of Y_0 is given by

$$P\left(Y_0|\hat{\beta},\hat{U},\hat{\sigma}^2,\hat{\varnothing}\right) = \int P\left(Y_0|\hat{\beta},U_0\right)P\left(U_0|\hat{U},\hat{\sigma}^2,\hat{\varnothing}\right)dU_0 \tag{3}$$

Where $\hat{\beta},\hat{\sigma}^2$ and $\hat{\phi}$ are the maximum likelihood estimates of the corresponding parameters. As part of the iterative estimation process, for penalized quasi-likelihood (PQL), $\hat{U}$ can be derived [35]. $P\left(Y_0\middle|\hat{\beta},U_0\right)$ is the Bernoulli-likelihood at new locations and $P\left(U_0\middle|\hat{U},\hat{\sigma}^2,\hat{\phi}\right)$ is the distribution of the spatial random effects U_0 at new sites, given $\hat{U}$ at observed sites and is assumed to follow the normal distribution that is

$$P\left(Y_0\middle|\hat{U},\hat{\sigma}^2,\hat{\phi}\right) = N\left(\sum_{01}\sum_{11}^{1}\hat{U},\sum_{00}-\sum_{01}\sum_{11}^{1}\sum_{10}\right) \tag{4}$$

With $\Sigma_{11} = E(UU')$, $\Sigma_{00} = E(U_oU'_0)$ and $\sum_{01} = \sum_{01}^{t} = E\left(U_oU_0'\right)$. The mean of the Gaussian distribution in (4) is the classical kriging estimator [20].

The Bayesian predictive distribution of $\boldsymbol{Y}_0$ is given by

$$P(Y_0|Y) = \int P(Y_0|\beta,U_0)P(U_0,|U,\sigma^2,\phi)xP(\beta,U,\sigma^2,\phi|Y) \\ \times d\beta dU_0 dU d\sigma^2 d\phi \tag{5}$$

Where $P(\beta, U, \sigma^2, \phi|Y)$ is the posterior distribution of the parameters obtained by the Gibbs sampler or the sampling importance re-sampling (SIR) approach. Simulation-based Bayesian spatial prediction is performed by consecutive draws of samples from the posterior distribution, the distribution of the spatial random effects at new locations and the Bernoulli-distributed predicted outcome. The maximum likelihood predictor (3) can be viewed or interpreted as the Bayesian predictor (5), with parameters fixed at their maximum-likelihood estimates. In contrast to Bayesian kriging, classical kriging does not account for uncertainty in estimation of β and the covariance parameters.

The data was analysed by fitting a generalized linear mixed model (GLMM) using SAS 9.2 PROC GLIMMIX.

Analysis and results

Using the identified thirteen main effects and six two-way and three-way interaction effects [17] several covariance structures including SP(EXP) (Exponential), SP(EXPA) (Anisotropic Exponential), SP(EXPGA)((2D Exponential, Geometrically Anisotropic), SP(GAU) (Gaussian), SP (GAUGA)((2D Gaussian), (Geometrically Anisotropic), SP(LIN) (Linear), SP(LINL) (Linear Log), SP(MATERN) (Matérn), SP(MATHSW)(Matérn (Handcock-Stein-Wallis)), SP(POW) (Power), SP(POWA) (Anisotropic Power), SP(SPH) (Spherical) and SP(SPHGA)((2D Spherical, Geometrically Anisotropic) were fitted but SP(GAU) (Gaussian) was found to be the best spatial covariance structure for the model [36].

The result presented in Figure 2 is a spatial scatter plot of the observed data. The scatter plot suggests distribution which is not indicative of a uniformly spread of the RDT measurements throughout the prediction area. No direct inference can be made about the existence of a surface trend in the data. However, the apparent stratification of RDT values might indicate a nonrandom trend. The Spatial Autocorrelation is an inferential statistic tool, which is important to test for randomness. This means that the results of the analysis are always interpreted within the context of its null hypothesis of a random occurrence of events. For the randomness test Moran's and Geary's C tests can be used [37-41]. Furthermore, the distribution of observed malaria infected households and distribution of observed malaria rapid diagnosis test is presented in Figures 3 and 4.

For these tests, the null hypothesis states that the spatial distribution of feature values is the result of random spatial processes. The result from Moran's (Z value = −40.4 and p – value < .0001) and Geary's c (Z value = −11.2 and P-value < .0001) tests indicate that the spatial distribution of feature values is not the result of random spatial processes. The Z values are negative for both Moran's and Geary's C tests. This indicates that the spatial distribution of high values and low values in the dataset is more spatially dispersed than would be expected if underlying spatial processes were random. A dispersed spatial pattern often reflects some type of competitive process, i.e., a feature with a high value repels other features with high values; similarly, a feature with a low value repels other features with low values. The observed spatial pattern of feature values could not very well be one of many possible versions of complete spatial randomness.

Figure 5 represents different semivariogram estimators using classical and robust estimators. The classical estimator was suggested by Matheron in 1963 [42]. The classical estimator can be calculated by

$$\hat{Y}(\mathrm{h}) = \frac{1}{|\mathrm{N(h)}|}\sum_{\mathrm{N(h)}}\left(\mathrm{Z(s_i)}-\mathrm{Z(s_j)}\right)^2,$$

where $(\mathrm{s_i})$ is the anscombe residual,

$$\mathrm{N(h)} = \left\{\left(\mathrm{s_i}-\mathrm{s_j}\right) : \left\|\mathrm{s_i}-\mathrm{s_j}\right\| = \mathrm{h} \pm \in\right\} \text{ and } |\mathrm{N(h)}|$$

is its cardinality. But, the classical estimator is sensitive to outliers. For this reason a robust estimator was proposed by Cressie and Hawkins in 1980 [43]. Among the different types of isotropic covariograms given above, Gaussian type was selected. Thus as discussed earlier, the best spatial covariance structure from all possible types was

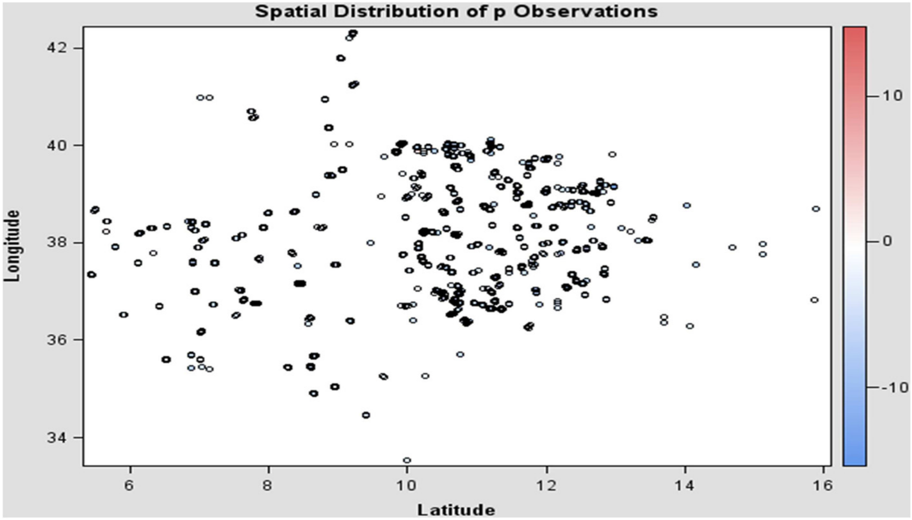


Figure 2 Scatter plot for the malaria prevalence.

found to be the SP(GAU) (Gaussian) covariance structure. Therefore, the Gaussian type of the variogram was used to perform variogram analysis. The figure (Figure 3) shows first a slow, then a rapid rise from the origin. Therefore, the shape of the graph suggests a Gaussian type form which is given by

$$y(t) = c_0 + c_1\left[1 - \exp\left(-\frac{t^2}{R^2}\right)\right].$$

In general, from Figure 3, it is possible to distinguish three main features. The first one is the Y-axis well above zero, indicating the possible presence of a nugget effect. Moreover, the shapes of the semivariogram up through distances in the low 40s have roughly the shape of a spherical covariance model. Besides these, the semivariogram values are extremely high for the largest distances.

Tables 1 and 2 presents the significant effects for the model which incorporate spatial variability using SP (GAU) (Gaussian) covariant structure. Among all significant effects namely family size, altitude, toilet facilities, availability of radio and television, number of rooms per person, main material of the room's wall, spraying of anti- mosquito, use of mosquito nets and number of nets per person, were not

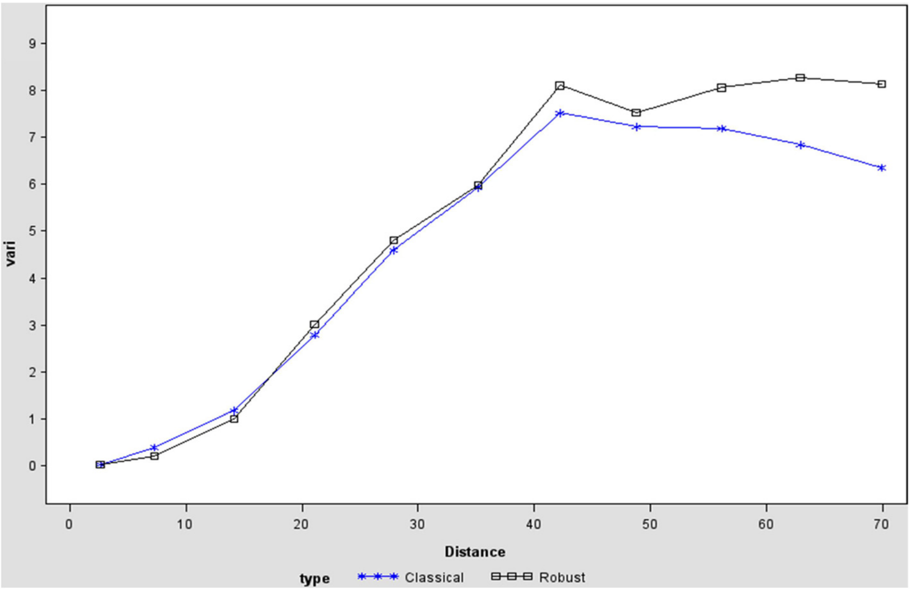


Figure 3 Distribution of observed malaria infected households.

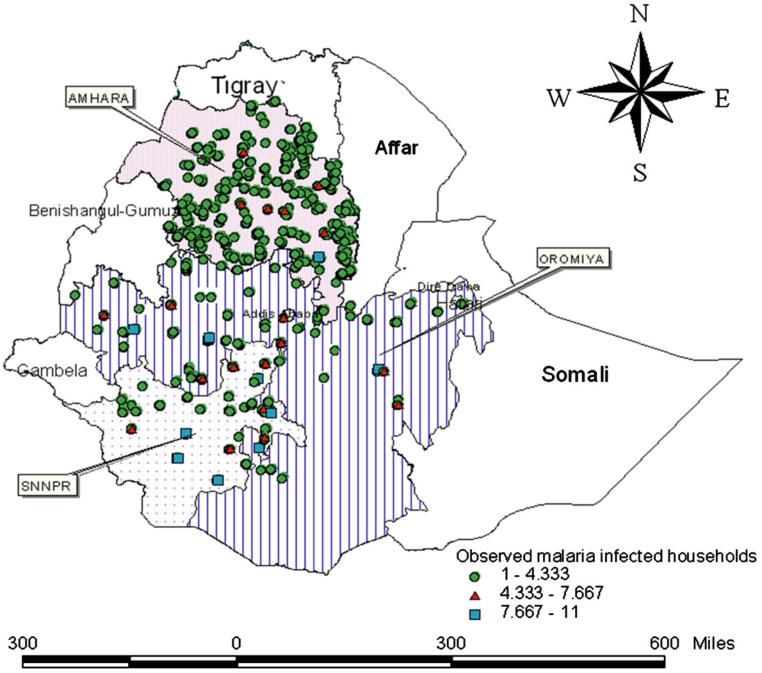


Figure 4 Distribution of observed malaria rapid diagnosis test.

involved in the interaction effects. The significant two-way and three-way interaction effects were found to be main source of drinking water and main material of the room's roof; time to collect water and main material of the room's floor; gender and main source of water; gender and main Material of the room's floor; age, gender and main source of drinking water; and age, gender and Availability of electricity. Based on these results for a unit increase in family size, the odds of positive rapid diagnosis test increases by 2.34% (OR = 1.0234, P-value < 0.0001). Furthermore, for a unit increase in altitude, the odds of positive rapid diagnosis test decreases by 1.4% (OR = 0.996, P - value <0.0001). With reference to individuals with no toilet facilities, the odds of a positive malaria rapid diagnosis test is lower for

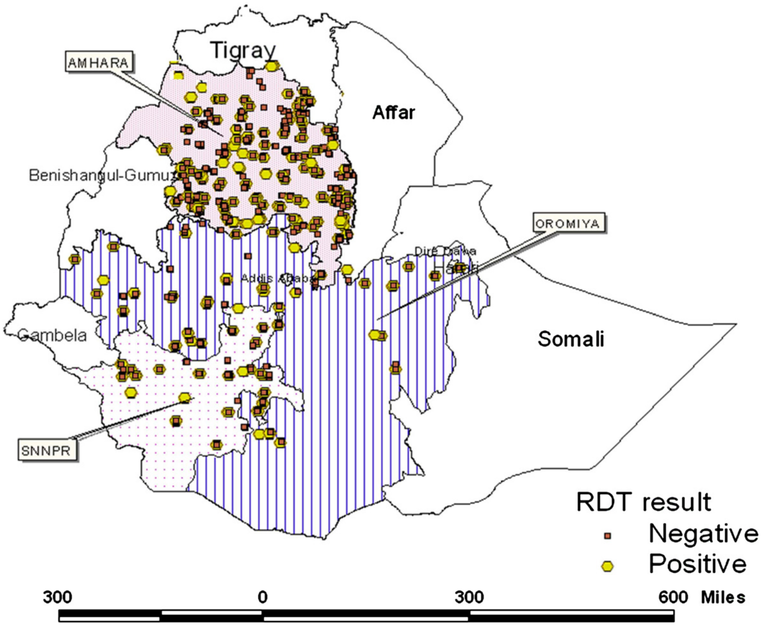


Figure 5 Classical and robust semivariogram for malaria prevalence.

Table 1 Socio-economic, demographic and geographic of effects on malaria RDT test for main effects

Parameters	Estimate	OR	SE	P -value
Intercept	−0.2460	0.7819	5.8100	0.9995
Age	0.0209	1.0212	0.0503	0.6772
Gender (ref. male)				
Female	−2.5463	0.0784	3.0804	0.4084
Family size	0.02311	1.0234	0.0527	<.0001
Region (ref. SNNP)				
Amhara	−0.6896	0.5018	0.4502	0.1256
Oromiya	−0.837	0.4330	0.5796	0.1487
Altitude	−0.0037	0.9963	0.0001	<.0001
Main source of drinking water (ref. protected water)				
Tap water	−0.5557	0.5737	0.722	<.0001
Unprotected water	0.6372	1.8912	0.6871	0.005
Time to collect water (ref. > 90 minutes)				
< 30 minutes	−0.7829	0.4571	0.252	0.0019
between 30 to 40 minutes	−0.603	0.5472	1.2666	0.6341
between 40–90 minutes	−4.0189	0.0180	2.8957	0.1652
Toilet facility (Ref. No facility)				
Pit latrine	−0.4403	0.6438	0.6433	<.0001
Toilet with flush	−0.9177	0.3994	0.6413	<.0001
Availability of electricity (ref. no)				
Yes	−3.1219	0.0441	1.0961	0.0044
Availability of television (ref. no)				
Yes	0.6991	2.0119	0.2121	0.001
Availability of radio (ref. no)				
Yes	−0.6991	0.4970	0.2121	0.001
Number of rooms/person	−0.4631	0.6293	0.0688	<.0001
Main material of room's wall (ref. cement block)				
Mud block/wood	−4.1691	0.0155	1.2646	0.038
Corrugated metal	−3.1196	0.0442	1.2576	0.004
Main material of room's roof (ref. corrugate)				
Thatch	1.5031	4.4956	1.6732	0.005
Stick and mud	0.454	1.5746	0.6726	0.0058
Main material of room's floor (ref. earth/Local dung plaster)				
Wood	−1.1407	0.3196	0.803	0.004
Cement	−0.9273	0.3956	0.114	0.028
Anti- mosquito spraying				
No	1.237	3.4453	0.1734	<.0001
Use of mosquito nets (ref. no)				
Yes	−0.8741	0.4172	0.1541	<.0001
Number of months room sprayed	−0.7626	0.4665	0.1274	<.0001
Number of nets/person	−0.9349	0.3926	0.0977	<.0001

Table 2 Socio-economic, demographic and geographic of effects on malaria RDT test for interaction effects

Parameters	Estimate	OR	SE	P-value
Gender and main source of drinking water (ref. Male & protected water)				
Female and tap water	−2.747	0.064	0.861	0.001
Female and unprotected water	1.224	3.402	1.064	0.250
Gender and material of room's floor (ref. Male and earth/Local dung plaster)				
Female and cement	−0.839	0.432	0.571	<.0001
Female and wood	0.762	2.143	0.387	<.0001
Age, gender and main source of drinking water (ref. Male & protected water)				
Female and tap water	−0.045	0.956	0.000	<.0001
Female and unprotected water	0.042	1.043	0.000	<.0001
Age, gender and availability of electricity (ref. Male & yes)				
Female and no	0.066	1.068	0.000	<.0001

those individuals using a flushing toilet to those who have septic tanks (OR = 0.399, P - value <0.0001) or pit latrine slabs (OR = 0.644, P - value <0.0001). Moreover, for a unit increase in the number of total rooms, the odds of malaria diagnosis test for an individual decreased by 37.07% (OR = 0.629, P - value <0.0001). Similarly, with a unit increase in the number of nets in the house, the odds of rapid diagnosis test of malaria for individuals decreased by 60.7% (OR = 0.392, P - value <0.0001). Furthermore, for a unit increase in the number of rooms in the household sprayed with anti- mosquito, the odds of a positive malaria diagnosis test decreased by 53.3% (OR = 0.467, P - value <0.0001).

Interaction effects

Figures 6 and 7 show the distribution of malaria rapid diagnosis test against age, main source of drinking water for both males and females respectively. As age increased, positive malaria diagnosis was less likely for males than females who were using protected, unprotected and tap water for drinking. Furthermore, as age of respondents increased, malaria rapid diagnosis test was less likely to be positive for individuals who use tap water for drinking for males and for females. More specifically, positive malaria diagnosis rate increases with age for females whereas it decreases as age increases for males (Figures 6 and 7). The figures further show that the gap in the rapid diagnosis test between respondents with unprotected, protected and tap water widens with increasing age.

The relationship between age, gender and availability of electricity is presented in Figure 8. As the figure indicates, positive malaria rapid diagnosis test decreases as age increases for both male and female respondents, whether or not they have access to electricity, except for females who responded to having electricity. However, the rate of decrease was not the same for males and females after controlling for other covariates in the model.

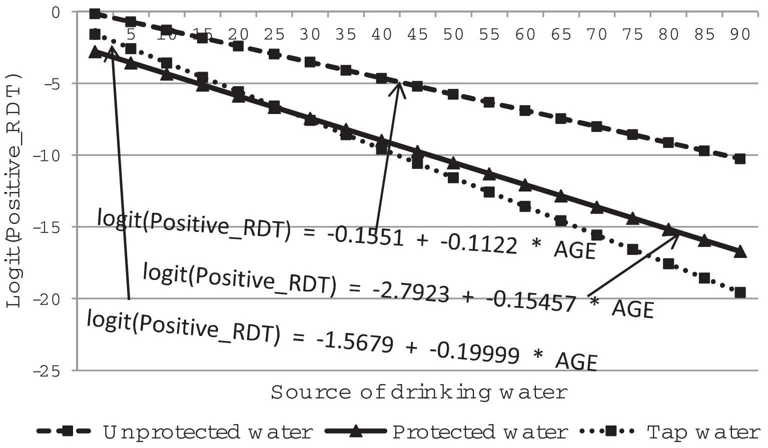


Figure 6 **Log odds associated with rapid diagnosis test and age for male respondents with source of drinking water.**

Interaction effects between the main source of water and the main material used for the room's roof is presented in Figure 9. From the figure, it is clearly seen that positive rapid diagnosis of malaria was significantly higher for households with a stick and mud roof followed by thatch and lastly a corrugated iron roof. This occurred with respondents who reported using tap water as well as protected and unprotected water for drinking (Figure 9). Furthermore, there was a significant difference in rapid diagnosis test between tap, protected and unprotected sources of drinking water for those who reported having thatch and stick and mud roofs. It is also shown that for corrugated iron roofs, the positive rapid diagnosis test was significantly lower for respondents who reported using tap water for drinking than for those who used protected and unprotected water for drinking.

The other significant two-way interaction effect was between the time taken to collect water and the main flooring material (Table 2). This result is presented graphically in Figure 10. A positive rapid diagnosis test was significantly higher in those rooms with earth and local dung plaster floors than for those with cement and wooden floors, for respondents who took < 30 minutes and >90 minutes to collect water. But, for respondents who took less than 30 minutes to collect water but had a cement floor, the positive rapid diagnosis was low. Furthermore, with respondents who took between 30 to 40 minutes to collect water, there was a lower positive rapid diagnosis test for those with earth and local dung plaster floors compared to wooden floors.

The relationship between the main source of drinking water and gender is presented in Figure 11. As the figure indicates, a positive rapid diagnosis test was significantly higher for female respondents than for male respondents who reported using unprotected water. There was however, no significant difference in a positive rapid diagnosis test between females and males who reported using protected and tap water for drinking.

The spatial model which is described above was used to produce a map of predicted prevalence of positive diagnosis malaria incidence rates for Amhara, Oromiya and SNNP regions of Ethiopia. When there is spatial data, the basic concern is the potential for spatial correlation in the observations. These spatial correlations could

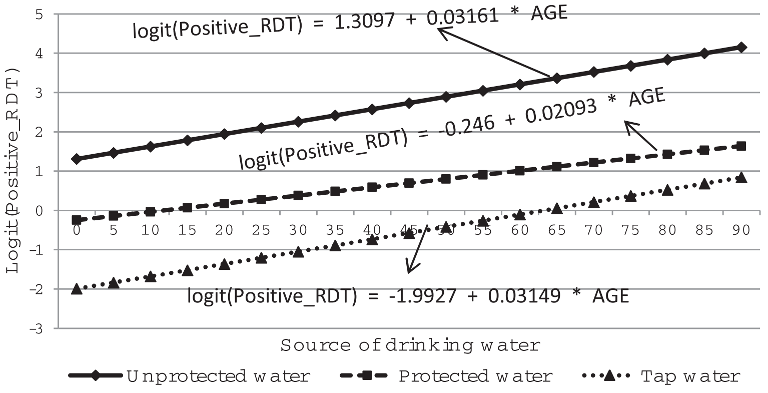


Figure 7 **Log odds associated with rapid diagnosis test and age for female respondents with source of drinking water.**

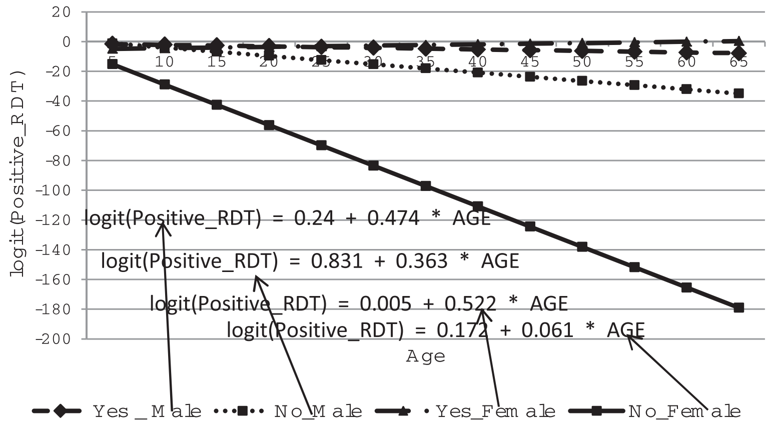


Figure 8 Log odds associated with rapid diagnosis test with age for male and female respondents with availability of electricity.

lead to incorrect estimates (estimates with underestimated standard errors). Spatial clustering of disease is almost to be expected since human populations generally live in spatial clusters rather than in a random distribution of space. An infectious disease that is highly associated with socio-economic, demographic and geographic factors is likely to be spatially clustered. This spatial clustering can occur even if the population distribution is not clustered. The model derived in this study explains some of the spatial patterns of the prevalence of malaria. The predicted prevalence of malaria is given in Figures 12 and 13. The prediction map (Figures 12 and 13) shows that the socio-economic, geographic and demographic factors are closely associated with the risk of malaria, mostly in the SNNP region followed by the Amhara and Oromiya regions. As can be seen from the map, the risk of transmission of malaria is of a moderately high intensity in almost all parts of the SNNP region. But, for the Oromiya region, the majority of households experience a lesser prevalence of malaria. Furthermore, from the map it can be seen that there is a high predicted value for the prevalence of malaria around the borders. This could be caused by cross-border migration of infected persons and the proximity of uncontrolled areas across the border, which may further add to the intensity of transmission in border areas.

Discussion

The first priority in the acute stage of a malaria epidemic is prompt and effective diagnosis and treatment. Having well-planned and timely vector control can significantly contribute to a reduction in the risk of infection and consequently in saving lives. Vector control must be proactive and should be implemented at an early stage of epidemic development. Timing depends on effective early warning and early detection. Because of this, the government of Ethiopia has developed strategies related to human resource development, monitoring, and evaluation to control malaria and reduce the hardship it causes. Based on this strategy, the main objective of the government is to make those areas with historically low malaria transmission, malaria free and a near zero malaria transmission in the remaining malarious areas of the country [44]. Based on some studies which were

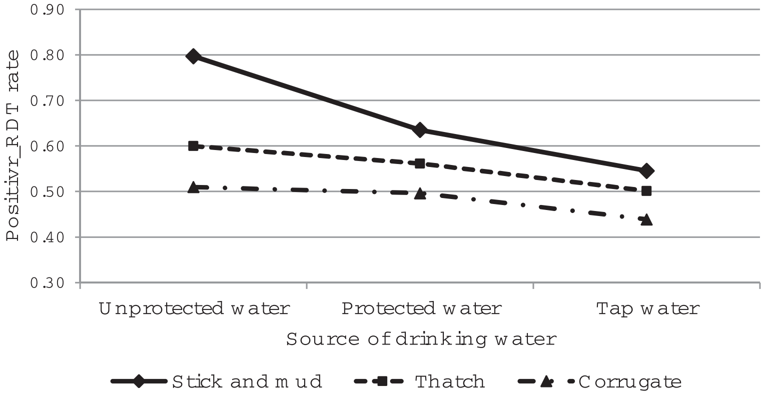


Figure 9 Log odds associated with rapid diagnosis test and source of drinking water with material of the room's roof.

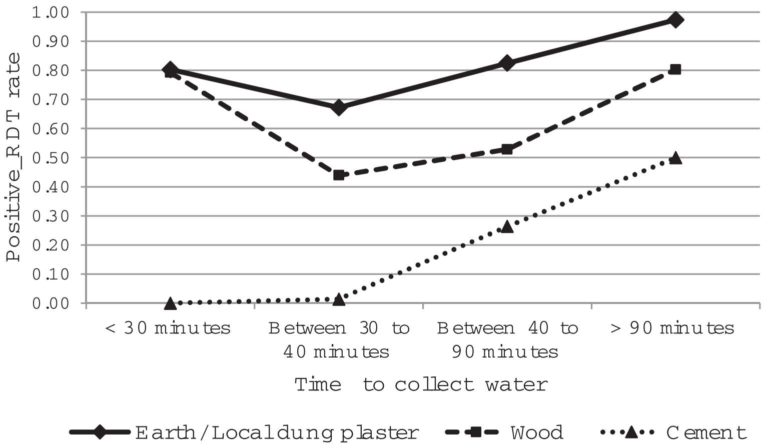


Figure 10 Log odds associated with rapid diagnosis test and time to collect water with material of the room's floor.

conducted previously, malaria was regarded as a disease of the poor or a disease of poverty [45]. Looking at the global distribution of malaria in the world suggested that the concentration of the disease is in the world's poorest continents and countries. Accurate information on the distribution of malaria in epidemic-prone areas on the ground permits interventions to be targeted towards the transmission and high-risk locations and households. Such targeting greatly increases the effectiveness of control measures but the inadvertent exclusion of these locations causes potentially effective control measures to fail. The computerized mapping and management of location data assists the targeting of interventions against malaria at the focal and household levels, leading to improved efficacy and cost-effectiveness of control.

As the distribution of malaria infection suggests, it is important to understand the relationship between malaria and poverty. This relationship is important to enable the design of coherent and effective policies and tools to tackle the problem [46,47]. As is already known, poverty is related to socio-economic factors. Therefore, it is important to identify those factors which are also related to the risk of malaria. Based on these facts, the findings from the current study show that the following socio-economic factors are related to the risk of malaria: main source of drinking water, time taken to collect water, toilet facilities, availability of radio, total number of rooms per person, main material of the room's walls, main material of the room's roof, main material of the room's floor, spraying of anti- mosquito, use of mosquito nets, total number of persons per net. Besides socio-economic factors, there are demographic and geographic factors which also have an effect on the risk of malaria. These include gender, age and family size. In addition to the main effects there were interactional effects between the socio-economic, demographic and geographic factors which also influenced the risk of malaria. Most notable of these were the interaction between main source of drinking water and main material of the room's roof,

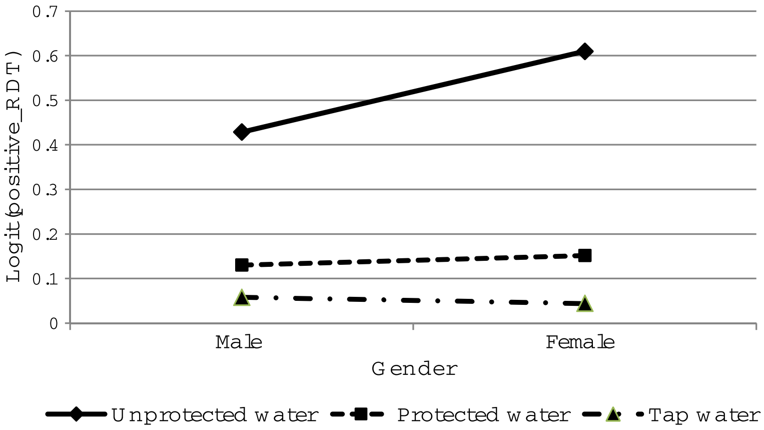


Figure 11 Log odds associated with rapid diagnosis test and main source of drinking water with gender.

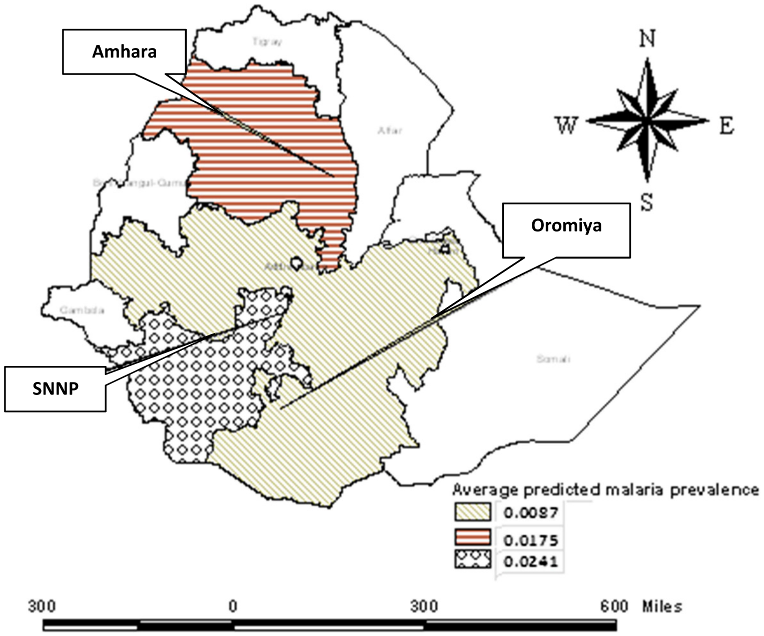


Figure 12 Predicted average spatial effects from the malaria prevalence model.

time taken to collect water and main material of the room's floor, age and gender, gender and availability of electricity, gender and main material of the room's floor, age, gender and main source of drinking water; and age, gender and availability of electricity.

Spatially correlated data cannot be regarded as independent observations. Therefore, ignoring the spatial variability might lead to an inaccurate estimation of parameters. Accordingly, unlike Ayele, et al. (2012), the spatial correlation structure was considered and the significance of the variables was checked and predictions of the malaria risk levels for the sampled areas were produced. A useful way of providing up to date information is in the use of GIS-based management systems. This method helps to address effective malaria vector control and management. Therefore, the spatial distribution of malaria incidence was one of the points which were important for such GIS studies.

Spatial clustering of malaria is almost predictable as human populations generally live in spatial clusters rather

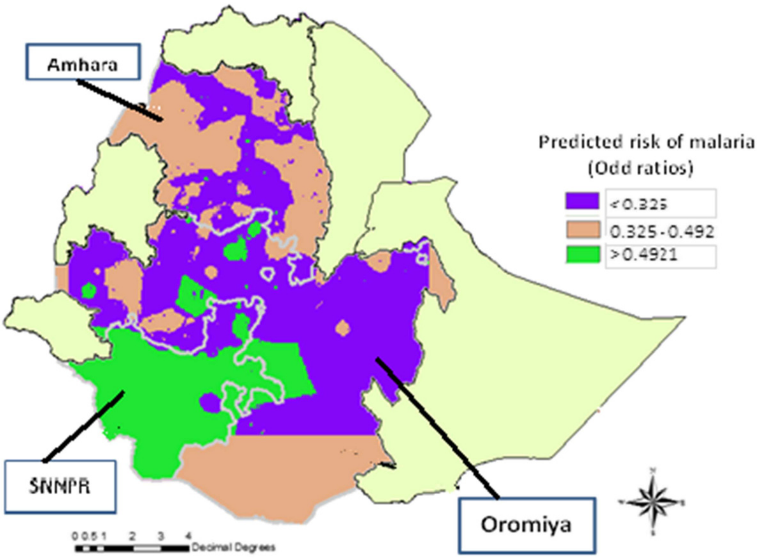


Figure 13 Predicted spatial effects from the malaria prevalence model.

than in random distributions of space. Disease which is highly correlated to socio-economic variables is likely to be spatially clustered. Therefore, the model explains some of the spatial patterns of malaria risk for Amhara, Oromiya and SNNP regions of Ethiopia. Moran’s and Geary's C tests were used to test for randomness [37-41]. The interest was to test if the spatial distribution of feature values is the result of random spatial processes. However, the test favors that the spatial distribution of feature values is not the result of random spatial processes. Moreover, the spatial distribution of high values and low values in the dataset is more spatially dispersed than would be expected. A dispersed spatial pattern often reflects some type of competitive process, i.e. a feature with a high value repels other features with high values; similarly, a feature with a low value repels other features with low values.

Malaria distribution is mainly related to the rainy seasons in Ethiopia. Therefore, understanding the nature of the Ethiopian climate is important. According to the Ethiopian National Meteorological Services Agency (NMSA), climates in Ethiopia can be divided into four climatic zones based on the pattern of rainfall. There are: the two-season type (the western half of Ethiopia) which is divided into district wet and dry seasons; bi-two season type (the south and southern of Ethiopia) is characterized by double wet seasons that occur between March to May and September to November with two dry seasons in between; the undefined season (dry northern part of the Ethiopian Rift Valley) mostly has irregular rainfall between July and February without any defined season; and the three-season type (central and south western Ethiopia). The average annual rainfall in the highlands of Ethiopia is above 1000 mm a year and it rises to 2000 mm and 3000 mm in the wet south western parts of Ethiopia. Therefore, the three regions have almost similar rainy months. Including the climate information into the analysis is important [48]. Since the climatic information is not included in the baseline household cluster malaria survey, this information will be included for future study.

Therefore, the results of this study provide evidence on the spatial distribution of socio-economic, demographic and geographic risk factors in the occurrence of malaria. This forms the basis for this research. Therefore, the utilization of socio-economic, demographic and geographic data on malaria rapid diagnosis test, including the information on the spatial variability, clarifies the effects of these factors. From the study it was observed that residents living in the SNNP region were found to be more at risk of malaria than those living in Amhara and Oromiya regions. Similarly, houses which were treated with anti- mosquito spray were less likely to be affected by malaria. However, a major challenge in the control of malarial infection was found to be in the type of toilet facilities available in the household. From the results, it was observed that individuals living in households which had no toilet facilities were more likely to be positive for malaria diagnosis tests. Furthermore, positive malaria diagnosis rates decreased with age and the risk of malaria increased per unit increase in family size. Generally, malaria parasite prevalence differed between age and gender, with the highest prevalence occurring in children and females.

From the findings of this study, it can be suggested that having toilet facilities, access to clean drinking water and the use of electricity offers a greater chance of knowing whether or not an individual in the household is at risk of malaria or not. In addition to this, using mosquito nets and spraying anti- mosquito treatment on the walls of the house were also found to be a way of reducing the risk of malaria. Similarly, having a cement floor and corrugated iron roof was found to be one means of reducing the risk of malaria. Based on the findings, different types of housing materials have an influence on the risk of malarial transmission with those houses constructed of poor quality materials having an increased risk. Moreover, the presence of particular structural features, such as bricks, that may limit contact with the mosquito vector, also helps to reduce infection. The risk of malaria therefore, is higher for households in a lower socio-economic bracket than for others who may enjoy a higher status and who are able to afford to take measures to reduce the risk of transmission. Therefore, with the correct use of mosquito nets, anti- mosquito spraying and other preventative measures, like having more rooms in a house, the incidence of malaria could be decreased. In addition to this, the study also suggests that the poor are less likely to use these preventative measures to effectively counteract the spread of malaria. To provide clean drinking water, proper hygiene and maintaining the good condition of a house is essential in controlling the transmission of malaria. With other control measures, including creating awareness about the use of mosquito nets, anti- mosquito sprays and malaria transmission, the number of malaria cases can be reduced. Furthermore, spatial statistics studies significantly contribute to the understanding of the distribution of malarial infections. The use of spatial statistics analysis is effective in monitoring and identifying high-rate malaria affected regions and helpful when implementing preventative measures. Finally, studies incorporating spatial variability are necessary for devising the most appropriate methodology for remedial action to reduce the risk of malaria.

Ethical clearance

The ethical protocol received approval from the Emory University Institutional Review Board (IRB 1816) and Amhara, Oromiya and SNNPR regional health bureaux. Informed consent was sought in accordance with the tenets of the declaration of Helsinki.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DGA acquired the data, performed the analysis and drafted the manuscript. TTZ and HGM designed the research. All authors discussed the results and implications and commented on the manuscript at all stages. All authors contributed extensively to the work presented in this paper. All authors read and approved the final manuscript.

Acknowledgements

We thank, with deep appreciation, Amhara RHB and The Carter Center, for providing and giving permission to use the data for this study.

Received: 17 December 2012 Accepted: 11 June 2013
Published: 17 June 2013

References

1. World Health Organization: *Malaria*. Geneva, Switzerland: WHO; 2011. http://www.who.int/mediacentre/factsheets/fs094/en/ Accessed Nov. 2011.
2. Snow RW, Peshu N, Forster D: **Environmental and entomological risk factors for the development of clinical malaria among children on the Kenyan coast.** *Trans R SocTrop Med Hyg* 1998, **92:**381–385.
3. Adugna A: *Malaria in Ethiopia.* Addis Ababa, Ethiopia: Ethiopian Demography and Health; 2011. http://www.ethiodemographyandhealth.org/MedVectoredDiseasesMalaria.pdf. Accessed Nov. 14, 2011.
4. Federal Ministry of Health (FMH): *Malaria and Other Vector-borne Diseases Control Unit.* Addis Ababa, Ethiopia: Federal Ministry of Health of Ethiopia; 1999.
5. The Carter Center (TCC): *Prevalence and risk factors for malaria and trachoma in Ethiopia.* Atlanta, USA: The Carter Center; 2007.
6. FMH: Federal Ministry of Health: *Guideline for malaria epidemic prevention and control in Ethiopia.* 2nd edition. Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2004.
7. World Health Organization: *Health action in crises: Horn of Africa Health Review.* Geneva, Switzerland: WHO. http://www.who.int/hac/crises/horn_of_africa/update_nov2011/en/.
8. WHO: World Health Organization: *Systems for the early detection of malaria epidemics in Africa: an analysis of current practices and future priorities, country experience.* Geneva, Switzerland: World Health Organization; 2006.
9. Zhou G, Minakawa N, Githeko A, Yan G: **Association between climate variability and malaria epidemics in the East African highlands.** *Proc Natl Acad Sci* 2004, **101:**2375–2380.
10. Adhanom T, Deressa W, Witten HK, Getachew A, Seboxa T: **Malaria.** In *The Eipdemiology and Ecology of Health and Disease in Ethiopia 1st edition.* 1st edition. Edited by Berhane Y, Hailemariam D, Kloos H, Shama PLC. Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2006:556–576. pp. 556–576.
11. FMH: *National five-year strategic plan for malaria prevention and control in Ethiopia 2006 – 2010.* Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2006.
12. Tulu NA: **Malaria.** In *The Ecology of Health and Disease in Ethiopia 2nd edition.* 2nd edition. Edited by Kloos H, Zein AZ. Geneva, Switzerland: Westview Press Inc; 1993:341–352.
13. FMH: Federal Ministry of Health: *Malaria: Diagnosis and Treatment Guidelines for Health Workers in Ethiopia.* Addis Ababa, Ethiopia: Federal democratic Republic of Ethiopia, Ministry of Health; 2004.
14. WHO: World Health Organization: *New Perspectives: Malaria Diagnosis, Report of a Joint WHO/USAID: Informal Consultation held on 25–27 October 1999.* Geneva, Switzerland: World Health Organization; 2000:4–48. 1999.
15. WHO: *Malaria Rapid Diagnostic Test Performance.* Geneva, Switzerland: WHO; 2009.
16. Wongsrichanalai C, Barcus MJ, Muth S, Sutamihardja A, Wernsdorfer WH: **A review of malaria diagnostic tools: microscopy and Rapid Diagnostic Test (RDT).** *The Am J Trop Med Hyg* 2007, **77**(Suppl 6):119–127.
17. Ayele DG, Zewotir T, Mwambi H: **Prevalence and risk factors of malaria in Ethiopia.** *Malar J* 2012, **11:**195.
18. Emerson PM, Ngondi J, Biru E, Graves PM, Yeshewamebrat E, Gebre T, Endeshaw T, Genet A, Mosher AW, *et al*: **Integrating an NTD with One of "The Big Three": combined Malaria and Trachoma Survey in Amhara Region of Ethiopia.** *PloS Negl Trop Dis* 2008, **2:**e197.
19. Shargie EB, Gebre T, Ngondi J, Graves PM, Mosher AW, Emerson PM, Ejigsemahu Y, Endeshaw T, Olana D, WeldeMeskel A, *et al*: **Malaria prevalence and mosquito net coverage in Oromia and SNNPR regions of Ethiopia.** *BMC Public Health* 2008, **8:**321.
20. Schabenberger O, Gotway CA: *Statistical Metods for Spatial Data Analysis.* Chapman and Hall/CRC; 2005.
21. Cressie N: *Statistics For Spatial Data.* New York: John Wiley & Sons; 1993.
22. Chiles JP, Delfiner P: *Geostatistics. Modelling Spatial Uncertainty.* Chichester: Wiley; 1999.
23. Goovaerts P: *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press; 1997.
24. Goovaerts P, Jacquez GM, Greiling D: **Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms.** *Geog Anal* 2005, **37:**152–182.
25. Lesaffre E, Spiessens B: **On the effect of the number of quadrature points in a logistic random-effects model: an example.** *Appl Stat* 2001, **50:**325–335.
26. Berridge DM, Crouchley R: *Multivariate Generalized Linear Mixed Models Using R Lancaster.* UK: CRC Press; 2011.
27. Zuur AF, Ieno EN, Walker N, Saveliev AA: *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer; 2009.
28. Zurr AF, Ieno EN, Smith GM: *Analysing Ecological Data.* New York: Springer; 2007.
29. Fox J: *Applied Regression Analysis and Generalized Linear Models.* 2nd edition. California: Sage Publications; 2008.
30. Madsen H, Thyregod P: *Introduction to General and Generalized linear models.* Baca Raton: CRC Press; 2010. Imprint of the Taylor & Francis Group.
31. Hengl T: *A Practical Guide to Geostatistical Mapping of Environmental Variables.* Italy: European Commission, Joint Research Centre, Institute for Environment and Sustainability; 2007.
32. Sacks J, Welch WJ, Mitchell TJ, Wynn HP: **Design and analysis of computer experiments.** *Statistical Science* 1989, **4:**400–423.
33. Gaetan C, Guyon X: *Spatial Statistics and Modeling.* New York: Springer; 2010.
34. Bivand RS, Pebesma EJ, Gomez-Rubio V: *Applied Spatial Data Analysis with R.* New York: Springer; 2008.
35. Breslow NE, Clayton DG: **Approximate Inference in Generalized Linear Mixed Models.** *J Am Stat Ass* 1993, **88:**9–25.
36. Kincaid C: *Guidelines for Selecting the Covariance Structure in Mixed Model Analysis.* Portage, USA: COMSYS Information Technology Services, Inc. http://www2.sas.com/proceedings/sugi30/198-30.pdf.
37. Cliff A, Ord J: **The choice of a test for spatial autocorrelation.** In *Display and Analysis of Spatial Data.* Edited by Davies JC, McCullagh MJ. London: John Wiley and Sons; 1975:54–77.
38. Cliff AD, Ord JK: *Spatial processes - models and applications.* London: Pion; 1981.
39. Geary R: **The contiguity ratio and statistical mapping.** *The Incorporated Statistician* 1954, **5:**115–145.
40. Moran PAP: **Notes on continuous stochastic phenomena.** *Biometrika* 1950, **37:**17–23.
41. Sokal RR, Oden NL: **Spatial autocorrelation in biology. 1. Methodology.** *Biological Journal of the Linnean Society* 1978, **10:**199–228.
42. Matheron G: **Principles of Geostatistics.** *Economic Geology* 1963, **58:**1246–1266.
43. Cressie N, Hawkins DH: **Robust estimation of the variogram.** *Mathematical Geology* 1980, **12**(2):115–125.
44. FMH: *Ethiopia National Malaria Indicator Survey 2007.* Addis Ababa, Ethiopia; 2008.
45. Abegunde D, Stanciole A: *An estimation of the economic impact of chronic noncommunicable diseases in selected countries.* Geneva, Switzerland: World Health Organization, Department of Chronic Diseases and Health Promotion (CHP); 2006.
46. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW: **The global distribution and population at risk of malaria: past, present, and future.** *Lancet Infect Disease* 2004, **4:**327–336.
47. Mendis K, Rietveld A, Warsame M, Bosman A, Greenwood B, Wernsdorfer WH: **From malaria control to eradication: the WHO perspective.** *Trop Med Inter Health* 2009, **14:**1–7.
48. NMSA: *Annual Climate Bulletin: For the year 2011.* Addis Ababa: Federal Democratic Republic of Ethiopia, Ministry of water and energy, National Meteorological Agency; 2011.

doi:10.1186/1475-2875-12-207
Cite this article as: Ayele *et al.*: Spatial distribution of malaria problem in three regions of Ethiopia. *Malaria Journal* 2013 **12**:207.