CO-MORBIDITY OF CHILDHOOD ANAEMIA AND MALARIA WITH DISTRICT-LEVEL SPATIAL EFFECTS

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Danielle Jade Roberts

October, 2021

Co-morbidity of childhood anaemia and malaria with district-level spatial effects

by

Danielle Jade Roberts

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
DOCTOR OF PHILOSOPHY
in
APPLIED STATISTICS

Thesis supervisor: Prof Temesgen Zewotir

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Danielle Jade Roberts, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
   (a) their words have been re-written but the general information attributed to them has been referenced, or
   (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

_____ _____

Danielle Jade Roberts (Student) Date

_____ _____

Prof Temesgen Zewotir (Supervisor) Date

Disclaimer

This document describes work undertaken as a PhD programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

DEDICATION

This thesis is dedicated to all of my family,

past and present.

ABSTRACT

Anaemia and malaria are the leading causes of sub-Saharan African childhood morbidity and mortality. This thesis aimed to explore the risk factors as well as the complex relationship between anaemia and malaria in young children across the districts or counties of four contiguous sub-Saharan African countries, namely Kenya, Malawi, Tanzania and Uganda. Nationally representative data from the Demographic and Health Surveys conducted in all four countries was used. The observed prevalence of anaemia and malaria was 52.5% and 19.7%, respectively, with a 15.1% prevalence of co-infection. Machine learning based exploratory classification methods were used to gain insight into the relationships and patterns among the explanatory variables and the two responses. The administrative districts are the level at which public health decisions are made within each of the countries. Accordingly, the best linear unbiased predictor (BLUP) ranking and selection approach was adopted to investigate the district-level spatial effects, while controlling for child-level, household-level and environmental factors. Further to the geoadditive model, a generalised additive mixed model with a spatial effect based on the geographical coordinates of the sampled clusters within the districts was applied. The relationship between the two diseases was further explored using joint modelling approaches: a bivariate copula geoadditive model and shared component model. The child's age, mother's education level, household wealth index and cluster altitude were found to be significantly associated with both the anaemia and malaria status of the child. The results of this study can help policy makers target the correct set of interventions or prevent the use of incorrect interventions for anaemia and malaria control and prevention. This aids in the targeted allocation of limited district health system resources within each of these countries.

Keywords: Adjusted odds ratios; Bayesian inference; Best linear unbiased predictor; Classification methods; Conditional autoregression; Copula model; Geoadditive model; Joint modelling; Spatial autocorrelation; Spline smoothing; Structured spatial effect; Unstructured spatial effect.

ACKNOWLEDGEMENTS

First and foremost, my thanks and appreciation goes to my supervisor, Prof Temesgen Zewotir. I am profoundly grateful for his unwavering support, encouragement and patience through this challenging journey of mine. I have grown and learnt a tremendous amount from his mentorship, which I will always be grateful for.

Many thanks goes to Prof Delia North, who has always believed in me, even when I did not. I wish to also extend special thanks to Ms Nombuso Zondo who has been on this PhD journey together with me. Knowing there was always a friend by my side going through similar challenges kept me strong and motivated. Appreciation also goes to all of my colleagues in the Statistics discipline at the University of KwaZulu-Natal, all of whom I consider my second family.

Thanks goes to Prof Glenda Matthews, Prof Benn Sartorius and Prof Robert Snow for their assistance during the preliminary stages of this thesis. I would also like to thank SACEMA (South African DST/NRF Centre for Epidemiological Modelling and Analysis) for the financial and academic support, as well as TDG (Teaching Development Grant) and UCDP (University Capacity Development Programme) for the financial support, with which I had the opportunity to attend conferences and workshops that aided in developing my ability to do this research.

Last but not least, thank you to by biggest fans; my family. I thank them for their constant love and support, and for their patience with me during the many hours that I was unavailable while working on this research.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

95% CrI	95% credible interval
95% CI	95% confidence interval
AIC	Akaike information criterion
ANN	Artificial neural network
AOR	Adjusted odds ratio
BIC	Bayesian information criterion
CART	Classification and regression trees
CDF	Cumulative distribution function
CI	Confidence interval
CrI	Credible interval
CV	Cross-validation
DHS	Demographic and Health Survey
DIC	Deviance information criteria
EVI	Enhanced vegetation index
g/dl	Grams per decilitre
GAM	Generalised additive model
GAMM	Generalised additive mixed model
GLM	Generalised linear model
GLMM	Generalised linear mixed model
GTS	Global Technical Strategy for Malaria 2016–2030
Hb	Hemoglobin
HBHI	High burden to high impact
HIC	High income countries
LMIC	Low and middle income countries
LST	Land surface temperature
MIS	Malaria Indicator Survey

RBF	Radial basis function
RDT	Rapid diagnostic test
SCM	Shared component model
SSA	sub-Saharan Africa
SVM	Support vector machine
WHO	World Health Organization

RESEARCH OUTPUT

Peer-reviewed Publications

The following papers have been published or submitted or are under preparation from this thesis:

1. **Roberts, D.J.** & Zewotir, Z. (2019) District Effect Appraisal in East sub-Saharan Africa: Combating Childhood Anaemia. *Anemia*, Volume 2019, Article ID 1598920, doi: https://doi.org/10.1155/2019/1598920

2. **Roberts, D.J.**, Matthews, G., Snow, R.W., Zewotir, Z. & Sartorius, B. (2020) Investigating the spatial variation and risk factors of childhood anaemia in four sub-Saharan African countries. *BMC Public Health*, 20, 126, doi: https://doi.org/10.1186/s12889-020-8189-8

3. **Roberts, D.J.** & Zewotir, Z. (2020) Copula geoadditive modelling of anaemia and malaria in young children in Kenya, Malawi, Tanzania and Uganda. *Journal of Health, Population and Nutrition*, 39, 8, doi: https://doi.org/10.1186/s41043-020-00217-8

4. **Roberts, D.J.** & Zewotir, Z. Shared component modelling of early childhood anaemia and malaria in Kenya, Malawi, Tanzania and Uganda. Under review with *Spatial and Spatio-temporal Epidemiology*.

5. **Roberts, D.J.** & Zewotir, Z. Machine learning to explore anaemia and malaria in young children. In preparation.

Conference Papers/Proceedings/Symposium

1. **Roberts, D.J.** & Matthews, GM. (2018). Geostatistical modelling of childhood anaemia in four sub-Saharan African countries. *Presented* at the 60th SASA An-

nual Conference, UNISA, Johannesburg, South Africa, 26-29 November 2018.

2. **Roberts, D.J.** & Zewotir, Z. (2019) Spatial heterogeneity of childhood anaemia in four sub-Saharan African countries. Department of Statistics Malaysia (DOSM). *Published* in the Proceeding of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 2, 239–248

3. **Roberts, D.J.** & Zewotir, Z. (2019). Spatial heterogeneity of childhood anaemia in four sub-Saharan African countries. *Presented* at the 62nd ISI World Statistics Congress, Kuala Lumpur, Malaysia, 18-23 August 2019.

4. **Roberts, D.J.** & Zewotir, Z. (2019). Risk factors and spatial heterogeneity of childhood anaemia in four sub-Saharan African countries. *Presented* at the Joint Conference of the Sub-Saharan Africa Network (SUSAN) of the International Biometrics Society (IBS) and DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB), Cape Town, South Africa, 8 - 11 September 2019.

5. **Roberts, D.J.** & Zewotir, Z. (2019). Joint modelling of anaemia and malaria in young children. *Presented* at the 61st SASA Annual Conference, Nelson Mandela University, Port Elizabeth, South Africa, 27 - 29 November 2019.

6. **Roberts, D.J.** & Zewotir, Z. (2021). Shared component modelling of early childhood anaemia and malaria in four sub-Saharan African countries. *Accepted for presentation* at the 62nd SASA Annual Conference, Stellenbosch University, South Africa, 1 - 3 December 2021.

CHAPTER 1

INTRODUCTION

Anaemia and malaria are major contributors of childhood morbidity and mortality, particularly in low-and middle-income countries in sub-Saharan African where health care resources are limited (Kuziga et al., 2017; WHO, 2018). Anaemia is defined as a significant reduction in hemoglobin (Hb) concentration which decreases the amount of oxygen reaching the tissues and organs of the body. According to the World Health Organisation (WHO) definitions for children aged 6-59 months, a child with an Hb level of 10-10.9 g/dL is considered as having mild anaemia, 7-9.9 g/dL as having moderate anaemia, and an Hb level of less than 7.0 g/dL as having severe anaemia (WHO, 2011). The adverse health consequences of childhood anaemia include altered cognitive function, impaired motor development and growth, poor school performance, poor immune function and susceptibility to infections, decreased responsiveness and activity, and increased body tension and fatigue. If left untreated, the long-term effects and consequences of anaemia in early childhood are irreversible, if mortality has not occurred (WHO, 2013). In 2019, the global prevalence of anaemia in children aged 6 to 59 months was 39.8%, which was equivalent to 269 million children with anaemia (WHO, 2021b). The greatest burden of childhood anaemia is experienced in Africa, where the prevalence stood at 60.2% in 2019 (WHO, 2021b). While the global prevalence of anaemia in young children has been on the decline since 2000 from 48%, it has become more stagnant in the last decade.

The causes of anaemia are multifactorial and interrelate in a complex way. Such causes include iron deficiency; other micronutrient deficiencies such as folate, vitamin B12 and vitamin A; intestinal parasites; malaria; HIV infection; and chronic diseases such as sickle cell disease. Iron deficiency is the most common cause of child-

hood anaemia in developed countries, however there are many other contributing factors in less developed countries. In countries that are highly malaria-endemic, particularly in sub-Saharan Africa, malaria accounts for the majority of cases of anaemia in young children (White, 2018). On the other hand, severe anaemia can increase a child's susceptibility to malaria in these regions (Adebayo et al., 2016). Malaria is caused by the *Plasmodium* parasite that is transmitted via the bite of infected *Anopheles* mosquitoes. The most severe form of malaria and the majority of malaria-related deaths are attributed to the *Plasmodium falciparum* parasite (WHO, 2016). Furthermore, *Plasmodium falciparum* infection contributes to the etiology as well as the severity of anaemia through several mechanisms, primarily through the direct destruction of infected red blood cells (Menendez et al., 1997). Iron absorption may also be affected by malaria infection, thus resulting in anaemia.

In 2019, the number of malaria cases globally in the general population was estimated at 229 million, with Africa experiencing a disproportionately high share of the burden at 94% of the cases and deaths (WHO, 2021a). As young children are yet to develop an immunity to malaria, they are considered most vulnerable. This is highlighted by children under 5 years of age accounting for 67% of all malaria deaths worldwide in 2019 (WHO, 2021a). However, in the last decade, significant progress has been made to scale up interventions and control measures for malaria. Increased partnerships between governments and other stakeholders in the fight against malaria have resulted in a substantial reduction in malaria morbidity and mortality in children with the proportion of infected children having halved in endemic areas of Africa since 2000 (Cibulskis et al., 2016). While malaria remains a major killer of young children, the number of malaria deaths in children has decreased from 723 000 globally in 2000 to 274 000 in 2019 (WHO, 2021a). These efforts in the reduction of malaria is largely attributed to the progress towards achieving the Millennium Development Goals (MDGs) set out by the United Nations (UN) and adhered to by all UN Member States. The MDGs ended in 2015 and were super-

seded by the Sustainable Development Goals (SDGs), which includes goals of ending epidemics of malaria and other communicable diseases by the year 2030 (WHO, 2015b). While it is up to all participating malaria-endemic countries to develop their own national framework in order to reach this goal, the WHO's Global Technical Strategy for Malaria 2016-2030 (GTS) has been developed with the aim of assisting countries in reducing their malaria burden. Adopted by the World Health Assembly in 2015, this strategy provides comprehensive technical guidance to countries and committed partners for the next 15 years, emphasising the importance of scaling up malaria control measures and moving towards elimination (WHO, 2015a).

While there are well-defined goals and targets for malaria elimination, there are no such direct goals and targets set for anaemia in children, thus it has not received adequate attention. Rather, goals for anaemia reduction in children coincide with Sustainable Development Goal 3 (Health), which includes ending preventable deaths of children under 5 years of age by 2030, as well as coincides with Sustainable Development Goal 2 (Hunger and Food Security), which includes ending all forms of malnutrition in children by 2030 (WHO, 2015b). It has been recommended by the WHO and UNICEF (2004) that strategies for anaemia control should be built into a country's primary health care system and existing programmes such as maternal and child health, integrated management of childhood illness, adolescent health, making pregnancy safer/safe motherhood, roll-back malaria, deworming (including routine anthelminthic control measures) and stop-tuberculosis. However, these strategies should be tailored to local conditions and take into account the specific etiology and prevalence of anaemia in a given setting and population group (WHO and UNICEF, 2004). Thus, it is of great importance to identify the groups of a population at risk of anaemia, as well as the significant contributing factors of anaemia in these groups. Targeting the correct interventions to the groups of populations most at risk would result in a more efficient delivery system of limited national resources.

As the causes of childhood anaemia are multiple and complex, identifying significant factors associated with an increased risk of anaemia in a child is relevant to developing appropriate and effective interventions. Such studies aid in identifying the subpopulations that are most at risk, thus assisting in creating a more targeted approach to anaemia control and prevention (Soares Magalhães & Clements, 2011). However, studies identifying these factors should account for spatial heterogeneity and spatial autocorrelation in the observations. Failure to do so may produce inaccurate estimates and thus misleading results and ineffective anaemia control programs (Mainardi, 2012; SoaresMagalhães & Clements, 2011).

Spatial autocorrelation arises when observations close in proximity tend to be more alike than those further apart and is present even if the observations have been recorded in a standardised way (Kneib et al., 2008). Spatial heterogeneity refers to the spatial variation or uneven distribution of attributes across a region (Wang et al., 2016). Climatic and environmental factors, such as temperature, rainfall, and proximity to waterbodies, among others, are largely responsible for such spatial heterogeneity as its effects are usually only partially explained by the covariates that are available in a model (Kneib et al., 2008). Indeed, many other factors that vary geographically can also contribute to spatial heterogeneity in observations, such as the availability and distance to quality child health care, access to a reasonable transport system, culture and the cost of living, all of which may not always be fully explained by the available covariates. Various methods of accounting for spatial autocorrelation and spatial heterogeneity have been well established due to the increased accessibility of spatially indexed data (Kazembe, 2007; Kneib et al., 2008; Besag et al., 1991).

There have been a considerable number of studies assessing the risk factors and determinants of anaemia in children (Gari et al., 2017; Kuziga et al., 2017; Moschovis et al., 2018; Phyllis Atta Parbey et al., 2019, and references therein), some of which

have also assessed the spatial variation of anaemia (Mainardi, 2012; Soares Magalhães et al., 2013a; Ngwira & Kazembe, 2015; Habyarimana et al., 2017). However, few studies have focused on countries in eastern sub-Saharan Africa which experiences a high burden of childhood anaemia (Stevens et al., 2013). In particular, there is a lack of studies on the risk factors and spatial variation of anaemia, as well as its relationship with malaria in young children in Kenya, Malawi, Tanzania and Uganda. Numerous individual studies on childhood anaemia have been carried out in Kenya (Ngesa & Mwambi, 2014), Malawi (Ngwira & Kazembe, 2015; Kazembe, 2007), Tanzania (Kejo et al., 2018) and Uganda (Kuziga et al., 2017), all of which differ in scope and coverage. However, the advantage of focusing on multiple countries that form contiguous regions is to be able to investigate the spatial heterogeneity between the countries. This assists in determining whether the significant drivers of childhood anaemia are country specific or whether they cross the borders of the countries and are thus shared between neighbouring countries.

In addition to the direct causes of anaemia, numerous studies have shown that individual, household and environmental factors have a significant effect on the risk of anaemia in young children. Various studies have evidenced an increased risk of anaemia among children whose caregivers are less educated, as well as among those who reside in low-income households and households with poor sanitation (Kuziga et al., 2017; Nambiema et al., 2019; Gayawan et al., 2014; Ngesa & Mwambi, 2014). The age and gender of the child has also been demonstrated to have a significant effect on their risk of anaemia (Habyarimana et al., 2017; Gari et al., 2017; Ngwira & Kazembe, 2015; Ngesa & Mwambi, 2014). Many of these risk factors of childhood anaemia overlap with those of malaria (White, 2018; Adebayo et al., 2016). In addition, such risk factors vary across geographical locations, which in turn contributes to the spatial patterns and variation in the prevalence and risk of both diseases. Thus, identifying the geographical locations associated with an increased risk would aid in formulating targeted interventions. This thesis therefore set out to

investigate anaemia and malaria and the relationship between the two diseases in children aged 6 to 59 months in Kenya, Malawi, Tanzania and Uganda. The specific objectives were:

- To perform exploratory data analysis to examine the patterns and relationships among numerous explanatory variables and the two diseases in order to identify the appropriate statistical techniques to be applied.
- To identify the significant risk factors associated with childhood anaemia and malaria in the four countries.
- To investigate the spatial variation of childhood anaemia and malaria across the four countries, with particular interest in the district-level spatial effect.
- To jointly model the spatial variation of childhood anaemia and malaria in young children across the districts of the four countries.

The remainder of this thesis is organised as follows: Chapter 2 introduces and describes the characteristics of the data that was used in the analyses. Chapter 3 presents the results of exploratory data analyses using various supervised machine learning techniques. Chapter 4 presents an overview of the geoadditive model and its application to investigate the spatial variation and risk factors of childhood anaemia. Chapter 5 provides a description of the best linear unbiased prediction (BLUP) technique that was applied in order to rank the performance of the districts on the likelihood of childhood anaemia. A review of the copula geoadditive model to jointly model childhood anaemia and malaria is presented in Chapter 6. Furthermore, the chapter discusses the results of the association between the two responses, which was set to vary according to the district of residence across the four countries. Chapter 7 provides an overview and the results of a child-level shared component model used to jointly model the residual spatial variation in the likelihood of childhood anaemia and malaria. Lastly, the discussion and conclusion is provided in Chapter 8.

CHAPTER 2

THE DATA

2.1 Study Regions

The four countries considered in this study consist of Malawi, Kenya, Tanzania and Uganda. These four countries are situated on the east of sub-Saharan Africa and together form one contiguous region as shown in Figure 2.1.

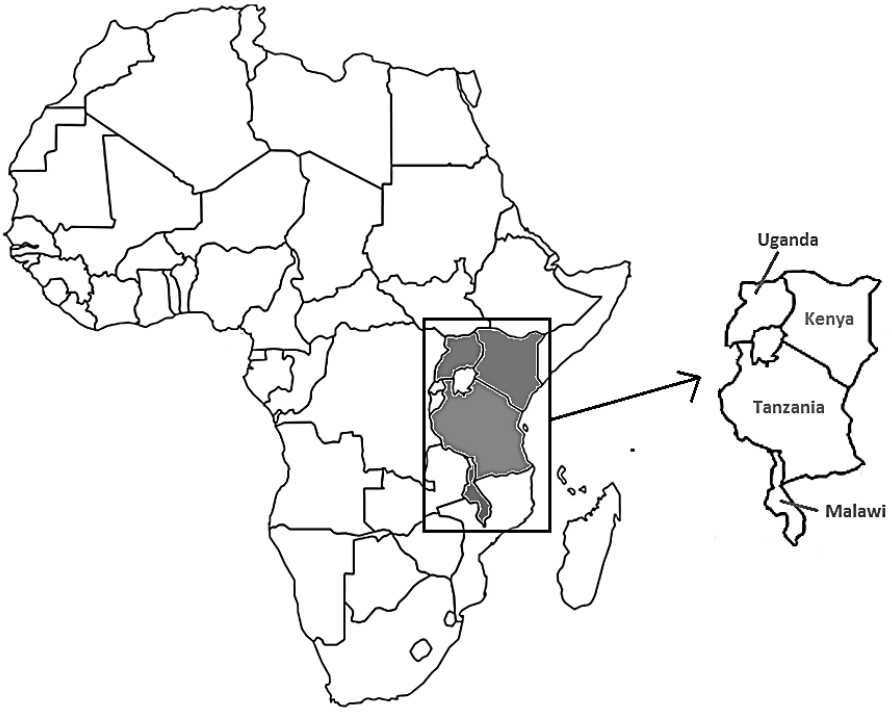


Figure 2.1: Study Regions

2.2 Data Sources

This thesis uses data collected in the Demographic and Health Surveys (DHS) and Malaria Indicator Surveys (MIS) carried out in Kenya, Malawi, Tanzania and Uganda between 2015 and 2017, namely the 2015 Kenya Malaria Indicator Survey (KMIS2015),

the 2017 Malawi Malaria Indicator Survey (MMIS2017), the 2015-2016 Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS2015) and the 2016 Uganda Demographic and Health Survey (UDHS2016).

2.3 Sampling Design and Data Collection

The DHS and MIS were designed to provide national, regional, urban and rural estimates of key health indicators (The DHS Program, 2017). These surveys were nationally represented and utilised a stratified two-stage cluster design. The samples from each country were not spread geographically in proportion to their respective populations, but rather equally across the regions. The first stage of the survey design involved selecting clusters from a list of enumeration areas (EA) which made up the primary sampling units (PSUs). Clusters were selected with a probability proportional to their size. The second stage of the selection process involved systematic sampling of households from the list of households in each cluster, with an equal number of households selected from the clusters. The selected households were visited and interviewed by trained staff. A thorough review of the sampling methodology is presented in the DHS Sampling Manual (ICF International, 2012).

Three questionnaires, the Household Questionnaire, Women's Questionnaire and Men's Questionnaires, were carried out in the selected households. These questionnaires were designed to collect information regarding the characteristics of the household and eligible women and men. The Household Questionnaire collected basic information on the characteristics of each member and recent visitors of the household, including age, sex, and relationship to the head of the household. This questionnaire also collected information on characteristics of the household's dwelling unit, such as source of water; type of toilet facilities; materials used for the floor, roof and walls of the house; and ownership of various durable goods. The Women and Men's Questionnaires were used to collect a range of information from all eligible women and men in the selected households.

With the consent of a parent or guardian in the household, all children between the ages of 6 and 59 months were tested for anaemia and malaria using blood specimens collected from a finger- or heel-prick. A child's Hb concentration was measured using a portable HemoCue analyser. Based on the Hb levels adjusted for altitude, a child was diagnosed as anaemic if their Hb level was less than 11 g/dl, and non-anaemic otherwise (WHO, 2011). At altitudes from 1000m above sea level, Hb levels increase to compensate for lower partial pressure of oxygen and reduced oxygen saturation of blood. This results in a compensatory increased production of red blood cells that enables sufficient supply of oxygen to the tissues (Centers for Disease Control and Prevention, 1989). Therefore, with this upward shift in Hb levels from higher altitudes, the Hb cut-off points for anaemia would be higher. However, rather than changing the cut-off points, for the purpose of determining a child's anaemia status in the demographic and health surveys, the Hb levels are adjusted for altitude using a calculation based on the original measured Hb level (Croft et al., 2018b).

Malaria in children was diagnosed using a rapid diagnostic test (RDT) which consisted of testing a drop of blood using the SD Bioline Pf/Pv RDT, which tests for the presence of the parasite *Plasmodium falciparum*, the most dangerous *Plasmodium* parasite, as well as tests for the presence of other *Plasmodium* species. The result of the test was available in 15 minutes. This type of test has become more widely used as a diagnostic test where a reliable microscopy test is not available (Nas et al., 2020).

2.4 Variables of Interest

The two child health outcomes of interest were the their anaemia status and malaria status, where both responses were binary. The explanatory variables considered were based on those found in literature to have some association with anaemia and/or malaria, as well as those expected to be determinants of each outcome.

These variables, which are displayed in Figure 2.2, comprised of a number of demographic, socio-economic and environmental factors, including the gender and age of the child, the mother's highest education level, the number of members in the household (size of the household), the type of place of residence (rural or urban), the household wealth index, the type of toilet facility, the age and gender of the head of the household, three environmental factors: cluster altitude, day land surface temperature and the enhanced vegetation index, as well as the country of residence.

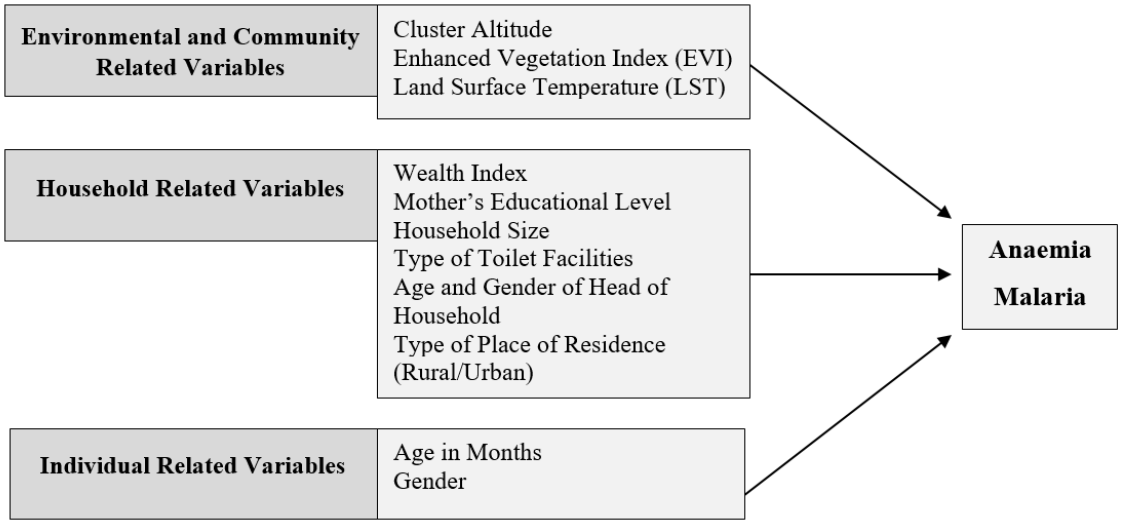


Figure 2.2: Potential risk factors of anaemia and malaria among young children

The household wealth index was based on the composite measure of a household's cumulative living standard and was calculated according to the ownership of various household assets (Croft et al., 2018a). The household was assigned a standardised score for each asset, the scores were then summed for each household to obtain a household wealth index Z-score, which is a continuous measure and the form of the wealth index used here. The DHS program has made available standardised files of the most commonly used geospatial covariates up to the year 2015, which can be linked to DHS datasets via the cluster ID (Mayala et al., 2018). Therefore, as no information regarding intestinal parasites (a known risk factor for anaemia (Alemu et al., 2017)) was collected in the surveys, selected spatially indexed envi-

ronmental covariates were considered as a proxy (Banhela et al., 2017; Michael et al., 2010). Specifically, the cluster level average day land surface temperature (LST) and the cluster level average Enhanced Vegetation Index (EVI) for 2015. These environmental factors also impact malaria transmission as they affect both the *Plasmodium* parasite and the host (the *Anopheles* mosquito). *Plasmodium* parasites are sensitive to changes in temperature where their development slows with a drop in temperature and stops at high temperatures (Weaver, 2014). However, rainfall expands the breeding ground of the mosquito and also indirectly contributes to the longevity of the adult mosquito by increasing relative humidity (Yamana TK, 2013). For the purpose of our study, we used the enhanced vegetation index as an indicator for rainfall, as it is correlated with rainfall (NASA Earth Observatory, 2020).

In addition, the spatial variation of childhood anaemia and malaria across the administrative levels of the countries will be investigated. These administrative levels were chosen based on the levels for which public health decisions are made within each country, which is represented by the districts/counties. We examine the spatial effect of all 47 counties or districts for Kenya; 26 out of 28 districts for which data was available for Malawi; 176 out of 184 districts for which data was available for mainland Tanzania; and 121 out of 122 districts for which data was available for Uganda. Thus, a total of 370 districts are considered.

2.5 Descriptive Statistics

The total sample size combined consisted of 18196 children from the four countries. Table 2.1 shows the distribution of the sample according to the categorical characteristics and Table 2.2 presents some descriptive measures of the continuous characteristics. Over 40% of the sample came from Tanzania, which is also the largest of the four countries (Table 2.1). The majority of the children in the sample had mother's with only a primary school level of education (53.6%) and resided in households with PIT Latrine toilet facilities (80.2%). Furthermore, the sample primarily con-

sisted of children from rural areas (74.7%). The average age of the children in the sample was 32.48 months (Table 2.2). The average household wealth index Z-score was -0.23, which is unsurprising considering that the majority of the children resided in rural areas.

Table 2.1: Sample size (%) according to categorical characteristics

Characteristic	Sample Size (%)
Country	
Kenya	3424 (18.8%)
Malawi	2270 (12.5%)
Tanzania	7819 (42.97%)
Uganda	4683 (25.7%)
Gender	
Male	9143 (50.2%)
Female	9053 (49.8%)
Mother's Highest Education Level	
No education	2893 (15.9%)
Primary	9757 (53.6%)
Secondary and Higher	3110 (17.1%)
Unknown	2436 (13.4%)
Type of Place of Residence	
Urban	4605 (25.3%)
Rural	13591 (74.7%)
Type of Toilet Facilities	
No Toilet Facility	2367 (13.0%)
PIT Latrine	14587 (80.2%)
Flush Toilet	1242 (6.8%)
Gender of Head of Household	
Male	13869 (76.2%)
Female	4327 (23.8%)

Table 2.2: Descriptive measures of continuous characteristics

Characteristic	Minimum	Maximum	Mean	SD
Child's age in months	6	59	32.48	15.55
Age of household head	13	95	41.20	14.17
Household size	2	48	6.61	3.53
Wealth index Z-score	-2.46	4.7	-0.23	0.90
Cluster altitude (in 100m)	0.05	29.9	11.52	4.93
EVI (in thousands)	1.36	5.3	3.25	0.66
LST	16.1	34.2	24.50	2.39

Table 2.3 presents the observed anaemia and malaria prevalences. The observed prevalence of anaemia was 52.5%, while the malaria prevalence was 19.7%, with a 15.1% prevalence of both anaemia and malaria in the children. The uncorrected Kendall's tau correlation between anaemia and malaria was estimated at 0.239, which was statistically significant at a 5% significance level. Based on Table 2.3, 76% (2750 out of 3592) of the children who tested positive for malaria, had anaemia as well. This is an indication of the contribution that malaria has on the burden of anaemia.

Table 2.3: Cross-tabulation of the sample according to anaemia and malaria status

		Result of malaria rapid test		Total
		Positive	Negative	
Anaemia Status	Anaemic	2750 (15.1)	6809 (37.4)	9559 (52.5)
	Non-anaemic	842 (4.6)	7795 (42.8)	8637 (47.5)
	Total	3592 (19.7)	14604 (80.3)	18196

Figure 2.3 displays the observed prevalences of anaemia (a), malaria (b) and both anaemia and malaria (c) according to the district of residence. This figure highlights the significant burden of anaemia compared to malaria. What is interesting to note, there were numerous districts with a high prevalence of anaemia but a low prevalence of malaria. This suggests that there were other contributing factors of anaemia in these districts. The patterns of the prevalence of malaria and the prevalence of both were fairly similar, therefore indicating that children with malaria are most likely to have anaemia as well.

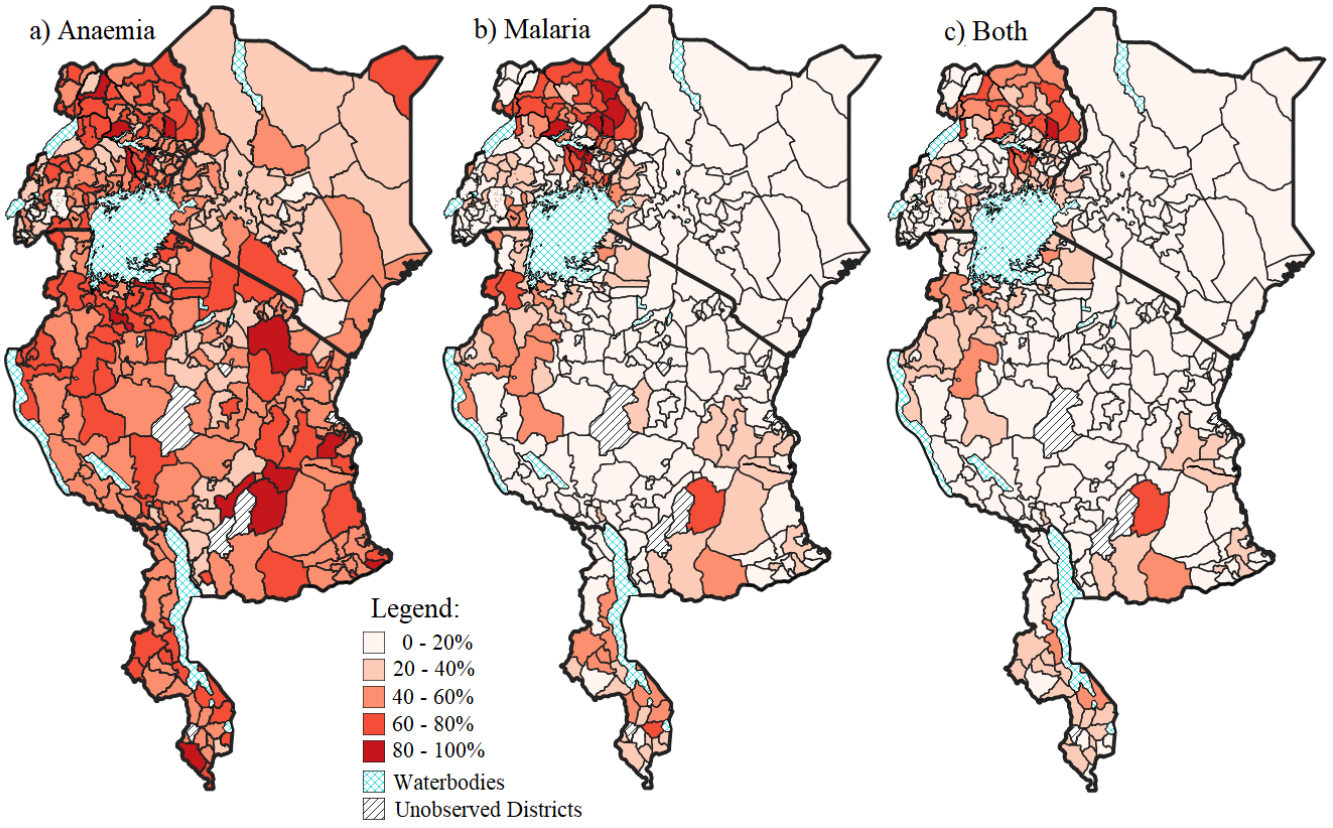


Figure 2.3: Observed prevalence of a) anaemia; b) malaria c) both anaemia and malaria according to district of residence

Figure 2.4 presents the observed prevalence of anaemia, malaria and both anaemia and malaria among the children residing in each country. To aid in the assessment of anaemia as a public health problem, the WHO categorises anaemia into four, where it is considered a severe health problem if the prevalence is 40% or more, moderate from 20% to 39.9%, mild from 5% to 19.9%, and no public health problem if the prevalence is less than or equal to 4.9% (Challa & Amirapu, 2016). According to these classifications, Malawi, Tanzania and Uganda have a severe public health problem. Kenya had the lowest observed prevalence of anaemia (38.3%), malaria (9.3%) and both (6%) in children. Uganda had the highest observed prevalence of malaria (33.6%) as well as both (25.3%) in children. The observed prevalences according to the gender of the child are displayed in Figure 2.5. No large differences were seen between male and female children.

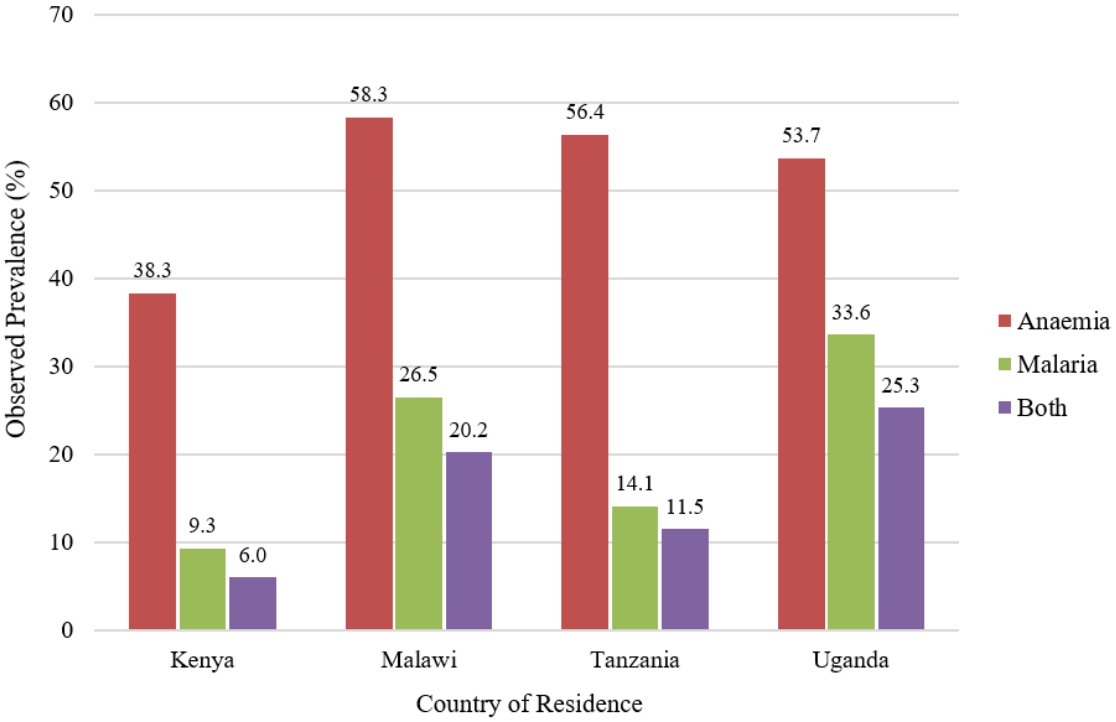


Figure 2.4: Observed prevalence of anaemia, malaria and both according to the country of residence

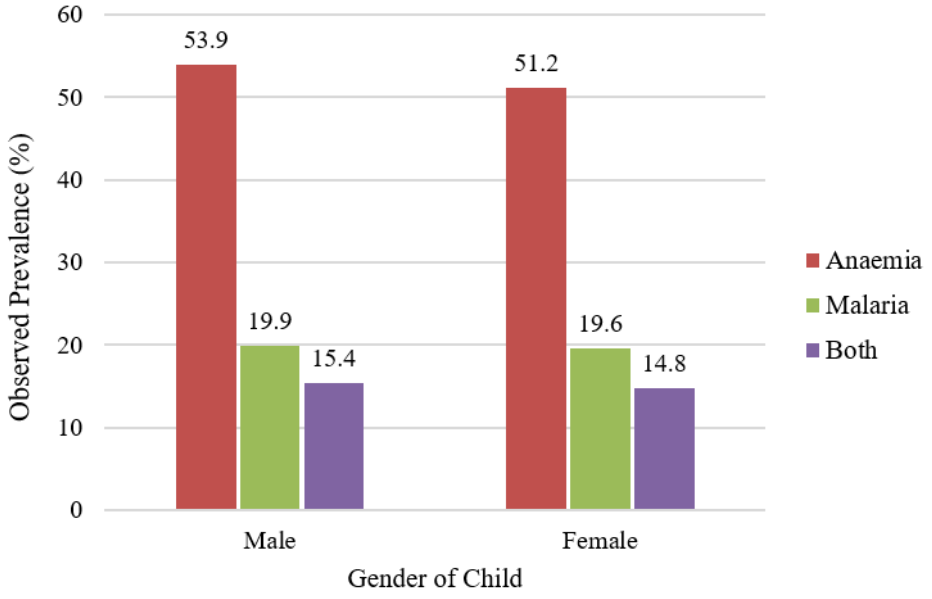


Figure 2.5: Observed prevalence of anaemia, malaria and both according to the child's gender

Figure 2.6 provides the observed prevalences according to the mother's highest education level as well as the gender of the head of household. The observed prevalence of anaemia, malaria and both anaemia and malaria decreased with an increase in the mother's education level. Very similar patterns were evident between children in households headed by males and females.

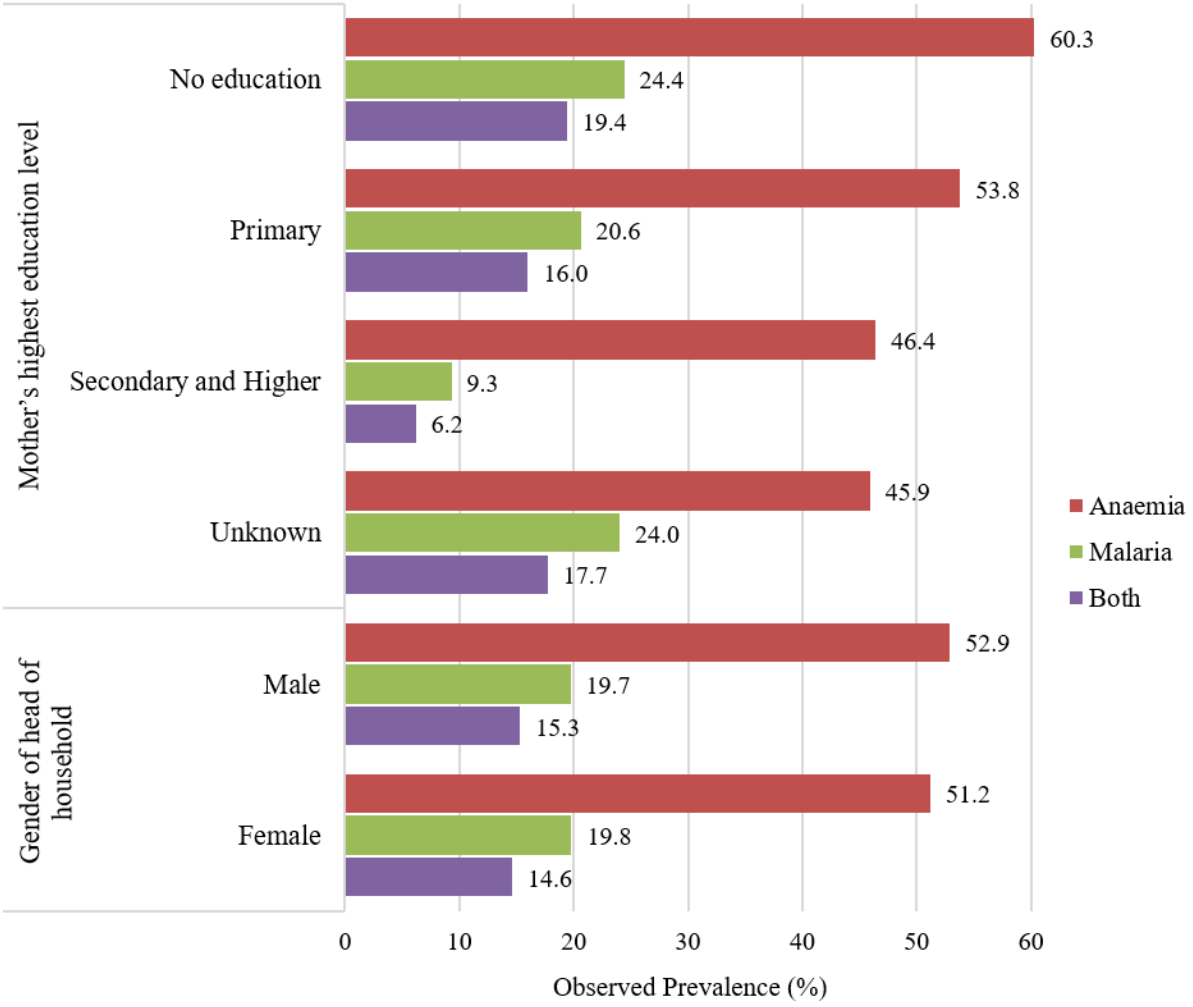


Figure 2.6: Observed prevalence of anaemia, malaria and both according to the mother's education level and gender of household head

The observed prevalence of anaemia, malaria and both according to the type of place of residence and type of toilet facility is displayed in Figure 2.7. While the observed prevalence of anaemia was lower among children residing in urban areas compared to those in rural areas, the difference was not substantial. However, a more promi-

nent difference was seen in the prevalences of malaria and both anaemia and malaria between children residing in urban and rural areas. Specifically, these prevalences were considerably higher among children residing in rural areas compared to those in urban areas. The observed prevalence of anaemia, malaria and both decreased with an improvement in the type of toilet facility.

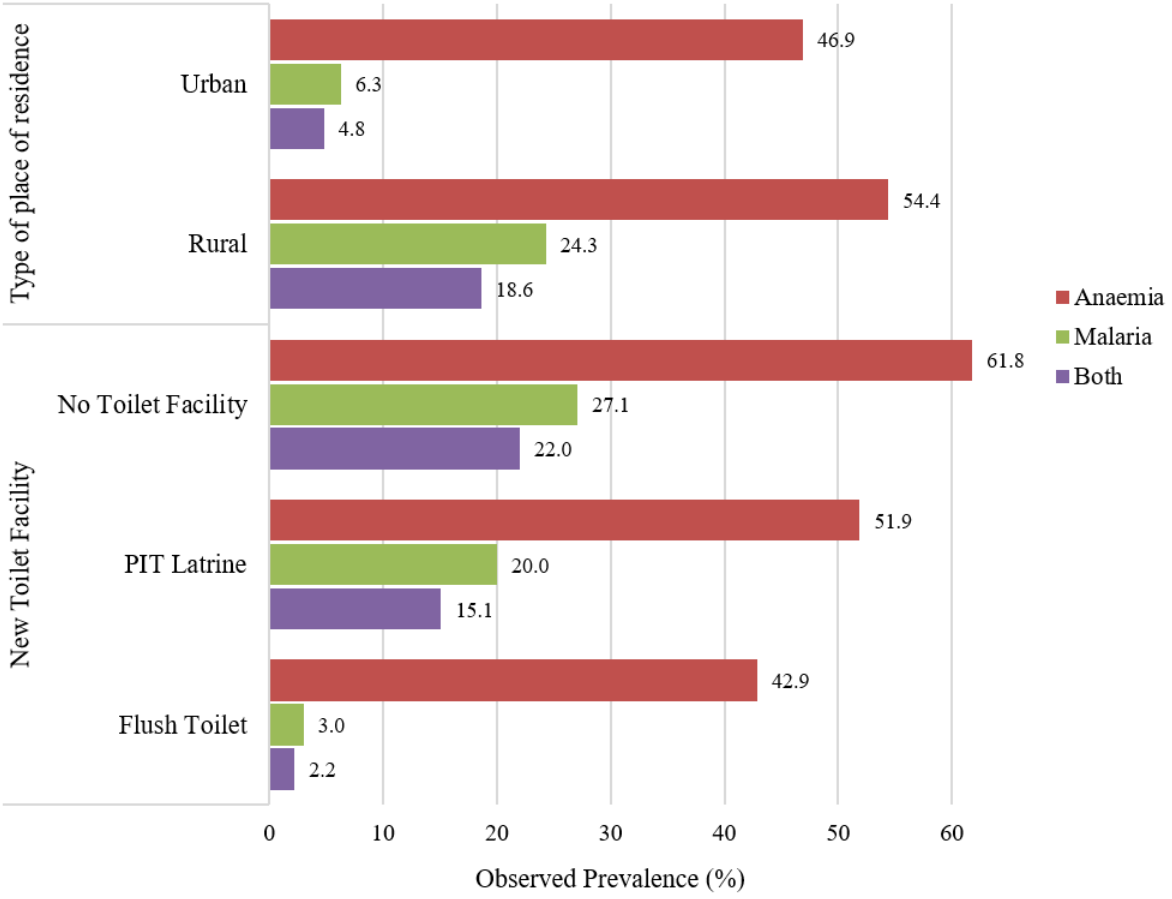


Figure 2.7: Observed prevalence of anaemia, malaria and both according to the type of place of residence and type of toilet facility

Boxplots for each of the continuous covariates are presented in Figure 2.8. These boxplots display the minimum, first quartile, median, third quartile, maximum and the mean of each covariate based on all the children in the sample, the children with anaemia, the children with malaria and the children with both anaemia and malaria. Children with anaemia had a lower age, on average, compared to those with malaria. Not much difference in the distributions of the age of the household head and the

household size was seen between the different samples of children. Children with malaria, on average, resided in clusters at a lower altitudes. On average, children with anaemia or malaria or both anaemia and malaria resided in households with a slightly lower wealth index compared to the full sample of children. The environmental factor EVI had the highest mean and median for those children with malaria. Not much difference in the mean or median of LST was evident between the samples.

2.6 Summary

This chapter introduced the data sources from the four countries and provided an overview of the sample as well as the categorical and continuous factors of interest. These factors were then explored in relation to the child's anaemia status, their malaria status and whether or not the child had both anaemia and malaria. Anaemia was considerably more prevalent in the children compared to malaria, and the majority of the children who tested positive for malaria had anaemia as well. The prevalences were notably lower among children residing in Kenya compared to the other three countries. In addition, the prevalence of anaemia was fairly heterogeneous across the districts of the four countries. The patterns of all three prevalences varied substantially across the child's mother's highest education levels as well as the type of toilet facilities. However, only the prevalence of malaria and the prevalence of both anaemia and malaria in a child were considerably higher among those residing in rural areas compared to urban areas. On average, children with anaemia had a lower age compared to those with malaria.

In the next chapter, we further explore the data and the relationships between the explanatory variables and the two child health outcomes using supervised machine learning techniques.

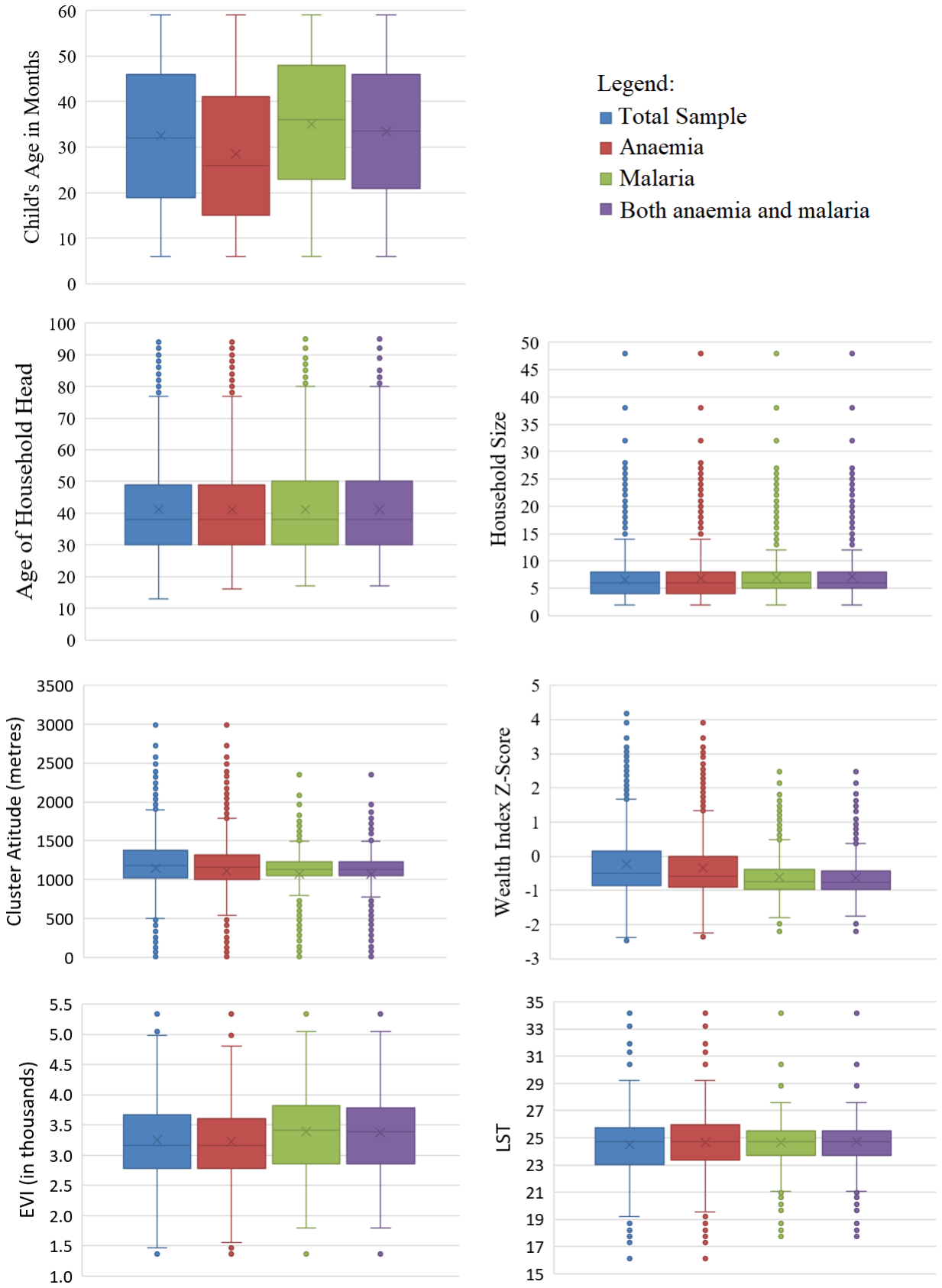


Figure 2.8: Boxplots for the continuous covariates by the outcome categories

CHAPTER 3

EXPLORATORY DATA ANALYSIS

This chapter sets out to apply a variety techniques to identify the patterns and relationships among the variables as well as to identify the most influential factors on the child health outcomes, namely anaemia and malaria. We first consider multiple correspondence analysis (MCA) to graphically explore the patterns among the categorical variables. Following MCA, classification techniques are applied as the outcomes considered are binary, consisting of two classes. Such techniques include logistic regression, classification and regression trees (CART), support vector machines (SVMs) and artificial neural networks (ANNs). These techniques are supervised machine learning methods where they model the relationships and dependencies between the target output (the response variable) and the input features/attributes (the explanatory variables) using a set of labelled training data. The trained models can be used to gain insight into which attributes contribute the most or the least in predicting each child health outcome. These models can further be used to predict the class label of the child health outcome for new, unseen data based on those relationships learned from the training data.

3.1 Multiple Correspondence Analysis

Correspondence analysis (CA) is another form of exploratory data analysis that provides a graphical representation of cross-tabular data. This statistical technique allows one to explore the structures and relationships among the categorical variables based on data given in a contingency table (Greenacre, 2017). The relationship between the categories given in the rows and columns of the table can be represented using correspondence analysis plots. In such plots, the distance between category points represent the relationships between the categories, where similar categories

are plotted closer to each other for each variable. Multiple correspondence analysis (MCA) is an extension of CA and allows the relationship of three or more categorical variables to be explored (Greenacre, 2017). This multivariate technique distributes values of a relative frequency table, known as a Burt table, in an n-dimensional space, and then establishes the similarity degree of the variables based on the distance between them in each dimension (Rodriguez-Sabate et al., 2017).

MCA is performed by applying simple CA on an indicator matrix (Greenacre, 2017). Suppose we have n observations with k categorical variables. Assume variable j has l_j distinct categories. Then, we define an $n \times l_j$ indicator matrix, $\boldsymbol{X}_j$. Concatenating the $\boldsymbol{X}_j$'s forms the $n \times l$ matrix $\boldsymbol{X}$, which is an observations-by-categories table that has as many rows as observations, and l is the sum of l_j (Greenacre, 2017). The elements of $\boldsymbol{X}$ are equal to 1 in the positions to indicate the categories of response of each observation and zero elsewhere. $\boldsymbol{X}$ can be divided up by its grand total nk to obtain a probability matrix $\boldsymbol{Z} = \frac{1}{nk}\boldsymbol{X}$, which contains the relative frequencies. This gives $\mathbf{1}_n'\boldsymbol{Z}\mathbf{1}_l = 1$, where $\mathbf{1}_i$ is an $i \times 1$ vector of ones. The vectors $\boldsymbol{r} = \boldsymbol{Z}\mathbf{1}_l$ and $\boldsymbol{c} = \boldsymbol{Z}'\mathbf{1}_n$ are called the row and column marginals, respectively. These marginals are collectively called masses for the rows and columns. Assume the diagonal matrices of the masses are defined by $\boldsymbol{D}_r = diag(\boldsymbol{r})$ for the rows and $\boldsymbol{D}_c = diag(\boldsymbol{c})$ for the columns. The factor scores are obtained from the following singular value decomposition

$$\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{Z} - \boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{Q}',$$

where $\boldsymbol{\Delta}$ represents the diagonal matrix of singular values, $\boldsymbol{P} = \boldsymbol{D}_r^{-\frac{1}{2}}$ and $\boldsymbol{Q} = \boldsymbol{D}_c^{-\frac{1}{2}}$, and $\boldsymbol{\Lambda} = \boldsymbol{\Delta}^2$ is the matrix of eigenvalues. The row and column factor scores are obtained respectively as

$$\boldsymbol{F} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{\Delta}, \tag{3.1}$$

and

$$\boldsymbol{G} = \boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Delta}. \tag{3.2}$$

Equations 3.1 and 3.2 will then lead to the calculation of the squared (χ^2) distance of the rows and columns to their respective barycentre, which is given in the following form

$$\boldsymbol{d}_r = diag\{\boldsymbol{F}\boldsymbol{F}'\},$$

and

$$\boldsymbol{d}_c = diag\{\boldsymbol{G}\boldsymbol{G}'\},$$

respectively. In CA, the total variance, referred to as inertia, of the factor scores is proportional to the independence Chi-square statistic of a cross-tabulation (Greenacre, 2017). Using this Chi-square distance between the row-points and that between the column-points, a graphical representation of the points in a reduced-dimension space can be obtained. The optimal space is based on that which maximizes the inertia, the measurement of the dispersion of the set of computed distances, between the points (Di Franco, 2016).

The contributions of the i^{th} row and the j^{th} column to factor m are obtained in the following manner, respectively

$$b_{i,m} = \frac{f_{i,m}^2}{\lambda_m},$$

and

$$b_{j,m} = \frac{g_{j.m}^2}{\lambda_m}.$$

These contributions of the rows and columns help locate the observations or variables that are of importance to a given factor. Supplementary elements can be projected onto the factors using the transition formula (Greenacre, 2017). Suppose we let $\boldsymbol{i}'_{sup}$ be the supplementary row and $\boldsymbol{j}_{sup}$ be the supplementary column to be projected, then the coordinates of supplementary functions $\boldsymbol{f}_{sup}$ and $\boldsymbol{g}_{sup}$ are given as

$$\boldsymbol{f}_{sup} = (\boldsymbol{i}'_{sup}\boldsymbol{1})\boldsymbol{i}'_{sub}\boldsymbol{G}\boldsymbol{\Delta}^{-1},$$

and

$$\boldsymbol{g}_{sup} = (\boldsymbol{j}'_{sup}\boldsymbol{1})\boldsymbol{j}'_{sub}\boldsymbol{F}\boldsymbol{\Delta}^{-1},$$

respectively.

Performing CA on the indicator matrix will provide the factor scores for the rows and columns, however MCA requires these scores to be re-scaled. The Burt matrix is the $l \times l$ table, which is obtained by $\boldsymbol{B} = \boldsymbol{X}'\boldsymbol{X}$. This matrix serves to give CA the same factors as the analysis of $\boldsymbol{X}$, however in a computationally easier way. The Burt matrix also plays an important role in providing eigenvalues which provide a better approximation of the of inertia described by the factors than the ones of $\boldsymbol{X}$ (Greenacre, 2017).

MCA uses an indicator matrix to create several binary columns for each variable by partitioning the indicator matrix, such that only one column contains the value 1. This method creates additional dimensions as one nominal variable is coded with multiple columns. The additional dimensions cause the solution space variance (inertia) to be artificially inflated, which causes the inertia described by the first dimension to be under-scaled (Greenacre, 2017). This leads to the variation produced by the first dimension to be under-estimated. However, this under-scaling can be corrected using correction formulae provided by Greenacre (2017), which allows for evaluation of the percentage of inertia respective to the average inertia of the off-diagonal blocks of Burt matrix.

3.1.1 Results of Multiple Correspondence Analysis

Here we applied MCA to visualise the associations between the categorical characteristics presented in Table 2.1 and the child's anaemia status, their malaria status and whether or not they had both anaemia and malaria. The CORRESP procedure in SAS Version 9.4 with the MCA option was used for this analysis. The inertiae were adjusted according to Greenacre (1984), which results in a more realistic percentage of the inertia explained along each axis. Figure 3.1 presents the Greenacre adjusted inertia decomposed into six components. The total inertia explained by these components was 72.03%. Note that a property of the Greenacre adjustment is that the total inertia is not 100% (Greenacre, 1984).

Greenacre Adjusted Inertia Decomposition

Principal Inertia	Adjusted Inertia	Percent	Cumulative Percent
0.25971	0.02795	59.40	59.40
0.16295	0.00340	7.23	66.63
0.14447	0.00141	2.99	69.62
0.14102	0.00113	2.41	72.03
0.11179	0.00000	0.00	72.03
0.11129	0.00000	0.00	72.03
	0.03389	72.03	

Figure 3.1: Inertia adjusted by Greenacre's correction

A multiple correspondence plot projects all categories onto a Euclidean space where the first two dimensions can be plotted to assess the association among categories. This plot is presented in Figure 3.2. The first dimension accounted for 16.7% of the variation in the data and the second dimension explains 10.48% of the variation. From Figure 3.2, it is observed that having both anaemia and malaria was mostly associated with testing positive for malaria. Furthermore, children with mothers who had no education were associated with residing in households with no toilet facilities. Similarly, those with mothers who had secondary or higher education levels were associated with residing in urban areas. Children in Uganda were most associated with having anaemia, and those in Kenya were most associated with not having anaemia.

3.2 Training and Evaluation of Classification Models

Model Training

Unlike model parameters, such as weights and biases which are learned during model training and cannot be set arbitrarily, hyperparameters are required to be set prior to training the model. Such hyperparameters include the cost complexity parameter in CART models, the cost function in SVMs and the regularisation parameter in ANNs. There is no rule or solution to finding optimum hyperparameters as

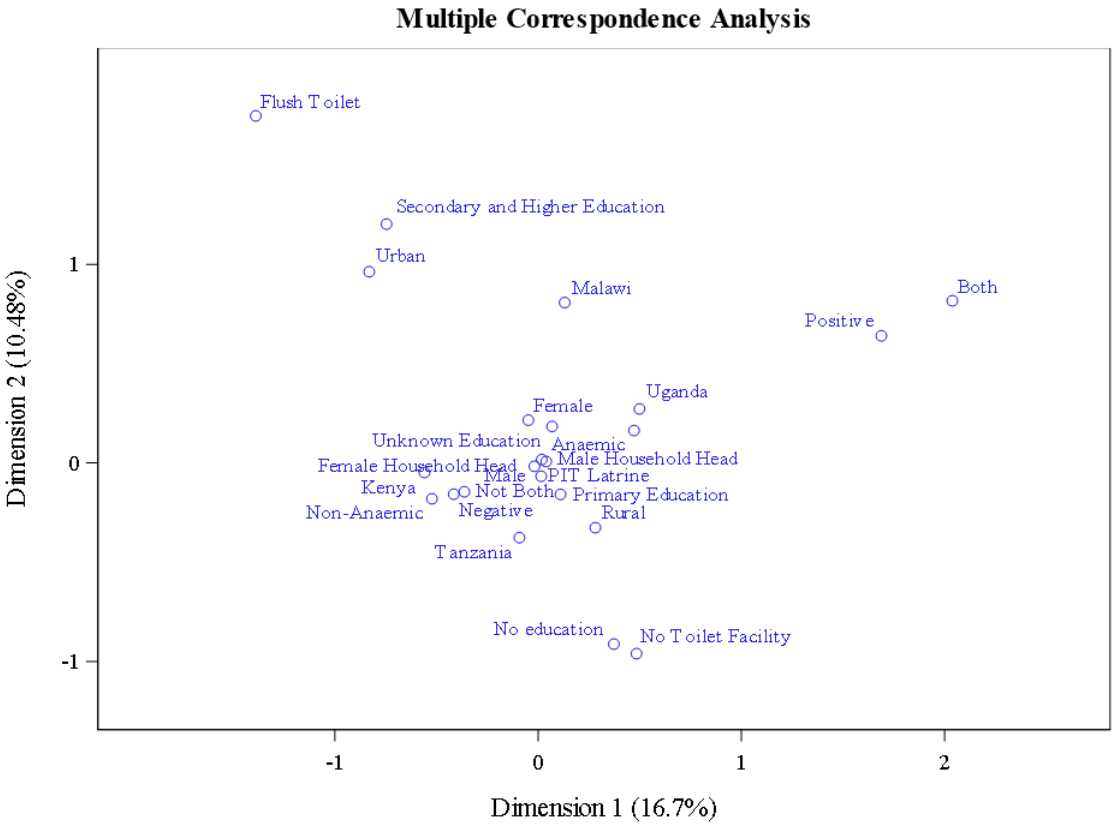


Figure 3.2: Multiple correspondence analysis for dimensions one and two

one model may perform well for one set of hyperparameters and poorly on another set. A possible way to select the ideal hyperparameter value is to train the model using different values, evaluate the model's performance on each and select the value of the hyperparameter that produces the best performing model. This process can be achieved automatically through a technique called K-fold Cross-Validation (CV). This method starts off by separating the training set into k subsets, where the model is trained using $k-1$ subsets and validated on the last one. This process is repeated k times, such that each time, one of the k subsets is used as the validation set and the other $k-1$ subsets are put together to form a training set. Each split is trained using different values of the hyperparameters, which are either based on a random search or a grid search across Cartesian products of sets of hyperparameters. Figure 3.3 illustrates the process of 5-fold CV. For our models, we use a 10-fold CV to tune the hyperparameters. The hyperparameter value that produced the highest accuracy

was used in training the final model.

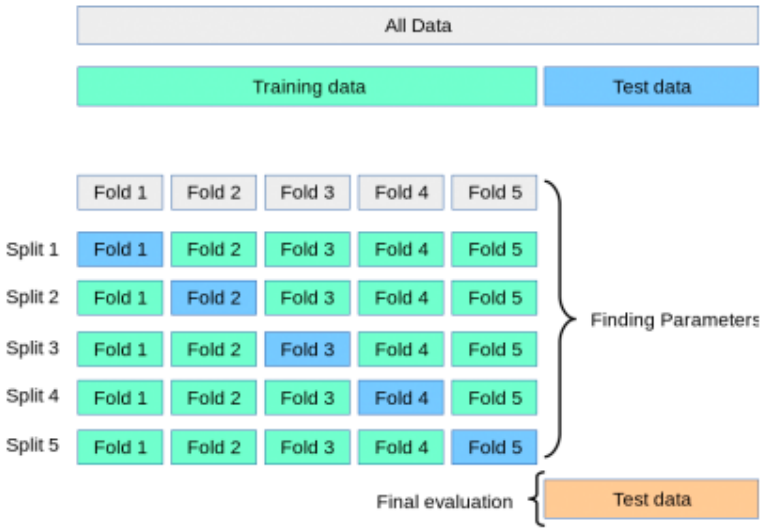


Figure 3.3: 5-fold Cross-Validation

Performance Measures

To evaluate the performance of machine learning models, it is customary to use the train-test data split technique. This involves dividing the full dataset into two subsets. The first subset is used to fit/train the model and is referred to as the training set. The second dataset is referred to as the test set. This test set is not used in the training of the model, thus it is considered as new, unseen data by the model. It is rather used to evaluate the performance of the trained model. In the application of each of the classification models to our data, a 75:25 split was considered, where 75% of the data was used for training and the remaining 25% was used for testing.

After a classification technique has been applied to the data, one can assess the goodness-of-fit of the classifier using a range of different measures. Such measures can be calculated from a confusion matrix given by

Table 3.1: Structure of a confusion matrix

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The entries in the confusion matrix are defined as follows

- True Positive (TP): the number of cases correctly identified for those that have the disease.
- False Positive (FP): the number of cases incorrectly identified for those that have the disease.
- True Negative (TN): the number of cases correctly identified as not having the disease.
- False Negative (FN): the number of cases incorrectly identified as not having the disease.

The following measures can then be calculated for the classifier:

- **Accuracy** measures the proportion of actual positives and negatives that are correctly identified as such.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Sensitivity** measures proportion of actual positives that are correctly identified as such.

$$Sensitivity = \frac{TP}{TP + FN}$$

- **Specificity** measures the proportion of actual negatives that are correctly identified as such.

$$Specificity = \frac{TN}{TN + FP}$$

- **Precision** measures the proportion of actual positives out of all those that are predicted to be positive.

$$Precision = \frac{TP}{TP + FP}.$$

Such performance measures can be obtained for both the training and test sets. A large discrepancy between the performance of the model on the training and test sets can be indicative of high variability and thus possible instability of the model. Furthermore, if a model performs well on the training set and poorly on the test set, the model is overfitted and should not be considered for making predictions on future, unseen data.

Each of the four classification models were applied to the child's anaemia status and their malaria status, separately, where the explanatory variables considered were based on those presented in Figure 2.2. These same models were fitted again, however the child's malaria status was included as an explanatory variable for anaemia, and the child's anaemia status was included as an explanatory variable for malaria. The accuracy of the models for anaemia improved by approximately 3%, and that of the models for malaria improved by approximately 1%. This highlights the relationship between the two diseases and how they may be important predictors of each other. The results for each technique considered in the subsequent sections are based on those models with anaemia and malaria included as predictors of each other.

The R *Caret*, *e1071* and *rminer* packages were used to fit the logistic regression, CART, SVM and ANN models to the data. The following sections provide an overview and the results of each technique.

3.3 Logistic Regression

The logistic regression model is a special case of a generalised linear model and is widely used to a model binary response. Suppose the response is defined as

$$Y_i = \begin{cases} 1 & \text{if the event has occurred, e.g. the child has anaemia or malaria,} \\ 0 & \text{if the event has not occurred, e.g. the child does not have anaemia or malaria.} \end{cases}$$

Thus, Y_i follows a Bernoulli distribution where $P(Y_i = 1) = \pi_i$ is the probability that the event occurs and $P(Y_i = 0) = 1 - \pi_i$ is the probability that the event does not occur. It therefore follows that the mean and variance of the response is respectively given by

$$E(Y_i) = \pi_i,$$
$$Var(Y_i) = \pi_i(1 - \pi_i).$$

As π_i is a probability, it is limited by $0 \leq \pi_i \leq 1$. Thus, a model for $E(Y_i)$ that restricts its values in this domain is required. Such a model is the logistic regression model, given by

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where the left hand side is referred to as the logit link function and represents the log of the odds of the event of interest occurring; $\boldsymbol{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of the unknown regression coefficients. The logit can take on any value from $-\infty$ to ∞ while restricting π_i between 0 and 1. The predictors, $\boldsymbol{x}_i$, have no restrictions and may consist of qualitative and quantitative variables.

The predicted probability, $\widehat{\pi}_i$, of the event occurring can be computed from the fitted model as follows

$$\widehat{\pi}_i = \frac{\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}.$$

This predicted probability can then be used to classify the class label of the response

for a given set of predictors, $\boldsymbol{x}_i$, based on some cut-off value, c. Thus

$$\widehat{Y}_i = \begin{cases} 1 & \text{if } \widehat{\pi}_i > c, \\ 0 & \text{otherwise.} \end{cases}$$

It is common to use a cut-off of $c = 0.5$. The parameters of the logistic regression model can be estimated using maximum likelihood estimation, a probabilistic framework. Thus, in addition to classification predictive modelling, the logistic regression model can also be used for statistical inference concerning the significance of the effect of the predictors on the likelihood of the event of interest occurring.

3.3.1 Results of Logistic Regression Models

Table 3.2 presents the results of the fitted logistic regression model for each response. This model produced an accuracy of 67.8% and 82.6% on the training sets for anaemia and malaria, respectively. A cut-off of 50% was used where those with a predicted probability greater than 0.5 were classified as having the disease, otherwise they were classified as not having the disease. This model revealed that anaemia and malaria were significant predictors of each other. The child's gender was not significantly associated with the likelihood of malaria, however it was significant factor for anaemia. In fact, the majority of the attributes considered were significant predictors of anaemia, whereas malaria had fewer significant predictors. The age and gender of the head of household did not have a significant association with either response.

While logistic regression can be used to assess the significance of the association between an attribute and the response, the parameter estimates cannot be compared and ranked between the different factors. Rather, the parameter estimates can only be used to make a comparison with respect to the reference category for a qualitative predictor or for one unit increase in a quantitative predictor. Thus, logistic regression cannot be used to determine the most or least influential predictors. For this purpose, we consider CART, SVM and ANNs in the subsequent sections.

Table 3.2: Parameter estimates, standard errors and p-values for the logistic regression models

Variable	Anaemia			Malaria		
	Estimate	St. Error	P-value	Estimate	St. Error	P-value
Anaemia status (ref = Negative)						
Positive		NA		1.495	0.056	<.001*
Malaria status (ref = Negative)						
Positive	1.503	0.056	<.001*		NA	
Gender (ref = Male)						
Female	−0.135	0.038	<.001*	−0.009	0.049	0.849
Type of place of residence (ref = Urban)						
Rural	−0.163	0.055	0.003*	0.686	0.087	<.001*
Mother's education level (ref = No Education)						
Primary	−0.198	0.058	0.001*	−0.025	0.068	0.707
Secondary and Higher	−0.239	0.076	0.002*	−0.202	0.105	0.055
Unknown	−0.171	0.075	0.022*	0.211	0.090	0.019*
Household head gender (ref = Male)						
Female	0.053	0.045	0.238	−0.077	0.059	0.189
Type of toilet facility (ref = No Facilities)						
PIT Latrine	−0.328	0.064	<.001*	−0.111	0.073	0.127
Flush Toilet	−0.475	0.111	<.001*	0.104	0.220	0.636
Country of residence (ref = Kenya)						
Malawi	0.785	0.076	<.001*	0.210	0.106	<.001*
Tanzania	0.704	0.056	<.001*	0.229	0.090	0.011*
Uganda	0.321	0.066	<.001*	1.521	0.094	<.001*
Child's age in months	−0.043	0.001	<.001*	0.026	0.001	<.001*
Age of the household head	−0.001	0.002	0.558	−0.002	0.002	0.267
Household size	0.020	0.006	0.001*	0.028	0.007	<.001*
Wealth index Z-score	−0.092	0.033	0.005*	−0.770	0.009	<.001*
Cluster altitude (in 100 metres)	−0.001	0.006	0.994	−0.046	0.009	<.001*
EVI	0.109	0.051	0.031*	0.335	0.067	0.001*
LST	0.048	0.016	0.003*	0.019	0.021	0.387

*significant at 5% level of significance

3.4 Classification and Regression Trees

Classification and regression trees (CART) were first introduced by Breiman et al. (1984). CART is a decision-tree procedure that is used to classify cases and make predictions, where it can be thought of as a flow chart of a sequence of questions and answers (Ma, 2018). A classification tree is produced if the response is categorical and a regression tree is produced if the response is quantitative. During the training procedure of such trees, the data is recursively split into mutually exclusive subgroups based on a set of 'yes/no' answers to questions pertaining to the state of the explanatory variables. CART is an effective exploratory procedure that can capture complex interactions and non-linear relationships in the data, which traditional statistical techniques cannot easily deal with (Ma, 2018). CART does not rely on any statistical models and does not contain any complex mathematical equations, thus it is easy to interpret and understand. In addition, the tree structure automatically indicates the most important explanatory variables and how they interact with one another to split the sample into the different subgroups pertaining to the different classes of the response variable.

Each box or group in a CART is called a node. The node on the top of a tree is called the *root node* as the analysis starts from this node and descends until it reaches the *terminal nodes* after the tree has concluded growing. To construct a tree, the CART algorithm starts at the root node where it partitions all of the observations into two mutually exclusive groups according to the best value of any explanatory variable. These groups form two *child nodes*, whereas the producing node is called the *parent node*. At each node, the observations are split such that those in the same group are as similar as possible and those in different groups are as different as possible. The split is based on determining the best explanatory variable to split as well as the best split point for that variable. This splitting criterion can either be based on a statistical test or based on maximising the decrease in node impurity. We will consider the splitting based on the node impurity.

The impurity of a parent node, $i(\tau)$, is defined as a non-negative number that is equal to zero for a pure node, which is a node that consists of observations that have the same value in the response variable. The impurity becomes large if an equal number of cases belong to different categories of the response variable (Ma, 2018). The objective when using an impurity measure as a splitting criterion is to produce the highest reduction in impurity, given by

$$\Delta i(s, \tau) = i(\tau) - \sum_{b=1}^{B} p(\tau_b|\tau) i(\tau_b),$$

where τ_b denotes the b^{th} child node, $p(\tau_b|\tau)$ is the proportion of observations in τ that are assigned to τ_b, and B is the number of branches after splitting τ. There are different impurity reduction criteria, such as entropy, Gini index and residual sum of squares. We will consider the Gini index criterion given by

$$i(\tau) = 1 - \sum_{j=1}^{J} p_j.$$

Here $i(\tau)$ is defined as the Gini index that corresponds to the average square error (ASE) of a class response. After all explanatory variables are considered, the variable with the largest reduction in impurity is selected to partition the root node into the two child nodes. The same procedure is then applied for partitioning each child node into two child nodes. Thus, the CART tree keeps growing new branches, each based on the reduction in a certain impurity measure (Ma, 2018).

A problem associated with an impurity measure is that it becomes smaller as the tree grows larger. In fact, any tree can have a zero impurity if the tree keeps growing a large number of terminal nodes with a single case in each terminal group. In this case, the number of terminal nodes in the tree will be equal to the number of observations in the sample. This is because increasing the size of the tree is monotonically related to decreasing the degree of the impurity at the terminal nodes (Ma,

2018). Thus, when using an impurity measure to grow a tree, the challenge is to do so while preventing the tree from growing too large. A solution to this is to obtain a smaller subtree from the full tree, where such a subtree produces a low error rate. However, the subtree must not be so small that it fails to capture important structural information. An optimal subtree is achieved by applying an approach called the cost-complexity pruning method (Breiman et al., 1984). Pruning the tree also aids in preventing overfitting, where the tree becomes too specific to the data that it has grown from and thus cannot generalize well to new data.

CART trees can also be used to determine the importance of an explanatory variable on the response. This measure is obtained by calculating the relative influence of each variable which is based on whether the variable was selected to split during the tree building process as well as how much the error was improved/decreased as a result.

3.4.1 Results of CART Models

To obtain the classification tree for each response, the Gini index criterion was used as the impurity function. At each non-terminal node, if the answer to the question used to split the node was affirmative, the case was assigned to the child node on the left, otherwise it was assigned to the child node on the right. As it is not known whether a node will become terminal during pruning, each node is assigned a class label according to the predominant class in that node. During pruning, a node becomes terminal if the change in the Gini index during the split of that node is less than the cost complexity (cp) parameter. The optimal value of cp was 0.00185 and 0.00296 for the anaemia and malaria classification trees, which produced an accuracy of 67.9% and 83.8% on the training sets for each model, respectively.

The optimal classification tree for anaemia is presented in Figure 3.4 and that for malaria is presented in Figure 3.5. The root node for the anaemia classification tree

was based on the child's age. If a child is at least 24 months old, they are assigned to a child node with a non-anaemic class label. The nodes in the second layer of the tree were based on the child's malaria status (Figure 3.4). However, considering Figure 3.5, the root node for the malaria classification tree was based on the child's anaemia status, where if the child does not have anaemia, they are immediately classified as being negative for malaria (a terminal node). However, if the child does have anaemia, they are further split based on the country of residence, specifically whether or not the they reside in Uganda (the attribute considered in the second layer of the tree).

Both classification trees are primarily split based on the environmental covariates (cluster altitude, EVI and LST) as well as the household's wealth index Z-score and the child's age. One of the properties of CART models is that it can reuse variables for splitting of the nodes, which aids in describing complex relationships among the variables. This is seen in the results of our classification trees for both responses, where many of the variables are reused.

Table 3.3 provides the variable importance measures for each response. Attributes not listed in the table resulted in a variable importance of 0. However, if the trees were allowed to grow larger, more predictor variables would have a chance to play a role in the tree construction process and thus would have non-zero importance measures. The child's age and a positive malaria status were of very high importance in growing the classification tree for anaemia. The variable importance measures of these two attributes by far exceeded any of the variable importance measures for the malaria classification tree. However, the environmental attributes contributed substantially more to the malaria classification tree compared to that of the anaemia classification tree. These results also revealed common important predictors of each response. Furthermore, the classification tree for anaemia revealed far fewer important predictors out of those considered compared to the classification tree for malaria.

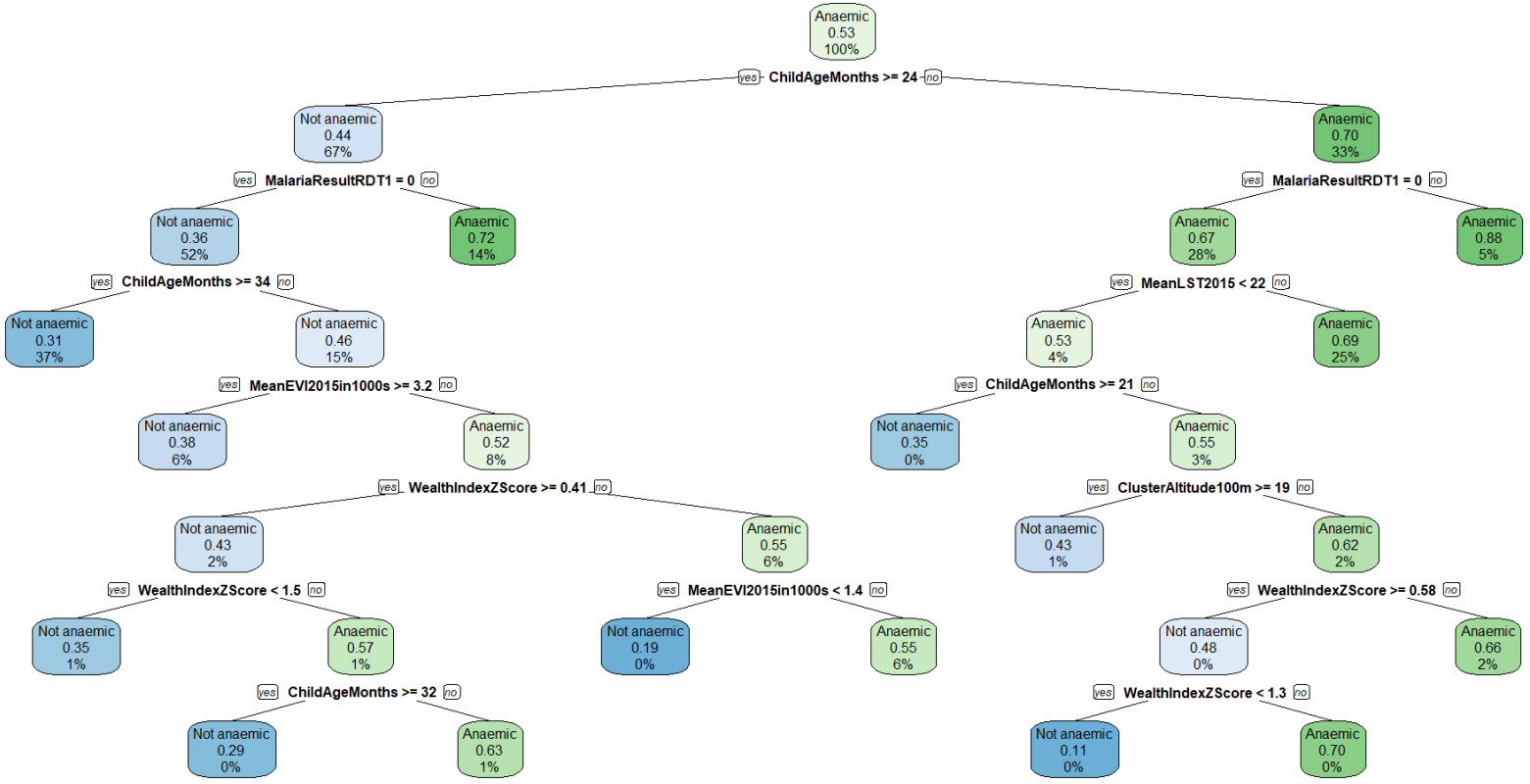


Figure 3.4: Optimal classification tree for anaemia

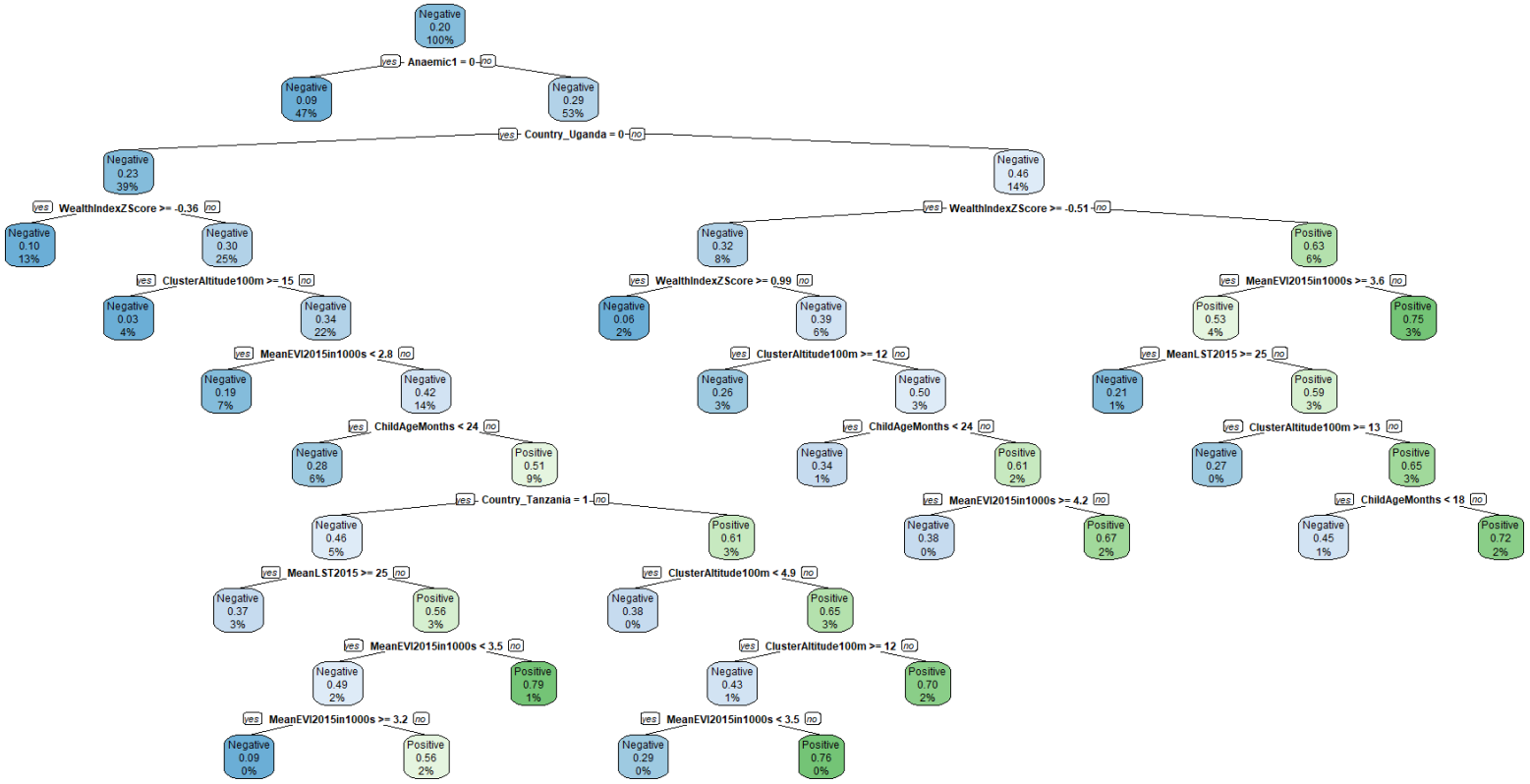


Figure 3.5: Optimal classification tree for anaemia

Table 3.3: Variable importance (VP) based on classification trees

Attribute	VP
Anaemia	
Child's age in months	512.16
Malaria status (Positive)	465.72
LST	36.83
EVI	25.65
Wealth index Z-score	22.46
Cluster altitude	16.47
Country of residence (Uganda)	7.85
Country of residence (Tanzania)	5.70
Type of toilet facility (None)	2.73
Type of place of residence (Rural)	2.73
Mother's highest education level (Secondary or higher)	2.11
Age of the household head	0.89
Malaria	
Anaemia status (Positive)	262.00
Wealth index Z-score	258.73
EVI	250.32
Cluster altitude	197.02
LST	174.29
Country of residence (Uganda)	156.86
Child's age in months	133.63
Type of place of residence (Rural)	48.04
Mother's Highest Education Level (Secondary or higher)	32.43
Type of toilet facility (None)	25.18
Type of toilet facility (PIT Latrine)	21.06
Country of residence (Tanzania)	12.50
Country of residence (Malawi)	10.56
Mother's highest education level (Unknown)	8.26
Age of the household head	3.60
Household size	3.03

3.5 Support Vector Machines

A support vector machine is a machine learning model that is used to perform classification by constructing a set of hyperplanes that maximizes the margin between two classes. SVMs have multiple advantages in that they can handle high dimensional data as well as data that are not linearly separable (Kantardzic, 2020). SVMs use a nonlinear mapping to transform the original training data into a higher dimension, then within this new dimension, it searches for the linear optimal separating hyperplane, a decision boundary separating the values of one class from another. This optimal separating hyperplane is called the maximum marginal hyperplane as its associated margin gives the largest separation between the two classes.

Consider a set of training examples consisting of n pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ is the input data with $y_i \in \{+1, -1\}$ as the corresponding class label. A separating hyperplane is defined by

$$\boldsymbol{w}'\boldsymbol{x}_i + b = 0, \tag{3.3}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)$ is a weight vector; p is the number of attributes and b is a scalar referred to as a bias. This bias can be considered as an additional weight given by w_0. Any set of attribute values $\boldsymbol{x}_i$ located along the separating hyperplane should satisfy Equation 3.3. The goal of SVMs is then to find a set of weights $\boldsymbol{w}$ and the bias b that specify two hyperplanes, given by Equation 3.4, such that the distance, known as the margin, between the two hyperplanes is maximised

$$\boldsymbol{w}'\boldsymbol{x}_i + b = \begin{cases} \geq 1 \quad \text{for} \quad y_i = +1, \\ \leq -1 \quad \text{for} \quad y_i = -1. \end{cases} \tag{3.4}$$

The conditions in Equation 3.4 impose the requirements that all training data points from class $y_i = +1$ fall above the first hyperplane and all the points of the class $y_i = -1$ fall below the second hyperplane, as long as the data are linearly separable.

Both inequalities in Equation 3.4 can be summarized in a more compact form as follows

$$y_i \times (\boldsymbol{w}'\boldsymbol{x}_i + b) \geq 1, \quad \forall \boldsymbol{x}_i,\ i = 1, \ldots, n. \tag{3.5}$$

Vector geometry defines the distance between the two planes as $\frac{2}{||\boldsymbol{w}||}$. Therefore, to maximise the margin between the two planes, the following optimisation problem needs to be solved

$$\min_{\boldsymbol{w},b} ||\boldsymbol{w}||^2,$$

subject to Equation 3.5 being satisfied. The training points that satisfy Equation 3.5 are called support vectors. Support vectors are the only data points required by the SVM to be trained. No data points are allowed in the margin between the two hyperplanes. This type of linear classification is known as hard margin classification (MLMath.io, 2021). However, this strict requirement can lead to a very narrow margin where the classifier will be sensitive to noisy data points, resulting in poor generalisation for new data. A solution around this is to allow for a more flexible classifier, where some data points are either allowed within the margin area or on the incorrect side of the decision boundary, which is a contrast to a hard margin classifier. This type of classification is referred to as soft margin classification where the constraints of Equation 3.5 are relaxed slightly by introducing a positive-valued slack variable, ξ_i, as follows

$$y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \forall \boldsymbol{x}_i,\ i = 1, \ldots, n, \tag{3.6}$$

which is subject to $\xi_i \geq 0$ for $i = 1, \ldots, l$ and $\sum_{i=1}^{l} \xi_i = C$, where C is a hyper-parameter that controls the trade-off between the width of the margin and the number of training data points misclassified. The optimal value of C is obtained using techniques such as cross-validation. When $C = 0$, no data points are allowed to be misclassified, which results in a hard margin classifier. For the soft margin classifier, the new optimisation problem that needs to be solved is given by

$$\min_{\boldsymbol{w},b} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{l} \xi_i, \tag{3.7}$$

subject to Equation 3.6 being satisfied. The estimated weights, $\widehat{\boldsymbol{w}}$, represent the relative importance of each corresponding attribute/covariate on the response.

In the case that the original data is not linearly separable, a kernel trick is applied where some non-linear transformation function, $\phi(\boldsymbol{x}_i)$, is used to map the original space to a higher dimensional feature space such that the data is separable. A kernel $K(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is a real valued function $K : G \times G \to \mathbb{R}$ for which there exists a function $\phi : X \to Z$, where Z is a real vector space, with the property $K(\boldsymbol{X}_1, \boldsymbol{X}_2) = \phi(\boldsymbol{X}_1)'\phi(\boldsymbol{X}_2)$. The kernel $K(\boldsymbol{X}_1, \boldsymbol{X}_2)$ acts as a dot product in the space of Z. G is referred to as the input space and Z is referred to as the feature space.

The optimal hyperplane using a kernel transformation is given by

$$\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b = 0.$$

This new form of the hyperplane can be obtained by solving the optimisation problem given in Equation 3.7, however, subject to $y_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1, \ldots, n$. Common kernel functions are

$$\textit{d^{th}-Degree polynomial: } K(\boldsymbol{X}_1, \boldsymbol{X}_2) = (1 + \langle \boldsymbol{X}_1', \boldsymbol{X}_2 \rangle)^d,$$
$$\textit{Radial basis: } K(\boldsymbol{X}_1, \boldsymbol{X}_2) = \exp(-\gamma ||\boldsymbol{X}_1 - \boldsymbol{X}_2||^2),$$

where γ is a parameter that sets the spread of the kernel and the polynomial kernel is valid for all positive integers $d \geq 1$ (Kantardzic, 2020).

After the SVM has been trained, the decision function to determine the predicted class label for a new set of attributes, $\boldsymbol{z}_i$, from unseen data is given by

$$\widehat{y}_i = \text{sgn}(\boldsymbol{w}'\phi(\boldsymbol{z}_i) + b).$$

3.5.1 Results of SVM Models

Some classification techniques are able to handle categorical explanatory variables, however some can only be applied to continuous numeric data. Among the four classification techniques applied here, only logistic regression and CART models can handle the categorical explanatory variables. Thus, in order to apply SVMs and ANNs, the data was first preprocessed where dummy variables were introduced to represent the different outcomes of the categorical variables. In addition, SVMs and ANNs also require numeric attributes to be normalised or standardised so that the data is on a common scale, without distorting the differences in the ranges of values. This is due to these techniques giving more emphasis to attributes with larger values. Thus, scaling all of the numeric attributes allows them to contribute equally to the analysis.

Two types of SVM models were considered for each response, one with a linear kernel and one with a radial basis function (RBF) kernel. The accuracy of the RBF SVM for anaemia was over 30% higher than that of the linear SVM for the training set. The linear SVM for malaria produced a sensitivity of 0 on the test set. This complete lack of ability of the trained model to predict positive malaria cases suggests that the model is highly inappropriate for classifying the child's malaria status. However, the RBF SVM for malaria performed much better, suggesting that it provides more sufficient fit for classifying a child's malaria status. In fact, the RBF kernal for SVM models are known for their good general performance (Meyer, 2021). Thus, only the results of the RBF SVM for both responses are presented and discussed.

In addition to the cost function (C) that is required to train a SVM model, the RBF SVM requires a value for γ, which decides how much curvature we want in a decision boundary. Generally, γ can range from 0.001 to 100, where a high value means more curvature. These two hyperparameters were tuned prior to training the models. For both responses, the optimal cost function was 4 and the optimal value of

γ was 0.5, which produced an accuracy of 93.9% and 96.5% on the training set for anaemia and malaria, respectively. Table 3.4 presents the relative variable importance for the RBF SVMs based on the estimated weights in the model. The higher the absolute value of these variable importance measures, the more important the predictor is. The attributes in Table 3.4 are listed in order of their importance. Similar to the classification tree, the child's age and malaria status were considered the most important predictors for anaemia. However, unlike the classification tree, EVI was among the least important variables for anaemia. For malaria, the child's anaemia status was one again among the top most influential variables, as well as the household's wealth index and the Uganda country of residence. The child's gender was the least influential factor for malaria.

The sign of the variable importance measures determines whether the attribute contributes to a positive or negative predicted class label. The results of the SVM models for both responses are generally in agreement with the exploratory data analysis. As one would expect, a positive malaria status contributed to a positive predicted anaemia status, and visa versa. Kenya was the only country of residence to contribute to a negative predicted anaemia status, which is unsurprising as it had the lowest observed anaemia prevalence. Uganda had the highest contribution to a positive predicted malaria status, which is also unsurprising as it had the highest observed prevalence on malaria.

Table 3.4: Variable importance (VP) based on RBF SVM models

Attribute	VP
Anaemia	
Child's age in months	-1505.76
Malaria status (Positive)	1066.13
Malaria status (Negative)	-1066.13
Wealth index Z-score	-730.06
Country of residence (Kenya)	-728.40
Type of toilet facility (Flush)	-432.72
Type of toilet facility (None)	432.00
Country of residence (Tanzania)	410.73
Household size	399.76
Cluster altitude	-398.86
Mother's highest education level (None)	381.53
Type of place of residence (Rural)	374.23
Type of place of residence (Urban)	-374.23
LST	355.27
Mother's highest education level (Unknown)	-301.26
Mother's Highest Education Level (Secondary or higher)	-293.05
Country of residence (Malawi)	220.39
Mother's Highest Education Level (Primary)	149.70
Gender (Female)	-93.31
Gender (Male)	93.31
Type of toilet facility (PIT Latrine)	-87.81
Age of the household head	80.57
EVI	-61.15
Household head gender (Male)	-21.94
Household head gender (Female)	21.94
Country of residence (Uganda)	19.04

Continued on the next page

Table 3.4 – *Continued from the previous page*

Attribute	VP
Malaria	
Wealth index Z-score	-900.59
Country of residence (Uganda)	680.99
Anaemia status (Negative)	-632.42
Anaemia status (Positive)	632.42
Type of place of residence (Rural)	609.10
Type of place of residence (Urban)	-609.10
Type of toilet facility (Flush)	-498.76
Country of residence (Kenya)	-440.42
Type of toilet facility (None)	438.38
EVI	423.75
Child's age in months	395.98
Country of residence (Tanzania)	-385.64
Mother's Highest Education Level (Secondary or higher)	-340.32
Cluster altitude	-290.11
Mother's highest education level (None)	268.63
Household size	214.16
Country of residence (Malawi)	198.44
Mother's highest education level (Unknown)	126.98
LST	111.64
Age of the household head	103.55
Household head gender (Male)	-65.70
Household head gender (Female)	65.70
Type of toilet facility (PIT Latrine)	-56.13
Mother's Highest Education Level (Primary)	-27.05
Gender (Female)	-2.39
Gender (Male)	2.39

3.6 Artificial Neural Networks

Artificial neural networks (ANNs) are multi-layered models where each layer consists of nodes called neurons. The capacity of an ANN to learn is rooted in its architecture. Although there are many forms of network architecture, three key characteristics differentiate them: the number of layers; whether information in the network is allowed to travel backward; and the number of nodes within each layer of the network. An ANN consists of at least two layers: an input layer and an output layer. The input layer contains nodes equal to the number of input features/attributes. The output layer contains nodes equal to the number of classes in the response. However, in the case of a binary response, one output node is sufficient.

To allow for more complex relationships between the attributes and response, hidden layers can be added in between the input and output layers. These hidden layers process the signals from the input nodes prior to reaching the output nodes. The number of nodes in the hidden layer is generally specified by the user prior to training the model. Most multilayer networks are fully connected, which means that every node in one layer is connected to every node in the next layer, however, this is not a requirement (Hastie et al., 2008). All nodes in an ANN are connected via weighted connections, w_j. These connection weights reflect the patterns observed over time. Training a ANN amounts to adapting the weights of the connections between the nodes until they fit the input-output relationships of the underlying data.

The inputs of an ANN are weighted according to their importance and then summed. A bias is then applied to this weighted sum, which is then passed through an activation function given by f. The output of this activation function becomes the input signal of the nodes in the next layer. This process is continued until the output node(s), the output signal of which represents the predicted response. A simple artificial neuron with p input attributes can be represented by the following formula

$$\widehat{y} = f\left(\sum_{i=1}^{p} w_i x_i + b\right).$$

The bias b can be considered as an additional weight, where $w_0 = b$ and $x_0 = 1$. Figure 3.6 presents common activation functions for ANNs. Such functions must be differentiable as this is a requirement in training the model. The goal of the ANN learning algorithm is to determine a set of weights $\boldsymbol{w}$ that minimises the total sum of squared errors

$$SSE = E(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2. \tag{3.8}$$

Equation 3.8 is referred to as a loss function. If the activation function f is linear, then the global minimum solution can be easily obtained. However, if the activation function f is some non-linear function, then the solution is not as straightforward. Thus, optimisation techniques are required. Specifically, the gradient descent method is used to iteratively update each weight as follows

$$w_j^{k+1} = w_j^k - \lambda\frac{\partial E(\boldsymbol{w})}{\partial w_j},$$

where λ is a hyperparameter, referred to as a regularisation parameter that controls the learning rate of the model and $\dfrac{\partial E(\boldsymbol{w})}{\partial w_j}$ is the gradient, which measures the change in the total error due to the change in the weight w_j. The backpropagation algorithm is used to compute this gradient, where it uses the derivative of each node's activation function to identify the gradient in the direction of each of the incoming weights. This backpropagation algorithm starts with forward propagation based on initial random starting values for the weights (and biases) as well as the values of the input attributes to obtain the predicted response. Forward propagation works from the input node through to the output node to obtain the predicted response. To calculate the updated weights based on the gradient descent method, backpropagation then works in the reverse direction, computing all the partial derivatives starting from the output node. The algorithm then iterates through many cycles of these two phases (forward and backward) and continues until a stopping criteria is met. The last iteration then determines the estimated weights and biases of the final neural network model, which can then be used in predicting/classifying future unseen classes (Hastie et al., 2008).

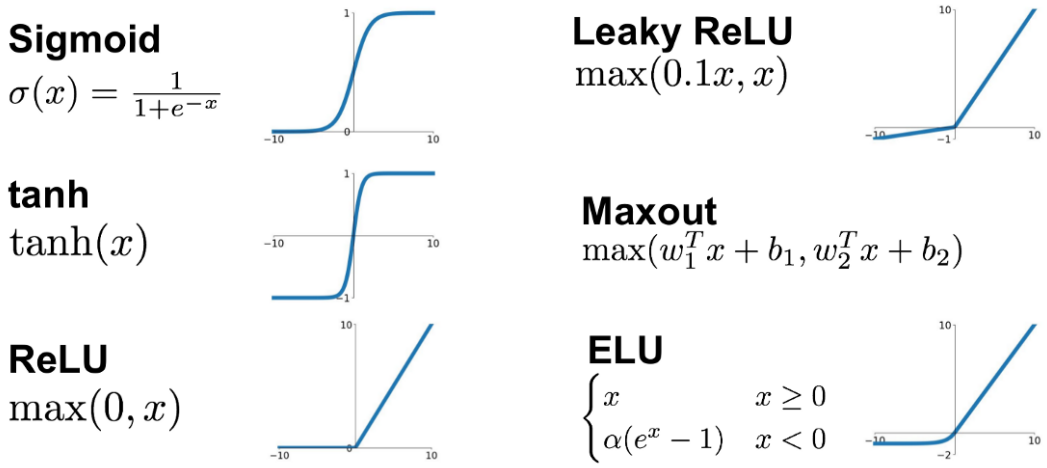


Figure 3.6: Common activation functions

3.6.1 Results of ANN Models

A single hidden layer ANN was fitted to each response. As the responses were binary, the sigmoid activation function was used. In training these models, two hyperparameters were required to be tuned; the number of units in the hidden layer and the value of the decay, which is a regularization parameter to avoid over-fitting. The optimal ANN model for anaemia was based on 3 nodes in the hidden layer and a decay of 0.1. This produced an accuracy of 68.9% on the training set. The optimal ANN model for malaria was based on 7 nodes in the hidden layer and a decay of 0.1, which produced an accuracy of 83.8% on the training set.

Table 3.5 displays the variable importance measures for the ANN models for each response. These measures are based on Gevrey et al. (2003), which uses combinations of the absolute values of the estimated weights. Once again, the child's age and malaria status ranked as the most influential predictors for anaemia. The child's anaemia status was of lower importance for malaria compared to the classification tree and SVM model. The household size and age of the head of household were the least influential attributes for both responses. Cluster altitude and EVI played larger roles in malaria compared to anaemia.

Table 3.5: Variable importance (VP) based on ANN models

Attribute	VP
Anaemia	
Child's age in months	11.49
Malaria status (Positive)	8.15
Malaria status (Negative)	7.97
Country of residence (Malawi)	7.66
Country of residence (Kenya)	7.57
Type of toilet facility (None)	6.16
LST	5.20
Mother's Highest Education Level (Secondary or higher)	4.92
Country of residence (Tanzania)	4.29
Gender (Female)	3.22
Type of toilet facility (Flush)	2.98
Wealth index Z-score	2.89
Type of place of residence (Rural)	2.62
Household head gender (Female)	2.56
Type of place of residence (Rural)	2.56
Type of toilet facility (PIT Latrine)	2.53
Household head gender (Male)	2.43
EVI	2.40
Country of residence (Uganda)	2.27
Mother's highest education level (None)	2.00
Cluster altitude	1.90
Gender (Male)	1.89
Mother's highest education level (Unknown)	1.66
Mother's Highest Education Level (Primary)	1.27
Household size	1.05
Age of the household head	0.36

Continued on the next page

Table 3.5 – *Continued from the previous page*

Attribute	VP
Malaria	
Country of residence (Uganda)	6.82
Country of residence (Kenya)	6.81
Cluster altitude	6.21
Anaemia status (Negative)	6.05
EVI	5.99
Country of residence (Malawi)	5.14
Anaemia status (Positive)	5.05
Type of place of residence (Rural)	4.64
LST	4.24
Country of residence (Tanzania)	4.11
Mother's Highest Education Level (Primary)	4.08
Type of toilet facility (Flush)	3.99
Child's age in months	3.79
Type of toilet facility (PIT Latrine)	3.74
Wealth index Z-score	3.68
Household head gender (Female)	3.29
Mother's Highest Education Level (Secondary or higher)	3.02
Type of toilet facility (None)	2.88
Gender (Female)	2.75
Gender (Male)	2.52
Type of place of residence (Rural)	2.52
Mother's highest education level (None)	2.37
Household head gender (Male)	2.21
Mother's highest education level (Unknown)	2.08
Age of the household head	1.20
Household size	0.83

3.7 Summary and Discussion

This chapter provided further insight and understanding into the relationships among the explanatory variables and the child's anaemia and malaria statuses. As our data consists of multiple categorical variables, MCA was used to explore the patterns and associations between them. In addition, four classification techniques were considered to explore the possible predictors of each response. Logistic regression is a popular technique for modelling a binary response. It allows for statistical inference concerning the association between the predictors and the response as well as classification. Further to the logistic regression model for classifying the outcome of the disease, we also considered classification trees, SVMs and ANNs, where such techniques allowed for the relative importance of the predictors to be obtained.

The results of the classification models revealed that the responses were important predictors of each other. However, the child's malaria status had a higher contribution to anaemia compared to the contribution of the anaemia status on malaria. In addition, the responses shared common important predictors in some of the models, such as the child's age, the household's wealth index Z-score and the environmental factors; cluster altitude, EVI and LST. Although the logistic regression model revealed that the cluster altitude had an insignificant effect on the child's anaemia status, the classification tree and SVM model ranked this predictor among the most important factors for anaemia. The country of residence ranked higher up in importance in the SVM and ANN models compared to the classification tree. While all three models consistently revealed the child's age and malaria status as the top most important predictors of anaemia, the results for malaria were less consistent, where the top most important variables differed among the three models. The child's gender was not significantly associated with their malaria status, in addition, this factor was among the least important predictors for malaria. The age of the head of household did not have a significant effect of the child's anaemia status or malaria status. This was also confirmed by the SVM and ANN models where this factor was among

the least important. While the gender of the household head also did not have a significant effect on either response, the ANN model ranked this factor slightly higher in importance for anaemia compared to the SVM model and classification tree.

Each of the trained models were applied to the test sets. The performance measures for each model on the test set are presented in Table 3.6. While the accuracy of the models for malaria were notably higher than those for anaemia, the sensitivity of the models were poor. This may be as a result of the lower observed prevalence of malaria in the sample, as the sensitivity of a predictive model is known to be affected when the proportion of events to non-events is less than 50% (Blagus & Goeman, 2016). A low sensitivity means that many children would be incorrectly classified as not having malaria, the effects of which could be detrimental.

All of the models for each response produced similar performance measures, with none substantially outperforming another. However, compared to the accuracy of the SVM model on the training set (93.9% and 96.5%, respectively for anaemia and malaria), the SVM model's accuracy was considerably lower on the test set (61.71% and 80.15%, respectively for anaemia and malaria). Thus, this model suffers from a large variance and poor generalisability on new data. The results of this model should therefore be considered with caution.

Table 3.6: Performance measures for the fitted classification models on the test set

Measure	**Anaemia**				**Malaria**			
	Logistic regression	CART	SVM	ANN	Logistic regression	CART	SVM	ANN
Accuracy	0.6876	0.6777	0.6171	0.6872	0.8206	0.8378	0.8015	0.8283
Sensitivity	0.7381	0.7021	0.6619	0.6987	0.2706	0.3486	0.3218	0.3998
Specificity	0.6381	0.6508	0.5674	0.6744	0.9559	0.9581	0.9195	0.9337
Precision	0.6900	0.6900	0.6288	0.7038	0.6015	0.6717	0.4957	0.5973

Each model for anaemia had a consistently poor accuracy compared to those for malaria. This suggests that the available predictors may not be sufficient in predicting the outcome of anaemia compared to malaria. This was also seen in the result of the classification tree for anaemia, which produced fewer important attributes for the growth of the tree compared to that of malaria. Thus, there may be numerous unmeasured factors that are contributing to anaemia in a child. A shortcoming of such classification models is that they are unable to account for spatial contributions to a response. They cannot assess and incorporate spatial effects which are surrogates for unmeasured factors that contribute to the response. This shortcoming may be influencing the poor predictive accuracy of the models, specifically for that of anaemia. Based on Figure 2.3 in Chapter 2, which showed evidence of how the observed prevalences of anaemia and malaria varied considerably across the districts of the four countries, it is reasonable to assume that there is spatial variation present that is influencing anaemia and malaria in children. This leads us to the next chapter which considers a statistical model to investigate the spatial variation and significant risk factors with a focus on anaemia, which was more prevalent among the children in the sample.

CHAPTER 4

SPATIAL VARIATION AND RISK FACTORS OF CHILDHOOD ANAEMIA

Statistical models to account for and investigate spatial variation and spatial autocorrelation have been well established, especially in the context of the mapping of disease prevalence and risk (Waller & Carlin, 2010). Common techniques in disease mapping make use of aggregated/areal data to model the relative risk based on the counts of the number of cases per administrative or sub-national region. Such techniques include the Poisson-Gamma model, the Poisson log-normal model, the conditional auto-regressive (CAR) model and the Besag, York and Mollie (BYM) model (Clayton & Kaldor, 1987; Lawson et al., 2000; Besag et al., 1991). However, a drawback to these techniques is that they do not permit the inclusion of individual- and household-level covariates. Thus, we consider a geoadditive model to model the likelihood of anaemia in a child while accounting for individual-level, household-level and environmental covariates. Such a model also allows for the inclusion of a spatial effect.

4.1 Geoadditive Model

A hierarchical multivariable geoadditive logit model to control for the confounding effects of the covariates was considered (Wand et al., 2011). This formulation is a structured additive regression model that includes a spatial effect and is based on the generalised linear model (GLM) and generalised additive model (GAM) frameworks (Umlauf et al., 2015). For this model, Y_{hijk} follows a Bernoulli distribution where

$P(Y_{hijk} = 1) = \pi_{hijk}$ is the probability that child k in household j within cluster i and district h is anaemic and $P(Y_{hijk} = 0) = 1 - \pi_{hijk}$ is the probability that the child is not anaemic. The hierarchical geoadditive model is given by

$$logit(\pi_{hijk}) = \boldsymbol{x}'_{hijk}\boldsymbol{\beta} + f_1(z_{hijk1}) + f_2(z_{hijk2}) + \ldots + f_p(z_{hijkp}) + f_{spat}(s_h), \quad (4.1)$$

where the left side of the Equation (4.1) is the logit link function and the right side is the geoadditive predictor. The parameter $\boldsymbol{\beta}$ is the vector of the linear fixed effects of the covariates that are modelled parametrically; and $f_r(.)$, $r = 1, \ldots, p$, are the unknown smooth functions that represent the non-linear effects of the continuous covariates which are modelled non-parametrically, thus Equation (4.1) is a semi-parametric model. The spatial effect of district s_h in which the child resides, $s \in (1, \ldots, 370)$, is given by $f_{spat}(s_h)$ which represents the effects of unobserved covariates that are not included in the model and also accounts for spatial autocorrelation (Kandala & Madise, 2004). This spatial effect may be partitioned into a spatially correlated (structured) and an uncorrelated (unstructured) effect as

$$f_{spat}(s_h) = f_{str}(s_h) + f_{unstr}(s_h).$$

The structured spatial effect $f_{str}(s_h)$ accounts for the assumption that neighbouring districts close are more likely to be correlated with regards to their outcomes. However, the unstructured spatial effect $f_{unstr}(s_h)$ accounts for the spatial variation due to effects of unmeasured district-level factors that are not spatially related (Ngwira & Kazembe, 2015).

In this analysis, inference was fully Bayesian, hence all parameters and functions were treated as random variables. The fixed effect parameters in $\boldsymbol{\beta}$ were assigned vague Gaussian priors $N(0, 1000)$, with precision $= 0.001 = 1/\text{variance}$. The Bayesian perspective of penalised splines (P-splines) was adopted for the unknown smooth functions f_r (Lang & Brezger, 2004). This approach assumes that the unknown func-

tions can be approximated by a polynomial spline of degree l with equally spaced knots $z_r^{\min} = \zeta_{r0} < \zeta_{r1} < \ldots < \zeta_{rn_r-1} < \zeta_{rn_r} = z_r^{\max}$ which are within the domain of the covariate z_r. In terms of a linear combination of $M_r = n_r + l$ B-spline basis functions, B_{rm}, the Bayesian spline can be written as

$$f_r(z_r) = \sum_{m=1}^{M_r} \alpha_{rm} B_{rm}(z_r).$$

Thus, $\boldsymbol{\alpha}_r = (\alpha_{r1}, \ldots, \alpha_{rM_r})'$ are unknown regression coefficients which are assigned first- or second-order random walk priors given by $\alpha_{rm} = \alpha_{r,m-1} + u_{rm}$ and $\alpha_{rm} = 2\alpha_{r,m-1} - \alpha_{r,m-2} + u_{rm}$, respectively, with Gaussian errors $u_{rm} \sim N\left(0, \frac{1}{\tau_r^2}\right)$ and diffuse priors α_{r1} or α_{r1} and α_{r2} as constants for initial values, respectively. The variance component τ_r^2 controls the smoothness of f_r. Here, we used second-order random walk smoothness priors and third degree splines.

For the structured spatial effect, $f_{str}(s_h)$, intrinsic Gaussian Markov random field (IGMRF) priors specified by Besag et al. (1991) were used (Besag et al., 1991). Two districts s_h and s_i are defined as neighbours if they share a common boundary. The spatial extension of random walk models leads to the conditional, normal distribution, spatially autoregressive specification

$$f_{str}(s_h)|f_{str}(s_i), h \neq i \sim N\left(\frac{1}{n_{s_h}} \sum_{s_i \in \delta_{s_h}} f_{str}(s_i), \frac{1}{n_{s_h}\tau_{str}^2}\right),$$

where n_{s_h} is the number of neighbours of district s_h, and $s_i \in \delta_{s_h}$ denotes that district s_i is a neighbour of district s_h. Therefore, the conditional mean of $f_{str}(s_h)$ is an average of the function evaluations $f_{str}(s_h)$ of neighbouring districts. Furthermore, the variance component τ_{str}^2 controls the smoothness of the spatial effect and accounts for spatial variation between the districts, it is also used to capture the amount of variation explained by the spatial structure. The unstructured spatial effect $f_{unstr}(s_h)$ was assigned i.i.d. Gaussian priors and specified as

$$f_{unstr}(s_h) \sim N\left(0, \frac{1}{\tau_{unstr}^2}\right).$$

The variance components, τ^2, of the random and spatial effects are unknown precision parameters that require estimation. Therefore, hyperpriors were assigned in a second stage of hierarchy. These hyperpriors are defined on a logarithmic scale and thus a log-gamma(a, b) distribution with hyper-parameters $a = 1$ and $b = 0.001$ was used. A sum-to-zero constraint was imposed on the non-linear and spatial effects to ensure model identifiability between the intercept and these effects.

Three types of models were fitted:
Model 1: GLM model: Linear fixed effects of all variables, categorical and continuous.
Model 2: GAM model: Linear fixed effects of categorical variables and some continuous variables, and non-linear effect of the child's age in months.
Model 3: Geoadditive Model: Model 2 with the inclusion of the spatial effects.

The posterior distributions of the parameters in the models were estimated using Integrated Nested Laplace Approximation (INLA) using the INLA package in R (http://www.r-inla.org/) (Rue et al., 2009). The final geoadditive model was selected based on the Deviance Information Criteria (DIC), where the model with the smallest DIC was considered a better fit (Spiegelhalter et al., 2002). The sensitivity to the choice of the hyper-parameter values a and b was investigated by fitting the model with different hyper-parameter values (Adebayo & Fahrmeir, 2005). However, the estimates had little sensitivity to these choices. QGIS 3.4 (https://qgis.org/en/site/index.html) was used to create maps displaying the posterior mean estimates of the spatial effects for the different districts of the countries. All of the maps created were based on the shapefiles of the four countries, each of which were partitioned into their respective districts. A shapefile is a geospatial vector data format for geographic information system (GIS) software. The shapefiles for each country were merged into one using QGIS. In addition, all of the islands in Tanzania and Malawi were removed as only mainlands were considered.

The estimation of the district-level structured spatial effect requires an adjacency matrix, $\mathbb{W}$, based on the neighbourhood structure of the districts. This adjacency matrix contains diagonal elements of zero and off-diagonal elements $w_{ij} = 1$ if district i and district j share a common boundary, and zero otherwise. The adjacency matrix used in this analysis was created based on the neighbourhood structure of the 370 districts across the four countries.

4.2 Inference from the Geoadditive Model

This analysis was based on the 18196 observations with the addition of 50 observations to include an additional category for the type of toilet facilities. The child's malaria status was incorporated as an explanatory variable to assess its effect on anaemia. Moreover, based on the results of the exploratory data analysis from Chapter 3, the age of the household head was not considered as a predictor of anaemia in this model. With the inclusion of the spatial effects at district level, the effect of each country can be obtained by systematic aggregation of the effects of the districts within the country. Thus, the country of residence was also excluded from the model. In addition, model diagnostics indicated that the model was of a reasonable fit.

Model selection

The variance inflation factor (VIF) was used to check for collinearity among the explanatory variables. All of the variables had a VIF < 4 and thus it was assumed that multicollinearity was not significantly present (Zuur et al., 2009). The non-linear effect of all continuous variables was investigated, however the only variable to display a significant non-linear effect on the log-odds of anaemia was the child's age in months. Thus, this was the only non-linear effect considered in the models fitted, while the remaining independent variables were included as linear fixed effects. Table 4.1 presents the results of the DIC and effective number of parameters, p_D, for

each of the fitted models. Model 3 (Equation (4.1)) produced the lowest DIC, and thus the results of this study are based on this model, which includes both linear and non-linear effects as well as the spatial effects.

Table 4.1: Model comparisons

	Model 1	**Model 2**	**Model 3**
DIC	22181.94	22086.94	21424.61
p_D	16.01	22.96	263.45

Fixed Effects

Table 4.2 displays the adjusted posterior odds ratio estimates (AOR) with their 95% credible intervals for the linear fixed effects included in the multivariable model. Female children had a significantly lower odds of anaemia compared to males (AOR = 0.873; 95% CrI: 0.818-0.932). Similarly, there was a significant decrease in the odds of anaemia with an increase in mother's education, cluster altitude and household wealth index. Furthermore, a significantly lower odds of anaemia was suggested for children living in households with improved toilet facilities (PIT latrine and flush toilet). Children residing in urban areas had a lower odds of anaemia compared to those residing in rural areas, however these odds were not significantly different (AOR = 0.926; 95% CrI: 0.835-1.027). Children with a positive malaria RDT result had a significantly higher odds of anaemia compared to those who had a negative malaria RDT result (AOR = 4.401; 95% CrI: 3.979-4.871), as did those children living in households with increasing number of residents (AOR = 1.019; 95% CrI: 1.008-1.030). While the odds of anaemia decreased with an increase in EVI (AOR = 0.987; 95% CrI: 0.927-1.051) and increased with an increase in LST (AOR = 1.008; 95% CrI: 0.994-1.022), these factors did not appear to be significantly associated with a child's anaemia status.

Table 4.2: Adjusted posterior odds ratio estimates (AOR) and 95% credible intervals (CrI)

Variable	AOR (95% CrI)
Individual and Household Level	
Gender (ref = Male)	
Female	0.873 (0.818, 0.932)*
Malaria RDT Result (ref = Negative)	
Positive	4.401 (3.979, 4.871)*
Household Size	1.019 (1.008, 1.030)*
Type of Place of Residence (ref = Urban)	
Rural	0.926 (0.835, 1.027)
Mother's Education Level (ref = No Education)	
Primary	0.857 (0.773, 0.950)*
Secondary and Higher	0.795 (0.694, 0.911)*
Unknown	0.845 (0.742, 0.963)*
Gender of Household Head (ref = Male)	
Female	1.003 (0.927, 1.086)*
Type of Toilet Facility (ref = No Facilities)	
PIT Latrine	0.813 (0.723, 0.914)*
Flush Toilet	0.749 (0.612, 0.916)*
Other	0.711 (0.382, 1.325)
Wealth Index	0.858 (0.807, 0.911)*
Cluster Level	
Cluster Altitude (in 100 metres)	0.974 (0.962, 0.987)*
EVI (in 1000s)	0.987 (0.927, 1.051)
LST	1.008 (0.994, 1.022)

*significant at 5% level of significance

Non-linear and Spatial Effects

Table 4.3 provides the posterior mean and 95% credible interval for the smooth term variance components (the precisions) for the non-linear and spatial effects. The precision of an effect is the inverse of its variance. Thus, the larger the precision, the smaller the variance of the effect. The precision corresponding to the structured spatial effect (853.58) was much higher compared to that of the unstructured spatial effect (3.84), thus suggesting that the unstructured spatial effect was more dominant (Kazembe et al., 2007).

Table 4.3: Posterior mean and 95% credible interval (CrI) for the smooth term variance components

Variable	Mean	95% CrI
Non-linear Effect		
Child's Age in Months (τ_r^2)	1648.49	(485.52, 3938.21)
Spatial Effect		
Structured Spatial Effect (τ_{str}^2)	853.58	(44.69, 3252.82)
Unstructured Spatial Effect (τ_{unstr}^2)	3.84	(3.041, 4.78)

Figure 4.1 shows the non-linear effect that a child's age in months has on the log-odds of being anaemic as well as the 95% credible interval. There was an increase in effect from 6 to 11 months, after which the effect declined. If a linear effect was used, it would have overestimated the effect of ages 30 to 50 months on anaemia.

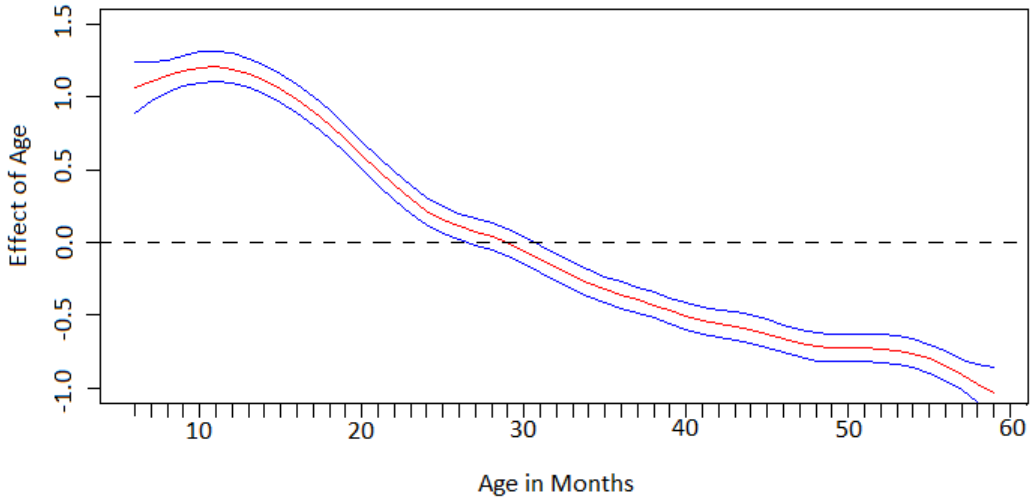


Figure 4.1: Estimated non-linear effects of child's age in months on the log-odds of anaemia. The posterior mean together with the 95% credible intervals are shown.

Figure 4.2 displays the estimated means of the structured and unstructured spatial effects on the log-odds of anaemia, where the blue districts have a negative spatial effect and are therefore associated with a lower odds of anaemia, and the red districts have a positive spatial effect and are therefore associated with a higher odds of anaemia. The structured spatial effect, which ranged from -0.0368 to 0.0316, was weak in comparison to the unstructured spatial effect, which ranged from -1.3061 to 0.9780. Furthermore, the 95% CrI of the log-odds for the structured spatial effect

in each district overlapped with the null of 0 (results not shown), thus the effects of spatially correlated factors contributing to childhood anaemia in all the districts were not statistically significant. However, 36 districts had a significantly positive unstructured spatial effect and 34 districts had a significantly negative unstructured spatial effect.

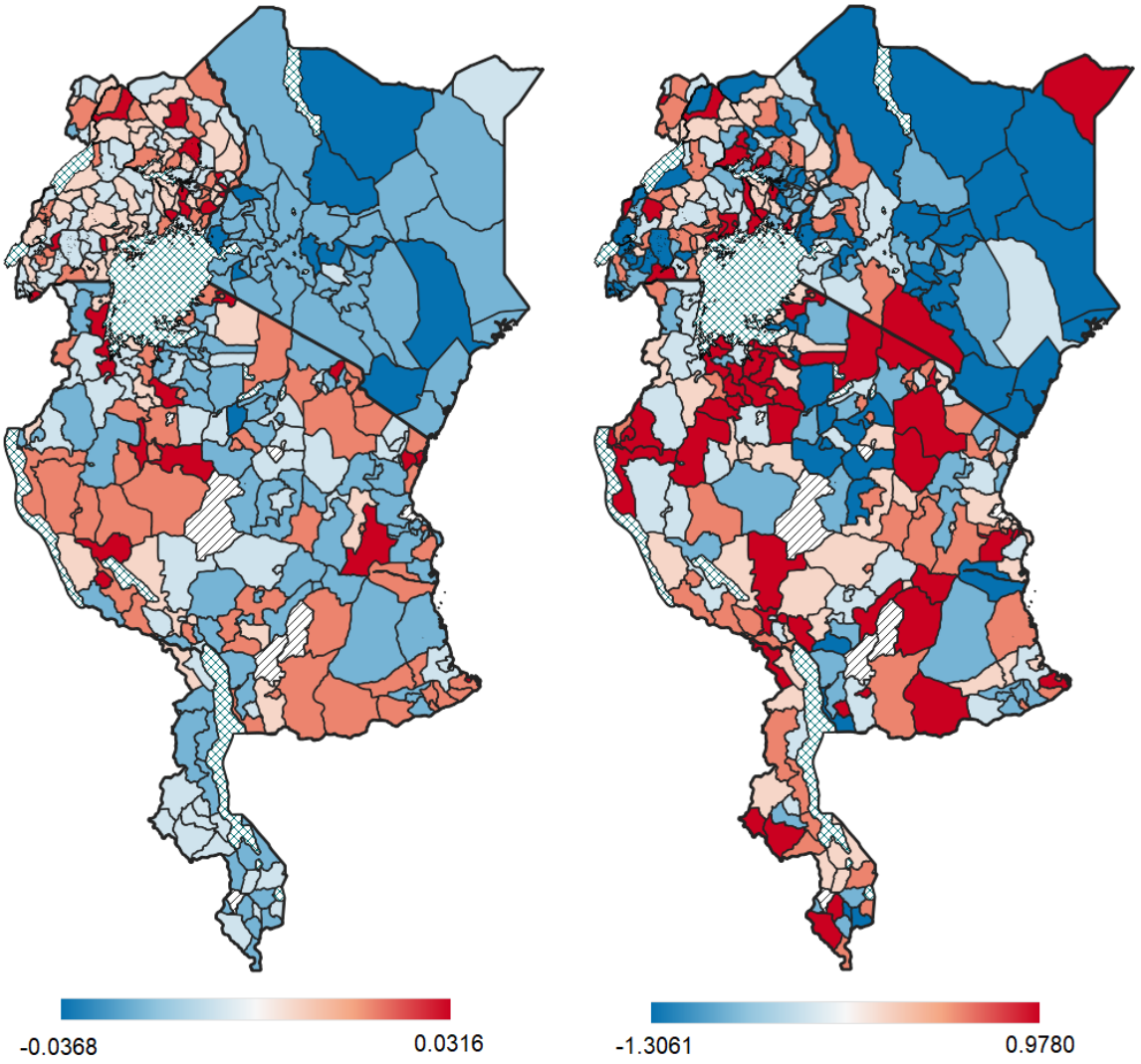


Figure 4.2: Estimated posterior means of the structured spatial effect (left) and the unstructured spatial effect (right) on the log-odds of anaemia (criss-cross pattern indicates water bodies; diagonal lines indicate districts with no data available).

4.3 Summary and Discussion

We utilised a hierarchical geoadditive logistic model to investigate the risk factors and spatial variation of anaemia in children aged 6 to 59 months in Kenya, Malawi, Tanzania and Uganda. This type of model allows one to assess and visualise the residual spatial effects on childhood anaemia while controlling for the effects of other covariates. Furthermore, it allows for the non-linear relationship of continuous covariates to be explored. In this analysis, incorporating the spatial effect in the model reduced the model's DIC.

The results of this analysis confirm that of other studies, where girls are less at risk of anaemia, and a child's risk decreases with an increase in mother's education level and wealth (Gayawan et al., 2014; Ngwira & Kazembe, 2015; Khan et al., 2015; Soares Magalhães et al., 2013a). This may be due to more educated individuals being more aware and having more of an understanding of health related issues. Similarly, this could be said of individuals with more wealth. However, a lack of wealth also restricts an individual's ability to access good health care and nutritional food sources. Having malaria was associated with a significantly higher risk of anaemia, thus suggesting much of the burden of childhood anaemia in these countries is contributed by malaria. The type of toilet facilities was significantly associated with a child's anaemia status. Poor sanitation is a known risk factor of the intestinal parasite hookworm which causes anaemia in infected children (Smith & Brooker, 2010). While a study by SoaresMagalhães & Clements (2011) found environmental factors LST and the normalized difference vegetation index (NDVI) to be significantly associated with an increased risk of anaemia in preschool-age children, the environmental factors LST and EVI considered in this study were not found to be significantly associated with anaemia. However, such environmental factors, especially EVI, are known to be highly correlated with malaria, and thus the inclusion of the child's malaria status may account for much of the effects that these environmental factors have on childhood anaemia (Cottrell et al., 2012; Nguyen et al., 2019).

The non-linear effect of the child's age on anaemia displayed an increase from 6 to 11 months, after which the effect declined. Multiple factors could be contributing to this increased risk of anaemia in children aged 6 to 11 months. Either these children are not receiving adequate nutrients or they are experiencing a decrease in their Hb concentrations due to other factors. Infants are born with a reserve of iron which is responsible for growth and protection from iron deficiency in the first 4 to 6 months of life (Ziegler et al., 2014). After 6 months of age, the iron store is depleted, and thus it is common for milk supplements to be introduced into a child's diet to complement breastfeeding as breast milk alone may not provide sufficient iron to meet the demand of the rapid growth experienced in children during this period (Gayawan et al., 2014; Miller, 2019). However, safe complementary feeding in children from 6 months is not always practised, where the feeding of unmodified cow's milk in children less than 12 months of age is common in some SSA countries despite evidence of increased risk of iron-deficiency anaemia and other adverse health outcomes (Ssemukasa & Kearney, 2014; Saldan et al., 2017). Wijndaele et al. (2009) found that low maternal education and low socio-economic status are associated with feeding of unmodified cow's milk in children less than a year old. In addition, malaria in mothers may also be a contributing factor to the increased risk of anaemia in children aged 6 to 11 months, where (White, 2018) states that the effects of maternal anaemia due to malaria can cause a physiological decline in Hb concentrations in infants from birth up to 9 months of age, after which there is a slow but steady rise in Hb concentrations. Other studies on young children from Kenya, Malawi, Tanzania and Uganda have also reported similar patterns of decreased Hb concentrations in children less than 11 months of age (Ngwira & Kazembe, 2015; Crawley, 2004; Schellenberg et al., 2003; McElroy et al., 2000; le Cessie et al., 2002).

The benefit of focusing on more than one country at a time is that one is able to consider whether factors that transcend boundaries are significantly contributing to

childhood anaemia, such as environmental and geographical factors. The results revealed that the structured spatially correlated effect was fairly weak in comparison to the unstructured spatial effect, suggesting that the contribution that a particular district has on the risk of anaemia is not similar among neighbouring districts. This is an indication that environmental and geographical factors that transcend boundaries of the districts may not play a significant role in childhood anaemia. With the unstructured spatial effect being more prominent in this study, it can be concluded that there are unmeasured district-specific factors that are not spatially structured (that are not correlated with that of neighbouring districts) contributing to childhood anaemia. In addition, there was a distinct pattern of variation in the spatial effects across the districts within each country, except for Kenya which was fairly homogeneous in both types of spatial effects. Kenya has made substantial progress in the reduction of malaria, however this has resulted in a heterogeneous risk of malaria across the country (Macharia et al., 2018). Thus, the homogeneous results of the spatial effects on childhood anaemia in Kenya could be due to the strong correlation between malaria and anaemia in the country, which is being accounted for by the inclusion of the child's malaria status. However, the spatial effects in Uganda, Tanzania and Malawi remain heterogeneous even after controlling for the child's malaria status, thus there are other significant drivers of childhood anaemia in these countries. On the whole, the spatial effects do not appear to transcend the borders between the countries as the pattern of effects differed around the borders, barring Longido district in Tanzania and Kajiado county in Kenya which share a border. This indicates that there are country-specific factors contributing to anaemia in children. Such factors may include the cost and quality of health care, and the cost of living, which can vary considerably between and within countries, the effects of which have been known to contribute to the spatial variation of other childhood diseases (Kandala & Madise, 2004).

CHAPTER 5

DISTRICT EFFECT APPRAISAL OF CHILDHOOD ANAEMIA

5.1 Introduction

Earlier, we considered the spatial effect at district-level, which was decomposed into a structured spatial effect and an unstructured spatial effect. The district-level structured spatial effect accounted for spatial autocorrelation in the observations between neighbouring districts. Now we consider spatial autocorrelation in the observations between the clusters *within* the districts, where we would expect clusters close in proximity to have similar responses. This leads to a cluster-level structured spatial effect, which allows us to assess the spatial variation within the districts. This cluster-level spatial effect is incorporated into the model based on the geographical coordinates of the sampled clusters. In addition, the performance of the districts is considered in order to assess and rank the 'best' and 'worst' performing districts with regard to their impact on childhood anaemia based on the Best Linear Unbiased Prediction (BLUP) technique. Such a district effect appraisal on anaemia can aid in providing a wider and richer insight in the effort to overcome childhood anaemia by prioritising the worst performing districts for action. Furthermore, it enables one to identify key differences between the best and worst performing districts compared to national and global levels. Identifying factors that contribute to these differences can aid in targeting the correct set of interventions in the districts where it is much needed. This BLUP technique is primarily used in animal and plant breeding for estimating and ranking genetic merit (Robinson, 1991; Soh, 1994; Bajetha et al., 2015), however to our knowledge, such a method has not been used for the appraisal of administrative levels in epidemiological studies. Once again in this analysis, we

utilise a statistical model that allows us to account for the effects of individual-level, household-level and environmental factors.

5.2 Best Linear Unbiased Predictor

Here we adopt a generalised additive mixed model (GAMM) for the hierarchical and spatially correlated data (Lin & Zhang, 1999). A GAMM is an additive extension of generalised linear mixed models (GLMMs) and uses additive nonparametric functions to model covariates and geospatial effects while accounting for correlation by adding random effects to the additive predictors (Wand et al., 2011; Umlauf et al., 2015). The fitted GAMM for $P(Y_{hijk} = 1) = \pi_{hijk}$ with a logit link function is

$$logit(\pi_{hijk}) = \boldsymbol{x}'_{hijk}\boldsymbol{\beta} + U_h + \sum_{r=1}^{p} f_r(z_{hijk}) + f_{spat}(\text{lon}_i, \text{lat}_i), \tag{5.1}$$

where π_{hijk} is the probability that child k in household j within cluster i and district h; $\boldsymbol{\beta}$ is the linear fixed effects; U_h is the district-level random effect modelled parametrically; $f_r(.)$, $r = 1, \ldots, p$, are the unknown smooth functions that represent the non-linear effects of the p covariates which are modelled non-parametrically; and the non-linear term $f_{spat}(\text{lon}_i, \text{lat}_i)$ is a function of the geographical coordinates of the i^{th} cluster where lon_i and lat_i are the longitude and latitude, respectively. Interactions between any of the terms in Equation 5.1 can also be explored.

Estimation of the smooth functions f_r was based on penalised splines (P-splines) (Eilers & Marx, 1996). This approach assumes that the unknown functions can be approximated by a polynomial spline of degree l with equally spaced knots $z_r^{\min} = \zeta_{r0} < \zeta_{r1} < \ldots < \zeta_{rn_r-1} < \zeta_{rn_r} = z_r^{\max}$ which are within the domain of the covariate z_r. The spline can be written in terms of a linear combination of $M_r = n_r + l$ B-spline basis functions, B_{rm}, and regression coefficients α_{rm} as

$$f_r(z_r) = \sum_{m=1}^{M_r} \alpha_{rm} B_{rm}(z_r). \tag{5.2}$$

The choice in the number of knots is important as too few results in a spline that may not be flexible enough to capture the variability in the data, however too many knots may result in estimated curves that overfit the data, which leads to functions that are too rough (Fahrmeir et al., 2004). To overcome this problem, a moderately large number of equally spaced knots of between 20 and 40 is used to ensure flexibility (Eilers & Marx, 1996). In addition, a roughness penalty is defined based on first or second order differences of adjacent B-Spline coefficients which guarantees sufficient smoothness of the fitted curves (Eilers & Marx, 1996). This leads to penalised likelihood estimation with penalty terms given by:

$$P(\lambda_r) = \frac{1}{2}\lambda_r \sum_{m=v+1}^{M_r} \left(\Delta^v \alpha_{rm}\right)^2, \quad v = 1, 2,$$

where λ_r is the smoothing parameter and Δ^v is the differencing operator of order v. First order differences penalise abrupt jumps $\alpha_{rm} - \alpha_{r,m-1}$ between successive parameters, while second order differences penalise deviations from the linear trend $2\alpha_{r,m-1} - \alpha_{r,m-2}$ (Fahrmeir et al., 2004). We used a choice of 20 knots and a spline of degree 3.

The effect of the i^{th} cluster location, given by $f_{spat}(\text{lon}_i, \text{lat}_i)$, $i = 1, \ldots, 1595$ was estimated based on a two-dimensional P-spline, which itself is based on the tensor product of 1 dimensional B-splines:

$$f_{spat}(\text{lon}_i, \text{lat}_i) = \sum_{m_1=1}^{1595} \sum_{m_2=1}^{1595} \alpha_{m_1 m_2} B_{m_1}(\text{lon}_i) B_{m_2}(\text{lat}_i).$$

The stochastic formulation of $f_{spat}(\text{lon}_i, \text{lat}_i)$ represents the realisation of a spatially correlated stochastic process, which assists in accounting for spatial correlations in the data (Kneib et al., 2008). The B-spline basis functions are now spatially aligned along the x and y axes, and thus a suitable difference penalty is then constructed based on squared deviations of $\alpha_{m_1 m_2}$ from the regression coefficients of the four nearest neighbours (Kneib et al., 2008).

Furthermore, in order to account for the correlation in the responses due to unmeasured district-specific factors, an independently and identically distributed random effect was included in the model based on the district in which the child resided. The function for this random effect in the model can also be approximated by a linear combination of B-spline basis functions given by Equation 5.2. However, the regression coefficients α_{rm} are i.i.d. random effects (Fahrmeir et al., 2004).

GAMMs can be represented as GLMMs after appropriate re-parameterisations of the smoothing splines (Fahrmeir et al., 2004). Based on the GLMM representation, regression parameters and variance components can be estimated using iteratively weighted least squares (IWLS) and restricted maximum likelihood (REML) estimation, respectively. The mixed model methodology permits the estimation of the fixed effects, as well as the prediction of the random effects using the BLUP procedure by solving the system of linear equations called mixed model equations (Furlani et al., 2005). BLUP values are realised values of the random effect (Zewotir, 2008, 2012). BLUP provides an unbiased method by adjusting for known sources of individual, household, cluster, geospatial and environmental variation (Henderson, 1975). Furthermore, BLUP is an appropriate technique for the ideal ranking or selection criteria that involve a random effect. It is well established on theoretical grounds that these properties can result in increased accuracy in ranking and selection (Henderson, 1975; Robinson, 1991; McCulloch et al., 2008). In other words, the ranking of the best linear predictors produces the same order as the true values of the random effects (Portnoy, 1982). Through this maximizing the probability of correct ranking, BLUP are appropriate values upon which to base selection and ranking. Thus, inclusion of the district-level random effect enables one to rank and select the 'best' and the 'worst' districts with regards to anaemia risk based on the obtained BLUP estimates for each district.

The estimation approach used in this study is referred to as an empirical Bayes ap-

proach. Empirical Bayes inference assumes that the regression and variance parameters are unknown constants, where the estimates are obtained by maximizing an objective function, thus the estimates can be interpreted as penalized likelihood estimates from a frequentist perspective (Fahrmeir et al., 2004). The model was fitted using the R2BayesX package in R (Umlauf et al., 2016). The estimates of the district-level random effect and cluster-level spatial effect were imported into ArcGIS 10.6 and mapped.

5.3 Fixed Effects Estimation

From the data consisting of 18247 observations that was used in the analysis in Chapter 4, only 18027 observations had valid geographical coordinates available for the cluster of residence. Thus, these results are based on only those 18027 observations. In this analysis, the explanatory variables comprised of the same demographic, socio-economic and environmental factors as that considered in Chapter 4, however the country of residence was also considered.

To avoid possible confounding effects, all two-way interactions of the fixed effects were explored. The only significant interaction was found between the type of place of residence (rural/urban) and the country, which is not a surprising result as the coverage and classification of rural/urban areas within a country differs from country to country. This significant interaction effect suggests that the effect that an urban or rural area has on anaemia in children differs across the four countries. Furthermore, the total effect that the place of residence and country has on the odds of anaemia is made up of their individual main effects as well as the simultaneous/interaction effect between the two variables.

Table 5.1 displays the results of the adjusted odds ratios and their 95% confidence intervals for the final model. Female children had a significantly lower odds of anaemia compared to males (AOR = 0.876; 95% CI: 0.820-0.935). The odds of anaemia

was significantly higher for children who tested positive for malaria based on the RDT result compared to those who tested negative (AOR = 4.315; 95% CI: 3.895-4.781).

Table 5.1: Adjusted odds ratio estimates (AOR) and 95% confidence intervals (CI) for the fixed effects

	AOR (95% CI)
Main Effects	
Gender (ref = Male)	
Female	0.876 (0.820, 0.935)*
Malaria RDT Result (ref = Negative)	
Positive	4.315 (3.895, 4.781)*
Household Size	1.014 (1.003, 1.025)*
Type of Place of Residence (ref = Urban)	
Rural	0.738 (0.582, 0.936)*
Mother's Education Level (ref = No Education)	
Primary	0.843 (0.760, 0.935)*
Secondary and Higher	0.794 (0.693, 0.911)*
Unknown	0.845 (0.742, 0.962)*
Gender of Household Head (ref = Male)	
Female	1.016 (0.939, 1.100)
Type of Toilet Facility (ref = No Facilities)	
PIT Latrine	0.780 (0.694, 0.877)*
Flush Toilet	0.725 (0.591, 0.889)*
Other	0.663 (0.357, 1.230)
Wealth Index	0.847 (0.797, 0.901)*
Country (ref = Malawi)	
Kenya	0.316 (0.160, 0.622)*
Tanzania	0.639 (0.355, 1.150)
Uganda	0.433 (0.214, 0.874)*
Cluster Altitude (in 100 metres)	0.986 (0.969, 1.004)
EVI (in 1000s)	1.026 (0.865, 1.217)
LST	1.015 (0.969, 1.063)
Interaction Effects	
Type of Place of Residence and Country (ref = Urban and Malawi)	
Rural and Kenya	1.376 (1.037, 1.825)*
Rural and Tanzania	1.237 (0.942, 1.624)
Rural and Uganda	1.119 (0.826, 1.514)

*significant at 5% level of significance

The odds of anaemia increased with an increase in household size (AOR = 1.014; 95% CI: 1.003-1.025), however the odds decreased with an increase in the household wealth index Z-score (AOR = 0.847; 95% CI: 0.797-0.901). Children whose mother had at least a primary level of education were associated with a lower odds of anaemia compared to those whose mother had no education (AOR = 0.843; 95% CI: 0.760-0.935 for primary level, AOR = 0.794; 95% CI: 0.693-0.911 for secondary or higher education level). Moreover, children in households with improved toilet facilities had a lower odds of anaemia compared to those in households with no toilet facilities (AOR = 0.780; 95% CI: 0.694-0.877 for PIT latrine, AOR = 0.725; 95% CI: 0.591-0.889 for flush toilet). The gender of the head of household, cluster altitude, EVI and LST were not significantly associated with the child's anaemia status.

While the adjusted odds ratios for the main and interaction effects of the type of place of residence and country of residence are presented in Table 5.1, they cannot be interpreted separately. Rather, their total effect on the log-odds of anaemia should be considered. Thus, Figure 5.1 presents the total estimated log-odds of anaemia for each type of place of residence across the four countries. This figure clearly displays a difference in the effect of the type of place of residence on the log-odds of anaemia between the four countries. Without the inclusion of this interaction effect, it would be assumed that the effect of the type of place of residence is constant for all the countries. While the log-odds of anaemia for children residing in rural areas was lower than that for children residing in urban areas in Malawi, Tanzania and Uganda, only Malawi displayed a considerable difference between urban and rural areas. Furthermore, Uganda and Kenya displayed a decreased log-odds of anaemia in both urban and rural areas, while Malawi and Tanzania displayed an increased log-odds in both urban and rural areas.

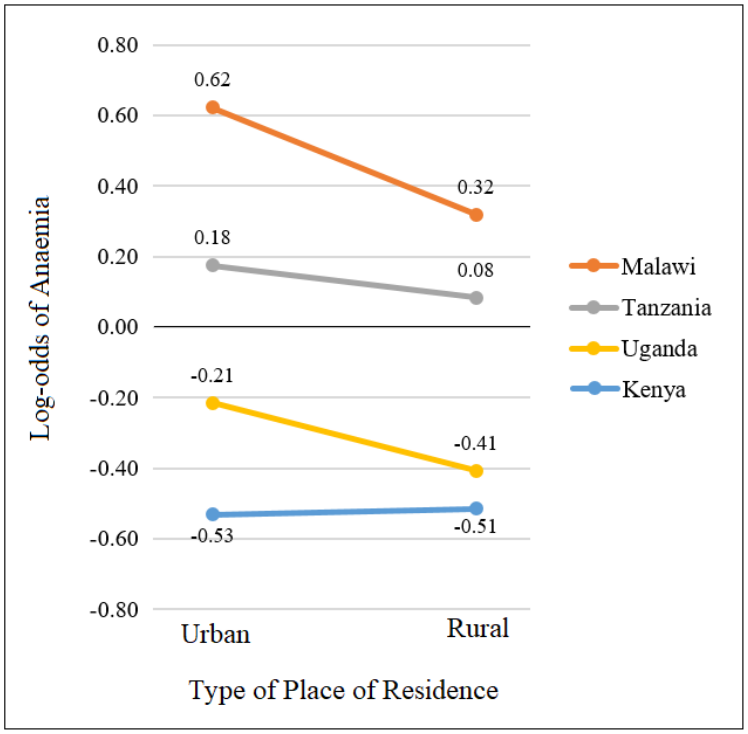


Figure 5.1: Log-odds of anaemia associated with the type of place of residence and country

5.4 Non-linear and Spatial Effects

The child's age in months had a fairly significant non-linear effect on the log-odds of anaemia with its non-zero variance estimate (Table 5.2). Similarly, the variance estimates for the district-level random effect and cluster-level spatial effect were non-zero.

Table 5.2: Variance estimates of non-linear terms

	Variance Estimate
Child's Age in Months	0.0127
District-Level Random Effect	0.1516
Cluster-Level Spatial Effect	0.6904

Figure 5.2 displays the non-linear effect of the child's age in months on the log-odds of anaemia. The effect increased from 6 to 11 months of age, after which there was a decline in the effect. Children from about 25 months of age displayed a negative

effect, and thus were associated with a lower risk of anaemia.

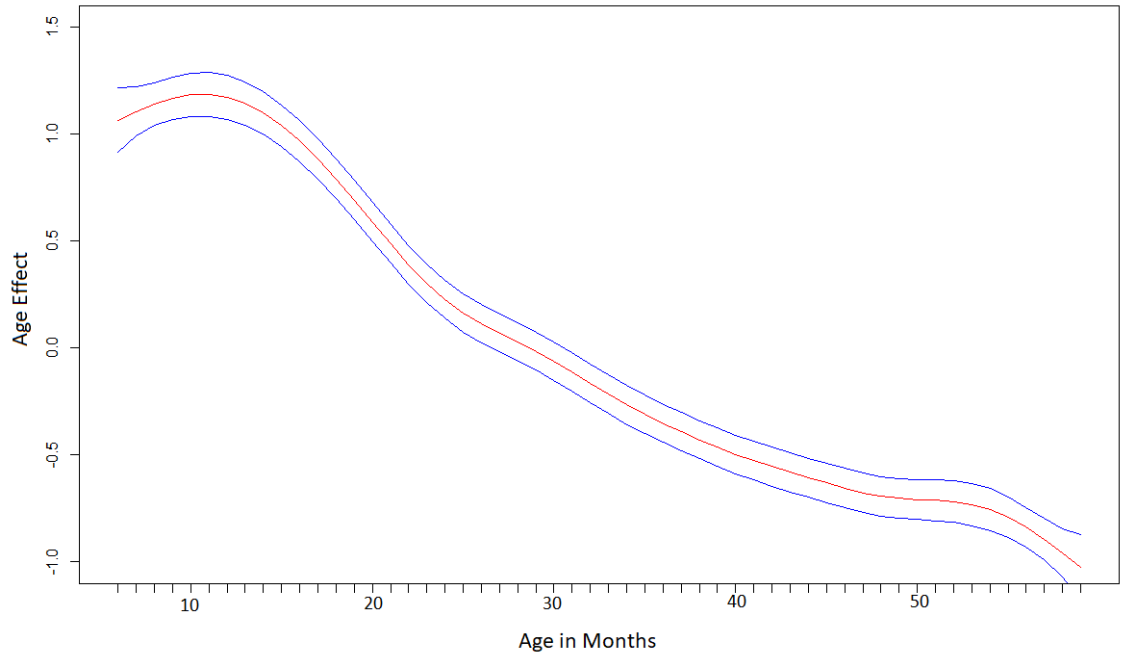


Figure 5.2: Estimated non-linear effect of the child's age in months on the log-odds of anaemia together with the 95% confidence interval

The estimated cluster-level spatial effect, which accounts for spatial autocorrelation, is presented in Figure 5.3. The clusters in shades of blue had a negative effect on the log-odds of anaemia and thus were associated with a decreased risk. Whereas, those in shades of yellow to red had a positive effect and were therefore associated with an increased risk of childhood anaemia. Uganda, which consisted of clusters with both positive and negative effects, displayed the largest spatial variation. Throughout all four countries, the majority of neighbouring clusters resulted in similar effects. In Kenya, Tanzania and Uganda, some areas displayed clusters with a positive effect and clusters with a negative effect within the same district. Clusters surrounding Lake Victoria, which lies across the border between Uganda and Tanzania, had a positive effect and thus were associated with an increased odds of anaemia. Malawi was fairly homogeneous as it consisted of clusters with only negative effects.

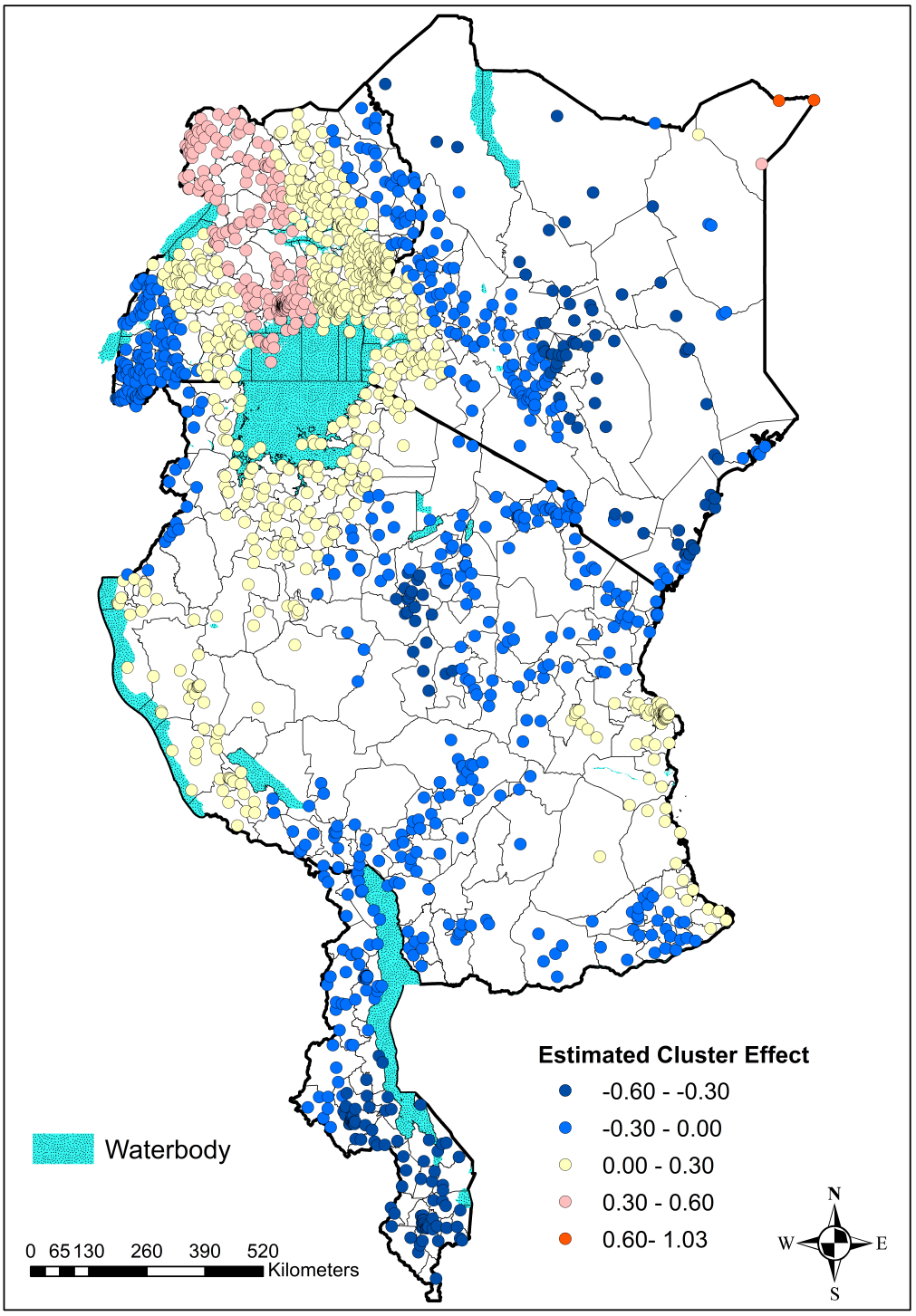


Figure 5.3: Estimated cluster-level spatial effect on the log-odds of anaemia

Figure 5.4 displays the estimated district-level random effect based on the BLUP estimates, where the shades of blue had a negative/decreased effect on the log-odds of anaemia and the shades of beige to red had a positive and therefore increased effect on the log-odds of childhood anaemia.

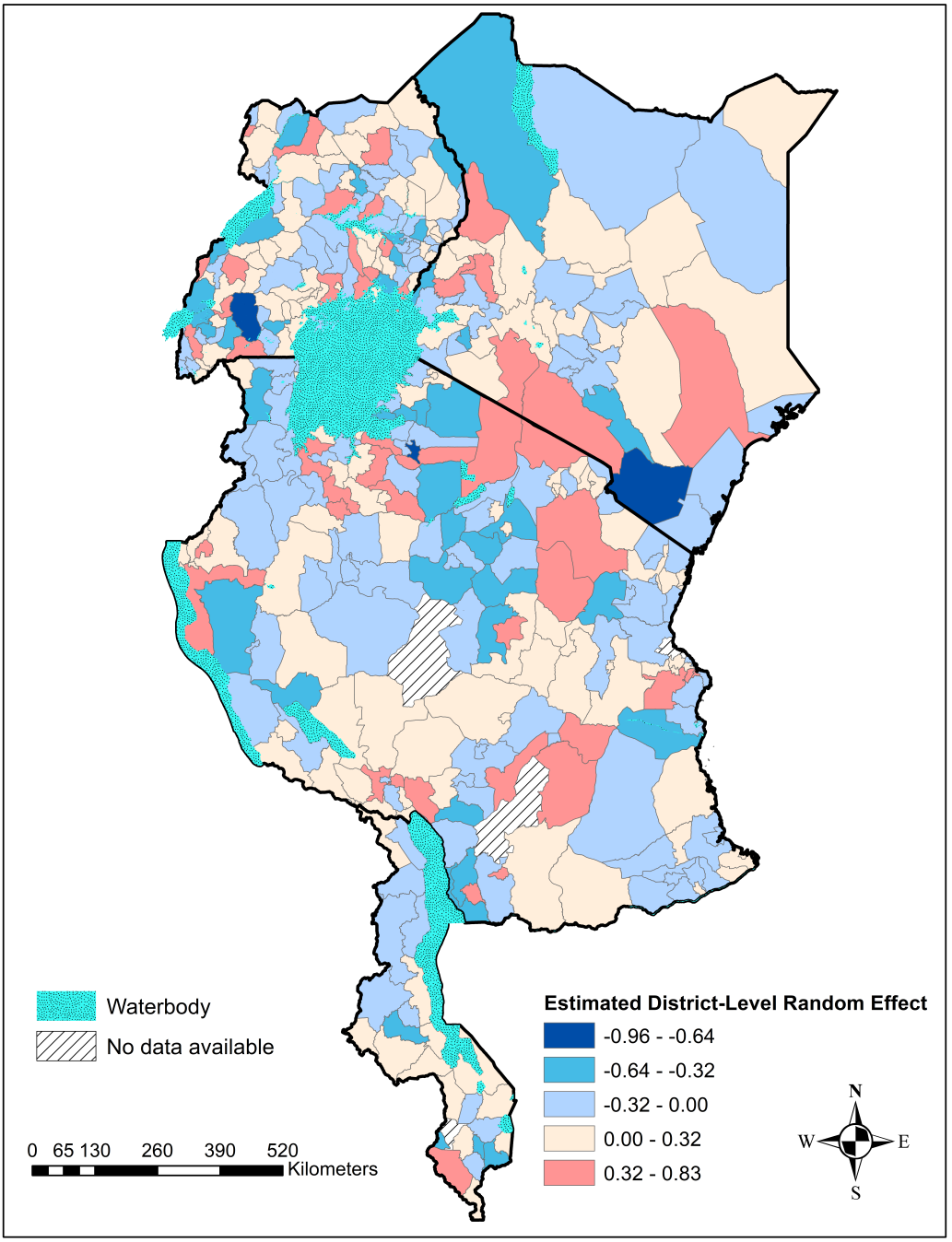


Figure 5.4: Estimated district-level random effect on the log-odds of anaemia

There was significant heterogeneity between and within the countries, with each country consisting of districts with both positive and negative effects. Kenya, Tanzania and Uganda each contained an isolated district with a considerably lower negative effect. Unlike the cluster-level spatial effect, Malawi displayed significant heterogeneity in this district-level random effect, with one district displaying a notably higher positive effect compared to the rest of the country.

5.5 Ranking and Selection of Districts

Based on the standardised BLUP estimates, the districts were ranked. A negative BLUP is associated with a decreased odds of anaemia in the district, while a positive BLUP is associated with an increased odds of anaemia in the district. The top 3 'best' performing districts (those with the lowest standardised BLUP values) and the top 3 'worst' performing districts (those with the highest standardised BLUP values) were determined for each country (Figure 5.5). The best performing district or county in Kenya was Taita-Taveta County, in Malawi was Mulanje, in Tanzania was Bariadi, and in Uganda was Kiruhura. However, the worst performing district or county in Kenya was West Pokot County, in Malawi was Chikwawa, in Tanzania was Ngorongoro, and in Uganda was Kyenjojo.

5.6 Summary and Discussion

Based on the structure of the surveys and data, a generalised additive mixed model was employed to assess the association between a child's anaemia status and potential individual, household and community level risk factors in Kenya, Malawi, Tanzania and Uganda while accounting for spatial heterogeneity of childhood anaemia. The results revealed significant spatial heterogeneity of childhood anaemia within and between the districts of the four countries. Two sources of spatial heterogeneity were accounted for, that due to spatial dependence of the observations between the sampled clusters, and that due to district-specific factors via the inclusion of a

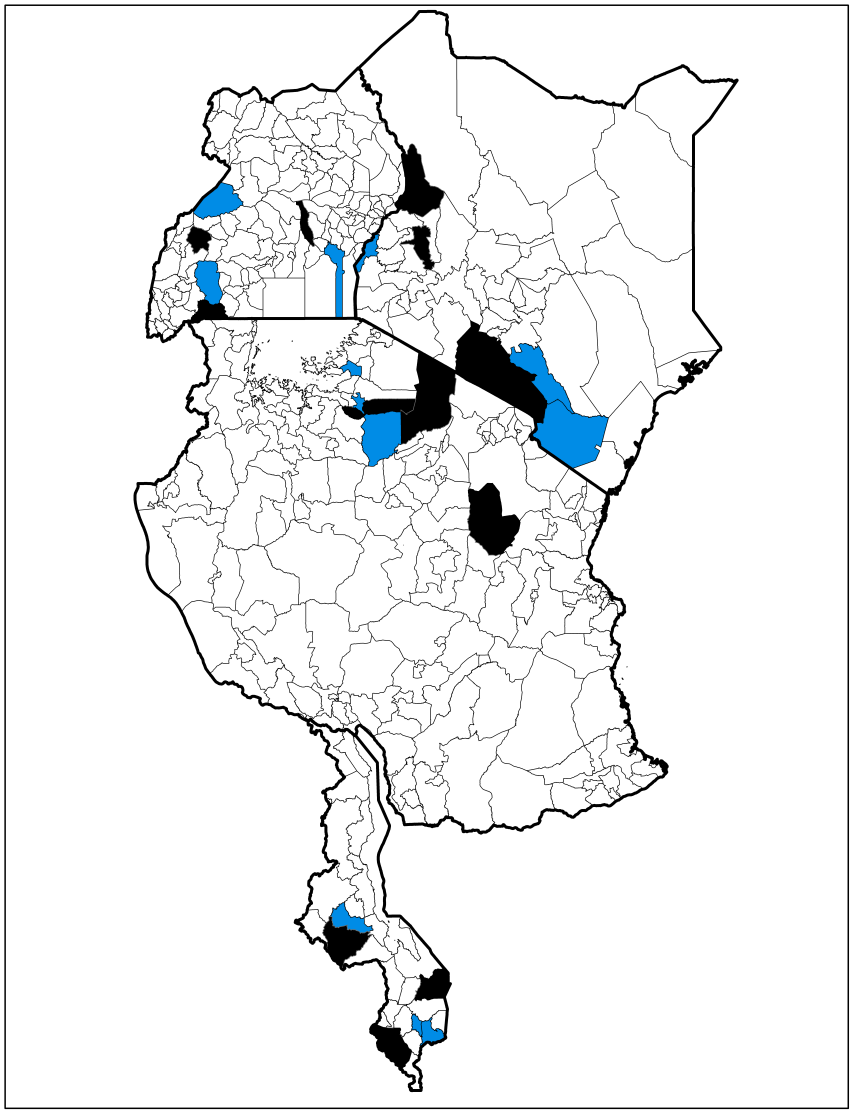


Figure 5.5: Top 3 districts within each country performing the best (in blue) and the worst (in black) with regards to the log-odds of anaemia in children

random effect based on the district of residence. The random and spatial effects are surrogates for influences of unmeasured factors, which may be local (district specific) or global (common between neighbouring clusters or districts), respectively (Ngwira & Kazembe, 2015).

Similar to the results of Chapter 4, the heterogeneity in the district-specific random effect suggests that there are local unobserved factors within each district contribut-

ing to anaemia in children. A further benefit of adding the district of residence as a random effect is that it allows for the ranking of the performance of the districts on the log-odds of anaemia based on the BLUP estimates, after controlling for potential risk factors of anaemia and spatial autocorrelation (Zewotir, 2012). In other words, the BLUP values can be regarded as the estimated effect that a district has on the log-odds of anaemia due to unmeasured factors. It would not have been possible to rank the performance of the districts if the district of residence was added as a fixed effect, which would have resulted in 369 indicator variables for the 370 districts in the model. Not only does this ranking procedure allow for the worst performing districts to be targeted in order to improve their anaemia control strategies, but it also allows for the best performing districts to be identified in order to further determine why they are performing better, and then to use these districts as examples in efforts to overcome childhood anaemia.

The cluster-level spatial effect allows one to observe any spatial dependence or heterogeneity within the districts of the countries, where many of the districts had more than one sampled cluster. An advantage of incorporating this spatial effect at a cluster level rather than at a district level, is that a district-level spatial effect aggregates the effect of spatial autocorrelation, which may result in missing some important information. This was evident by some districts within Kenya, Uganda and Tanzania containing both clusters associated with a lower as well as a higher risk of childhood anaemia. This is further indication that strategies for anaemia control should be tailored to what's happening within a specific district.

After accounting for the apparent spatial heterogeneity, child level characteristics (gender, malaria RDT result, and mother's highest education level), household level characteristics (household size, household's wealth index Z-score, the type of toilet facility available, and the type of place of residence) as well as the country of residence were found to be significantly associated with the child's anaemia sta-

tus. These findings are generally in agreement with that in the literature (Moschovis et al., 2018; Zhao et al., 2012; Al-Qaoud et al., 2015; Gari et al., 2017; Khan et al., 2015; Soares Magalhães et al., 2013a; Habyarimana et al., 2017; Mainardi, 2012) as well as the results of the geoadditive model from Chapter 4.

So far we have incorporated the child's malaria status as a covariate into the models. The results indicated a highly significant association between anaemia and malaria, where children with malaria were over 4 times more likely to have anaemia. Many other studies have considered the determinants of anaemia and malaria in children separately (Kuziga et al., 2017; Roberts & Matthews, 2016; Kateera et al., 2015), and others have considered them as determinants of each other where children who tested positive for malaria were more than 3 times as likely to have anaemia. On the other hand, researchers have reported that those with anaemia were more than twice as likely to have malaria (Ugwu & Zewotir, 2018; Wirth et al., 2016; Kweku et al., 2017). This demonstrates the association between the two outcomes, however modelling the two jointly would reveal more about their relationship. Next, we consider joint modelling of anaemia and malaria in young children to further gain insight into this relationship.

CHAPTER 6

MODELLING OF CHILDHOOD ANAEMIA AND MALARIA JOINTLY

The relationship between malaria and anaemia can be confounded by several factors, including nutritional deficiencies (specifically iron deficiency) and intestinal parasites, all of which contribute to anaemia in children (White, 2018). In addition, childhood anaemia as an indicator for the burden of malaria would only be appropriate and useful in regions where malaria is the primary driver of anaemia. Therefore, investigating the relationship between anaemia and malaria and how the relationship changes according to a geographical location would be valuable in this regard. Structural equation modelling is a common multivariate technique used to explore the relationship among numerous variables. However, such a technique cannot incorporate spatial effects and thus would not be suitable for this study. We thus consider a joint model approach.

Joint modelling has several advantages over univariate analyses, which include improved control over Type I error rates during multiple testing and efficiency in estimating parameters (Ayele et al., 2014). Further, through joint modelling, the correlation between the outcomes can be quantified and controlled for. Several approaches to joint modelling exist. The most common approach is the use of a multivariate model, where the univariate models for each response are combined through the specification of a joint multivariate distribution for the random effects Ayele et al. (2014); Habyarimana et al. (2016). Copula regression is another approach to simultaneously modelling multiple outcomes, where a copula function is used to separate the marginal distributions from the dependence structure of a given multivariate distribution Nelsen (2006).

Joint modelling can also be extended into disease mapping through spatial modelling. This aids in gaining more insight into the geographical variation of each of the multiple diseases, while accounting for the association between them. Spatial mapping of single diseases is a well established method for identifying the geographical locations that are most at risk, thus creating a more effective delivery system of limited resources (Kazembe et al., 2009; Besag et al., 1991; Roberts et al., 2020; Gayawan et al., 2014; Habyarimana et al., 2017). Such an approach for joint spatial modelling includes the multivariate conditional autoregressive (MCAR) model (Gayawan & Fadiji, 2020; Adeyemi et al., 2019). This approach allows one to assess and visualise the residual spatial effect of the geographical location on each response, while controlling for the correlation between the responses. However, this MCAR approach does not allow one to assess how the correlation between the responses changes based on the geographical location.

The spatial extension of the copula approach to joint modelling of multiple responses can aid in answering questions about how the association between the responses varies according to the geographical location. Thus, a joint copula regression model is used to explore the correlation between anaemia and malaria in young children across the districts of Kenya, Malawi, Tanzania and Uganda, while accounting for the effects of socio-economic, demographic and environmental factors as well as spatial variation in the two responses. The association parameter between the two responses is varied according to the district of residence across the four countries. The results are then mapped to visualise the relationship between the two responses across the districts. To our knowledge, no studies have jointly modelled anaemia and malaria in children in these four countries. Thus, this analysis contributes to a better understanding of the relationship between anaemia and malaria in children in these regions of sub-Saharan Africa.

6.1 Copula Geoadditive Model

Here we make use of a bivariate copula regression model to jointly model anaemia and malaria. The model is based on a pair of responses and a copula specification for the dependence structure between the two responses (Klein et al., 2019). Copulas are functions that enable the separation of the marginal distributions from the dependence structure of a given multivariate distribution (Nelsen, 2006). The application of copula regression is diverse. McNeil et al. (2015) demonstrated its use in quantitative risk management, Smith et al. (2010), Madsen & Fang (2011), and Kürüm et al. (2018) extended the application of copula regression to longitudinal data, where the approach used by Kürüm et al. (2018) allowed for the model parameters to vary with time. de Leon & Chough (2013) discuss further applications of copula regression to jointly model discrete as well as mixed outcomes. In addition, copula regression is commonly used in finance and insurance (Nelsen, 2006; Umberto, 2011; Kolev et al., 2006, and references therein).

Bivariate Copula Regression

Suppose Y_{i1} is the anaemia status of the i^{th} child and Y_{i2} is the malaria status of the i^{th} child. In this study, each response is binary where $Y_{ij} = 1$ if the child had anaemia or malaria, otherwise $Y_{ij} = 0$, $j = 1, 2$. The joint probability of event $(Y_{i1} = 1, Y_{i2} = 1)$, conditional on a set of covariates $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$, is defined as

$$P(Y_{i1} = 1, Y_{i2} = 1|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}) = C(P(Y_{i1} = 1|\boldsymbol{x}_{i1}), P(Y_{i2} = 1|\boldsymbol{x}_{i2}); \theta).$$

$C : [0, 1]^2 \rightarrow [0, 1]$ is a two-place copula function and θ, known as the copula parameter, is an association parameter which measures the dependence between the two random variables (Marra & Radice, 2017a). If Y_{i1} and Y_{i2} were both continuous, the copula C would be unique. However, in the case of both outcomes being binary, the copula is no longer uniquely defined (Klein et al., 2019). As such, we make use of the latent (unobserved) variable representation of binary models where we define a

continuous latent variable $Y_{ij}^* = \eta_{ij} + \varepsilon_{ij}$, where η_{ij} is the linear predictor consisting of fixed and random effects as well as non-linear and spatial effects, and ε_{ij} is an error term. Therefore, Y_{ij} can be regarded as an indicator variable such that

$$\begin{aligned} P(Y_{ij} = 1|\boldsymbol{x}_{ij}) &= P(Y_{ij}^* > 0|\boldsymbol{x}_{ij}) \\ &= P(\eta_{ij} + \varepsilon_{ij} > 0|\boldsymbol{x}_{ij}) \\ &= P(\varepsilon_{ij} > -\eta_{ij}|\boldsymbol{x}_{ij}) \\ &= 1 - F_j(-\eta_{ij}), \end{aligned} \tag{6.1}$$

where $F_j(\cdot)$ is the cumulative distribution function (CDF) of a standardised univariate distribution (Marra & Radice, 2017a). The copula approach allows for the specification of different families for each marginal distribution. We used the standard normal distribution for the marginal distribution of each latent response variable Y_{ij}^*, leading to a probit model. Although using a logit link would not lead to different conclusions, we selected the probit specification as it is computationally less demanding. Equation 6.1 can be represented as

$$P(Y_{ij} = 1|\boldsymbol{x}_{ij}) = \Phi(\eta_{ij}),$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution. Therefore, a unit increase in the covariate x_{ijk} leads to a β_{jk} increase in the Z-score for the probability of $Y_{ij}^* = 1$. Thus, higher values of the estimated coefficients mean that the event is more likely to happen.

Marginal Model Specification

For each marginal model, we considered the non-linear effects of the continuous covariates. We incorporated an independently and identically distributed random effect based on the district in which the child resided. This random effect, also referred to as an unstructured spatial effect, accounts for the correlation in the observations due to unmeasured district-specific factors. In other words, it accounts for

the possibility that children residing in the same district would be more alike than those from different districts. In addition, we further accounted for spatial variation and spatial autocorrelation in the observations by incorporating a structured spatial effect, which accounts for the assumption that children residing in neighbouring districts are more likely to have correlated observations. We also incorporated fixed effects of all the categorical variables as well as the continuous covariates that did not display a strong non-linear effect on each response. The resulting model for each response takes the form of a geoadditive mixed model, which is an extension of a generalised additive mixed model (GAMM) (Lin & Zhang, 1999). Each marginal model can consist of different effects. The non-linear effects were estimated by smooth functions using a regression spline approach, and the structured spatial effect was estimated using a Markov random field smoother, which was based on the neighbourhood structure of the districts across the four countries. Two districts are considered neighbours if they share a border. More information on the specification and estimation of each marginal model can be found in Klein et al. (2019).

Copula Specification

An advantage of the copula approach to joint modelling is that the selection of the copula for modelling the dependence between the outcomes is independent of the choice of the marginal distributions (Brunner et al., 2019). Several different types of copulas exist, of which the most common are discussed in Nikoloulopoulos & Karlis (2008) and Marra & Radice (2017b). To choose the most appropriate copula, information criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used, where the copula producing the lowest of these values is selected. In our study, the Frank copula produced the smallest AIC value and thus was selected to jointly model our responses. The Frank copula is of the Archimedean class and has the following form

$$C(F_1(Y_{i1}), F_2(Y_{i2}); \theta) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta \times F_1} - 1)(e^{-\theta \times F_2} - 1)}{e^{-\theta} - 1} \right].$$

The copula parameter, θ, is not straightforward to interpret. Therefore, it can be converted into the Kendall correlation coefficient, or Kendall's Tau ($\tau \in [-1, 1]$), which is a measure of the degree of concordance (Marra & Radice, 2017a). For the Frank copula, τ can be obtained by solving

$$\frac{D_1(\theta) - 1}{\theta} = \frac{1 - \tau}{4},$$

where

$$D_1(\theta) = \frac{1}{\theta} \int_0^\theta \frac{t}{e^t - 1} dt.$$

If $\tau = 0$, then Y_{i1} and Y_{i2} are independent. The Frank copula is comprehensive, which means it covers the full spectrum of possible values of τ, which is not the case for all copulas (Winkelmann, 2011).

The copula parameter, θ, may also vary according to different groups of observations. Therefore, θ can be specified as a function of a linear predictor, such as $\theta_i = m(\eta_{i3})$, where m is a one-to-one transformation that ensures that θ_i lies in its range, and η_{i3} is the linear predictor associated with the copula parameter (Marra & Radice, 2017a). The transformation applied depends on the specified copula function. This framework allows one to explore the association between the two outcomes according to the levels or categories of certain factors. In this study, we varied the copula parameter according to the district of residence to enable us to determine the districts in which there is a strong association between anaemia and malaria. Conversely, we are also able to determine the districts in which the association is weak, therefore suggesting that there are other significant drivers of anaemia in children in those districts.

We used the R package *GJRM* (Generalised Joint Regression Modelling) for the analysis (Marra & Radice, 2017c). The mapping of the results was done in QGIS 3.4 (https://qgis.org/en/site/index.html) and all the maps created were based on our

results by making use of shapefiles freely available from the DHS Program's Spatial Data Repository (https://spatialdata.dhsprogram.com/boundaries).

6.2 Results of the Copula Geoadditive Model

For this analysis, data from two districts were merged. These included Songea Urban and Songea DC in Tanzania, where Songia Urban is encompassed by Songia DC. Therefore, we considered a total of 369 districts. In addition, observations corresponding to 'unknown' toilet facilities were removed due to the sparsity of malaria and anaemia cases for this category. Thus, these results are based on 18196 observations. Following the results of the exploratory data analysis from Chapter 3, we did not consider the age of the household head as a predictor of anaemia as well as malaria. Furthermore, the child's gender and household head's gender were not considered as predictors of malaria. Once again, as we considered the spatial effects at district-level, the country of residence was not incorporated as a predictor.

Table 6.1 presents the results of the fixed effects for each marginal model. Based on these results, children residing in rural areas had a lower likelihood of malaria compared to those residing in urban areas, however there was no significant difference in the likelihood of anaemia between these children (Rural estimate = -0.020, p-value = 0.535 for anaemia; Rural estimate = 0.299, p-value $<.001$ for malaria). The likelihood of each outcome significantly decreased with an increase in the mother's highest education level.

Table 6.1: Parameter estimates, standard errors and p-values of the fixed effects for the bivariate copula regression model for anaemia and malaria

Variable	Anaemia			Malaria		
	Estimate	St. Error	P-value	Estimate	St. Error	P-value
Gender (ref = Male)						
Female	−0.083	0.019	<.001*		NA	
Type of Place of Residence (ref = Urban)						
Rural	−0.020	0.032	0.535	0.299	0.047	<.001*
Mother's Education Level (ref = No Education)						
Primary	−0.115	0.031	<.001*	−0.125	0.039	0.001*
Secondary and Higher	−0.164	0.042	<.001*	−0.250	0.057	<.001*
Unknown	−0.095	0.039	0.016*	0.012	0.049	0.802
Gender of Household Head (ref = Male)						
Female	0.011	0.024	0.633		NA	
Type of Toilet Facility (ref = No Facilities)						
PIT Latrine	−0.158	0.035	<.001*	−0.078	0.043	0.072
Flush Toilet	−0.165	0.062	0.008*	0.102	0.114	0.366
Household Size	0.009	0.003	0.006*	0.001	0.004	0.705
Wealth Index	−0.158	0.019	<.001*	−0.503	0.029	<.001*
Cluster Altitude (in 100 metres)	−0.016	0.005	0.002*	−0.089	0.009	<.001*
EVI	0.068	0.057	0.229	0.405	0.121	0.001*
LST	0.011	0.015	0.452	0.019	0.033	0.563

NA: Not applicable as the factor was not incorporated into the marginal model for that response
*significant at 5% level of significance

Type of toilet facilities was significantly associated with a child's anaemia status, but not their malaria status, where the likelihood of anaemia decreased with an improvement of the toilet facility type (PIT Latrine estimate = -0.158, p-value $<.001$; Flush Toilet estimate = -0.165, p-value = 0.008 for anaemia). An increase in the number of household members resulted in a significantly higher likelihood of anaemia, however it had no significant effect on a child's malaria status (Household size estimate = 0.009, p-value = 0.006 for anaemia; Household size estimate = 0.001, p-value = 0.705 for malaria). A unit increase in the household's wealth index Z-score was associated with a significant decrease in the likelihood of each anaemia and malaria (Wealth index estimate = -0.158, p-value $<.001$ for anaemia; Wealth index estimate = -0.503,

p-value $<.001$ for malaria). Cluster altitude was significantly associated with each response, where the likelihood of each decreased with an increase in altitude (Cluster altitude estimate = -0.016, p-value = 0.002 for anaemia; Cluster altitude estimate = -0.089, p-value $<.001$ for malaria). EVI was significantly associated with only malaria, where an increase resulted in an increased likelihood of malaria (EVI estimate = 0.405, p-value = 0.001 for malaria). LST was not significantly associated with either response.

Table 6.2 displays the significance of the non-linear and spatial effects for both responses. Both the structured spatial effect and unstructured spatial effect (the district-level random effect) had a significant effect on the likelihood of each response. Further, the child's age in months had a significant non-linear effect on the likelihood of each response.

Table 6.2: Approximate significance for the non-linear and spatial effects

Variable	Anaemia		Malaria	
	χ^2 Value	P-value	χ^2 Value	P-value
Child's age in months	1472.50	$<.001$*	138.49	$<.001$*
Unstructured spatial effect	357.70	$<.001$*	34.75	$<.001$*
Structured spatial effect	183.80	$<.001$*	1412.17	$<.001$*

*significant at 5% level of significance

Figure 6.1 displays this non-linear effect that a child's age in months had on anaemia and malaria. The likelihood of anaemia decreased with an increase in age. However, there was a reverse effect of age on malaria, where the chance of malaria increased with an increase in age.

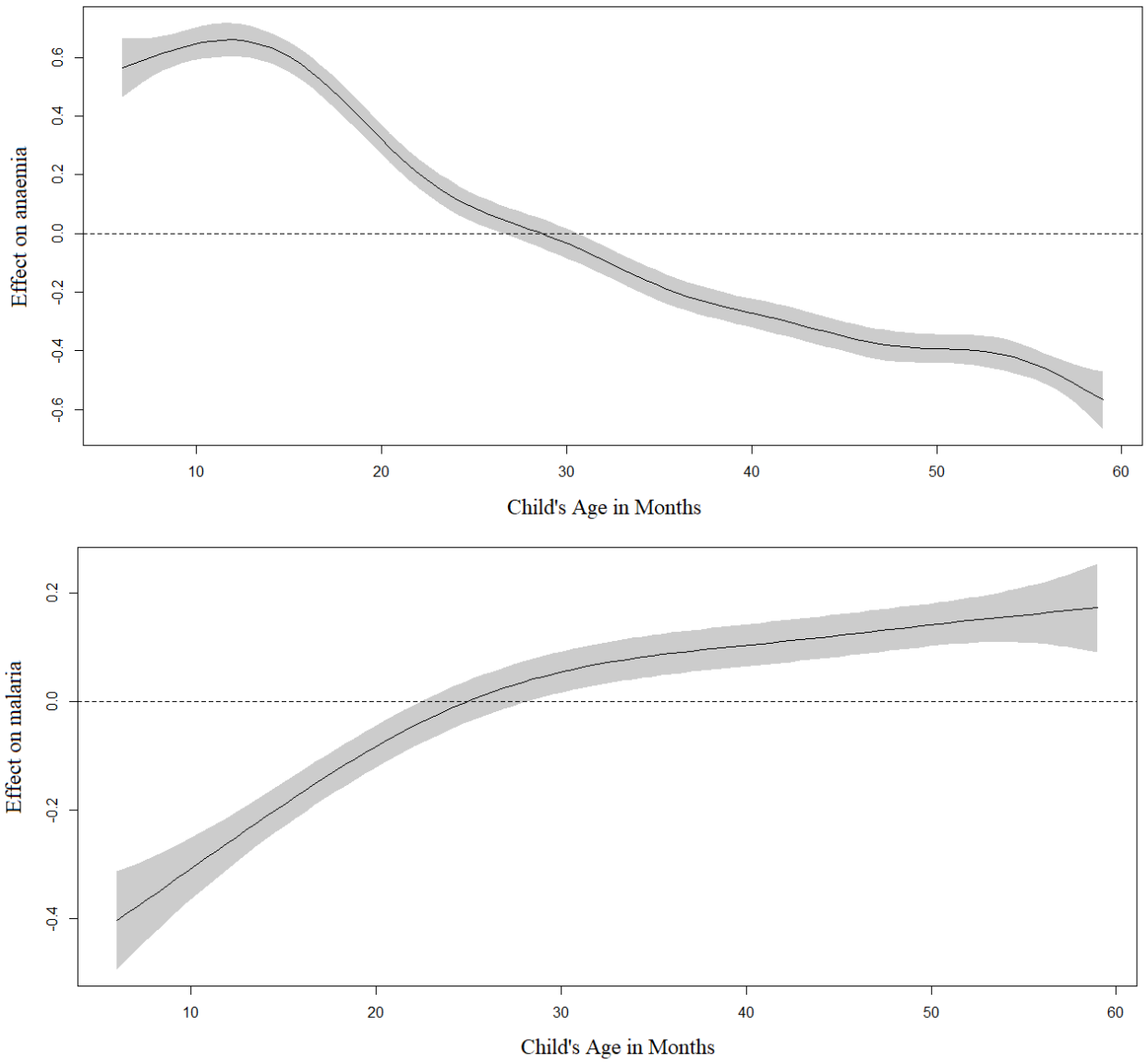


Figure 6.1: Estimated non-linear effect of the child's age on anaemia (top) and malaria (bottom) together with the 95% confidence intervals

The district-level structured spatial effect for both anaemia and malaria is presented in Figure 6.2. The districts in shadings of blue correspond to a negative estimated effect and were therefore associated with a lower likelihood of the event. However, districts in shadings of red correspond to a positive estimated effect and were therefore associated with a higher likelihood of the event. There was a lot less variation observed in the structured spatial effect for anaemia compared to that for malaria. The structured spatial effect for malaria revealed that Tanzania consisted of districts

associated with a lower likelihood of malaria as well as districts associated with a higher likelihood of malaria. This apparent spatial variation suggests that it was important to control for as failure to do so would reduce the statistical power of inference in the model and therefore lead to inaccurate results (Mainardi, 2012).

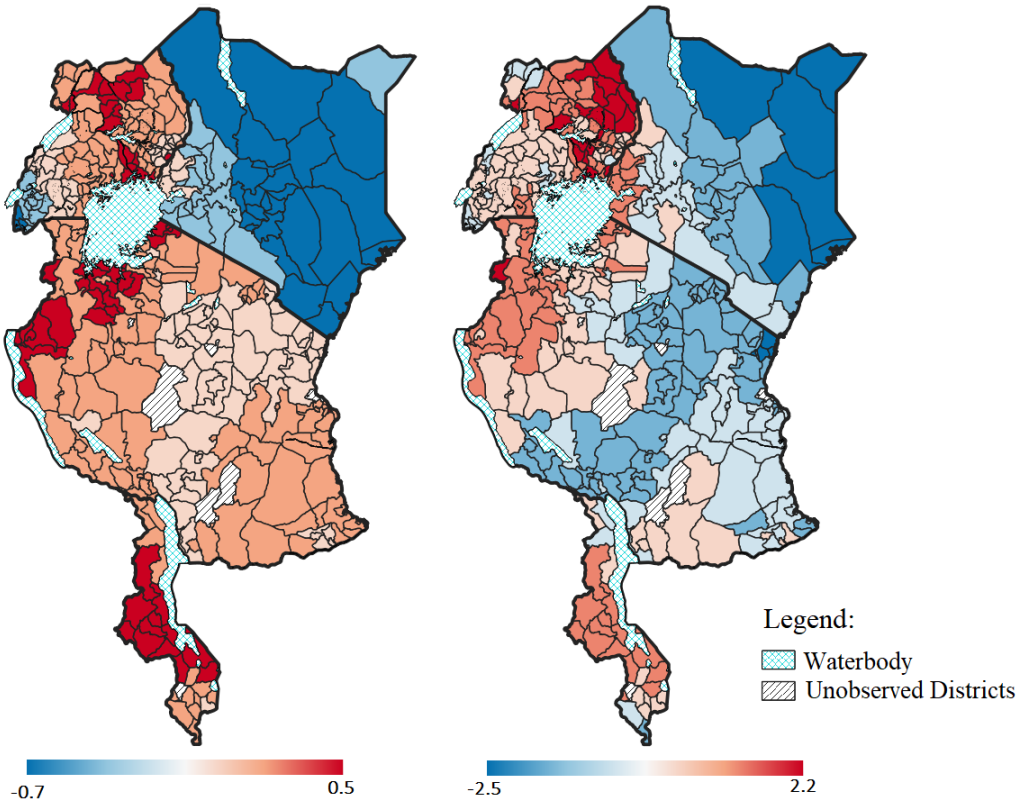


Figure 6.2: Estimated effect of the structured spatial effect on anaemia (left) and malaria (right).

6.3 Conditional Dependence of Anaemia and Malaria

The copula parameter was set to vary according to the district/county of residence across the four countries. This was done by linking the additive predictor for the copula parameter to a Markov random field term based on these districts of residence. The estimated value of the copula parameter, averaged out over the districts, was 3.07 with a 95% confidence interval of (1.56, 4.61). This copula parameter, which was estimated conditioned on the observed covariates and spatial variation, was then used to estimate Kendall's τ for each district as shown in Figure 6.3.

This figure displayed a fairly heterogeneous, non-zero association between anaemia and malaria in young children across the districts. With using the Frank copula, we allowed for positive and negative associations between anaemia and malaria. However, Kendall's τ ranged between 0.09 and 0.41, with an average of 0.31 and a 95% confidence interval of (0.16, 0.42). Thus, there was a positive association between malaria and anaemia. A stronger association was observed in some districts compared to others. Kenya depicted more districts with the highest association.

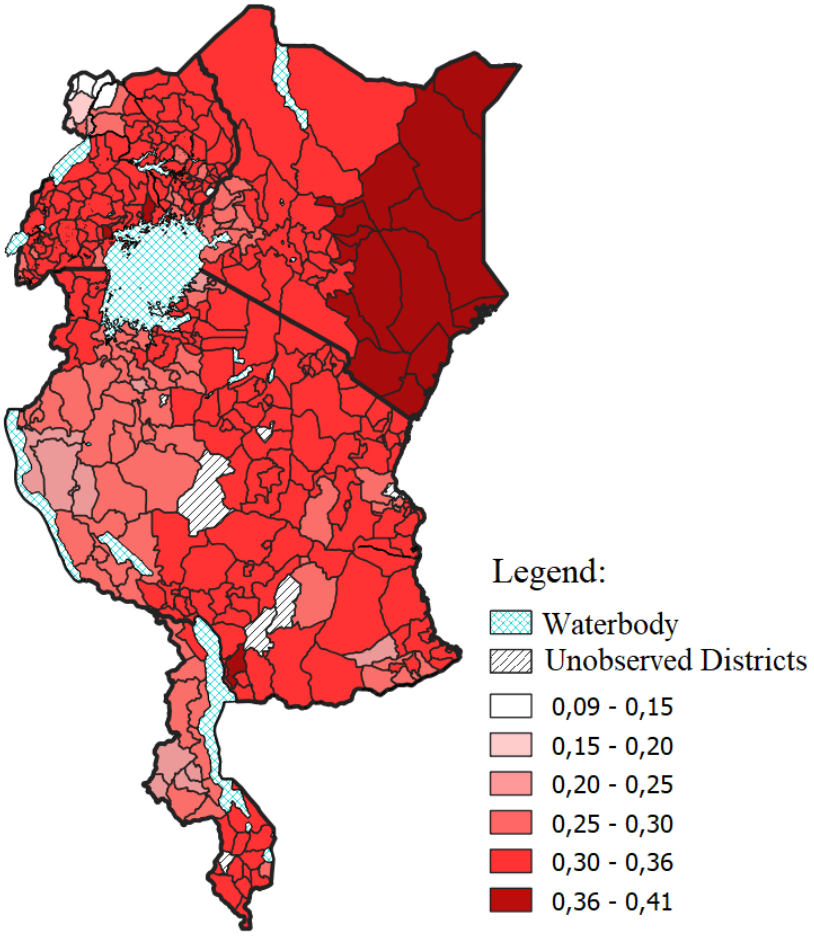


Figure 6.3: Estimated Kendall's τ according to district of residence.

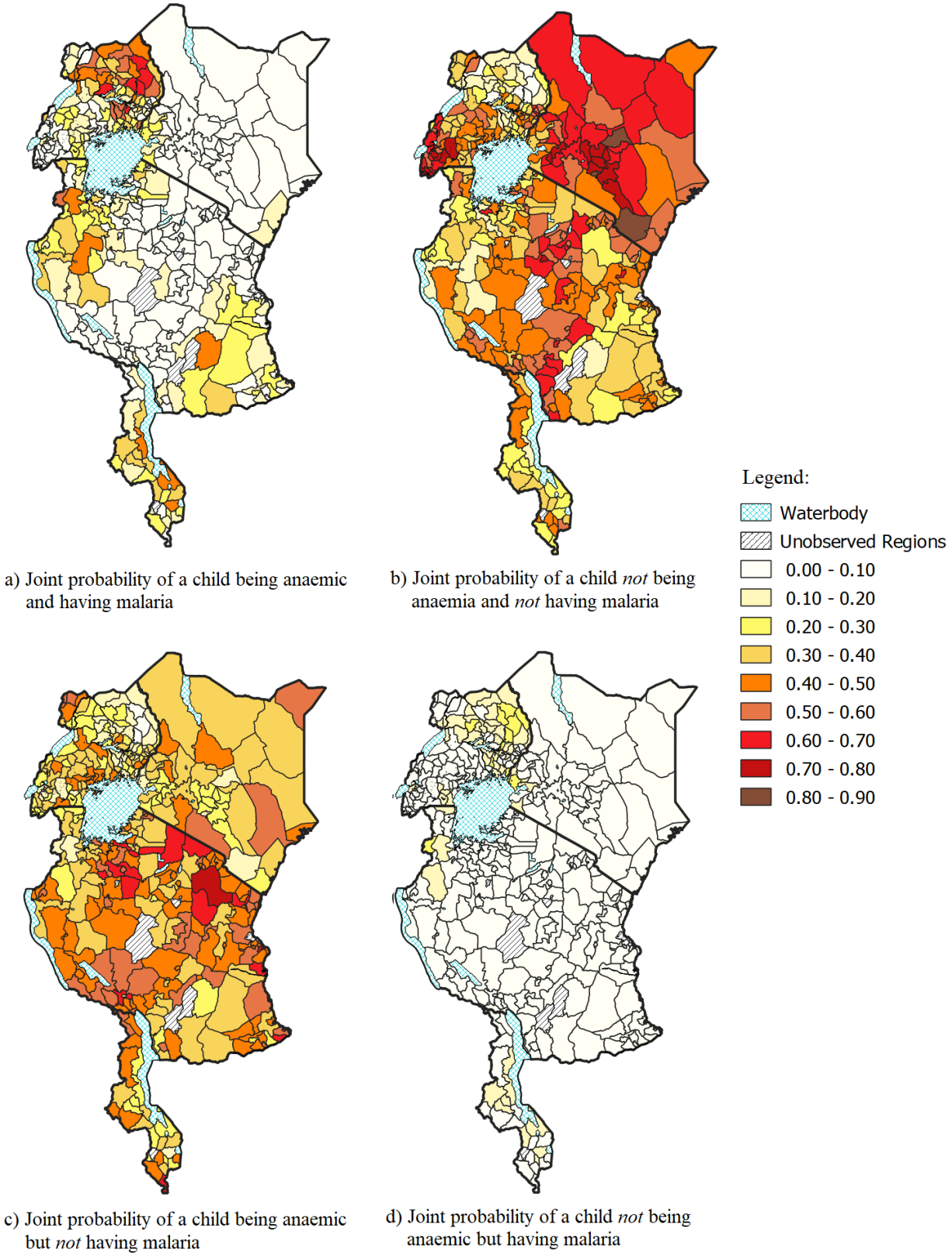
Figure 6.3 suggests that the probability of a child being anaemic or having malaria in a particular district should be based on the joint probability from the bivariate model rather than each independent univariate model. These joint probabilities can further reveal more about the relationship between anaemia and malaria in children

across the districts of the four countries.

Estimated Joint Probability of Anaemia and Malaria

Based on the fitted bivariate copula regression model, the estimated joint probabilities were extracted and averaged over the districts. Figure 6.4 shows these joint probabilities for each combination of outcome for anaemia and malaria in young children. On the whole, these joint probabilities were generally heterogeneous within each country.

Considering image a) in Figure 6.4, a large number of districts in Uganda showed a considerably high joint probability of a child having anaemia and malaria, particularly in the north/north east of the country. Kenya was homogeneous in these probabilities, which were also all fairly low (all were below 0.20). Malawi had a few districts with a relatively high probability of both anaemia and malaria in children. From image b), we can observe that the majority of districts in Kenya had a high probability of a child not having anaemia nor malaria. This is unsurprising as Kenya also had the lowest observed prevalence of anaemia and malaria. Paying particular attention to image c) in Figure 6.4, throughout the districts considered in each country, there were a fair number that displayed a high chance of a child having anaemia but not malaria. In these districts, it would be inaccurate to use anaemia as an indicator for malaria as this image suggests that there are other significant drivers of anaemia in children in these districts. Image d) revealed very low probabilities of a child having malaria but not anaemia throughout the majority of the districts. In other words, it is highly unlikely for a child to have malaria but not anaemia in these districts. Thus, it is clear that there is a high likelihood of a child developing anaemia when they have malaria. Based on images a) and d), districts in the northern part of Uganda had a relatively high probability of a child having malaria, regardless of anaemia status. This is also supported by Uganda having the highest observed prevalence of malaria.



a) Joint probability of a child being anaemic and having malaria

b) Joint probability of a child *not* being anaemia and *not* having malaria

c) Joint probability of a child being anaemic but *not* having malaria

d) Joint probability of a child *not* being anaemic but having malaria

Figure 6.4: Estimated joint probabilities based on the bivariate copula regression model.

6.4 Summary and Discussion

Here we aimed to explore the relationship between anaemia and malaria in young children across the districts/counties of Kenya, Malawi, Tanzania and Uganda by making use of a joint bivariate copula regression model. This approach allows the correlation between the two responses to be estimated while controlling for the linear and non-linear effects of independent variables, as well as the effect of spatial variation. The copula framework allows the dependency structure between the responses to be isolated from their marginal distributions. The advantage of copula regression over multivariate analysis is that normality and linearity of the dependence between the responses is not assumed. In fact, in general, dependence in copulas is non-linear (Winkelmann, 2011). Further, the appeal of the copula approach is that one is able to vary the association between the responses according to the different levels of certain factors, rather than obtaining one estimated value for the correlation as is the case with a joint multivariate model (Gari et al., 2017).

We varied the association according to the district of residence. This revealed a positive association between anaemia and malaria throughout the districts, however the strength of which varied across the districts of the four countries. Some districts had a stronger association between the two responses compared to other districts. While we are interested in the likelihood of a child having both anaemia and malaria, considering the likelihood of all combinations of outcomes of these events can further aid in better understanding the relationship between anaemia and malaria. Therefore, we made use of the estimated joint probabilities for the combination of outcomes, which we mapped across the districts. These maps generally indicated a variation in the joint probabilities within each country. This suggests that any approach to anaemia or malaria control should be targeted rather than a country wide approach. Districts in the north to north east part of Uganda displayed high probabilities of a child having malaria, for both those with or without anaemia. These districts need an up-scaled targeted approach to malaria control. Districts in Kenya

showed the least amount of variation in some of the joint probabilities and also had the lowest joint probability associated with a child having malaria, for those with or without anaemia. This is as a consequence of the major progress that Kenya has made in the fight against malaria, which is most likely owed to the recent malaria prevention measures that have been tailored to local needs (WHO, 2017).

If anaemia is to be used as an indicator for the success of malaria control programs, in any country, it would only be useful in areas where there is a strong correlation between anaemia and malaria as well as a high probability of the two. Thus, the maps created in this study aid in identifying such areas. In addition, based on the map of the joint probability of a child having anaemia but not malaria, a high likelihood of this event was revealed in many of the districts. In such districts, it would be reasonable to assume that there are other drivers of anaemia in children, other than malaria. Therefore, applying malaria interventions in these districts to aid in the reduction of the prevalence of childhood anaemia would be ineffective. Further investigation into the drivers of childhood anaemia in these districts is therefore required.

The results of the effects considered in this study are consistent with those from other studies that modelled anaemia and malaria separately, where the child's age, mother's education level, household wealth index and cluster altitude were significantly associated with both anaemia and malaria status (Roberts & Matthews, 2016; Kateera et al., 2015; Khan et al., 2015; Gayawan et al., 2014). The child's gender, the household size and type of toilet facility were further significantly associated with anaemia in children, as seen in other studies (Goswmai & Das, 2015; Zhao et al., 2012).

Very few studies have jointly modelled anaemia and malaria. The studies that have done so, have also utilised different techniques and thus answered different ques-

tions (Seyoum, 2018; Adebayo et al., 2016). A bivariate probit model was used to jointly model anaemia and malaria in individuals between the ages of 15 and 60 in Alaba District, Southern Ethiopia, the result of which showed a positive correlation between malaria and anaemia, however the magnitude of the correlation was not explored (Seyoum, 2018). Similar to our study, (Adebayo et al., 2016) jointly modelled anaemia and malaria in children under five in Nigeria and found substantial geographical variations in the likelihood of malaria, however the association between anaemia and malaria was not directly explored.

As multiple factors were significantly associated with both anaemia and malaria, accordingly we propose further varying the association parameter by the levels of these factors. For example, the additive predictor for the copula parameter can include the effects of the mother's education level in addition to the district-level structured spatial effect. The correlation and joint probabilities can then be estimated according to the levels of the additional factors, which will further reveal more about the relationship between anaemia and malaria.

The copula technique applied here provides an alternative to joint modelling of anaemia and malaria in young children which assists in understanding more about their relationship compared to techniques of multivariate modelling. This approach aids in visualising the relationship through mapping of their correlation and joint probabilities. However, a short-coming of the copula geoadditive model is that it is not able to inform us which geographical locations contribute to a higher or lower likelihood of both diseases simultaneously. This leads to a shared component model (SCM), in which the spatial effect is decomposed into a shared and disease-specific spatial effect.

CHAPTER 7

SHARED COMPONENT MODELLING OF CHILDHOOD ANAEMIA AND MALARIA

We considered a shared component model for the joint spatial analysis of anaemia and malaria in children in the four countries, where both the shared and disease-specific district-level spatial effects are estimated while controlling for known risk factors. This allows the districts of high risk of one or the other, or both diseases to be identified for a more targeted approach to anaemia and malaria control and prevention as well as for a targeted allocation of limited district health system resources.

7.1 Shared Component Model

The shared component model (SCM) was originally proposed by Knorr-Held & Best (2001) to jointly model the spatial variation of rates of several diseases with common risk factors. The SCM allows for the underlying risk surface of the diseases to be decomposed into two: shared and disease-specific variation. The SCM has been used in a wide variety of applications, such as to identify shared patterns among chronic related preventable hospitalizations (Ibáñez-Beroiz et al., 2011), for joint spatial modelling of common morbidities of childhood fever and diarrhoea in Malawi (Kazembe et al., 2009), and for joint modelling of brain cancer incidence and mortality rates in two regions in the north of Spain (Etxeberria et al., 2018). Recently, the SCM was used to identify crime-general and crime-specific hotspots in a region in Canada (Law et al., 2020).

The SCM is typically used when interest is on the relative risk of two or more diseases in a particular region, where regional level covariates can be incorporated in the model. In this case, the response represents the disease counts for the region. However, we consider the SCM to model the probability, π_{ijk}, of child j residing in district i having anaemia ($k = 1$) or malaria ($k = 2$).

Thus, we make use of logistic regression models given by

$$logit(\pi_{ij1}) = \alpha_1 + \boldsymbol{x}'_{ij1}\boldsymbol{\beta}_1 + \delta u_i + v_{i1},$$
$$logit(\pi_{ij2}) = \alpha_2 + \boldsymbol{x}'_{ij2}\boldsymbol{\beta}_2 + \frac{u_i}{\delta} + v_{i2},$$

where α_k, $k = 1, 2$, are the disease specific intercepts; $\boldsymbol{\beta}_k$ is the vector of regression parameters corresponding to the covariates $\boldsymbol{x}'_{ijk}$ for the k^{th} disease, where such covariates comprise of child-level, household-level and environmental factors; u_i is the disease-general shared spatial component common to both diseases; and v_{ik} is the disease-specific spatial component which captures the spatial patterns that deviate from the shared spatial component. Both the shared and specific spatial components were considered at district level, where a total of 369 districts across the four countries were included. δ is referred to as the partitioning weight and allows for a different odds gradient of the shared component. The advantage of our approach to the SCM is that it enables one to explore the individual-, household-, and community-level risk factors for each disease. Thus, such risk factors are well accounted for in the model.

Figure 7.1 presents a schematic representation of the shared component model for this study. The shared component captures the spatial pattern common to both diseases, where δ allows each disease to have a unique association with this spatial pattern. A value of δ close to one indicates that anaemia and malaria have a similar magnitude of association with the shared spatial pattern, whereas a smaller positive

value of δ indicates that anaemia has a weaker association with the shared spatial pattern compared to malaria (Law et al., 2020). It should be noted that estimating a partitioning weight (δ) for one disease and assigning the inverse to the second disease improves model identifiability compared to estimating separate partitioning weights for each disease (Knorr-Held & Best, 2001; Law et al., 2020).

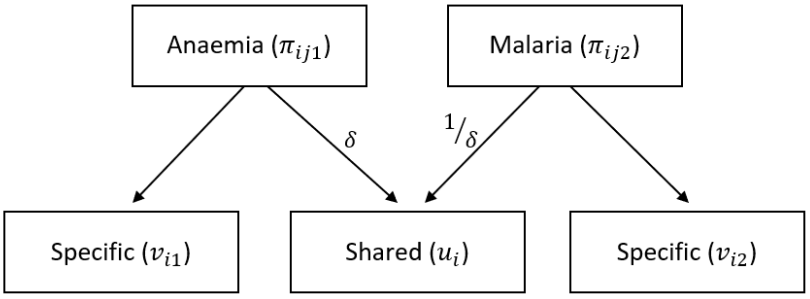


Figure 7.1: Schematic representation of the shared component model for this study

Each of the shared and disease-specific spatial components, u_i and v_{ij}, can be decomposed as

$$u_i = u_{str_i} + u_{unstr_i},$$

$$v_{ij} = v_{str_{ij}} + v_{unstr_{ij}},$$

where u_{str_i} and $v_{str_{ij}}$ are spatially structured effects and u_{unstr_i} and $v_{unstr_{ij}}$ are the random heterogeneity (spatially unstructured) effects. These spatial effects are due to unmeasured factors that have not been controlled for in the model, where such factors may be common among neighbouring districts (and thus contribute to the structured spatial effect) or specific to a district (and thus contribute to the unstructured spatial effect).

A Bayesian approach was used to fit the model, where each of the parameters were assigned a prior distribution. Weakly informative $N(0, 10000)$ priors for the regression coefficients β_k were assumed. The spatial components followed a Besag framework (Besag et al., 1991), where the structured spatial effects were assigned intrinsic Gaussian Markov random field (IGMRF) priors, also known as conditional autoregressive (CAR) priors. This prior assumes that the structured spatial effect of the

districts follow a normal distribution with a conditional mean equal to the average of the neighbouring districts' effects and a conditional variance inversely proportional to the number of neighbours. Two districts are considered neighbours if they share a border. The unstructured spatial effects were assigned assigned i.i.d. Gaussian priors with a mean of zero. The variance components of these spatial effects comprised of unknown precision (inverse variance) parameters that were assigned a Gamma (1, 0.001) hyperprior distribution. The intercepts α_k were assigned flat priors as recommended for a model that includes a CAR random effect (Thomas et al., 2004). In addition, a sum-to-zero constraint was imposed on the spatial effects to allow for model identifiability. The partitioning weight δ was assigned a log-normal distribution with a mean of 0 and variance of 0.169 (Knorr-Held & Best, 2001). This prior then assumes that both δ and $1/\delta$ are both positive, which is a reasonable assumption as there is a positive correlation between anaemia and malaria, as demonstrated in the results of the copula regression from Chapter 6. This prior also assumes that the ratio of δ and $1/\delta$ (i.e. $\delta/(1/\delta)$) is between 0.2 and 5 with a 95% probability, regardless of which disease is labelled 1 or 2 (Knorr-Held & Best, 2001).

The models were fitted using Markov Chain Monte Carlo (MCMC) simulations in WinBUGS version 1.4.3 (Thomas et al., 2003). The WinBUGS program for the area-level SCM was adapted for our child-level SCM. Three parallel MCMC chains with varying starting values were run for a total of 50 000 iterations each. After a burn-in period of 50 000, every 10th sample was retained for posterior inference. Convergence was assessed using the Brooks and Gelman statistic and autocorrelation plots. A sensitivity analysis with various prior and hyperprior specifications was performed. The estimates and their significance remained largely the same. The models were compared using the deviance information criterion (DIC), where the results presented are based on the model with the lowest DIC. The estimated spatial effects were extracted and mapped in QGIS 3.20 (https://qgis.org/en/site/index.html). All of the maps created were based on the results of this study and made use of shapefiles freely available from the DHS Program's Spatial Data Repository (https://spatialdata.dhsprogram.com/boundaries).

7.2 Fixed Effects Results

A restriction of the SCM is that the same predictors need to be considered for both responses, unlike the copula approach which can incorporate different effects into the model for each response. Thus, these results include the effect of the child's gender and household head's gender on both responses, even though they were not considered as predictors of malaria in Chapter 6.

Table 7.1 presents the adjusted posterior odds ratios (AOR) and corresponding 95% credible intervals for the fixed effects. The results of the previous chapters revealed a non-linear effect of the child's age in months on the likelihood of anaemia, where there was an increase in the likelihood for children younger than 12 months followed by a decrease in the likelihood for children aged 12 months and older. However, due to the limitations of WinBUGS, the effect of age was incorporated as a linear fixed effect where it was categorised accordingly (under 12 months versus 12 months and older). The child's age had a significant effect on the likelihood of anaemia as well as malaria. However, while the odds of anaemia were substantially lower for those aged 12 months and older (AOR = 0.316; 95% CrI: 0.285-0.351), the odds of malaria were higher for children in this age group compared to those younger than 12 months (AOR = 2.166; 95% CrI: 1.850-2.531). The odds of anaemia were significantly lower for female children compared to male children (AOR = 0.879; 95% CrI: 0.826-0.936). However, there was no significant difference in the odds of malaria between male and female children (AOR = 0.977; 95% CrI: 0.892-1.068). The type of place of residence only had a significant effect on the likelihood of malaria in children, not anaemia, where those residing in rural areas were 1.797 times more likely to have malaria compared to those residing in urban areas (95% CrI: 1.514-2.136). The mother's highest educational level had a significant impact on the odds of anaemia as well as malaria, where the odds of each decreased with an increase in education level. Likewise, there was a significant decrease in the odds of either disease with an increase in the household's wealth index Z-score (AOR = 0.769; 95% CrI: 0.725-0.816 for anaemia, AOR = 0.400; 95% CrI: 0.360-0.443 for malaria).

In addition, there was a significant decrease in the odds of anaemia with an improvement in toilet facilities. However, the type of toilet facility had no significant effect on the odds of malaria in children. The cluster altitude had a decreased effect on both the odds of anaemia and malaria (AOR = 0.969; 95% CrI: 0.953-0.985 for anaemia, AOR = 0.847; 95% CrI: 0.819-0.874 for malaria). The odds of malaria significantly increased with an increase in the environmental factor EVI (AOR = 2.074; 95% CrI: 1.380-3.304), however it had no significant effect on the odds of anaemia (AOR = 1.050; 95% CrI: 0.889-1.268). The gender of the head of household, household size and the environmental factor LST did not have any significant effects on the odds of anaemia or malaria.

Table 7.1: Adjusted posterior odds ratio estimates (AOR) and 95% credible intervals

Variable	**Anaemia**	**Malaria**
	AOR (95% CrI)	**AOR (95% CrI)**
Gender (ref = Male)		
Female	0.879 (0.826, 0.936)*	0.977 (0.892, 1.068)
Age in Months (ref = Under 12 months)		
12 months and older	0.316 (0.285, 0.351)*	2.166 (1.850, 2.531)*
Type of Place of Residence (ref = Urban)		
Rural	0.948 (0.859, 1.047)	1.797 (1.514, 2.136)*
Mother's Education Level (ref = No Education)		
Primary	0.874 (0.793, 0.964)*	0.803 (0.703, 0.915)*
Secondary and Higher	0.852 (0.748, 0.972)*	0.609 (0.498, 0.749)*
Unknown	0.742 (0.654, 0.838)*	1.119 (0.946, 1.329)
Gender of Household Head (ref = Male)		
Female	1.005 (0.931, 1.082)	0.927 (0.828, 1.038)
Type of Toilet Facility (ref = No Facilities)		
PIT Latrine	0.779 (0.697, 0.869)*	0.878 (0.757, 1.017)
Flush Toilet	0.763 (0.624, 0.929)*	1.051 (0.667, 1.620)
Household Size	1.008 (0.998, 1.019)	1.003 (0.990, 1.016)
Wealth Index	0.769 (0.725, 0.816)*	0.400 (0.360, 0.443)*
Cluster Altitude (in 100 metres)	0.969 (0.953, 0.985)*	0.847 (0.819, 0.874)*
EVI	1.050 (0.889, 1.268)	2.074 (1.380, 3.304)*
LST	1.011 (0.969, 1.055)	1.037 (0.922, 1.180)

*significant at 5% level of significance

7.3 Spatial Effects Results

Figure 7.2a presents the estimated effect of the shared spatial component on the log-odds of anaemia and malaria. The districts in blue shadings correspond to a negative estimated log-odds and were therefore associated with a lower likelihood of the disease. Whereas, those in red shadings correspond to a positive estimated log-odds and were therefore associated with a higher likelihood. Notably, there were distinct patterns of clustering among neighbouring districts. In particular, there were clusters associated with increased likelihoods of both diseases in the west of Tanzania and throughout Uganda and Malawi. Kenya primarily consisted of districts/counties associated with decreased likelihoods of both diseases. This shared spatial effect presented a non-random pattern, as suggested by Moran's I statistic of 0.758 (p = 0.001). The partitioning weight (δ) was estimated at 0.626 (95% CrI: 0.429-0.952) (Table 7.2). Thus, malaria had a stronger association with this shared spatial pattern compared to anaemia, as is evident from the shared spatial effects in Figures 7.2b and 7.2c for anaemia and malaria, respectively. Table 7.2 indicates that 82.70% of the spatial variation in the likelihood of anaemia was captured by the shared spatial component, while only 62.43% of the spatial variation in the likelihood of malaria was captured by this component.

Table 7.2: Partitioning weight posterior estimate (95% CrI) and empirical variances

	Anaemia	**Malaria**
Partitioning weight (δ)	0.626 (0.429, 0.952)	1.597 (1.050, 2.331)
Empirical variance of shared component	0.182	1.183
Empirical variance of disease-specific component	0.083	0.712
% of total variation explained by shared component	82.70	62.44

The disease-specific spatial effects for anaemia and malaria are displayed in Figures 7.3a and 7.3b, respectively. Similar to the shared spatial component, this disease-specific spatial effect was more prominent for malaria than for anaemia. In addition, this component explained a higher proportion of the spatial variation in the

likelihood of malaria (37.56%) compared to that for anaemia (17.30%). Once again, both spatial patterns consisted of clusters of increased likelihoods (positive values) and decreased likelihoods (negative values). These patterns were non-random, as confirmed by Moran's I statistic of 0.258 for anaemia and 0.866 for malaria, both of which were significant at a 5% level of significance. Unlike the shared component, there were fewer clusters in the west of Tanzania and in Uganda for the anaemia-specific spatial effect. Multiple districts across the four countries had contrasting effects on the likelihood of anaemia and malaria. More specifically, many of the districts that had a decreased likelihood of malaria, had an increased likelihood of anaemia.

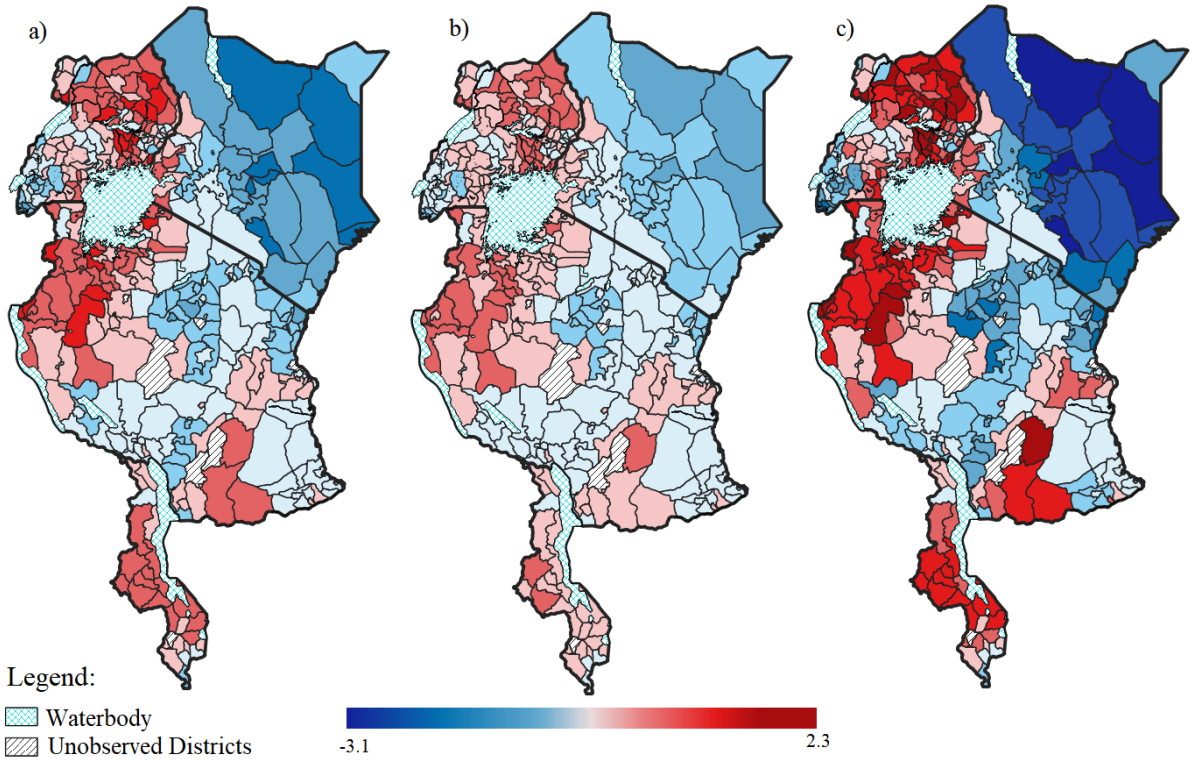


Figure 7.2: Estimated effect of the overall shared spatial component (a); shared spatial component for anaemia (b); and shared spatial component for malaria (c)

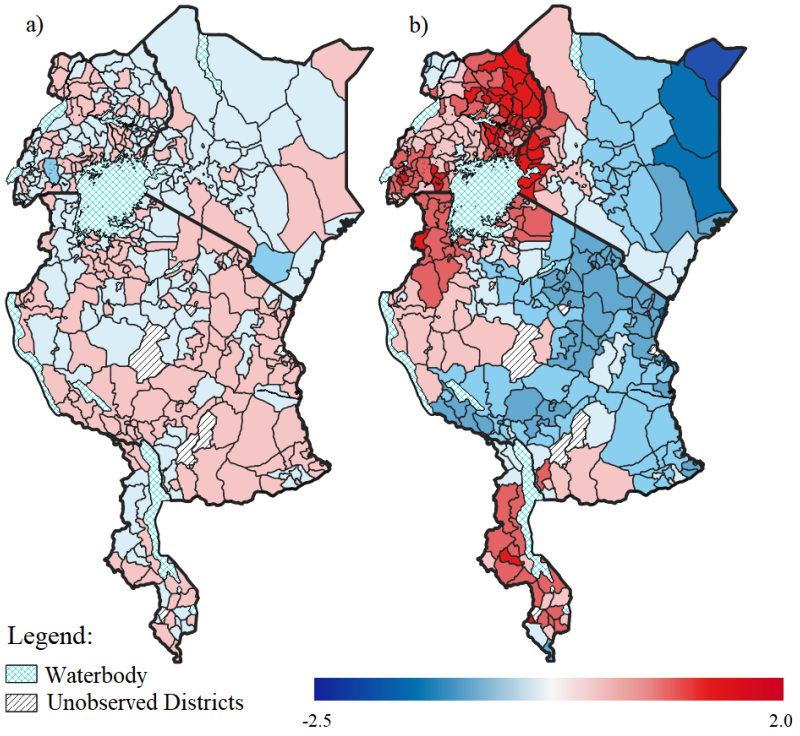


Figure 7.3: Estimated effect of the disease-specific component for anaemia (a) and the disease-specific component for malaria (b)

7.4 Summary and Discussion

We sought to jointly model the residual spatial variation in the likelihood of anaemia and malaria in young children across the districts of Kenya, Malawi, Tanzania and Uganda using a shared component model, while controlling for child-level, household-level and environmental characteristics. The district-level spatial effect for each disease was partitioned into a shared spatial component and a disease-specific spatial component. These spatial components can be considered as proxies for variations in unmeasured factors that contribute to both (shared) or only one (specific) of the diseases (Law et al., 2020). In this study, each of the shared and disease-specific spatial components were further partitioned into structured and unstructured spatial effects to account for unmeasured factors that are shared among neighbouring districts or that are district-specific, respectively.

Malaria had a stronger association with the shared spatial component compared to anaemia, which suggests that the unmeasured risk factors common to both diseases had a higher impact on the likelihood of malaria. The shared spatial pattern revealed significant hotspots of increased likelihoods of each disease in the west of Tanzania and throughout the majority of the districts in Uganda and Malawi. This shared spatial component had a higher contribution to the spatial variation in the likelihood of both diseases compared to the disease-specific spatial component. This suggests that if programs for control and prevention of one of the diseases are targeted in the high risk districts, they should also make an impact on the other disease.

The disease-specific component was more prominent for malaria as well as contributed to a higher proportion of the spatial variation in the likelihood of malaria compared to that of anaemia. This indicates that there are additional unmeasured risk factors relevant to malaria only. One of the consequences of malaria is anaemia (White, 2018). However, while severe anaemia can exacerbate malaria, it does not lead to malaria (Adebayo et al., 2016). It is therefore reasonable to hypothesize that there are other drivers of anaemia in children in the districts that are associated with an increased likelihood of anaemia but a decreased likelihood of malaria based on the disease-specific spatial component. This study identified multiple such districts throughout all four countries.

Of note from the malaria-specific spatial pattern is that the districts with increased likelihoods were clustered around many of the water bodies in the countries, such as Lake Victoria shared by Tanzania, Kenya and Uganda; Lake Malawi; and Lake Turkana in Kenya. It has been suggested that the lake environments, specifically wetlands along the lakeshore, may maintain a high number of malaria vectors (Minakawa et al., 2012). In particular, several vector breeding sites have been found to be associated with Lake Victoria and Lake Malawi (Minakawa et al., 2012; Frake

et al., 2020). Thus, efforts for malaria vector control, such as insecticide-treated nets and indoor residual spraying, should be continued and upscaled in these high risk districts. Such control measures have been noted as the primary driver of the significant reductions in the burden of malaria in sub-Saharan Africa over the past two decades (Guerra et al., 2020). This clustering pattern of increased likelihood of malaria around the water bodies differed for the anaemia-specific spatial component, which had less distinctive patterns. Anaemia is likely driven more by demographic, socioeconomic, and dietary-related factors than environmental factors, as suggested by the fixed effects results, as well as highlighted in other studies which found malnutrition and intestinal parasites to also play a role in childhood anaemia (Rahman et al., 2019; Roba et al., 2016; Soares Magalhães et al., 2013b).

While the focus here was not on determining the significant risk factors of each disease, the SCM allowed us to identify as well as control for such. However, the findings of this analysis regarding the child-level, household-level and environmental factors largely agreed with that of the copula regression from Chapter 6. The only contrasting result was concerning the effect of the household size on the likelihood of anaemia in a child. While the copula regression found this effect to be significant, this current analysis did not. This may be attributed to the different models as well as estimation procedures applied. In summary, the child's age, the mother's education level, the household wealth index, and cluster altitude had a significant effect on the likelihood of both anaemia and malaria. While the type of place of residence was not significantly associated with a child's anaemia status, those residing in rural areas had a significantly higher likelihood of having malaria. This common finding has resulted in malaria being considered predominantly as a rural disease in Africa (Donnelly et al., 2005). In rural areas, poor-quality household construction materials are common, which have been shown to be associated with a higher incidence of malaria due to increased mosquito entry (Snyman et al., 2015; Wanzirah et al., 2015).

CHAPTER 8

DISCUSSION AND CONCLUSION

Childhood anaemia does not receive the attention it deserves, however its effects in young children can be considerable, so much so that it has been termed a silent killer (Medix Team, 2021). In malaria endemic countries, such as Kenya, Malawi, Tanzania and Uganda, it is believed that the majority of childhood anaemia is due to malaria (White, 2018). In this study, we examined the relationship between anaemia and malaria in young children in these four countries using nationally representative Demographic and Health Survey data. The final sample comprised of 18196 children from the four countries, with an observed prevalence of 52.5% and 19.7% of anaemia and malaria, respectively. The prevalence of co-infection was 15.1%. In addition, out of those who tested positive for malaria, 76% had anaemia as well. This is an indication of the contribution that malaria has on the burden of anaemia. The observed prevalences were mapped across the districts of the countries, which revealed heterogeneity within and between the countries. Anaemia was considerably more prevalent than malaria in the majority of the districts. In addition, the patterns of prevalence of anaemia and malaria co-infection in children were markedly similar to that of malaria, suggesting that children with malaria are most likely to have anaemia as well. This means that children are not being treated for malaria timeously.

Numerous explanatory variables were considered, which comprised of demographic, socio-economic and environmental factors. In exploring the data, graphical techniques were used to examine these factors in relation to the child's anaemia status, their malaria status and whether or not they had both anaemia and malaria. Various patterns were divulged. Kenya had a lower prevalence of all three outcomes compared to the other three countries, with Uganda having the highest. The difference in

the prevalence of anaemia and prevalence of malaria was greatest in Tanzania, highlighting possible other causes of childhood anaemia in the country. A decreasing trend in the prevalences was observed with an increase in the mother's education level as well as with an improvement in the type of toilet facility. The observed prevalence of malaria as well as that of both diseases substantially differed according to the type of place or residence. Children with anaemia had a lower age, on average, compared to those with malaria.

In addition to the graphical techniques used to explore the data, further exploratory analysis was performed using supervised machine learning techniques with the aim of discovering relationships among the explanatory variables and the two responses. In particular, a multiple correspondence analysis plot was considered to examine the relationships among the categorical factors and the outcomes. This plot highlighted that having both anaemia and malaria was mostly associated with testing positive for malaria. Four techniques were applied to classify the child's anaemia and malaria statuses, which included logistic regression, CART models, support vector machines and artificial neural networks. The results revealed that the responses were important predictors of each other. Moreover, the child's age, the household's wealth index Z-score, cluster altitude, EVI and LST were among the most important predictors for both responses, with the age of the household head being among the least important. The child's gender and the gender of the household head were also established as the least important predictors of malaria.

The results of the classification models were valuable in paving the way for further modelling, where it provided guidance on the appropriate statistical models for the data, as well as the predictors to be incorporated into the models. However, the accuracy of the classification models for anaemia was poor, suggesting that there are unmeasured important predictors of anaemia. This poor predictive accuracy may also be attributed to the inability of the classification models to incorporate the

effect of the district in the form of a spatial effect. Such a spatial effect is a surrogate for unmeasured factors that may have an influence on the response as well as contribute to spatial heterogeneity in the likelihood of the disease. Such factors include the quality of health care, distance to nearest health clinic, or climatic factors. Thus, we adopted a geoadditive model to investigate the spatial variation and risk factors of childhood anaemia across the districts of the four countries. The child's malaria status was incorporated as a predictor to evaluate its effect on the likelihood of anaemia in a child. The geoadditive model allowed us to assess and visualise the district-level residual spatial effects on the likelihood of childhood anaemia while controlling for the effects of individual-level, household-level and environmental factors. The spatial effects were considered at district-level as the districts represent the level for which public health decisions are made within each country. The geoadditive model also enabled us to explore possible non-linear effects of the continuous covariates.

The district-level structured spatial effect, which accounts for spatial autocorrelation among neighbouring districts, was weak in comparison to that of the district-level unstructured spatial effect. This unstructured spatial effect is due to factors that are specific to the district and uncorrelated with neighbouring districts. Through the inclusion of such spatial effects in the model, the environmental factors; EVI and LST, no longer had significant effects on the likelihood of childhood anaemia. As the spatial effects are as a result of unmeasured factors, these results confirm that there are other important predictors of anaemia that have not been considered in this study, where such predictors are predominantly specific to the district and not shared among neighbouring districts. These spatial effects were mapped to reveal numerous districts associated with an increased likelihood of anaemia in children. Further, these maps highlighted the heterogeneity in the likelihood of anaemia across the districts of the four countries.

The child's age in months had a significant non-linear effect on the likelihood of anaemia. Children under 12 months of age had an increased likelihood of anaemia. This result supported the data exploration which revealed a lower average age of children suffering with anaemia. The child's malaria status had a highly significant effect on the likelihood of anaemia, where those with malaria were over 4 times more likely to have anaemia compared to those who did not have malaria. These findings were consistent with that of the classification models, which indicated that the child's age and malaria status were the most important predictors of anaemia. The geoadditive model also revealed that the child's gender had a significant effect on their anaemia status. Female children were less likely to be anaemic compared to males. There was an increased odds of anaemia in children with less educated mothers. Besides education being associated with one's earning potential, educated individuals are also associated with the ability to have more awareness and understanding of health and nutritional related issues. Furthermore, the odds of anaemia also increased with a decrease in the household's wealth index Z-score. Individuals with low wealth are often subjected to economic constraints where they are not able to afford the dietary and sanitation needs of themselves and their family. Moreover, children in households with no toilet facilities were associated with a significantly higher odds of anaemia. Poor sanitation can aid in the development of a number of infectious and parasitic diseases, which indirectly contribute to childhood anaemia (Muchie, 2016).

We adopted a generalised additive mixed model with a district-level unstructured spatial to assess and rank the performance of the districts with regard to their contribution to the burden of anaemia. This district-level spatial effect was incorporated as an i.i.d. random effect, which enabled us to utilise the BLUP technique, an appropriate technique for the ranking of the districts' performances. The top best and worst performing districts were identified and mapped. In addition, a cluster-level structured spatial effect was incorporated into the model to account for spatial auto-

correlation between the clusters within the districts. This revealed multiple districts consisting of clusters with increased likelihoods as well as clusters with decreased likelihoods of anaemia.

The results of the exploratory data analysis, the geoadditive model and the generalised additive mixed model all highlighted the relationship between anaemia and malaria in young children. This relationship was further examined via joint modelling of the two responses. A copula geoadditive model was utilised, where the association between the two responses was varied according to the district of residence. This revealed a positive association between anaemia and malaria throughout the districts, which varied in strength. Some districts had a stronger association between the two responses compared to other districts. The estimated joint probabilities for each combination of the outcomes were mapped across the districts to further aid in better understanding the relationship between anaemia and malaria. A considerable number of districts had a high joint probability of a child being anaemic but not having malaria, further highlighting the existence of other significant drivers of childhood anaemia in these districts. The copula geoadditive model identified the same significant risk factors for anaemia as the geoadditive model. In addition, the significant risk factors for malaria were also identified. Children residing in rural areas had significantly higher odds of malaria compared to those residing in urban areas. The odds of malaria significantly decreased with an increase in the mother's education level, the household's wealth index and the cluster altitude. However, the odds significantly increased with an increase in EVI. Furthermore, there was an increased likelihood of malaria with an increase in the child's age, which is contrasting to the effect of the child's age on anaemia as anaemia is predominantly in early childhood.

The copula geoadditive model allowed the district-level spatial effects to be incorporated for each response. However, a shortcoming of this approach is that we were

not able to decompose such spatial effects into shared and disease-specific components. Thus, a shared component model was adopted to jointly model the spatial variation in the likelihood of anaemia and malaria in young children across the districts of the four countries. Unlike the majority of other studies which apply this model at aerial level to model the relative risk of diseases in a particular region, we adapted this approach to model the child-level likelihood of anaemia and malaria. The shared and disease-specific district-level spatial effects were estimated while accounting for individual-level, household-level and environmental characteristics. These spatial effects were mapped to visualise the geographical locations which contributed to a higher or lower likelihood of both diseases simultaneously as well as individually. The results indicated that the spatial variation in the likelihood of malaria was more prominent compared to that of anaemia, for both the shared and specific spatial components.

The findings of this thesis have the potential to contribute to achieving SDG 3 (Health) as well as SDG 2 (Hunger and Food Security) by providing guidance to policy makers. The benefit of focusing on contiguous countries was to be able to determine whether the spatial effects transcend the borders of the countries. While there was stronger evidence of the spatial effect for malaria transcending the country borders compared to that of anaemia, it would be advantageous for the countries that share borders to develop joint policies for malaria and anaemia control and prevention. The ranking of the performance of the districts in Kenya, Malawi, Tanzania and Uganda with regard to their impact on childhood anaemia allows the worst-performing districts to be identified and targeted for further research to improve their anaemia control strategies. Moreover, it allows for the best-performing districts to be identified to further determine why they are performing better and then to use these districts as role models in efforts to overcome childhood anaemia. The maps created in the analyses in Chapters 4 to 7 can go hand in hand, providing tools to allow for more targeted action in malaria and anaemia control and prevention,

as well as for the appropriate allocation of limited district health system resources. Control measures in Kenya, Malawi, Tanzania and Uganda need to account for the spatial variation in the two diseases, as a one-size fits all strategy may not work in such a setting. Specifically, strategies should be tailored to local conditions at district level, by accounting for child, household and environmental characteristics. Furthermore, as it is more common for co-infection of the two diseases to start with malaria, we recommend that programs and interventions for malaria in children be targeted in high malaria risk districts as identified by both the shared and malaria-specific spatial components in the maps produced in Chapter 7. Such a targeted approach for malaria would likely also make a positive impact on anaemia. In addition, further investigation into those districts with simultaneous high anaemia risk and low malaria risk should be considered in order to identify the significant drivers of anaemia in children within those districts. This would aid in applying the appropriate control measures and interventions for childhood anaemia in those districts, while saving on resources for malaria control and prevention which should be directed to the districts most in need.

We further recommend that awareness and educational programs about the symptoms and multiple complex causes of anaemia in children be aimed at parents and caregivers, especially those with children in the younger, more vulnerable age group of 6 to 11 months. Furthermore, programs that ensure the introduction of safe and adequate complementary foods in a child's diet from the age of 6 months should be considered. These types of programs would be beneficial as these children are more susceptible to anaemia due to the rapid growth during that stage of their lives.

This research contributes to the literature, where very few studies have focused on childhood anaemia and its relationship with malaria in multiple contiguous sub-Saharan African countries, none of which have considered Kenya, Malawi, Tanzania and Uganda simultaneously. A further strength lies in the novelty of applying the

BLUP technique to appropriately rank the performance of the districts with regard to their impact on childhood anaemia, and thus in the performance of eradication and prevention goals set nationally, regionally and globally. An additional novelty includes the adoption of a child-level shared component model with district-level shared and disease-specific spatial effects to model the likelihood of anaemia and malaria in a child, which, to our knowledge, has not been considered for these two diseases. Such a SCM allows for the effects of the individual-level, household-level and environmental factors to be estimated and controlled for.

A limitation of this study is that it was based on cross-sectional survey data, therefore a causal relationship cannot be established. In addition, no information on iron levels in the children was available, however iron deficiency plays a major role in childhood anaemia (Thejpal, 2015). Furthermore, while this study could not assess the contribution of intestinal parasites to the burden of anaemia in children directly, proxies for this factor were used instead. However, individual level malaria RDT results were used rather than estimates or indicators of malaria.

In this study, the child's anaemia status was considered in its binary form. Therefore, a possible future direction is to consider the child's anaemia status in terms of a four-level ordinal variable (non-anaemic, mild, moderate and severe), where the effect of the child's malaria status on the severity of anaemia can be explored. In addition, as the DHS programs are conducted every 3 to 5 years in each country, the temporal effect in addition to the spatial effect can be considered. Such a spatio-temporal analysis will enable us to examine the change in the spatial variation in the likelihood of the diseases across the districts of the four countries. This will aid in identifying which districts are moving closer to or further away from the elimination goals.

REFERENCES

Adebayo, S. B., & Fahrmeir, L. (2005). Analysing child mortality in Nigeria with geoadditive discrete-time survival models. *Stat Med*, *24*(5), 709–728.

Adebayo, S. B., Gayawan, E., Heumann, C., & Seiler, C. (2016). Joint modeling of Anaemia and Malaria in children under five in Nigeria. *Spatial and Spatio-temporal Epidemiology*, *17*, 105–115.

Adeyemi, R., Zewotir, T., & Ramroop, S. (2019). Joint spatial mapping of childhood anemia and malnutrition in sub-Saharan Africa: a cross-sectional study of small-scale geographical disparities. *J Afri Health Sci*, *19*(3), 2692–2712.

Al-Qaoud, N. M., Al-Shami, E., & Prakash, P. (2015). Anemia and associated factors among Kuwaiti preschool children and their mothers. *Alexandria J. Med.*, *51*(2), 161–166.

Alemu, M., Kinfe, B., Tadesse, D., Mulu, W., Hailu, T., & Yizengaw, E. (2017). Intestinal parasitosis and anaemia among patients in a Health Center, North Ethiopia. *BMC Res Notes*, *10*(1), 632.

Ayele, D., Zewotir, T., & Mwambi, H. (2014). Modelling the joint determinants of a positive malaria Rapid Diagnosis Test result, use of mosquito nets and indoor residual spraying with insecticide. *Occupational Health Southern Africa*, *20*(4), 20–27.

Bajetha, G., Singh, C. V., & Barwal, R. S. (2015). Sire Evaluation on The Basis of First Lactation Traits Using Best Linear Unbiased Prediction (BLUP) Method in Sahiwal and Crossbred Cattle. *3*(4), 85–88.

Banhela, N., Taylor, M., Zulu, S. G., Sund, L., Kjetland, E. F., & Gundersen, S. (2017). Environmental factors influencing the distribution and prevalence of Schistosoma haematobium in school attenders of ILembe and uThungulu Health Districts , KwaZulu-Natal Province , South Africa. *South African J Infect Dis*, *32*(4), 132–137.

Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*, *343*, 1–20.

Blagus, R., & Goeman, J. J. (2016). What (not) to expect when classifying rare events. *Briefings in Bioinformatics*, *19*(2), 341–349.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth Statistics/Probability.

Brunner, M. I., Furrer, R., & Favre, A.-C. (2019). Modeling the spatial dependence of floods using the Fisher copula. *Hydrol Earth Syst Sci*, *23*, 107–124.

Centers for Disease Control and Prevention (1989). Criteria for anemia in children and childbearing-aged women. *CDC WONDER*, *MMWR Morb Mortal Wkly Rep 38*(22), 400–404.

Challa, S., & Amirapu, P. (2016). Surveillance of anaemia: Mapping and grading the high risk territories and populations. *J Clin Diagnostic Res*, *10*(6), 1–6.

Cibulskis, R. E., Alonso, P., Aponte, J., Aregawi, M., Barrette, A., Bergeron, L., Fergus, C. A., Knox, T., Lynch, M., Patouillard, E., Schwarte, S., Stewart, S., & Williams, R. (2016). Malaria: Global progress 2000 - 2015 and future challenges. *Infectious Diseases of Poverty*, *5*(1), 61.

Clayton, D., & Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*, 671–681.

Cottrell, G., Kouwaye, B., Pierrat, C., le Port, A., Bouraïma, A., Fonton, N., et al. (2012). Modeling the Influence of Local Environmental Factors on Malaria Transmission in Benin and Its Implications for Cohort Study. *PLoS ONE*, *7*(1), e28812.

Crawley, J. (2004). Reducing the burden of anemia in infants and young children in malaria-endemic countries of Africa: From evidence to action. *Am J Trop Med Hyg*, *71*(2 Suppl.), 25–34.

Croft, T. N., Marshal, A. M. J., Allen, C. K., et al. (2018a). Guide to DHS Statistics. Rockville, Maryland, USA: ICF.

Croft, T. N., Marshall, A. M. J., Allen, C. K., et al. (2018b). *Guide to DHS Statistics*. Rockville, Maryland, USA: ICF.

de Leon, A. R., & Chough, K. C. (2013). *Analysis of Mixed Data Method & Application*.

Di Franco, G. (2016). Multiple correspondence analysis: one only or several techniques? *Qual Quant*, *50*, 1299–1315.

Donnelly, M.J., McCall, P., Lengeler, C., et al. (2005). Malaria and urbanization in sub-Saharan Africa. *Malar J*, *4*(12), 684–693.

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

Etxeberria, J., Goicoa, T., & Ugarte, M. D. (2018). Joint modelling of brain cancer incidence and mortality using Bayesian age- and gender-specific shared component models. *Stochastic Environmental Research and Risk Assessment*, *32*(10), 2951–2969.

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Stat Sin*, *14*, 731–761.

Frake, A. N., Peter, B. G., Walker, E. D., & Messina, J. P. (2020). Leveraging big data for public health: Mapping malaria vector suitability in malawi with google earth engine. *PLOS ONE*, *15*(8), 1–21.
URL https://doi.org/10.1371/journal.pone.0235697

Furlani, R. C. M., de Moraes, M. L. T., de Resende, M. D. V., Junior, E. F., et al. (2005). Estimation of variance components and prediction of breeding values in rubber tree breeding using the REML/BLUP procedure. *Genetics and Molecular Biology*, *28*(2), 271–276.

Gari, T., Loha, E., Deressa, W., Solomon, T., Atsbeha, H., Assegid, M., et al.

(2017). Anaemia among children in a drought affected community in south-central Ethiopia. *Int Health*, *12*(3), e0170898.

Gayawan, E., Arogundade, E. D., & Adebayo, S. B. (2014). Possible determinants and spatial patterns of anaemia among young children in Nigeria: A bayesian semi-parametric modelling. *Int Health*, *6*(1), 35–45.

Gayawan, E., & Fadiji, F. A. (2020). Joint Spatial Modelling of Childhood Morbidity in West Africa Using a Distributional Bivariate Probit Model. *Stat Biosci*, (p. 56–76).

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, *160*, 249–264.

Goswmai, S., & Das, K. K. (2015). Socio-economic and demographic determinants of childhood anemia. *Jornal de Pediatria*, *91*, 471–477.

Greenacre, M. (2017). *Correspondence Analysis in Practice*. Chapman & Hall Interdisciplinary Statistics Series, 3 ed.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London (UK) Academic Press.

Guerra, C. A., Tresor Donfack, O., Motobe Vaz, L., Mba Nlang, J. A., Nze Nchama, L. O., Mba Eyono, J. N., et al. (2020). Malaria vector control in sub-Saharan Africa in the time of COVID-19: no room for complacency. *BMJ Global Health*, *5*(9).
URL https://gh.bmj.com/content/5/9/e003880

Habyarimana, F., Zewotir, T., & Ramroop, S. (2016). Joint Modeling of Poverty of Households and Malnutrition of Children Under Five Years from Demographic and Health Survey Data: Case of Rwanda. *JEBS*, *8*(2), 108–114.

Habyarimana, F., Zewotir, T., & Ramroop, S. (2017). Structured additive quantile regression for assessing the determinants of childhood anemia in Rwanda. *Int J Environ Res Public Health*, *14*, 652.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, 2 ed.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423–447.

Ibáñez-Beroiz, B., Librero-López, J., Peiró-Moreno, S., & Bernal-Delgado, E. (2011). Shared component modelling as an alternative to assess geographical variations in medical practice: gender inequalities in hospital admissions for chronic diseases. *BMC Med Res Methodol*, *11*(1), 1–10.

ICF International (2012). *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International.

Kandala, N.-B., & Madise, N. (2004). The Spatial Epidemiology of Childhood Diseases in Malawi and Zambia. *African Population Studies*, *Supplement B*.

Kantardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, Inc., 3 ed.

Kateera, F., Mens, P. F., Hakizimana, E., Ingabire, C. M., Muragijemariya, L., et al. (2015). Malaria parasite carriage and risk determinants in a rural population: a malariometric survey in Rwanda. *Malar J*, *14*, 16.

Kazembe, L. N. (2007). Spatial modelling and risk factors of malaria incidence in northern Malawi. *Acta Trop*, *102*(1), 126–137.

Kazembe, L. N., Appleton, C. C., & Kleinschmidt, I. (2007). Spatial analysis of the relationship between early childhood mortality and malaria endemicity in Malawi. *Geospat Health*, *2*(1), 41–50.

Kazembe, L. N., Muula, A. S., & Simoonga, C. (2009). Joint spatial modelling of common morbidities of childhood fever and diarrhoea in Malawi. *Health and Place*, *15*, 165–172.

Kejo, D., Petrucka, P. M., Martin, H., Kimanya, M. E., & Mosha, T. C. (2018). Prevalence and predictors of anemia among children under 5 years of age in Arusha District, Tanzania. *Pediatric health, medicine and therapeutics*, *9*, 9–15.

Khan, J. R., Awan, N., & Misu, F. (2015). Determinants of anemia among 6–59 months aged children in Bangladesh: evidence from nationally representative data. *BMC Pediatr.*, *16*(1), 3.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S., & McGovern, M. (2019). Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Stat Med*, *38*, 413—-436.

Kneib, T., Müller, J., & Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environ Ecol Stat*, *15*, 343–364.

Knorr-Held, L., & Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *164*(6), 73–85.

Kolev, N., Dos Anjos, U., & Mendes, B. V. D. M. (2006). Copulas: A review and recent developments. *Stochastic Models*, *22*, 617–660.

Kürüm, E., Hughes, J., Li, R., & Shiffman, S. (2018). Time-varying copula models for longitudinal data. *Statistics and its Interface*, *11*, 203–221.

Kuziga, F., Adoke, Y., & Wanyenze, R. K. (2017). Prevalence and factors associated with anaemia among children aged 6 to 59 months in Namutumba district, Uganda: a cross- sectional study. *BMC Pediatr*, *17*, 25.

Kweku, M., Takramah, W., Axame, W. K., Owusu, R., Takase, M., et al. (2017). Prevalence and risk factors of malaria among children under five years in High and Low altitude rural communities in the Hohoe Municipality of Ghana. *J Clin Immunol Res*, *1*, 1–8.

Lang, S., & Brezger, A. (2004). Bayesian P-Splines. *J Comput Graphical Statist*, *13*, 183–212.

Law, J., Quick, M., & Jadavji, A. (2020). Annals of GIS A Bayesian spatial shared component model for identifying crime-general and crime-specific hotspots. *Annals of GIS*, *26*(1), 65–79.

Lawson, A., Biggeri, A., Boehning, D., et al. (2000). Disease mapping models: An empirical evaluation. Disease mapping collaborative group. *Statistics in medicine*, *19*(17-18), 2217–2241.

le Cessie, S., Verhoeff, F., Mengistie, G., Kazembe, P., Broadhead, R., & Brabin, B. J. (2002). Changes in haemoglobin levels in infants in Malawi: effect of low birth weight and fetal anaemia. *Arch Dis Child - Fetal Neonatal Ed*, *86*(3), F182–F187.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *JRSSB*, *55*(2), 381–400.

Ma, X. (2018). *Using Classification And Regression Trees: A Practical Primer*. Information Age Publishing.

Macharia, P. M., Giorgi, E., Noor, A. M., Waqo, E., Kiptui, R., Okiro, E. A., & Snow, R. W. (2018). Spatio-temporal analysis of Plasmodium falciparum prevalence to understand the past and chart the future of malaria control in Kenya. *Malar J*, *17*(1), 1–13.

Madsen, L., & Fang, Y. (2011). Joint Regression Analysis for Discrete Longitudinal Data. *Biometrics*, *67*, 1171–1175.

Mainardi, S. (2012). Modelling spatial heterogeneity and anisotropy: child anaemia, sanitation and basic infrastructure in sub-Saharan Africa. *Int J Geogr Inf Sci*, *26*(3), 387–411.

Marra, G., & Radice, R. (2017a). A joint regression modeling framework for analyzing bivariate binary data in R. *Depend Model*, *5*, 268–294.

Marra, G., & Radice, R. (2017b). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, *112*, 99–113.

Marra, G., & Radice, R. (2017c). *GJRM: Generalised Joint Regression Modelling*. R package version 0.1-2. [Available on CRAN].

Mayala, B., Fish, T. D., Eitelberg, D., & Dontamsetti, T. (2018). *The DHS Program Geospatial Covariate Datasets Manual*. Rockville, Maryland, USA: ICF, 2 ed.

McCulloch, C., Sear, l. S., & Neuhau, J. (2008). *Generalized, Linear, and Mixed Models*. New York, John Wiley, 2 ed.

McElroy, P., ter Kuile, F., Lal, A., Bloland, P., Hawley, W., Oloo, A., et al. (2000). Effect of Plasmodium falciparum parasitemia density on hemoglobin concentrations among full-term, normal birth weight children in western Kenya, IV. The Asembo Bay Cohort Project. *Am J Trop Med Hyg*, *62*, 504–512.

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2 ed.

Medix Team (2021). Anaemia: the invisible disorder that can become a silent killer. https://www.medix-global.com/eu-en/content/articles/view/?ContentID=3045. [Online; accessed August 2021].

Menendez, C., Kahigwa, E., Hirt, R., Vounatsou, P., Aponte, J. J., Font, F., et al. (1997). Randomised placebo-controlled trial of iron supplementation and malaria chemoprophylaxis for prevention of severe anaemia and malaria in Tanzanian infants. *The Lancet*, *350*(9081), 844–850.

Meyer, D. (2021). *Support Vector Machines: The Interface to libsvm in package e1071*. FH Technikum Wien, Austria.

Michael, E., Spear, R., & Remais, J. (2010). Modelling environmentally-mediated infectious diseases of humans: transmission dynamics of Schistosomiasis in China. In E. Michael, & R. C. Spear (Eds.) *Modelling parasite transmission and control*, (pp. 79–98). New York: Springer.

Miller, R. E. (2019). Anaemia. https://kidshealth.org/en/parents/anemia.html. [Online; accessed January 2019].

Minakawa, N., Dida, G., Sonye, G., Futami, K., & Njenga, S. (2012). Malaria vectors in Lake Victoria and adjacent habitats in western Kenya. *PLoS One*, *7*(2), e32725.

MLMath.io (2021). Math behind SVM (Support Vector Machine). https://ankitnitjsr13.medium.com/math-behind-svm-support-vector-machine-864e58977fdb. [Online; accessed August 2021].

Moschovis, P. P., Wiens, M. O., Arlington, L., Antsygina, O., Hayden, D., Dzik, W., et al. (2018). Individual, maternal and household risk factors for anaemia among young children in sub-Saharan Africa: A cross-sectional study. *BMJ Open*, *8*(5), e019654.

Muchie, K. (2016). Determinants of severity levels of anemia among children aged 6–59 months in Ethiopia: further analysis of the 2011 Ethiopian demographic and health survey. *BMJ Open*, *2*, 51.

Nambiema, A., Robert, A., & Yaya, I. (2019). Prevalence and risk factors of anemia in children aged from 6 to 59 months in Togo: analysis from Togo demographic and health survey data, 2013–2014. *BMC Public Health*, *19*(215), 1–9.

Nas, F. S., Yahaya, A., Mu'azu, L., Ali, M., & Abdullahi, S. I. (2020). Comparative Assessment of Microscopy and Rapid Diagnostic Test (RDT) As Malaria Diagnostic Tools in Kano, Northern Nigeria. *Sumerianz Journal of Biotechnology*, *3*(6), 38–42.

NASA Earth Observatory (2020). Vegetation & Total Rainfall. https://earthobservatory.nasa.gov/global-maps/MOD_NDVI_M/TRMM_3B43M. [Online; accessed February 2020].

Nelsen, R. B. (2006). *An introduction to Copulas (Springer Series in Statistics)*.

Ngesa, O., & Mwambi, H. (2014). Prevalence and risk factors of anaemia among children aged between 6 months and 14 years in Kenya. *PLoS ONE*, *9*(11), e113756.

Nguyen, M., Howes, R. E., Lucas, T. C. D., Battle, K. E., Cameron, E., Gibson, H. S., et al. (2019). Mapping malaria seasonality: a case study from Madagascar. *arXiv e-prints*, (p. arXiv:1901.10782).

Ngwira, A., & Kazembe, L. N. (2015). Bayesian random effects modelling with application to childhood anaemia in Malawi. *BMC Public Health*, *15*, 161.

Nikoloulopoulos, A. K., & Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Stat Med*, (pp. 6393–6406).

Phyllis Atta Parbey, E. M., Elvis Tarkang, et al. (2019). Risk Factors of Anaemia among Children under Five Years in the Hohoe Municipality, Ghana: A Case Control Study. *Anemia*, *2019*(Article ID 2139717), 1–9.

Portnoy, S. (1982). Maximizing the probability of correctly ordering random variables using linear predictors. *J Multivariate Analysis*, *12*, 256–269.

Rahman, M., Mushfiquee, M., Masud, M., & Howlader, T. (2019). Evidence from Bangladesh Demographic and Health Survey 2011. *PLoS One*, *14*(7), e0219170.
URL doi:10.1371/journal.pone.0219170

Roba, K., O'Connor, T., Belachew, T., & O'Brien, N. (2016). Anemia and undernutrition among children aged 6–23 months in two agroecological zones of rural Ethiopia. *Pediatric Health Med Ther*, *7*, 131–140.
URL https://doi.org/10.2147/PHMT.S109574

Roberts, D., & Matthews, G. (2016). Risk factors of malaria in children under the age of five years old in Uganda. *Malar J*, *27*, 246.

Roberts, D., Matthews, G., Snow, R., Zewotir, T., & Sartorius, B. (2020). Investigating the spatial variation and risk factors of childhood anaemia in four sub-Saharan African countries. *BMC Public Health*, *20*(126), 126.

Robinson, G. (1991). That blup is a good thing: The estimation of random effects. *Stat Sci*, *6*(1), 15–32.

Rodriguez-Sabate, C., Morales, I., Sanchez, A., & Rodriguez, M. (2017). The Multiple Correspondence Analysis Method and Brain Functional Connectivity: Its Application to the Study of the Non-linear Relationships of Motor Cortex and Basal Ganglia. *Frontiers in Neuroscience*, *11*, 345.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J Royal Stat Soc*, *71*(2), 319–392.

Saldan, P. C., Venancio, S. I., Saldiva, S. R. D. M., Daniele Gonçalves Vieira, & Débora Falleiros de Mello (2017). Milk Consumption in Infants Under One Year of Age and Variables Associated with Non-Maternal Milk Consumption. *Rev Paul Pediatr*, *35*(4), 407–414.

Schellenberg, D., Armstrong Schellenberg, J. R. M., Mushi, A., Savigny de, D., Mgalula, L., Mbuya, C., & Victoria, C. (2003). The silent burden of anaemia in Tanzania children:a community-based study. *Bull World Health Organ*, *81*.

Seyoum, S. (2018). Analysis of Prevalence of Malaria and Anemia Using Bivariate Probit Model. *Annals of Data Science*, *5*, 301–312.

Smith, J. L., & Brooker, S. (2010). Impact of hookworm infection and deworming on anaemia in non-pregnant populations: A systematic review: Systematic Review. *Trop Med Int Heal*, *15*(7), 776–795.

Smith, M., Min, A., Almeida, C., & Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, *105*, 1467–1479.

Snyman, K., Mwangwa, F., Bigira, V., Kapisi, J., Clark, T. D., Osterbauer, B., et al. (2015). Poor housing construction associated with increased malaria incidence in a cohort of young Ugandan children. *Am J Trop Med Hyg*, *92*(6), 1207–1213.

Soares Magalhães, R. J., & Clements, A. C. A. (2011). Mapping the risk of anaemia in preschool-age children: The contribution of malnutrition, malaria, and helminth infections in West Africa. *PLoS Med*, *8*(6), e1000438.

Soares Magalhães, R. J., Langa, A., Pedro, J. M., Sousa-Figueiredo, J. C., Clements, A. C. A., & Nery, S. V. (2013a). Role of malnutrition and parasite infections in the

spatial variation in children's anaemia risk in northern Angola. *Geospat Health*, *7*(2), 341–354.

Soares Magalhães, R. J., Langa, A., Pedro, J. M., Sousa-Figueiredo, J. C., Clements, A. C. A., & Nery, S. V. (2013b). Role of malnutrition and parasite infections in the spatial variation in children's anaemia risk in northern Angola. *Geospatial Health*, *7*(2), 341–354.

SoaresMagalhães, R. J. S., & Clements, A. C. A. (2011). Spatial heterogeneity of haemoglobin concentration in preschool-age children in sub-Saharan Africa. *Bull World Health Organ*, *89*, 459–468.

Soh, A. C. (1994). Ranking parents by best linear unbiased prediction (BLUP) breeding values in oil palm. *Euphytica*, *76*(1-2), 13–21.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J R Statist Soc*, *B*(64), 583–639.

Ssemukasa, E., & Kearney, J. (2014). Health and food safety concerns of early dietary introduction of unmodified cow milk to infants in developing countries. *African J Food, Agric Nutr Dev*, *14*(1), 8504–8517.

Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., et al. (2013). Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995-2011: A systematic analysis of population-representative data. *Lancet Glob Heal*, *1*(1), 16–25.

The DHS Program (2017). What We Do. https://dhsprogram.com/What-We-Do/index.cfm. [Online; accessed November 2017].

Thejpal, R. (2015). Iron deficiency in children. *S Afr Med J*, *105*(7), 607.

Thomas, A., Best, N., Lunn, D., Arnold, R., & Spiegelhalter, D. (2003). *BUGS: Bayesian inference using Gibbs sampling*. Version 1.4. Cambridge, United Kingdom: MRC Biostatistics Unit.

Thomas, A., Best, N., Lunn, D., Arnold, R., & Spiegelhalter, D. (2004). *GeoBUGS User Manual*. Version 1.2.

Ugwu, C. L. J., & Zewotir, T. (2018). Using mixed effects logistic regression models for complex survey data on malaria rapid diagnostic test results. *Malaria J*, *17*, 453.

Umberto, C. (2011). *Copulas in Finance*, (pp. 305–309). Springer Berlin Heidelberg.

Umlauf, N., Adler, D., Kneib, T., Lang, S., & Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *J Stat Softw*, *63*(21), 1–46.

Umlauf, N., Kneib, T., Lang, S., & Zeileis, A. (2016). Package 'R2BayesX'. [Online; accessed November 2017].

Waller, L., & Carlin, B. (2010). Disease mapping. *Chapman & Hall/CRC handbooks of modern statistical methods*, *2010*, 217–243.

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

Wang, J. F., Zhang, T. L., & Fu, B. J. (2016). A measure of spatial stratified heterogeneity. *Ecol Indic*, *67*, 250–256.

Wanzirah, H., Tusting, L. S., Arinaitwe, E., Katureebe, A., Maxwell, K., Rek, J., et al. (2015). Mind the gap: house structure and the risk of malaria in Uganda. *PLoS One*, *10*(1), e0117396.

Weaver, H. (2014). *Climate change and human parasitic disease*. CABI Nosworthy Way Wallingford UK: ©CAB International.

White, N. (2018). Anaemia and malaria. *Malar J*, *17*, 371.

WHO (2011). *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. Vitamin and mineral nutrition information system*. Geneva: World Health Organization.

WHO (2013). *Essential nutrition actions: improving maternal, newborn, infant and young child health and nutrition.*. Geneva: World Health Organization.

WHO (2015a). *Global technical strategy for malaria 2016-2030*. Geneva: World Health Organization.

WHO (2015b). *Health in 2015: From MDGs, Millennium Development Goals to SDGs, Sustainable Development Goals*. Geneva: World Health Organization.

WHO (2016). Malaria fact sheet. http://www.who.int/mediacentre/factsheets/fs094/en/. [Online; accessed January 2017].

WHO (2017). In Kenya, the path to elimination of malaria is lined with good preventions. https://www.who.int/news-room/feature-stories/detail/in-kenya-the-path-to-elimination-of-malaria-is-lined-with-good-preventions. [Online; accessed March 2020].

WHO (2018). Malaria in children under five. https://www.who.int/malaria/areas/high_risk_groups/children/en/. [Online; accessed February 2020].

WHO (2021a). Malaria. https://www.who.int/news-room/fact-sheets/detail/malaria. [Online; accessed June 2021].

WHO (2021b). WHO Global Anaemia estimates, 2021 Edition. https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children. [Online; accessed June 2021].

WHO and UNICEF (2004). Focusing on anaemia. http://www.who.int/nutrition/publications/micronutrients/WHOandUNICEF_statement_anaemia/en/. [Online; accessed November 2019].

Wijndaele, K., Lakshman, R., Landsbaugh, J. R., Ong, K. K., & Ogilvie, D. (2009). Determinants of Early Weaning and Use of Unmodified Cow's Milk in Infants: A Systematic Review. *J Am Diet Assoc*, *109*(12), 2017–2028.

Winkelmann, R. (2011). Copula Bivariate Probit Models: With an Application to Medical Expenditures. *Health Econ*, *21*, 1444–1455.

Wirth, J. P., Rohner, F., Woodruff, B. A., Chiwile, F., Yankson, H., Koroma, A. S., Russel, F., Sesay, F., Dominguez, E., Petry, N., Shahab-Ferdows, S., De Onis, M., & Hodges, M. H. (2016). Anemia, micronutrient deficiencies, and malaria in children and women in Sierra Leone prior to the Ebola outbreak - Findings of a cross-sectional study. *PLoS ONE*, *11*(5), 1–21.

Yamana TK, E. E. (2013). Incorporating the effects of humidity in a mechanistic model of *Anopheles gambiae* mosquito population dynamics in the Sahel region of Africa. *Parasit Vectors*, *6*, 235.

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in Statistics — Theory and Methods*, *37*(7), 1071–1084.

Zewotir, T. (2012). On employees' performance appraisal: the impact and treatment of the raters' effect. *SAJEMS*, *15*(1), 44–54.

Zhao, A., Zhang, Y., Peng, Y., Li, J., Yang, T., Liu, Z., Lv, Y., & Wang, P. (2012). Prevalence of anemia and its risk factors among children 6-36 months Old in Burma. *Am J Trop Med Hyg*, *87*(2), 306–311.

Ziegler, E. E., Nelson, S. E., & Jeter, J. M. (2014). Iron stores of breastfed infants during the first year of life. *Nutrients*, *6*(5), 2023–2034.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., & Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science, New York, USA.

PUBLICATIONS

Roberts *et al. BMC Public Health* (2020) 20:126
https://doi.org/10.1186/s12889-020-8189-8

BMC Public Health

RESEARCH ARTICLE Open Access

Investigating the spatial variation and risk factors of childhood anaemia in four sub-Saharan African countries

Danielle J. Roberts[1*], Glenda Matthews[2], Robert W. Snow[3,4], Temesgen Zewotir[1] and Benn Sartorius[5]

Abstract

Background: The causes of childhood anaemia are multifactorial, interrelated and complex. Such causes vary from country to country, and within a country. Thus, strategies for anaemia control should be tailored to local conditions and take into account the specific etiology and prevalence of anaemia in a given setting and sub-population. In addition, policies and programmes for anaemia control that do not account for the spatial heterogeneity of anaemia in children may result in certain sub-populations being excluded, limiting the effectiveness of the programmes. This study investigated the demographic and socio-economic determinants as well as the spatial variation of anaemia in children aged 6 to 59 months in Kenya, Malawi, Tanzania and Uganda.

Methods: The study made use of data collected from nationally representative Malaria Indicator Surveys (MIS) and Demographic and Health Surveys (DHS) conducted in all four countries between 2015 and 2017. During these surveys, all children under the age of five years old in the sampled households were tested for malaria and anaemia. A child's anaemia status was based on the World Health Organization's cut-off points where a child was considered anaemic if their altitude adjusted haemoglobin (Hb) level was less than 11 g/dL. The explanatory variables considered comprised of individual, household and cluster level factors, including the child's malaria status. A multivariable hierarchical Bayesian geoadditive model was used which included a spatial effect for district of child's residence.

Results: Prevalence of childhood anaemia ranged from 36.4% to 61.9% across the four countries. Children with a positive malaria result had a significantly higher odds of anaemia [AOR = 4.401; 95% CrI: (3.979, 4.871)]. After adjusting for a child's malaria status and other demographic, socio-economic and environmental factors, the study revealed distinct spatial variation in childhood anaemia within and between Malawi, Uganda and Tanzania. The spatial variation appeared predominantly due to unmeasured district-specific factors that do not transcend boundaries.

Conclusions: Anaemia control measures in Malawi, Tanzania and Uganda need to account for internal spatial heterogeneity evident in these countries. Efforts in assessing the local district-specific causes of childhood anaemia within each country should be focused on.

Keywords: Adjusted odds ratio, Anaemia, Bayesian, Child, Haemoglobin level, Hierarchical geoadditive model, Spatial effect

Background

Anaemia, which is a condition in which the haemoglobin (Hb) concentration is lower than that required by the body to meet its physiological needs, is a major cause of morbidity and mortality among pregnant women and young children in most Low and Middle Income countries (LMIC), particularly those in sub-Saharan Africa (SSA) [1]. Anaemia contributes to adverse health problems in children, and affects their cognitive, behavioural and physical development [2, 3]. If left untreated, the long-term effects and consequences of anaemia in early childhood are irreversible, if mortality has not occurred [3]. According to the most recent estimates of the World Health Organization (WHO), the highest anaemia prevalence of 42.6% in 2011 occurred in children under the age of five years old, which translated to just over 273 million

*Correspondence: danjader@gmail.com
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban South Africa
Full list of author information is available at the end of the article

Hindawi
Anemia
Volume 2019, Article ID 1598920, 10 pages
https://doi.org/10.1155/2019/1598920

Research Article

District Effect Appraisal in East Sub-Saharan Africa: Combating Childhood Anaemia

Danielle J. Roberts and Temesgen Zewotir

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Correspondence should be addressed to Danielle J. Roberts; robertsd@ukzn.ac.za

Received 25 July 2019; Revised 26 September 2019; Accepted 23 October 2019; Published 13 November 2019

Academic Editor: Duran Canatan

Background. Anaemia in children is a significant health problem that receives little attention. This study aimed at determining the factors significantly associated with anaemia in children aged 6 to 59 months in Kenya, Malawi, Tanzania, and Uganda while accounting for the spatial heterogeneity within and between the districts of the four countries. In addition, the performance of the districts with regard to their impact on anaemia was assessed and ranked. *Methods*. A generalised additive mixed model with a spatial effect based on the geographical coordinates of the clusters was used. A district-level random effect was included to further account for the heterogeneity as well as to rank the performance of the districts based on the best linear unbiased prediction (BLUP). *Results*. The results depicted significant spatial heterogeneity between and within the districts of the countries. After accounting for such spatial heterogeneity, child-level characteristics (gender, malaria test result, and mother's highest education level), household-level characteristics (household size, household's wealth index Z-score, the type of toilet facility available, and the type of place of residence), and the country of residence were found to be significantly associated with the child's anaemia status. There was a significant interaction between the type of place of residence and the country of residence. Based on the BLUP for the district-level random effect, the top 3 best- and worst-performing districts within each country were identified. *Conclusion*. The ranking of the performance of the districts allows for the worst-performing districts to be targeted for further research in order to improve their anaemia control strategies, as well as for the best-performing districts to be identified to further determine why they are performing better and then to use these districts as role models in efforts to overcome childhood anaemia.

1. Introduction

Identifying significant factors associated with an increased risk of anaemia in children is relevant to developing appropriate and effective interventions. Such studies aid in identifying subpopulations that are most at risk, which assists in creating a more efficient delivery system of limited national resources [1]. However, studies identifying these factors should account for spatial heterogeneity and spatial autocorrelation in the observations. Failure to do so may produce inaccurate estimates and thus misleading results and ineffective anaemia control programs [2, 3].

Spatial autocorrelation arises when observations close in proximity tend to be more alike than those further apart and is present even if the observations have been recorded in a standardised way [4]. Spatial heterogeneity refers to the spatial variation or uneven distribution of attributes across a region [5]. Climatic and environmental factors, such as temperature, rainfall, and proximity to waterbodies, among others, are largely responsible for such spatial heterogeneity as its effects are usually only partially explained by the covariates that are available in a model [4]. Indeed, many other factors that vary geographically can also contribute to spatial heterogeneity in observations, such as the availability and distance to quality child health care, access to a reasonable transport system, culture, and the cost of living, all of which may not always be fully explained by the available covariates. Various methods of accounting for spatial autocorrelation and spatial heterogeneity have been well

Roberts and Zewotir *Journal of Health, Population and Nutrition* (2020) 39:8
https://doi.org/10.1186/s41043-020-00217-8

Journal of Health, Population and Nutrition

RESEARCH ARTICLE Open Access

Copula geoadditive modelling of anaemia and malaria in young children in Kenya, Malawi, Tanzania and Uganda

Danielle J. Roberts* and Temesgen Zewotir

Abstract

Background: Anaemia and malaria are the leading causes of sub-Saharan African childhood morbidity and mortality. This study aimed to explore the complex relationship between anaemia and malaria in young children across the districts or counties of four contiguous sub-Saharan African countries, namely Kenya, Malawi, Tanzania and Uganda, while accounting for the effects of socio-economic, demographic and environmental factors. Geospatial maps were constructed to visualise the relationship between the two responses across the districts of the countries.

Methods: A joint bivariate copula regression model was used, which estimates the correlation between the two responses conditional on the linear, non-linear and spatial effects of the explanatory variables considered. The copula framework allows the dependency structure between the responses to be isolated from their marginal distributions. The association between the two responses was set to vary according to the district of residence across the four countries.

Results: The study revealed a positive association between anaemia and malaria throughout the districts, the strength of which varied across the districts of the four countries. Due to this heterogeneous association between anaemia and malaria, we further considered the joint probability of each combination of outcome of anaemia and malaria to further reveal more about the relationship between the responses. A considerable number of districts had a high joint probability of a child being anaemic but not having malaria. This might suggest the existence of other significant drivers of childhood anaemia in these districts.

Conclusions: This study presents an alternative technique to joint modelling of anaemia and malaria in young children which assists in understanding more about their relationship compared to techniques of multivariate modelling. The approach used in this study can aid in visualising the relationship through mapping of their correlation and joint probabilities. These maps produced can then help policy makers target the correct set of interventions, or prevent the use of incorrect interventions, particularly for childhood anaemia, the causes of which are multiple and complex.

Keywords: Joint modelling, Joint probabilities, Kendall's tau, Spline smoothing

*Correspondence: danjader@gmail.com
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Shared component modelling of early childhood anaemia and malaria in Kenya, Malawi, Tanzania and Uganda

Danielle J. Roberts*, Temesgen Zewotir

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Abstract

Malaria and anaemia contribute substantially to child morbidity and mortality. Using a child-level shared component model, we sought to jointly model the residual spatial variation in the likelihood of these two correlated diseases, while controlling for individual-level, household-level and environmental characteristics. This shared component model allowed the district-level spatial effect to be partitioned into a shared and disease-specific spatial component. The results indicated that the spatial variation in the likelihood of malaria was more prominent compared to that of anaemia, for both the shared and specific spatial components. In addition, multiple districts associated with an increased likelihood of anaemia but a decreased likelihood of malaria were identified. This suggests that there are other drivers of anaemia in children in these districts, which warrants further investigation. The maps of the shared and disease-specific spatial patterns provide a tool to allow for more targeted action in malaria and anaemia control and prevention, as well as for the targeted allocation of limited district health system resources.

Keywords: Adjusted posterior odds ratios, Bayesian inference, Conditional autoregressive, Joint modelling, Spatial modelling

*Corresponding author
Email address: danjader@gmail.com (Danielle J. Roberts)