MAPPING ANAEMIA PREVALENCE IN CHILDREN UNDER-FIVE YEARS OLD IN MALAWI

MASTERS OF SCIENCE IN BIOSTATISTICS THESIS

JEREMOT MASOAMBETA

UNIVERSITY OF MALAWI

MAY 2023

MAPPING ANAEMIA PREVALENCE IN CHILDREN UNDER-FIVE YEARS OLD IN MALAWI

MSc (Biostatistics) Thesis

By

JEREMOT MASOAMBETA

BSc (Statistics) - University of Malawi

BEd (Science) - Domasi College of Education

Submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

MAY 2023

DECLARATION

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

Full Legal Name

Signature

Date

CERTIFICATE OF APPROVAL

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

Signed:

____

Michael G. Chipeta, PhD (Supervisor)

Date:

____

Signed:

____

Tsirizani Kaombe, PhD (Programme coordinator)

Date:

____

DEDICATION

I dedicate this thesis to my family and my siblings for all the love and support. MacConnell, Loreen, Annie, Tissy, Dennis, Jane, Griffin, Patrick, Clement, Love, Bridget, Gift, Millium, Chrissy, Peace, Zoe, Prince and Bertha.

ACKNOWLEDGEMENTS

I am very grateful to our Lord Jesus Christ for his life and for taking me through my entire academic journey.

I wholeheartedly respect and thank my supervisor Micheal Chipeta, PhD, for his untiring support, encouragement, constructive ideas and guidance throughout this project. You stood by me and ensured I had courage and zeal to keep going with my work.

Special thanks to my loving wife Peace Ngawe Sachira for her endless support and encouragement an inspirer and energiser throughout this project.

To all my friends, thanks for the encouragement and all timely support that you rendered to me throughout this project, of special mention are Nicholas Siame Adam and Glory Mshali. I have learnt alot from you. You have been my pillars in this academic journey.

Finally, I am grateful to DELTAS Africa Initiative for supporting this work.

ABSTRACT

Anaemia in children under-five years of age remains a major public health problem in Malawi. Effective anaemia reduction control programmes requires evidence-based targeting and optimum utilization of resources. Child anaemia has adverse consequences on physical growth and mental development and is associated with long-term health and economic consequences. Timely identification of locations highly impacted by anaemia is a key to optimise resources to fight against the burden.

The study aimed to map child anaemia prevalence to identify hotspot areas and assessed determinants of anaemia in children under-five years using 2015-2016 Malawi Demographic and Health Survey (MDHS) data. Generalised Linear Geostatistics Model (GLGM) was fitted to estimate and predict Malawi's child anaemia prevalence at a high spatial resolution of 5 $\times$ 5 km pixel level. A total of 4, 601 children aged 6-59 months were assessed. Out of these children, 2, 877 (62.5%) were anaemic.

Chikwawa, Nsanje and Salima were anaemia hotspot areas. At exceedance probability of 75%, these districts had anaemia prevalence above 62.5%. Child age, child fever, child stunting, number of children under-five years in a household, and household wealth index were significantly associated with child anaemia. Elevation also called the altitude of a place above sea level, had inversely association with child anaemia. Areas along water bodies were more prone to high child anemia prevalence. Chikwawa, Nsanje and Salima districts need priority in terms of anaemia reduction control programmes and interventions. Multisectoral approaches at all levels and nutrition programmes are needed in order to reduce child anaemia.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criteria
DALYs	Disability Adjusted Life Years
EP	Exceedance Probabilities
GDP	Gross Domestic Product
GLGM	Generalized Linear Geostatistics Model
GLM	Generalized Linear Model
GNT	Global Nutrition Targets
Hb	Haemoglobin
MBG	Model Based Geostatistical
MCML	Monte Carlo Maximum Likelihood
MDGs	Millenium Development Goals
MDHS	Malawi Demographic Health Survey
MPHC	Malawi Population and House Census
NSO	National Statistical Office
PHC	Primary Health Care
SDG	Sustainable Development Goal
SEAs	Standard Enumeration Areas
SSA	Sub-Saharan Africa
UNICEF	United Nations Children Fund
VIF	Variance inflation factor
WHA	World Health Assembly
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Background

Anaemia in children under-five years remains a significant public health challenge globally, and in sub-Saharan Africa (SSA) in particular (Ogunsakin, Babalola, & Akinyemi, 2020). It has adverse consequences on children such as impairment of cognitive development and is associated with long-term health and economic consequences (World Health Organization, 2017) both in developed and developing countries. World Health Organisation (WHO) data revealed anaemia as one of the ten most serious health problems globally (WHO, 2012). According to WHO, in 2011, 43% (273.2 million) of children aged 6–59 months globally were estimated to be anaemic (Amugsi, 2020). In Africa, approximately 60% of preschool children are anaemic (Amugsi, 2020). In 2019, approximately 62.3% (84.5 million) in SSA were anaemic Ogunsakin et al. (2020). The anaemia prevalence in SSA region ranges from 42% in Swaziland to 91% in Burkina Faso (Amugsi, 2020). The World Health Organization considers anaemia prevalence over 40% as a major public health problem, between 20% and 40% as a medium-level public health problem, and between 5% and 20% as a mild public health problem (WHO, 2001). Based on criteria by WHO, anaemia is therefore a severe public health problem in almost all the developing countries. In 2016, Malawi Demographic Health Survey (MDHS) findings indicated that prevalence of anaemia among children aged 6-59 months was at 63.0% which is high (National Statistical Office, 2017). High prevalence of anaemia and its consequences is so challenging to children's health particularly in growth and development (Kawo, Asfaw, & Yohannes, 2018).

In spite of several interventions and control programs such as iron supplementation and insecticide-treated bed nets distribution to curb the menace and provision of antimalarial medicine Ministry of Health (2022); Ogunsakin et al. (2020), anaemia is still a severe public health problem in Malawi. Existing studies such as, Calis et al. (2016); Khulu and Ramroop (2020); Tony, Ramroop, and Habyarimana (2021) on child anaemia did not look at spatial heterogeneity of childhood anaemia to better quantify childhood anaemia in space (geography) to identify disease hotspots that could benefit from targeted interventions. Therefore, knowledge on local spatial heterogeneity of child anaemia is essential for planning and evaluation of anaemia interventions. Timely identification of locations highly impacted by anaemia is key to optimise usage of resources to fight against anaemia.

1.1.1 Definition of anaemia

Anaemia is defined as a health condition characterised by insufficient haemoglobin (Hb) concentration in a human body (WHO, 2012). Haemoglobin is a basic unit of red blood cells responsible for oxygen transportation to the body's tissues. Anaemia results when haemoglobin concentration falls below accepted levels, due to either compromised production, excessive destruction or excessive loss of red blood cells (Macdonald, Alison, Mike, Rose, & Miriam, 2010). Severity of anaemia is decided by measurement of blood haemoglobin concentration. According to (World Health Organization, 2017), children under-five years old are anaemic if Hb $<$ 11.0 gram/decilitre (g/dL) and classified as mildly anaemic (Hb value of 10 – 10.9 g/dL), moderately anaemic (Hb value of 7 – 9.9 g/dL), and severely anaemic (Hb value of $<$ 7 g/dL).

1.1.2 Causes of anaemia

Anaemia arises from multifaceted factors and are classified as nutritional, non-nutritional, and genetic bases (Molla, Egata, Mesfin, Arega, & Getacher, 2020). Iron deficiency (ID) is estimated to contribute to approximately one-half of anaemia cases worldwide. Low iron content in the diet and low iron absorption are major risk factors for anaemia (National Statistical Office, 2017; Ngwira & Kazembe, 2016; Parbey et al., 2019). In the developing world, infectious diseases such as malaria, Helminth infections, Human Immune Virus (HIV) and tuberculosis (TB) are other important causes of anaemia (Gebreweld, Ali, Ali, & Fisha, 2019; National Statistical Office, 2017). Iron is the main component for the haemoglobin production (Rakanita, Sinuraya, Suradji, Suwantika, & Syamsunarno, 2020). Though there are limited studies on the aetiology of severe anaemia, malaria is frequently identified as a principal cause of severe anaemia, particularly in African children (Chaparro & Suchdev, 2019). Existing evidence suggests that severe anaemia, accounting for most anaemia related deaths, mostly occurs among children under-five years old, and generally in the rainy season when the incidence of malaria is at its peak (Simo et al., 2020). Other important causes of anaemia worldwide include infections, other nutritional deficiencies (especially folate and vitamins B_{12}, A and C) and genetic conditions (including sickle cell disease, thalassaemia – an inherited blood disorder – and chronic inflammation) (WHO, 2012). Additionally, insufficient safe drinking-water, inadequate hygiene and sanitary conditions, and poverty also contribute to the development of anaemia (Rakanita et al., 2020; Semedo, Santos, Baião, Luiz, & Da Veiga, 2014; WHO, 2012).

1.1.3 Burden of anaemia

Anaemia burden is so huge to the society and in children under-five years in particular. Anaemia was quantified to account for close to 9% of the total global disability burden from all conditions (World Health Organization, 2017). It, therefore, has significant consequences for human health as well as social and economic development. For instance, anaemia impairs cognitive development and is associated with both short and long-term health and economic consequences. The consequences have a significant impact on the growth and development of children in the early stages of life. It poses a significant public health issue leading to an increased risk of child mortality (Ogunsakin et al., 2020). Worldwide, anaemia accounted for 591, 000 perinatal deaths (World Health Organization, 2017).

Anaemia has impact on work productivity in adults as well. Instead of concentrating to their productive work, parents and guardians nurse the sick children. The phenomenon impacts the household economically in terms of income or wage losses from decreased productivity (World Health Organization, 2017). Worse still, exorbitant costs are incurred to cure anaemia in these infected children, for example in India where anaemia is very prevalent, the lifetime costs of iron-deficiency anaemia between the ages of 6 and 59 months amounted to 8.3 million disability-adjusted life-years (DALYs) and annual production losses of US$ 24 billion in 2013 (corresponding to 1.3% of GDP) (World

Health Organization, 2017).

1.2 Factors associated with anaemia in children under-five years

Broad range of factors contribute to prevalence and distribution of anaemia in children under-five years. These factors fall into biological, socio-economic and contextual determinants, with many acting simultaneously (World Health Organization, 2017). Examples of factors influencing anaemia in children include individual level factors; child age, child sex, child morbidities (i.e., fever, diarrhea, cough, malaria), household level factors; number of children under-five years in the household, number of children in the household, poor sanitation, maternal education and wealth index of the household and context level factors (Rainfall, elevation, temperature and vegetation cover) (Harding, Aguayo, Namirembe, & Webb, 2018; Ogunsakin et al., 2020; Rahman, Mushfiquee, Masud, & Howlader, 2019; Simo et al., 2020; Sorsa, Habtamu, & Kaso, 2021). Interventions to tackle anaemia must, therefore, integrate a range of potential and unique risk factors at play in a particular setting and address their independent and overlapping effects.

Observed association between child's age and anaemia has been reported in several studies (Harding et al., 2018; Ogunsakin et al., 2020; Rahman et al., 2019; Simo et al., 2020; Sorsa et al., 2021). Among children under-five years, the younger the child is, the more vulnerable the child is to anaemia. The risk of child anaemia decreases with increasing age (Rahman et al., 2019). At young age, children experience high rates of growth which increases demand for micronutrients such as iron, folate and vitamin

B_{12}. Ngwira and Kazembe (2016) reported that the chance of having anaemia is much higher in children aged 5 to 20 months and decreases thereafter. A community-based study in Ethiopia showed that child anaemia varies with the age group (Sorsa et al., 2021). Among the age group between 9 - 12 months, anaemia prevalence was 54.6%, and the lowest prevalence (15%) was reported among the older age group of 18 - 24 months.

Sicknesses in children like fever, diarrhoea, and cough are predisposing factors for anaemia as they may lead to loss of blood. In addition, sickness reduces bodily immunity, nutrient absorption, and appetite for food, and thus further predisposes such children to anaemia. This points to the need for measures that reduce or prevent the risk of such sicknesses in children. Measures may include sleeping under insecticide-treated mosquito nets, improving household sanitation, regular medical check-ups, and timely treatment for all childhood infectious illnesses.

Socio-economic status which accounts for household wealth index is another factor that influences variation in anaemia in children under-five years. It affects the prevalence of anaemia through several pathways. Households which are not well to do have poverty which is a major determinant of health outcomes (World Health Organization, 2017). Such households are associated with poor living and working conditions. In return, they are susceptible to poor water, sanitation, hygiene and inadequate infrastructure which can lead to increased disease. Poverty is also linked to inadequate access to health-care services including limited access to anaemia prevention and treatment services (iron supplements, de-worming, insecticide-treated bed nets, as well as reproductive care). In Malawi, the MDHS 2015-16 report on nutritious status of children under five indicated that the lowest wealth quintile were highly malnourished at 15% than in the highest wealth quintile at 6% (National Statistical Office, 2015). In the study of (Ogunsakin et al., 2020), It was revealed that less wealthy households have greater odds of being anaemic than children from wealthy households. This entails that poverty is a significant determinant of childhood anaemia.

Education is another major determinant of health outcomes. Studies by National Statistical Office (2015); Ogunsakin et al. (2020) have shown that mothers with higher educational status are more likely to provide a healthy and hygienic balanced diet, resulting in better health outcomes for both mothers and their children. Low maternal education level may affect mothers' ability to access and understand health and nutrition information, and ultimately negatively affect their children's quality of diet. Mothers' education level may also influence decision-making and compliance with recommended health practices such as iron supplementation or reproductive health practices, as well as care taking practices including feeding and hygiene behaviours. A study conducted by Sorsa et al. (2021), in Dodota district, Southeast Ethiopia, revealed that children whose mothers did not have formal education were at a higher risk of developing anaemia; 1.5-fold odds than children from educated mothers.

Anaemia and malnutrition remain a concerning health problem. Improving both the nutritional and anaemic status in children younger than 5 years is critical to ensure high quality of life to future contributors and leaders of a country. Insufficient folate, vitamin B_{12}, protein deficiencies, nutrients can also increase the risk of anaemia.

Malnutrition develops with either over- or under- consumption of food but herein, it is defined as an insufficient intake of nutrients and/or other minerals. In developing countries, a low nutritional status of a child is usually an indicator of health problems (Gaston, Habyarimana, & Ramroop, 2022). Consequences of malnutrition in children include, poor performance at school, delayed psychomotor development, lower capacity for work and reduced quality of life in adulthood. Determinants of nutritional status in children is based on anthropometric indicators (stunting, wasting and being underweight) in accordance to WHO growth standards. Stunting (height-for-age) indicates chronic or long-term malnutrition, wasting (low-weight-for-height) is linked to low food intake and/or illness and is described as acute malnutrition, while an underweight child (weight-for-age) can be either stunted, wasted or both.

1.3 Problem statement

Ending child anaemia is in tandem with SDG 3.1 and SDG 3.2 which strives at ending all forms of malnutrition and preventable deaths of children under-five years by 2030 and reducing under-five mortality to as low as 2.5% (Simo et al., 2020). WHO and United Nations Children Fund (UNICEF) recommend strategies for anaemia control to be integrated in a country's primary health care (PHC) system and existing programmes such as maternal and child health, integrated management of childhood illness (Roberts, Matthews, Snow, Zewotir, & Sartorius, 2020). Malawi like many SSA nations is a low-resourced nation. Identification of *hotspot* areas (i.e., areas with above average prevalence) is paramount to optimise utilization of available resources to fight child anaemia. Existing studies, Khulu and Ramroop (2020); Tony et al. (2021) etc on child anaemia did not consider spatial effects. Better understanding of child anaemia prevalence spatial effects is a fundamental key in reducing child mortality and morbidity due to anaemia. Geostatistic models are therefore best fit in low-resource settings where comprehensive disease registries do not exist. GLGM within a model-based geostatistical (MBG) framework are used for modelling and predicting prevalence, both at observed and unsampled locations, for spatially correlated data. Through risk mapping of the diseaseYankson, Anto, and Chipeta (2019), hotspot areas are of great value in tracking and guiding anaemia control efforts.

1.4 Objectives

1.4.1 Main objective

To map anaemia prevalence of children under five in Malawi, model-based geostatistical model.

1.4.2 Specific objectives

Specifically, the study aimed to:

- Assess factors of anaemia in children under-five years
- Identify areas that need targeted interventions in order to control anaemia in children under-five years

1.5 Justification of the study

Reliable and accurate estimation of the prevalence of anaemia is essential in planning, monitoring and targeting effective interventions. Malawi being low-resourced, it is even more imperative to come up with cost-effective and efficient methods of estimating prevalence of anaemia in children under-five years(National Statistical Office, 2015).

Spatial mapping is one of the effective ways to achieve the above as these spatial maps help identify hotspots and so make targeted interventions possible.The fact is that causes of child anaemia are multiple and complex and may not be fully addressed within a short term (Harding et al., 2018). For example, how do you address the educational levels of mothers in a particular area in the short term? Or the cultural practice of having many children, even under-five ones at the same time? Or religious beliefs that prevent adherents from accessing medical interventions? Culture, religion and education levels take time to tackle.

At all levels, both at local and international, interventions need to target high burden areas if fighting against anaemia is to be a success. Anaemia maps are of help to policy makers to formulate correct targets and interventions to reduce anaemia in children under-five years. years(Roberts & Zewotir, 2020).

1.6 Thesis structure

The thesis has six chapters, introduction through conclusion. Chapter 2 presents literature review about anaemia prevalence modelling. This chapter reviews the commonly applied generalised linear modelling techniques and approaches which have been used to model anaemia prevalence, giving a background in their formulation and development. Chapter 3 presents methodology employed for mapping anaemia prevalence disease in children under-five years. Chapter 4 presents results, such as exploratory analysis and GLGM modelling. Chapter 5 presents a discussion of results presented in Chapter 4. Finally, Chapter 6 presents the major findings, recommendations and limitations of the study.

1.7 Conclusion

Anaemia in children under-five years is a public health problem globally. It needs attention in order to curb the menace to human health which impacts social and economic development of individuals. In 2016, Child anaemia prevalence in Malawi was at 63% which is high and severe health problem according to WHO. In such a low-resourced nation, timely identification of locations highly impacted by anaemia is key in optimising usage of resources to fight against anaemia.

CHAPTER 2

LITERATURE REVIEW: ANAEMIA PREVALENCE MODELLING

2.1 Introduction

This chapter reviews the commonly applied generalised linear modelling techniques and approaches which have been used to model anaemia prevalence, giving a background in their formulation and development.

2.2 Generalised linear models

The class of generalised linear models (GLMs) extends the linear regression modelling framework (Nelder & Wedderburn, 1972). Response variables are not necessarily continuous and normally distributed. There is appropriate link function that link the outcome variable to the independent variables. Binomial and Poisson regression models are widely used to analyse counts of disease cases. Observations Y_i are assumed to be independent explanatory variables measured without error.

GLM consists of three components:

a) random component

Specifies the conditional distribution of response variable, $Y_i, i = 1, \ldots n$ which is independently sampled observations and a member of an exponential family. The basic form of exponential family is as follows:

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{2.1}$$

where

- $p(y;\theta,\phi)$ is the probability function for the discrete random variable Y, or the probability density function for continuous Y.
- $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions that vary from one exponential family to another
- θ is the canonical parameter for the exponential family
- $\phi > 0$ is a dispersion parameter

b) linear predictor

A linear function of regressors, X_{ij}.

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \tag{2.2}$$

c) A smooth and invertible linearizing link function $g(\cdot)$

Transforms the expectation of the response variable, $\mu_i = E(Yi)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \tag{2.3}$$

GLM model is used in many disciplines including medical research (Khulu, 2019). According to Khulu (2019), logistic regression is a special case of GLM. Commonly used regression model when the research does not want to handle the sample design into the analysis is simply logistic regression. However, survey logistic regression is used to incorporate sample design into the analysis. Survey logistic regression is given by

$$ln = \left[\frac{\eta_{ij}}{1-\eta_{ij}}\right] = \beta_0 + \beta_1 X_{1ij} + \ldots + \beta_P X_{pij} \tag{2.4}$$

2.2.1 Stepwise regression

In GLM modelling, individual variables are tested if they have any significance in a model. Thus stepwise regression aims to select a model step by step, adding or deleting one predictor at a time based on the statistical significance (Wang & Chen, 2018). Stepwise regression is a combination of both the forward and backward selection techniques. Every individual variable in the model is checked to see if its significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model. Stepwise regression requires two significance levels: one for adding variables and one for removing variables(Wang & Chen, 2018). The cut-off probability for adding variables should be less than the cut-off probability for removing variables so that the procedure does not get into an infinite loop.

In multiple linear regression there is one output variable but many input variables.

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_ix_i \tag{2.5}$$

where y is the response variable, x_i are covariates and b_i are coefficients that are to be generated by the linear regression algorithm. Instead, a subset of those features needs to be selected which can predict the output accurately. Stepwise regression is crucial in child anaemia modelling, many factors associated with child anaemia cannot be studied at the same time, stepwise regression is therefore done to reduce compute time and to remove redundant variables. This also reduces the complexity of the problem.

2.2.2 Correlation matrix

A correlation matrix is a simple way to summarize the correlations between all variables in a dataset (Komiyama, 2008). A correlation matrix can help us quickly understand the correlations between each pair of variables.

One key assumption of multiple linear regression is that no independent variable in the model is highly correlated with other variables (Hadavand-Siri & Deutsch, 2012). Multicollinearity in regression analysis occurs when two or more predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. Highly correlated variables cause problems when fitting and interpreting the regression model. Consequences of high multicollinearity is that it increases standard error of estimates of the β's and in return may lead to misleading results. Predictors whose correlation are close to 1 or -1, one of the two correlated predictors need not to be included in the model (Hadavand-Siri & Deutsch, 2012).

2.2.3 Variance inflation factor (VIF)

The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model (Shrestha, 2020).

The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and *p*-values in the regression output are likely unreliable.

Variance inflation factor is therefore used to ensure child anaemia modelling had no multicollinearity predictors.

2.2.4 Information criteria and model selection

Information criteria is a tool which is vital for identifying a model that best fits the data. Model selection refers to the problem of using the data to select one model from the list of competing models (de Graft Acquah, 2010). Model selection is a fundamental part of the statistical modelling process, and it has been an active research area since the 1970s (Xue, Luo, & Liang, 2017). Model selection using information criteria has been developed to summarise data evidence in favour of a model. Specifically, information criteria techniques emphasise minimising the amount of information required to express the data and model. For stepwise regression, the function step is called and the direction is set to "both" so that the algorithm can add and drop predictors in every iteration. Once it is called, the iterating process will proceed by itself. Just as what happens in the Information criteria, the process adds and/or subtracts the predictors till the best model that fits the data is yielded.

Akaike information criterion

The Akaike Information Criterion (AIC) is used to compare competing nested and non-nested models (Xue et al., 2017). The idea is to select the model that minimises the negative likelihood penalised by the number of parameters.

$$AIC = -2\log\ p(L) + 2p \tag{2.6}$$

where

- L refers to the likelihood under the fitted model
- p is the number of parameters

Bayesian information criteria (BIC)

BIC is another popular model selection principle. It selects the model that minimizes Equation (2.7),.defined as

$$BIC = -2\log\ p(L) + p\log(n) \tag{2.7}$$

where

- L refers to the likelihood under the fitted model
- p is the number of parameters
- n is the sample size

The AIC and the BIC do have the same aim of identifying good models even if they differ in their exact definition of a "good model"(Ding, Tarokh, & Yang, 2018). In both criteria, the best model that is selected is one with the minimum value.

While GLM methods are flexible and incorporate non-normal binary response variables as opposed to linear models, they are unable to measure the spatial effects of predictors. Geostatistics models provide a robust framework to understand the spatial variation of the burden (Noor, 2011).

2.3 Geostatistics models

The term geostatistics encompasses statistical methods relevant to the analysis of geo-located data whose aim is to study geographical variation throughout a region of interest but the available data are limited to observations from a finite number of sampled locations (Diggle & Giorgi, 2019). Geostatistics deals with statistical models and methods associated with spatially discrete data relating to an unobserved spatially continuous phenomenon (Diggle & Giorgi, 2016). Geostatistics data refers to data gathered at a discrete set of points in an area of interest say A, with an intention of understanding the behaviour of an unobserved, spatially continuous phenomenon that exists throughout A and could, in principle if not in practice, be observed at any point x in A. This phenomenon is typical in low-resource settings where comprehensive disease registries do not exist. Statistical theories for the analysis of geostatistical data propose model selection of variables for spatial estimation and prediction. For this reason it is important to think of what sampling points to use in order to obtain accurate predictions. The spatial estimation is the inference about the spatial process and prediction at new locations is based upon partial realization (Banerjee, Gelfand, & Carlin, 2003).

2.3.1 Data structure

The canonical geostatistical problem, expressed in the language of model-based geostatistics, is the following:- Data,

$$\{(y_i, x_i) : i = 1, \ldots, n\} \tag{2.8}$$

are realised values of random variables Y_i associated with pre-specified locations $x_i \in A \subset \mathbb{R}^2$. The Y_i are assumed to be statistically dependent on an unobserved stochastic process, $\{S(x) : x \in \mathbb{R}^2\}$, as expressed through a statistical model $[S,Y] = [S][Y|S]$, where $[\cdot]$ means "the distribution of," $Y = (Y_1, \ldots, Y_n)$ and $S = \{S(x_1), \ldots, S(x_n)\}$

The formal model-based solution is the conditional distribution, $[S|Y]$, which follows as a direct application of Bayes' theorem, $[S|Y]$.

$$[S|Y] = [S][Y|S] / \int [S][Y|S] dS \tag{2.9}$$

2.3.2 Geostatistical model formulation

Basically geostatistical prevalence survey consists of visiting communities at sampling locations $x_i : i = 1, \ldots, n$ distributed over a region of interest A and, in each community, sampling m_i individuals and recording whether each tests positive or negative for the disease of interest. If $p(x)$ denotes prevalence at location x, the standard sampling model for the resulting data is binomial

$$Y_i \sim \mathrm{Bin}(m_i, p_i) \tag{2.10}$$

Linkage of the $p(x_i)$ at different locations is usually desirable, and is essential if we wish to make inferences about $p(x)$ at unsampled locations x (Diggle & Giorgi, 2016).

Thus Y_i follows a binomial distribution with mean $E[Y_i|S(x_i), Z_i] = m_i p_i$ such that

$$Y_i = \log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = d(x_i)'\beta + S(x_i) + Z_i \tag{2.11}$$

where $d_i = d(x_i)$ is the set of explanatory variables and Z_i are independent N(0, τ^2) variates.

Model $S(x)$ is a stationary isotropic Gaussian process with variance σ^2 and Matern correlation function given by

$$\rho(u,\phi,\kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}\kappa(u/\phi), u > 0,$$

where $\phi > 0$ is a scale parameter, $\kappa_\kappa(.)$ is the modified Bessel function of the second kind of order $\kappa > 0$ and u is the distance between two sampling locations. The shape parameter κ determines the smoothness of $S(x)$, in the sense that $S(x)$ is $\lceil\kappa\rceil - 1$ times mean-square differentiable, with $\lceil\kappa\rceil$ denoting the smallest integer greater than or equal to κ.

2.3.3 Parameter estimation

A long-standing and important problem in geological mapping and modelling is the calculation of estimates at unsampled locations (Diggle & Giorgi, 2016). The central idea is to calculate an estimate that minimizes the expected squared error between the unknown true value and the estimate. In general, all data that are related to the unsampled location should have an opportunity to influence the estimate. Another central feature of estimation is that the estimates should be constructed with a clearly defined measure of optimality. In classical and non-Bayesian approaches parameter estimation precede spatial prediction (Diggle & Giorgi, 2016).

2.3.4 Generalised linear geostatistical model

The Generalised linear model is an extension of the linear regression model for analysing non-Gaussian data under the assumption that measurements at different locations are statistically independent of each other (Diggle & Giorgi, 2019). This section describes the extension of generalised linear models (GLMs) to the geostatistical models called Generalised Linear Geostatistical Models (GLGMs).

2.3.5 Motivation for GLGM modelling

The theory of generalized linear models and quasi-likelihood provides a flexible framework for analysing non-normal data (Nelder & Wedderburn, 1972). GLGM within a model-based geostatistical (MBG) framework are used for modelling and predicting prevalence, both at observed and unsampled locations, for spatially correlated data. A typical feature of most geostatistical problems is a focus on prediction rather than on parameter estimation. GLGM improves prediction of outcome of interest compared to random sampling techniques (Ngwira & Kazembe, 2015) and is therefore applied more generally to scientific problems that involve predictive inference about an unobserved spatial phenomenon $S(x)$ using any form of incomplete information (Diggle & Giorgi, 2016). GLGMs model prevalence mapping of disease in low-resource countries where registry data are lacking. Data predictions are vital in such areas. It informs public health action yielding to early interventions.

2.3.6 GLGM Model Formulation

The hierarchical representation of the joint distribution of a spatial process S and data $Y = (Y_1, \ldots, Y_n)$, is given as

$$[Y, S, \theta] = [S; \theta][Y|S; \theta] \tag{2.12}$$

where condition on S, Y_i are independent Normally distributed with means $d(x_i)^T\beta + S(x_i)$ and common variance τ^2.

Class of GLGMs is a modification of Equation (2.12) and the following are adhered to:

- Y_i is a non-Normal
- Incorporation of new set of independent random effects $U_i \sim N(0, v^2)$

The resulting model has the hierarchical form

$$[Y, S, U; \theta, v^2] = [S; \theta][U; v^2][Y|S; U; \theta] \tag{2.13}$$

Equation (2.13) is the GLGM model form. The observed responses $Y = (Y_1, \ldots, Y_n)$ are conditionally independent given the realisations of an unobserved Gaussian process $S(x)$ and a set of independent Normally distributed random variables U_i. Expectation of Y_i is $g(\eta_i)$, where

$$\eta_i = d(x_i)^T\beta + S(x_i) + U_i \tag{2.14}$$

where;

- η_i is the linear predictor
- $g(\cdot)$ is the link function of the model
- U_i is analogous to the nugget effect in the linear model

The non-hierarchical representation of Equation (2.13) is

$$Y_i = d(x_i)^T \beta + S(x_i) + U_i + Z_i \tag{2.15}$$

where the Z_i are mutually independent $N(0, \tau^2)$.

Most widely used examples of GLGMs use binomial logistic and Poisson log-linear models because it is possible to estimate both the covariance structure of the spatial process $S(x)$ and the variance of the independent random variables U_i (Diggle & Giorgi, 2019). In both sampling distributions the variance of the conditional sampling distribution of Y_i is a specified function of its mean.

Binomial sampling

Let x_i designate the location of a sampled community where n_i children are selected at each x_i to ascertain whether they have anaemia or not. If the number testing positive is Y_i such that $d(x)$ denote the explanatory variables associated with a location x. Then Y_i is a Binomial distribution with n_i trials and probability of a positive test $p(x_i)$, where

$$Y_{ij} = \log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = \alpha + d(x_i)^T \beta + S(x_i) + Z_i \tag{2.16}$$

Where α is the intercept parameter, $S(x)$ is unobservable random effect which is Gaussian process with zero mean and a constant variance σ^2 and Z_i are mutually independent zero-mean Gaussian random variables with variance τ^2. The index i represents the household and the index j represents an individual within the household.

Poisson sampling

This is widely used to model sampling distribution of outcome Y that it is open-ended count. Poisson distribution with mean $\lambda = np \sim Bin(n, p)$ whose n number of trials is large and probability of success p is small; mostly applicable for rare disease in a large population over a fixed spatial region over a fixed time-interval. The assumption is that cases occur independently in a spatial or temporal continuum (Diggle & Giorgi, 2019).

Poisson log-linear geostatistical model for random variable Y_i associated with location x_i is defined as

$$\log\{\lambda(x_i)\} = d(x_i)^T \beta + S(x_i) + U_i \tag{2.17}$$

where $S(x)$ is a Gaussian process with zero-mean, U_i is a set of independent zero-mean Normally distributed random variables and $\lambda(x_i)$ is conditional expectation.

In this study, distribution of child anaemia Y_i was not an open-ended count on the cluster x_i but rather the sampled individuals at each location x_i. Thus Binomial sampling was preferred over Poisson sampling.

2.3.7 Spatial prediction

In spatial prediction we firstly define predictive target T^*, location from the realisation of the spatial component of the linear predictor, $d(x)^T \beta + S(x)$ for all values of x in the region of interest A (Diggle & Giorgi, 2016). The focus is on the unexplained component of the spatial variation, $S(x)$. Extreme values in a predictive map of $S(x)$ hot-spot and cold-spot areas are very critical in spatial prediction. They provide the clear direction for evidence-based decisions to policy makers and programs implementers. Summary statistics of predictive distribution depends on probability distribution of T^* conditional on the observed data.

For the geostatistical Binomial model, a natural predictive target is the prevalence over the region of interest A given by

$$T^* = \{p(x) = \exp\{T(x)\}/(1+\exp\{T(x)\}) : x \in A\} \tag{2.18}$$

where $T(x) = d(x)^T\beta + S(x)$

In Equation (2.18), the point prediction takes the form of a map. Thus, the predictive target map shows a particular feature of the map, for example, its maximum or an indicator of whether the average value over a sub-region exceeds a policy-relevant threshold. The most commonly used summaries of the predictive distribution of a prevalence surface are maps of its means, standard errors and selected quantiles for each spatial unit (Giorgi, Diggle, Snow, & Noor, 2018)

In spatial prediction, we draw a number, B, of random samples from the predictive distribution of the complete spatial surface $\{S(x) : x \in A\}$. Values of the specific target from each sample, $T_1^*, \ldots, T_B^*$. Suitable summaries of the resulting empirical distribution of the T_i^*. In the prediction process, region A is approximated by a regular grid $\mathscr{X} = \{x_1^*, \ldots, x_q^*\}$. This regular grid, $\mathscr{X}$ have q prediction locations that cover A.

To make inference on T^*, we obtain samples from its predictive distribution, $[T^*|y]$. Since any target T^* is calculated directly from the fitted model parameters and the spatial $S(x)$, the problem reduces to sampling from the predictive distribution of $S^* = \{S(x) : x \in \mathscr{X}\}$. Hence

$$[S^*|y] = \int [S^*, S|y]dS = \int [S|y][S^*|S]dS \tag{2.19}$$

where $[S^*|S,y] = [S^*|S]$. To sample from $[S^*|y]$, we sample from $[S|y]$ and then from $[S^*|S]$ to obtain our sample s_h^* for $h = 1, \ldots, B$. $[S^*|S]$ is a multivariate Gaussian distribution with mean vector and covariance matrix given by

$$E[T^*|y] = \mu^* + \sigma^2 C^T \Sigma^{-1}(y - \mu) \tag{2.20}$$

where C is the n by q matrix with i^{th} column

$$c_i = (\rho(||x_i^* - x_1||;\phi), \ldots, ||x^* - x_n||;\phi)$$

and

$$Cov[T^*|y] = \Sigma^* - \sigma^4 C^T \Sigma^{-1} C \tag{2.21}$$

where the matrix Σ^* has (i, j)-th element $\Sigma_{ij}^* = \sigma^2 \rho(||x_i^* - x_n^*||;\phi)$.

In spatial prediction the unobserved data is interpolated from the observed data which together are used to have fine maps with exceedance probabilities showing extreme areas of disease prevalence.

2.3.8 Exceedance probabilities

In prevalence estimation analysis, it is worthy to identify *hotspot* areas for sound decision making (Yankson et al., 2019). To identify hotspot areas, defined as areas with prevalence above average, or some policy relevant threshold, we use what is called an exceedance probability. Areas whose anaemia prevalence is above a set threshold, say c, may be considered to have high anaemia prevalence. Exceedance probabilities are important when assessing the localised spatial behaviour of a phenomenon and the assessment of unusual clustering or aggregation of disease (Ghosh, 2009). The simplest case of an exceedance probability is

$$EP = Pr(x > c) \tag{2.22}$$

where $Pr(x)$ is a probability which estimates how frequently the relative risk exceeds the null risk and can be regarded as an indicator of how unusual the risk is in that area. This leads to assessment of hotspot communities. The closer the EP is to 1 the higher the likelihood that it is to be above the threshold c. If EP is closer to 0, the prevalence is likely to be below the threshold c. In cases where it is closer to 0.5, prevalence is equally likely to be above or below the threshold c.

2.4 Spatial model diagnostics

2.4.1 Variograms

A variogram is an exploratory tool for spatial data. It is widely used in geostatistical analysis for both exploratory analysis and model validation for parameter estimation and formal model comparison. It is mostly used in likelihood based methods, whether non-Bayesian or Bayesian (Diggle & Giorgi, 2019). In spatial data analysis variograms are often used instead of covariance functions in other analyses. The variogram is based on second-order moments, and therefore gives a very natural way to describe the dependence structure in a Gaussian model (Brockwell & Davis, 2006). Variograms are computed to explore spatial correlation in the data. The variogram of a spatial stochastic process $S(x)$ is the function

$$V(x,x') = \frac{1}{2}Var\{S(x) - S(x')\} \tag{2.23}$$

where x and x' are any two points in $\mathbb{R}^2$. If the process $S(x)$ is stationary with variance σ^2 and correlation function $\rho(u)$, then

$$\begin{aligned} V_S(x,x') &= \frac{1}{2}(Var\{S(x)\} + Var\{S(x')\} - 2Cov\{S(x),S(x')\}) \\ &= \sigma^2\{1-\rho(u)\} \end{aligned}$$

For a geostatistical data set, $(x_i,Y_i) : i = 1,\ldots,n$, if Y_i is imprecise, the value of $Y(x)$ is not unique. For instance two measurements Y_1 and Y_2 at the same location x, may well have different values.

For data value

$$Y_i = S(x_i) + Z_i : i = 1,\ldots,n \tag{2.24}$$

where Z_i are mutually independent with mean zero and variance τ^2. If $S(x)$ in Equation (2.24) has a variance σ^2 and correlation function $\rho(u)$, where u, is the distance between the data-locations x_i and x_j, then $Corr\{Y_i,Y_j\} = \sigma^2\rho(\nu)/(\tau^2+\sigma^2)$, which approaches $\sigma^2/(\tau^2+\sigma^2)$ as u approaches zero. The equivalent expression as a variogram is

$$V(u) = \tau^2 + \sigma^2\{1-\rho(u)\} \tag{2.25}$$

The measurement error variance τ^2 is also called the nugget variance or simply nugget. The variance of $S(x)$ is sometimes called the sill.

Computation of variograms

A variogram can be computed through the equation Equation (2.26).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{ij \in N(h)} (z_i - z_j)^2 \tag{2.26}$$

Where $N(h)$ is a number of pair observations (i, j) separated by a spatial distance h. Terms z_i and z_j are the attribute values of observations i and j respectively.

The function in Equation (2.26) calculates the attribute difference between neighbouring observations separated by a lag h to evaluate if these observations display the same information. Semi-variance increases with increase in the distance between observations because near observations share more characteristics than distant ones.

Variogram parameters

- nugget

  It represents the small-scale spatial variation within the field. An indicator of how noisy the spatial structure is. For instance, inside the community, there might exist children exposed to different factors that may cause anaemia. When the minimal distance between the said children is very large, the nugget might be found higher than it should be.

- partial still

  Represents the magnitude of variation of the variable of interest. Intuitively, the higher the partial sill compared to the nugget, the stronger the spatial structure. The sill is the variance of the dataset and can be computed as the sum of the partial sill and nugget.

- range

  This is the distance beyond which observations are no longer spatially correlated. On average, above a specific spatial distance and whatever the pair of points examined, observations are too dissimilar and do not share any relationship.

2.5 Model validation

Validation of a fitted linear geostatistical model is based on the empirical variogram of the residuals from an ordinary least squares fit to the fixed effects component of the model. In GLGM the variation in the outcome, is partitioned into explained and unexplained variation on the scale of the linear predictor. Assessment is therefore based on how the unexplained component is spatially correlated (Diggle & Giorgi, 2019). Based on the GLM framework we first assume that Y_i conditionally on a set of independent Gaussian variables Z_i, with mean zero and variance τ^2, belongs to the family of exponential distribution, with link function $g(\cdot)$ and linear predictor

$$\eta_i = d(x_i)^T \beta + Z_i \tag{2.27}$$

In the geostatistical setting, unexplained variation, Z_i in Y_i might also include a spatially structured component, which would manifest itself in the form of residual spatial correlation. Evidence of spatial dependence in the data is through testing of independence of Z_i in Equation (2.27). A point predictor $\tilde{Z}(x_i)$ is used since Z_i are not observed. The point predictor is a suitable summary of the conditional distribution of Z_i given the data, called the predictive distribution of Z_i. By the application of Baye's theorem,

$$[Z_i|y_i] = [Z_i, y_i] / \int_{\mathbb{R}} [Z_i][y_i|Z_i] \, dZ_i \tag{2.28}$$

Common choices for the point predictor are the mean, median or mode of $[Z_i|y_i]$. Mean is however favoured since it minimises the mean square error, $E[(\tilde{Z}_i - Z_i)^2]$ (Diggle & Giorgi, 2019). This distribution $[Z_i|y_i]$ depends on the model parameters.

To establish whether the apparent patterns of Z_i are or are not compatible with random fluctuations about a constant value, we use the Monte Carlo strategy to simulate the behaviour of empirical variograms under the assumption of spatial independence.

The first step in the model validation is to randomly permute the labelling of the $\tilde{Z}_i$ while holding fixed the location x_i. Then we compute the empirical variogram using the permuted $\tilde{Z}_i$. The first two steps are repeated B times. The B empirical variograms are used to compute pointwise 95% tolerance interval at each of the pre-specified distance bins under the hypothesis of spatial independence. Residual spatial correlation is validated if the empirical variogram falls within the generated 95% tolerance band.

2.6 Conclusion

To identify anaemia hotspot areas, geostatistics model is crucial to study and understand the behaviour of an unobserved spatially continuous phenomenon in the study area.

CHAPTER 3

METHODS

3.1 Introduction

This chapter presents the methodology employed for mapping anaemia prevalence disease in children under-five years. It focuses on the study area and data sources, cluster locations and how the data was managed and analysed. It further provides information on the statistical modelling and model selection procedure.

3.2 Study area

Malawi is a landlocked country South of the equator in sub−Saharan Africa (SSA). It is bordered to the North and North-East by the United Republic of Tanzania; to the East, south and South-West by Mozambique and to the North West by the Republic of Zambia. The country is 901 kilometres long and ranges in width from 80 to 161 kilometres. It has a total area of 118,484 square kilometres of which 94,276 square kilometres is land area. The remaining area is mostly composed of Lake Malawi, which is about 475 kilometres long and runs down Malawi's Eastern boundary with Mozambique. Malawi's most striking topographic feature is the Rift Valley that runs the entire length of the country, passing through Lake Malawi in the Northern and Central regions to the Shire Valley in the South. The Shire River drains the water from Lake Malawi into the Zambezi River in Mozambique. To the West and South of Lake Malawi lie fertile plains and mountain ranges whose peaks range from 1,700 to 3,000 metres above sea level. See Figure 1.

The country is divided into three regions: the Northern, Central and Southern regions which are further divided into districts. In total, there are 28 districts in the country. Administratively, the districts are subdivided into Traditional Authorities (TAs), presided over by chiefs. Traditional Authorities are composed of villages, which are the smallest administrative units and are presided over by village headmen and headwomen. Demographically, the country has a population of 17,563,749 people of which 9,042,289 (51.5%) are female while men account for 8,521,460 (48.5%) (National Statistical Office, 2019). The population is largely rural based with only 16.0 percent residing in the urban areas.

Figure 1: Geographical position of Malawi

3.3 Data sources

The study used 2015-2016 Malawi Demographic Health Survey (MDHS) data whose survey took place between October 2015 and February 2016 (National Statistical Office, 2015). It captures information in such areas as nutrition of children, births from women aged 15-49, women's characteristics, children's and household characteristics, among others. For the prediction and projecting of WHA Global Nutritional Targets, 2004 MDHS, 2010 MDHS and 2019-20 MICS data were also used. To obtain 2015-2016 MDHS data, a nationally representative sample of households was selected using two staged sampling technique. First stage was to select 850 clusters (standard enumeration areas, SEAs) from a master sampling frame constructed from the 2008 Malawi population and housing census (MPHS). Clusters were selected from 173 urban and 677 rural areas using a probability proportional to size selection. In the second stage, households listed from the selected clusters were used as a sampling frame for selecting households into the final sample. All women aged 15 to 49 and children aged 6-59 months were eligible to participate in the survey and with the parent's or guardian's consent, children aged 6-59 months were also tested for anaemia. The data on children were obtained through face-to-face interviews with their mothers/caretakers. In total 5,245 children under-five years were tested for anaemia from which 4,601 were included in the study.

A total representative sample of 27,516 households was selected, and 26,564 households were considered to be occupied in the 2015-16 MDHS. Data collection was done through questionnaires. Total households that were successfully interviewed were 26,361, yielding a response rate of 99.2%. Out of 25,146 eligible women, 24,562 were successfully interviewed, yielding a response rate of 97.7%. The data set used in the analysis was child record data set, which was based on woman and household questionnaires. Data management in terms of extracting and generation of variables from child record data set was done in STATA Version 14 (StataCorp, 2015) and in R (R Core Team, 2021).

3.4 Cluster locations

Figure 2 shows sampled cluster locations within the 28 districts of Malawi. Many clusters were situated in the Southern part of Malawi followed by the Central region (Table 1).

Table 1: Cluster allocation

Region	Clusters No	Individuals sampled
Northen	161	819
Central	284	1599
Southern	402	2183
Total	**847**	**4601**

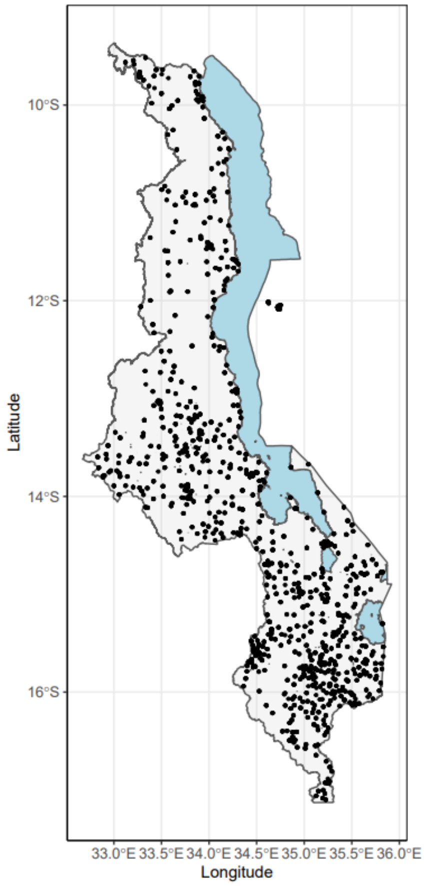


Figure 2: MDHS sampled cluster locations.

3.5 Statistical analysis

Bivariate analysis was employed to examine association between the response variable (anemia status) and the explanatory variables. Statistical significance was dependent on the p-values and was set at $p = 0.05$ representing 5% level of significance. Stepwise regression model selection using GLM was used to identify the covariates; identified covariates were carried forward to fit geostatistics binomial logistic regression models and spatial predictions using Monte Carlo Maximum Likelihood (MCML) approach in order to predict anaemia outcomes. The results in GLM logistic regressions were used to examine the relationship between anaemia status and the determinant variables.

3.5.1 Measure of variables

Response variable

In the study, the binary response variable, anaemia status, was determined by haemoglobin level. Children whose haemoglobin values were less than 11.0 g/dl were anaemic and were coded as 1. Children with values of haemoglobin values of 11.0 g/dl or higher were categorised as not anaemic and were coded as 0.

Let Y_i be anaemia status of children aged 6–59 month. Then;

$$y_{ij} = \begin{cases} 1\text{: Hb} < 11.0 \text{ g/dL} & \text{anaemic} \\ 0\text{: Hb} \geq 11.0 \text{ g/dL} & \text{not anaemic} \end{cases}$$

Explanatory variables

In the analysis, covariates from the community, household, child levels and environmental factors were included. Community level variables included residence and region. Household level factors included maternal education, age of household head, sex of household head, number of children under-five years, wealth index, water source and birth order. Individual level variables included were child related variables. Child specific variables include child age, child fever, child birth weight, child diarrhoea, child cough, child stunting, child sex and haemoglobin level. Environmental factors included, population count in 2015, rainfall amount in 2015, land surface temperature, altitude, mean temperature in 2015, proximity to water, annual precipitation 2015, built population in 2014, day land surface temperature in 2015, diurnal temperature range 2015, enhanced vegetation index 2015, insecticide treated bed-net use coverage in 2015, land surface temperature 2015, malaria incidence and prevalence in 2015 and nightlights (composite). The covariates are presented in a summarised form in Table 2.

Table 2: Covariates used

Individual	**Household**	**Community**	**Environment**
child age in months	wealth index	region	Elevation
child fever	no of u5 children	residence	rainfall
child stunt	maternal education	GPS locations	malaria incidence
child sex	age of household head		population count
child birth weight	sex of household head		land surface temperature
child diarrhoea	water source		mean temperature
child cough	birth order		proximity to water
child wasting	toilet share		annual precipitation
child overweight	toilet type		day land surface temp
haemoglobin level			diurnal temperature range
			enhanced vegetation index
			itn coverage
			malaria prevalence
			nightlights

Two GLM models (Model 1, Model 2) were fitted: Model 1 had the following covariates: child age, elevation, wealth index, number of under five children, child stunting, child fever and age of the household head.

Model 2 had the following covariates: child age, elevation, wealth index, number of under five children, child stunting and child fever. The BIC at 95% confidence interval was used to select the best model. The two models were fitted to find one model that best fitted the model. The model with least value of BIC was regarded as the best model.

3.5.2 Geostatistical model for child anaemia

The formulation of the geostatistical model followed standard geostatistical model for prevalence surveys we defined earlier in Equation (2.11). For the visited enumeration areas at different villages x_i, in the study region $A \subset \mathbb{R}^2$ and sample $m_i : i = 1, \ldots, n$ individuals at risk in each community. y_i designate the number of children who tested anaemia positive out of m_i children at location x_i in the community $A \subset \mathbb{R}^2$, and a vector of associated covariates $d(x_i) \in \mathbb{R}^p$. The standard geostatistical model for $Y_i \sim Bin(m_i, p(x))$ where $p(x_i)$ measures the child anaemia prevalence at location x_i. Adopting the logistic link function, the model assumes that:

$$\log\left\{\frac{p(x)}{1-p(x)}\right\} = \alpha + d(x)'\beta + S(x) \tag{3.1}$$

where;

- α is the intercept parameter
- $S(x)$ is an unobserved random effect which is Gaussian process with zero mean and a constant variance σ^2
- $d(\cdot)$ is a vector of observed spatial explanatory variables associated with the response Y
- β is a vector of spatial regression coefficients for the covariates.

The empirical logit transform is defined as follows:

$$Y_{ij}^* = \log\left\{\frac{(Y_{ij} + 0.5)}{m_{ij} - Y_{ij} + 0.5}\right\} \tag{3.2}$$

The underlining assumption is that:

$$Y_{ij}^{*} = \alpha + d(x_{ij})'\beta + S(x_i) + Z_i \tag{3.3}$$

where;

- Z_i are mutually independent zero-mean Gaussian random variables with variance τ^2
- Index i represents the household and the index j represents an individual within the household.

All the anaemia severity levels; namely mildly anaemic (10–10.9 g/dL), moderately anaemic (7–9.9 g/dL), and severely anaemic (<7 g/dL) fell under status "anaemia presence" coded as 1 and 0 otherwise. The covariate vector $d(x_i)'$ represents DHS covariates in the model.

Spatial model was fitted. The Matérn shape parameter κ and relative variance parameters τ^2 were fixed at 1.5 and 0, respectively. We fix the shape parameter because it is generally very difficult to estimate it from data. It was assumed that the true surface was a realisation of a stationary Gaussian process. The resulting estimated model parameters in the model were used to make spatial predictions over a fine grid of 5 $\times$ 5 km. Anaemia prevalence in children under-five years for all the unsampled locations was mapped. All the analysis and mapping were carried out using the R statistical software environment version 3.6.0 (R Core Team, 2021).

3.6 Ethical considerations

To ensure research ethics were not compromised, the survey protocol was reviewed and approved by the National Health Sciences Research Committee in Malawi and the ICF Institutional Review Board National Statistical Office (2015). Data collectors were also trained on how to conduct themselves during the data collection in order to respect the rights of individuals participating in the study. Participants were informed of their right to determine whether they wanted to participate in the study or not. Blood specimens for anaemia testing were collected from women age 15-49 who voluntarily consented to be tested and from all children age 6-59 months for whom consent was obtained from their parents or the adult responsible for the children. They were also informed of their right to abstain or withdraw at any time without reprisal. The risks and benefits of the study were adequately explained to study participants. Parents of children with a haemoglobin level below 7 g/dl were instructed to take the child to a health facility for follow-up care National Statistical Office (2015). Suitably trained investigators conducted the study, using an approved protocol. Written informed consent was obtained from participants before the survey. For children aged 6-59 months, consent was obtained from their parents or the adults responsible for the children. In this study, DHS Program, USA, granted permission to use the data after application was made to seek for the permission. No further ethical clearance was therefore sought.

Data collected was protected and was strictly used by NSO for analytical purposes for three years from the time of data collection. De-identified was available to the public after release of the 2015- 16 MDHS (National Statistical Office, 2015).

3.7 Conclusion

The study used 2015/16 MDHS data. Sampled clusters were within the 28 districts. In total, 4, 601 individuals were used in the analysis. Anaemia status of children aged 6-59 months was dependent on haemoglobin level (Hb), Hb $<$ 11.0 g/dL being anaemic and Hb $>$ 11.0 g/dL not anaemic. In the bivariate analysis, association between the anemia status and other explanatory variables was sought. Stepwise regression model selection using GLM was used to identify variables to best fit GLGM model.

CHAPTER 4

RESULTS

4.1 Introduction

The first part of this Chapter 4 presents results based on exploratory data analysis from the bivariate analysis and results of non-spatial binary logistic regression(GLM model). The remainder of the Chapter presents result from geostatistics binomial logistic regression models and spatial predictions.

4.2 Characteristics of study population

A total of 4,601 children aged 6–59 months (3,370 males and 1,231 females) were included in the study. Out of these children; 2,877 were anaemic representing 62.5% anaemia prevalence national-wide. Table 8 in appendix 1 summarises the proportions of children with anaemia. Children from Southern part of Malawi had higher anaemia prevalence at 63.5% followed by Central region and Northern region with anaemia prevalence of 62.4% and 60.2% respectively. Children from the rural areas were more vulnerable to anaemia. Out of 3,865 cases examined, 2,460 were anaemic representing 63.6% prevalence compared to 56.7% of children from the urban residence. Anaemia prevalence was high in children from households whose water sources were unimproved, at 63% against 61.3%. Children from households which had unimproved toilet facilities were at high risk of child anaemia, they had anaemia prevalence of 68.1% compared to 62.0% prevalence of anaemia from households with improved toilet. Children from male headed households had slightly less proportion of anaemia, at 62.2% as compared to female headed households, 63.4%.

Children from households with shared toilets recorded high anaemia prevalence, 64.5% than children from households which unshared toilets, 61.4%. Highest anaemia prevalence was registered in children from poorest households, 69.3% than in children from the richest households, 56%. The trend entails that child anaemia was less likely in well-to-do households. Children who were breastfed during the first 6 months of birth had 62.5% proportion of child anaemia compared to their counterpart, at 63.6%. Anaemia prevalence was high in children who were stunted at 66.4% than in children who were not stunted 60.3%. Child anaemia was higher in children who were wasting, at 64.8% compared to students who were not wasting, 62.5%. Prevalence was shown to be higher in children under weight at 64.8% compared to children not under weight, 62.5%. Male children anaemia prevalence was slightly higher at 63.5% than their female counter part, 61.6%.

Children of mothers with no education were shown to have the highest anaemia prevalence at 69.4% compared to children of mothers with higher education level (45.5%). Increasing levels of education were associated with decrease of anaemia prevalence in children. Children who had fever during the survey had higher prevalence at 67.6% compared to children who had no fever. Anaemia prevalence in children who had diarrhoea was high at 68.1% than children who had no diarrhoea, 61.4%. Prevalence was shown to be the same in children who had cough (62.4%) and children without cough (62.6%).

The average number of household members for the households which had anaemic children was 5.54 ± 2.06. Average age of the household head for anaemic households was 35.85 ± 12.11 years. The average number of anaemic children in a household was 1.67 ± 0.72. The average age for anaemic child was 35.05 ± 15.61 months. The average birth weight for the anaemic children were 3947.98 ± 2101.28 grams

4.3 Bivariate test of individual covariates

The bivariate analysis was used to identify variables which were associated with anaemia in children under five See Table 8 in appendix 1. All covariates with p-value less than 0.05 were significantly associated with child anaemia. Table 3 is a summary of covariates significantly associated with anaemia and were retained to be part of the GLM model.

Table 3: Statistically significant covariates in bivariate test

Covariate	**p-value**
residence	<0.001
wealth index	<0.001
maternal education	<0.001
number of children under five	<0.001
child age in months	<0.001
child fever	<0.001
child stunting	<0.001
child diarrhea	<0.001

4.4 Generalized linear model

The stepwise regression method using glm function was used to select explanatory variables. Two GLM models were fitted, as presented in Table 4; Model 1 had a BIC of 5,916, and Model 2 had a BIC of 5,911. Model 2 was selected for having least BIC. In the model selection process, non-significant covariates (i.e., those with p-value $>$ 0.05) in Model 1 were dropped to yield Model 2. All the covariates in Model 2 were statistically significant.

Table 4: Fitted GLM models

Covariates	Model 1		Model 2	
	Estimate	p-value	Estimate	p-value
(Intercept)	1.882	0.000	1.722	< 0.001
Child age (mth)	-0.023	0.000	-0.023	< 0.001
Elevation	-0.001	0.000	-0.001	< 0.001
Wealth-poor	-0.158	0.102	-0.162	0.094
Wealth-middle	-0.449	0.000	-0.466	< 0.001
Wealth-richer	-0.304	0.003	-0.330	0.001
Wealth-richest	-0.489	0.000	-0.516	< 0.001
No children u5	0.240	0.000	0.232	< 0.001
Child stunted	0.243	0.000	0.239	< 0.001
Child fever	0.214	0.003	0.209	0.004
HH age	-0.005	0.060	-	-
BIC	5916		5911	

4.4.1 Correlation matrix and variance inflation factor

The likelihood ratio test (BIC values of 5911) suggests that model 2 in Table 4 provides an excellent fit to the data. For the multicollinearity, the VIF was employed to check for multicollinearity among the selected covariates, none of the VIF values was up to 5. This implies that there was no collinearity in the model. Correlation matrix for all variables have Pearson correlation coefficient of less than $|0.75|$ and it was assumed that multicollinearity was not present (Table 5).

Table 5: Correlation matrix and variation inflation factor

Covariate	Child age	Child fever	Child stunted	Children u5	Wealth index	Elevation
Child age (mth)	1.000					
Child fever	-0.137	1.000				
Child stunted	0.023	0.015	1.000			
No children u5	0.192	-0.068	0.066	1.000		
Wealth index	-0.046	-0.045	-0.130	-0.080	1.000	
Elevation	-0.008	-0.023	0.039	-0.036	0.026	1.000
VIF	1.074	1.025	1.020	1.071	1.032	1.006

Table 4 is a GLM model (Model 2) which best fitted the 2015-2016 MDHS under five anaemia prevalence data.

Child age was found to inversely associate with anaemia in children under-five years, -0.023 (p; < 0.001). An increase in age is associated with a reduction in anaemia outcomes. Child stunting was found to directly associate with childhood anaemia, 0.239 (p; < 0.001). Stunting children were more vulnerable to anaemia than non stunting children. Child fever was shown to be directly associated with anaemia in children under-five years, 0.209 (p; $= 0.004$). More children 881 (67.6%) who suffered from fever had anaemia. Number of children under-five years in a household was found to be positively associated with anaemia, 0.232 (p; < 0.001). Children from households with more children under-five years were at higher risk of having anaemia. Child anaemia was found to inversely associate with the wealth index of the household, (poor, middle, richer and richest, with -0.162 (p; $= 0.094$), -0.466 (p-value < 0.001), -0.330 (p; $=$

0.001) and -0.516 (p; < 0.001) respectively. The trend entails that children from well-to-do households were at less risk of having anaemia. Land elevation was found to be inversely associated with child anaemia, -0.001(p; < 0.001). Children living in highlands had less anaemia prevalence.

4.5 Spatial modelling of anaemia using GLGM model

Table 6: Monte Carlo maximum likelihood estimates and 95% confidence intervals for the binomial logistic model fitted to 2015-16 MDHS under five years anaemia prevalence data.

Covariate	Odds ratio	Lower Bound	Upper Bound
Intercept	5.842	4.063	8.398
Child‘s age	0.976	0.972	0.981
Child stunted	1.285	1.124	1.470
Fever	1.223	1.055	1.416
U5 children	1.250	1.138	1.374
Wealth - Middle	0.692	0.582	0.823
Wealth - richer	0.800	0.670	0.957
Wealth - Richest	0.655	0.541	0.793
Elevation	0.999	0.999	1.000
σ^2	1.184	0.303	4.618
ϕ	1.124	0.065	19.356

A unit increase in child age reduced anaemia in children under-five years by 2.4%, i.e., an estimate of 0.976 with (95% CI[0.972, 0.981]). Children who were stunting had 29% risk of being anaemic than non stunting children, i.e. adjusted odds ratio (aOR) 1.285 (95% CI[1.124, 1.470]). The study found that children with fever were at high risk by 22% compared to children who had no fever, aOR 1.223 (95% CI[1.055, 1.416]). A unit increase in number of children under-five years in the household increased anaemia risk by 25%, aOR 1.250 (95% CI [1.138, 1.374]). Children from middle, richer and richest households had reduction in risk of anaemic by 30.8%, aOR 0.692 (95% CI [0.582, 0.823]), 20%, aOR 0.800 (95% CI [0.670, 0.957]) and 34.5%, aOR 0.655 (95% CI [0.541, 0.793]), respectively, as compared to the children from the poor households. Children from households at upper altitudes were less likely to have anaemia than children from households at lower altitudes. Elevation was, however, not statistically significant in explaining child anaemia (Table 6).

4.6 Diagnostic test for spatial modeling

To test for the validity of independence assumption of the spatial process, empirical variogram was used.

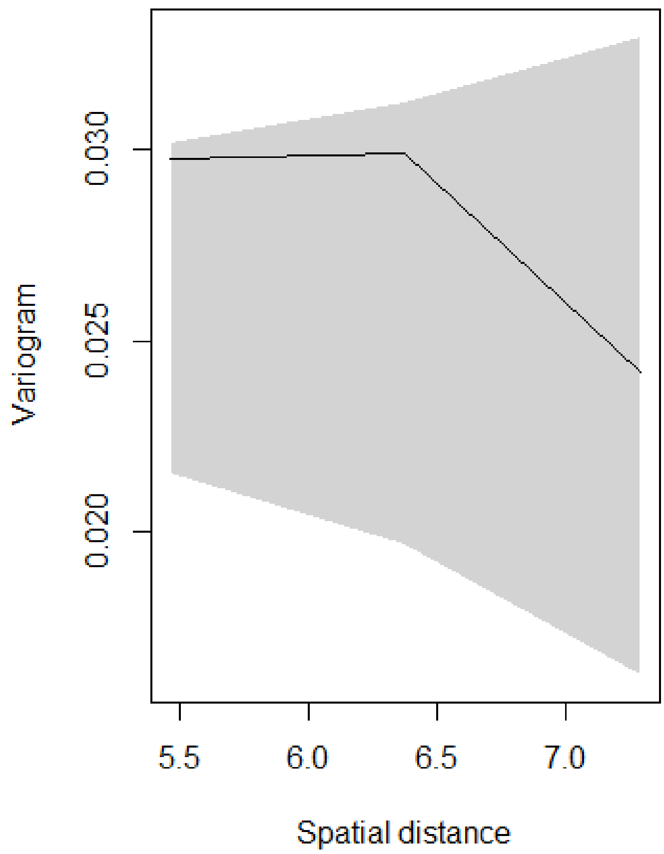


Figure 3: Empirical variogram for the residuals

The solid line is the empirical variogram of the residuals from a standard linear regression model. The shaded area is a 95% pointwise tolerance band generated under the fitted geostatistical model. Empirical variogram in Figure 3 entails compatibility of the adopted spatial correlation function for the child anaemia data. The empirical variogram is falling within the shaded area.

4.7 Anaemia prevalence in children

Model-based geostatistical methods allowed the mapping of prevalence at a fine-scale resolution of 5×5 km. Based on Figure 4, anaemia prevalence in children under 5 years was generally high in Malawi at 62.5% as of 2016, characterised by several hotspots.

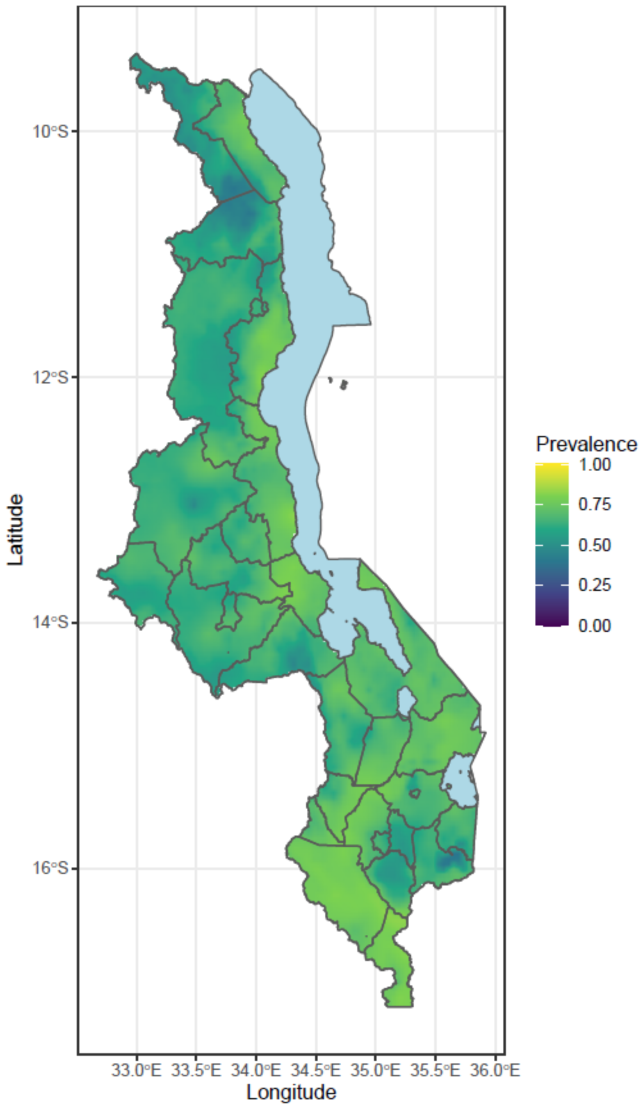


Figure 4: Anaemia prevalence among children aged under 5 years in Malawi

Anaemia prevalence at district level showed that anaemia prevalence was very high in Nsanje, Chikwawa and Salima districts, prevalence of greater than 75%. Mangochi, Balaka, Mwanza, Nkhotakota, Neno, Machinga, Likoma, Blantyre and Salima had anaemia prevalence between 70% to 75%. Chiradzulu, Rumphi and Chitipa had anaemia prevalence of less than 60% (see Table 9 in appendix 2 and Figure 9 in appendix 3).

Most districts in Malawi had anaemia prevalence of greater than 62.5%. This showed that Malawi remains a high anaemia prevalence country for child anaemia (Figure 4 and Figure 5).

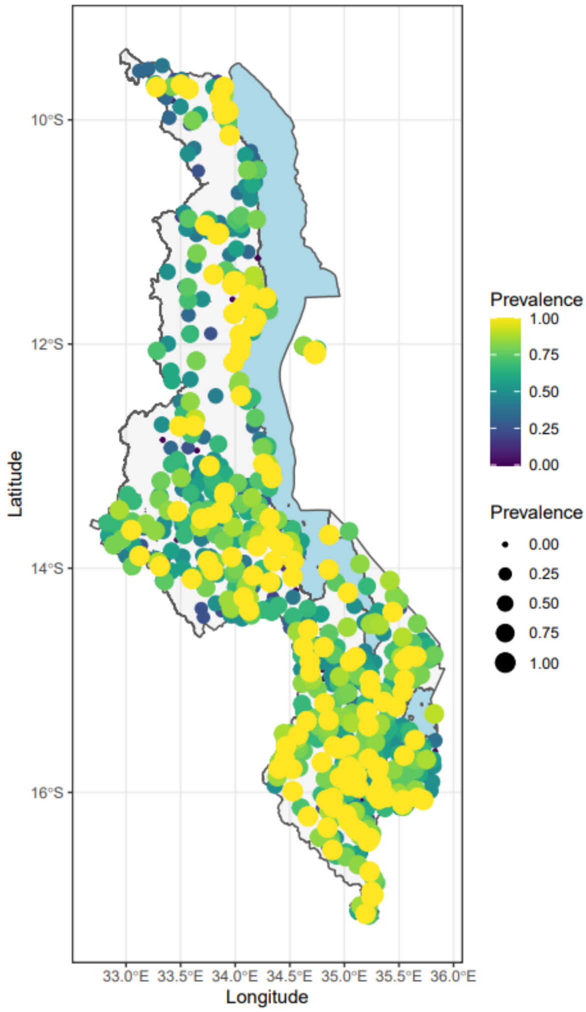


Figure 5: Emprical prevalence of child anaemia in sampled locations

Prevalence quantiles indicate that, at lower quantile (0.25), almost half of the clusters were affected by anaemia. At a 0.75 quantile almost entire country had high anaemia prevelance (Figure 6).

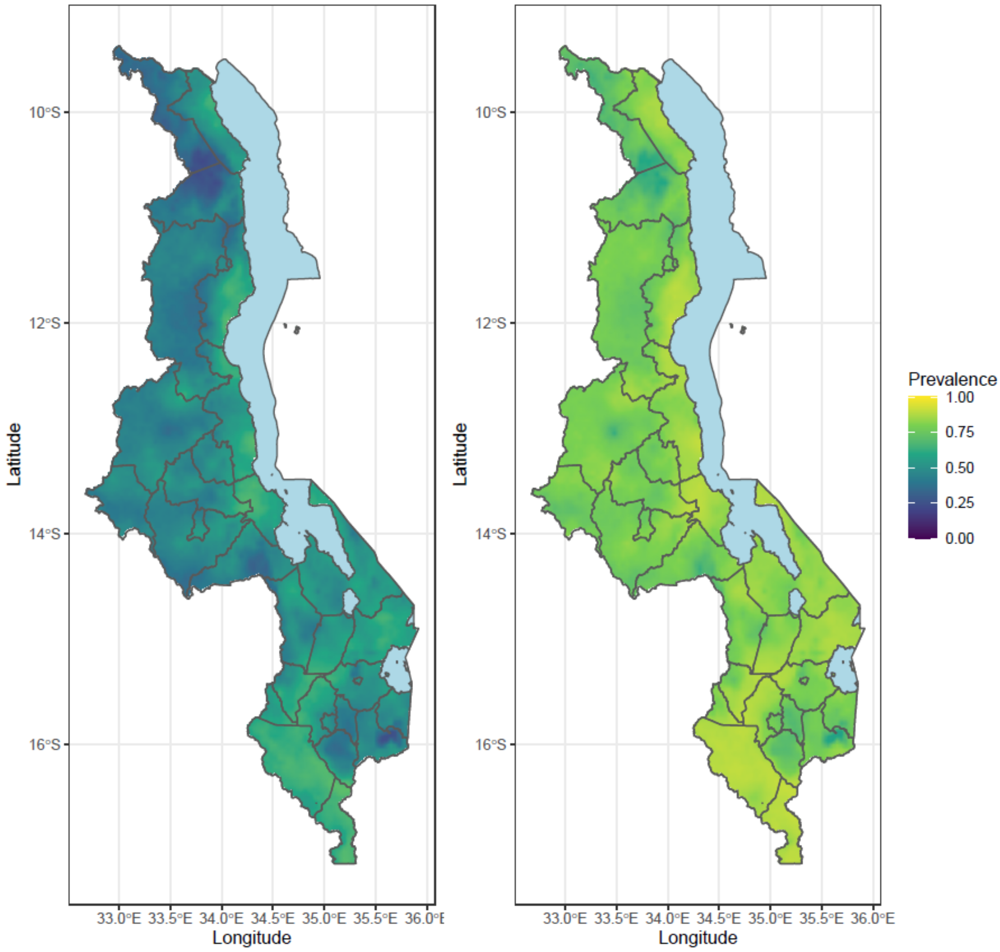


Figure 6: prevalence quantiles of child anaemia

In Figure 7, where standard errors for child anaemia prevalence are presented, the result showed relatively smaller values in areas closest to the sampled locations as presented in Figure 2. The small margin of standard error indicates that data was adequately observed, owing to the relatively large number of data points available for model estimation.

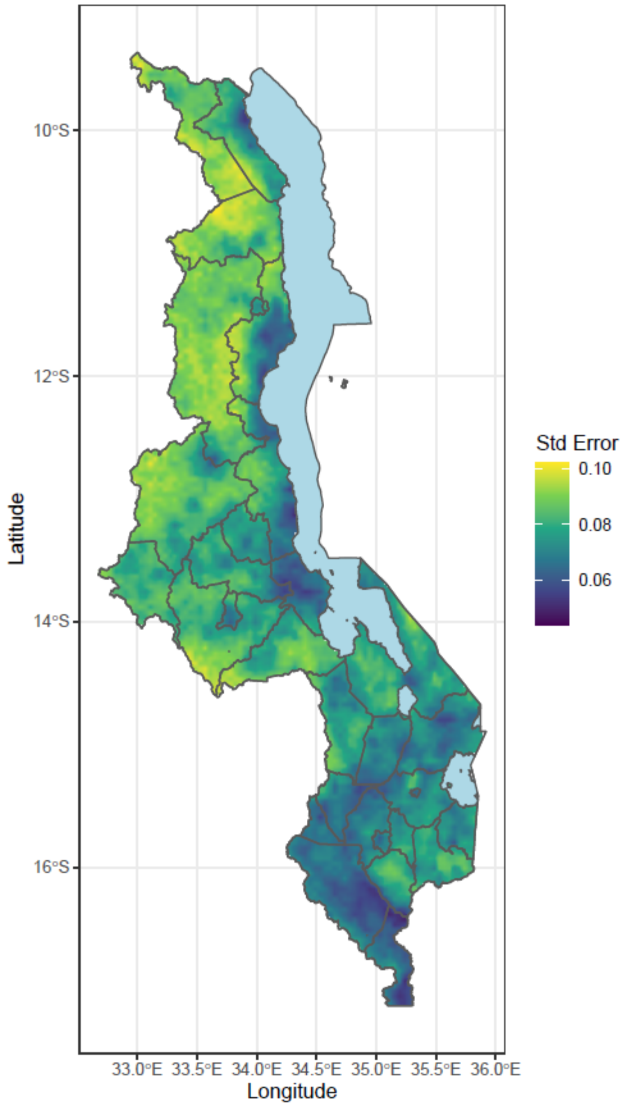
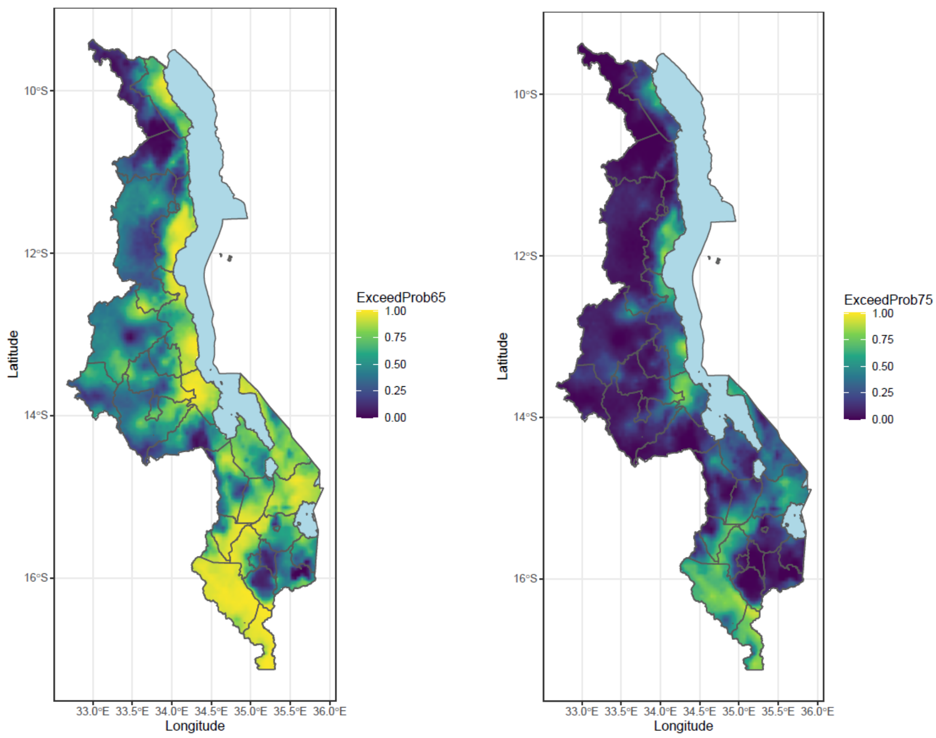


Figure 7: Standard errors for child anaemia prevalence

4.8 Exceedance probability maps

Model-based geostatistical methods allowed the mapping of prevalence at a fine-scale resolution of 5×5 km. Based on Figure 4, anaemia prevalence in children under 5 years was generally high in Malawi at 62.5% as of 2016, characterised by several hotspots. Nsanje, Chikwawa, Salima, Blantyre, Likoma, Machinga, Neno, Nkhotakota, Mwanza, Balaka and Mangochi were districts with high anaemia prevalence of $\geq 70\%$.

Figure 8 is a map for the anaemia exceedance probabilities showing areas where $p(x) \geq 0.625 \mid data$ with 95% certainty in both cases. The predictive maps for anaemia prevalence at threshold of 0.625 was selected since the child anaemia prevalence in Malawi was at 62.5%. Several areas in Malawi including Nsanje, Chikwawa, Zomba, Mangochi, Nkhota Kota and Karonga districts have locations with predicted prevalence above 0.625 see Figure 4. The yellow areas show locations where prevalence is above 65%, and 75% (Figure 8). The identified areas, marked in yellow, are potential areas the policymakers need to focus on when formulating programmes and interventions to control anaemia in children under five.



(a) 65% EP

(b) 75% EP

Figure 8: Exceedance probabilities at 65% and 75%

Higher anaemia prevalence was registered in Southern part of Malawi. Eight districts namely, Nsanje, Chikwawa, Blanyre, Machinga, Neno, Mwanza, Balaka and Mangochi had prevalence ≥70%. In the central region, Salima and Nkhotakota districts had a prevalence of ≥70%. Only Likoma district in the Northern region had prevalence of ≥70%. See Table 7.

Table 7: Child anaemia prevalence ≥ 62.5% and above

Northern	Prevalence	Central	Prevalence	Southern	Prevalence
Likoma	73.5	Salima	74.9	Nsanje	78.5
Nkhatabay	68.8	Nkhotakota	71.3	Chikwawa	77.6
Karonga	67.8	Ntcheu	66.6	Blantyre	74.9
Mzuzu city	63.4	Dowa	66.1	Machinga	73.1
		Ntchisi	65.8	Neno	72.4
		Lilongwe city	65	Mwanza	71.3
		Lilongwe	64.1	Balaka	70.7
		Kasungu	64.0	Mangochi	70.0
		Mchinji	62.7	Zomba	68.4
				Zomba city	65.4
				Phalombe	64.6
				Blantyre city	64.2

4.9 Conclusion

Majority of children under-five years old in Malawi were anaemic, a prevalence of 62.5%. Stepwise model selection using GLM found the following explanatory variables; child age, elevation, wealth index, number of children under-five years, child stunting and child fever to be statistically significant in explaining child anaemia. There was no element of multicollinearity among the selected covariates. None of the VIF values exceeded 5. Child anaemia data for the selected variables was compatible with the GLGM model since the empirical variogram fell within the 95% tolerant band.

CHAPTER 5

DISCUSSION

5.1 Introduction

The study aimed to asses factors of anaemia in children under-five years old as well as identifying hotspot areas for child anaemia. The factors and hotspot areas for child anaemia were assessed and identified respectively using the GLGM model. The chapter represents study findings on child anaemia determinants and anaemia *hotspot* areas.

5.2 Factors associated with anaemia in children under-five years

The study established child age, child stunting, child fever, number of children-under five years and household wealth index as significant determinants of child anaemia. Child age was found to inversely associate with anaemia in children under-five years. This could be because iron demand by younger children to facilitate physical growth during the first year of life is high. Children who are getting older are able to supplement diet richer in iron. In a similar study, Gayawan, Arogundade, and Adebayo

(2014) reported that from a higher likelihood of being anaemic around the age of 6 months, the likelihood reduces as the child advances in age. The findings reveal that, younger children were observed to be at higher risk of childhood anaemia compared to older children. Similar outcomes have been observed in studies conducted and reported in Malawi and Togo by Roberts and Zewotir (2020) and Nambiema, Robert, and Yaya (2019), in which older children were negatively associated with childhood anaemia . Sufficient intake of iron prevents the occurrence of anaemia in older children. According to World Health Organization (2017) report, infants and toddlers from 1 to 3 years are encouraged to take foods rich in iron namely, cereal, red meats and vegetables This is to ensure the body has enough hemoglobin levels for physical and psychomotor functioning. In summary, cognitive development in children during early years of life depends on child age. This could be a reason enough why child age is negatively associated with child anaemia.

The relationship between number of children under-five years in the household and anaemia in children is well established. The higher the number of children under-five years the more vulnerable they are to child anaemia. Due to increases in demand for the amount of household resources available per child, household feeding patterns are likely affected. In so doing, foods consumed end up not being sufficient to supply the required nutrients such as Iron to the body for each under five child. Elmardi et al. (2020) in a similar findings argued that a family with a large number of household members could compete in available resources including foodstuff. This renders children in such households to have less haemoglobin levels the body requires for growth and development. In such setting under five children are more likely to have child anaemia.

Child fever was found to be positively associated with anaemia which also agrees with the study by Ngwira and Kazembe (2016), where fever had a positive effect. According to Ngwira and Kazembe (2016), fever is a common symptom of acute and chronic inflammatory diseases, mostly infections, which have been associated with lower Hb levels. This is probably due to the fact that inflammatory diseases are a source of low production of iron and is a cause of distortion of blood cells. Additionally, fever is an agent of loss of food appetite, in the process, it hinders children from consuming foodstuff which could be source of iron and other nutritious food which are supplements to the production of heamoglobin like folate, vitamins B_{12}, A and C as is also discussed in the literature (section 1.1.2) . Such children are, therefore, at high risk to anaemia.

Child anaemia was found to inversely associate with the wealth index of the household. A possible explanation for the high prevalence of anaemia in children from the poor households could be that such families have low income and are less likely to purchase nutrient-rich foods (like iron, vitamins, etc). They are even food insecure and they cannot afford better health service during illness for their children. Amugsi (2020) in his study also reported that children from well-to-do households, the middle and richest households had a normal Hb concentration in relation to children from poor households. In this view, children from well-to-do households have easy access to nutritious food and better caring practices essential for optimal child health. The higher prevalence of anaemia among children in low-income families is, therefore, inevitable (Gebreweld et al., 2019).

Land elevation was found to be inversely associated with child anaemia. A possible explanation is that, hemoglobin increases as altitude of residence increases. In one study by Alfonso and Leon-abarca (2020), it was revealed that children from low altitudes had higher aneamia prevalence (8.5%) than those from high altitudes (1.2%, $p<0.001$). Elevation has therefore possible negative influence on anaemia. Children from households at higher altitudes are associated with less anaemia prevalence than children from households at lower altitudes. In a similar study, Ngwira and Kazembe (2016), highlighted that highland areas are associated with lower temperatures and hence less risk to malaria which is also a cause of anaemia. According to Nambiema et al. report, children with malaria were at high risk of anaemia. Increased hemolysis or decreased red blood cell production rates, could explain the higher risk of anemia among malaria patients compared to non-malaria patients (Nambiema et al., 2019). Thus this study findings agrees with literature of inverse association between land elevation and child anaemia.

5.3 Child Anaemia Hotspot Areas

Exceedance probability (EP) maps were used to identify anaemia hotspot areas for accurate and timely interventions and effectiveness in monitoring and evaluation of anaemia control programmes.

Malawi anaemia prevalence was at 62.5%. Thus exceedance probabilities showed areas such that

$$p(x) = 0.625 \mid data \text{ with 95\% certainty}$$

At 65% EP, most districts in Malawi had child anaemia prevalence. These include Nsanje, Chikwawa, Salima, Blantyre, Likoma, Machinga, Neno, Nkhotakota, etc . See (Table 7 and Figure 8a). At 75% EP, districts like Nsanje, Chikwawa and Salima had high prevalence of anaemia See (Table 7 and Figure 8b). Several factors lead to high child anaemia prevalence in these districts. The observed spatial heterogeneity were due to unobserved factors not captured by the covariates in the GLGM model. The geographical variation in anaemia-causing infections, such as malaria, hookworms, and helminths, could be among influences of such spatial variations. All anaemia hotspot districts namely, Nsanje, Chikwawa and Salima are at lower attitudes and are associated with low Hb levels, high temperatures and are near water bodies. According to the study findings (Table 6) and the literature(Khulu, 2019; Ocas-co, Tapia, & Gonzales, 2018), altitude is inversely associated with anaemia; it increases with decrease in altitude. In addition water bodies and high temperatures ($>21°$C) are favourable conditions for mosquitoes to breed, resulting in increased malarial transmission and let alone malaria anaemia which is a cause of anaemia.

On the contrary, areas at higher altitudes are associated with higher Hb levels, low temperatures, and are at highlands (Ocas-co et al., 2018). No wonder districts like Chitipa, Rumphi, Chiradzulu, Mulanje, Mzimba, Thyolo, Dedza and Mchinji had anaemia prevalence of $\leq 63\%$. See Table 9. This difference is due to the lower oxygen concentration at higher altitude than at lower altitude so that an individual at high altitude requires relatively a large number of Hb cells to carry enough oxygen needed by the body.

Geographical nutritional variation also explains the spatial heterogeneity of childhood anaemia in Malawi (Ngwira & Kazembe, 2016). According to Ngwira and Kazembe, the cause of regional nutritional differences can be natural disasters such as floods and difference in climatic conditions. Flooding of shire river which annually destroys crops in Nsanje and Chikwawa affects nutrition of these areas. Even population densities in these areas are This explains why Nsanje and Chikwawa are the most top hotspot districts for child anaemia in Malawi.

Additionally, anaemia hospot districts are reported to have high population densities (National Statistical Office, 2019). This entails high households size and inadequate social services to support the rapid growing populations. Henceforth, there is high likelihood of low levels of education which renders most homes to be in the lower wealth index category. This even make it hard for parents and guardians to give the desired health care support like provision of nutritious food, balanced diet foods and even in a time of need.

Accelerating progress in anaemia reduction for children under-five years in Malawi is possible. In situations, where resources are not enough, the policy direction is to target interventions and control programmes in anaemia highly affected areas like Nsanje, Chikwawa and Salima. See Table 7.

5.4 Conclusion

Child age, child stunting, child fever, number of children-under years and household wealth index were revealed to be significant determinants of child anaemia. At 75% EP, Nsanje, Chikwawa and Salima were reported to have high child anaemia prevalence. A range of factors attributed to high anaemia prevalence in these areas such as being along water bodies and high temperatures, lower altitudes and adverse climate change variations. To optimise utilisation of resources in anaemia reduction, prevalence maps are crucial to prioritise hotspot areas (Nsanje, Chikwawa and Salima districts).

CHAPTER 6

CONCLUSION

6.1 Introduction

This chapter represents the study overview, in terms of research topic, research problem, research objectives as well as theoretical framework underlying Geostatistics models and analysis. It also represents conclusions, recommendations and limitations of the study.

6.2 Conclusions

The study was set to map anaemia prevalence in children under-five years old in Malawi using the GLGM model. Ending child anaemia entails achieving SDG 3.1 and SDG 3.2 which strives at ending all forms of malnutrition and preventable deaths of children under-five years by 2030 and reducing under-five mortality to as low as 2.5%. Malawi as a low-resourced, identification of hotspot areas is critical to optimise utilization of available resources to fight child anaemia. Existing studies on child anaemia did not consider prevalence spatial effects which are fundamental key in identifying areas with above average anaemia prevalence. GLGM model was therefore a best fit in such a low-resource settings where comprehensive disease registries do not exist. It has the ability to estimate and predict prevalence both at observed and unsampled locations.

In 2016, MDHS survey indicated child anaemia prevalence of 63% which was high and a severe public health problem according to WHO. High prevalence of anaemia and its consequences is so challenging to children's health, growth and development. This study aimed at assessing factors of child anaemia and identifying areas that need targeted interventions in order to control anaemia in children under-five years old.

Geostatistical model formulation considered visited communities at sampled EAs $x_i : i = 1,\ldots,n$ distributed over the boundary of Malawi, in each community, sampling m_i individuals and recording whether each tests positive or negative for the anaemia. If $p(x)$ denotes prevalence at location x, the standard sampling model for the resulting data was a binomial

$$Y_i \sim \mathrm{Bin}(m_i, p_i) \tag{6.1}$$

Linkage of the $p(x_i)$ at different locations is usually desirable, and is essential when making inferences about $p(x)$ at unsampled locations x. Adopting the logistic link function, the model assumed that:

$$\log\left\{\frac{p(x)}{1-p(x)}\right\} = \alpha + d(x)'\beta + S(x)$$

To identify hotspot areas, exceedance probabilities were used. Areas whose anaemia prevalence was above a set threshold of 0.625, was considered to have high anaemia prevalence. The simplest case of an exceedance probability is

$$EP = Pr(x > c) \tag{6.2}$$

where $Pr(x)$ is a probability which estimates how frequently the relative risk exceeds the null risk and can be regarded as an indicator of how unusual the risk is in that area. c is the set threshold and was at 0.625 since the child anaemia prevalence was at 62.5%.

Bivariate analysis was employed to examine association between child anemia status and other explanatory variables. Statistical significance level of 5% was assumed. Stepwise regression model selection using GLM was used to identify covariates to fit GLGM. MCML spatial approach was used to predict anaemia outcomes.

The study revealed anaemia prevalence of 62.5% in children under-five years. Thus a severe public health problem based on WHO criteria. Child age, child fever, child stunting, number of children under-five years in a household and household wealth index were strongly associated with child anaemia. Elevation was found to be inversely associated with child anaemia.

Spatial variations were marked and anaemia hotspot areas were identified. At 65% EP, majority of districts in Malawi had anaemia prevalence of $\leq$65%. These included, Lilongwe, Zomba, Ntchisi, Dowa, Ntcheu, Karonga etc, see Table 7. Thus, child anaemia was severe in most parts of Malawi. At 75% EP, anaemia hotspot districts were identified to be Nsanje, Chikwawa and Salima.

6.3 Recommendations

In order to reduce the burden of anaemia in children under-five years in Malawi the study recommends the following:

- Set nutritional targets at the country level, including the desired average annual reduction rate and country-level baselines

- All established interventions should aim at maintaining sufficient iron levels among children by having in place programmes like sanitation and hygiene, disease control, and reproductive health to ensure effective, safe, and wide delivery to those at high risk.
- Ensure development policies and programmes should include nutrition, by integrating nutrition outcomes across multiple sectors, such as health, food systems, water, sanitation and hygiene, and delivery platforms for improved nutrition across the population.
- Timely identification of anaemia hotspot through mapping to optimise usage of resources to fight against the condition.

6.4 Limitation of study

The following were the study's limitations:

- No information on iron levels in the children was established. Iron deficiency plays a major role in childhood anaemia.
- Heterogeneity of child anaemia is both in space and time. To include time component, more data spanning several years was needed.The study did not therefore consider time factor since only used one cross sectional data set for 2015/16 MDHS.
- The study did not model the maternal anaemia, which is usually the biggest predictor of child anaemia. Maternal modelling was beyond the scope of current work but a future work.

REFERENCES

Alfonso, R., & Leon-abarca, J. A. (2020). Age and altitude of residence determine anemia prevalence in Peruvian 6 to 35 months old children. *PLoS ONE*, 1–18.

Amugsi, D. A. (2020). Determinants of normal haemoglobin concentration among children in Ghana: a positive deviance analysis of nationally representative cross-sectional survey data. *Scientific Reports*, *10*(1), 1–9. doi: 10.1038/s41598-020-64072-7

Banerjee, S., Gelfand, A. E., & Carlin, B. P. (2003). *Hierarchical Modeling and Analysis for Spatial Data (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)* (No. 1960).

Brockwell, P. J., & Davis, R. A. (2006). *Springer Series in Statistics Advisors :.*

Calis, J. C., Phiri, K. S., Faragher, E. B., Brabin, B. J., Bates, I., Cuevas, L. E., ... van Hensbroek, M. B. (2016). Severe anemia in Malawian children. *Malawi Medical Journal*, *28*(3), 99–107. doi: 10.1056/NEJMoa072727

Chaparro, C. M., & Suchdev, P. S. (2019). Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries. *Annals of the New York Academy of Sciences*, *1450*(1), 15–31. doi: 10.1111/nyas.14092

de Graft Acquah, H. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, *2*(1), 001–006.

Diggle, P. J., & Giorgi, E. (2016). Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings. *Journal of the American Statistical Association*, *111*(515), 1096–1120. doi: 10.1080/01621459.2015.1123158

Diggle, P. J., & Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health : Methods and Applications*. Taylor and Francis Group.

Ding, J., Tarokh, V., & Yang, Y. (2018). Model Selection Techniques: An Overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34. doi: 10.1109/MSP.2018.2867638

Elmardi, K. A., Adam, I., Malik, E. M., Ibrahim, A. A., Elhassan, A. H., Kafy, H. T., ... Abdin, M. S. (2020). Anaemia prevalence and determinants in under 5 years children : findings of a cross- sectional population-based study in Sudan. *BMC Pediatrics*, 1–14.

Gaston, R. T., Habyarimana, F., & Ramroop, S. (2022). Joint modelling of anaemia and stunting in children less than five years of age in Lesotho : a cross - sectional case study. *BMC Public Health*, 1–11. Retrieved from `https://doi.org/10.1186/s12889-022-12690-3` doi: 10.1186/s12889-022-12690-3

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. *BMC Public Health*(January), 35–45. doi: 10.1093/inthealth/iht034

Gebreweld, A., Ali, N., Ali, R., & Fisha, T. (2019). Prevalence of anemia and its associated factors among children under five years of age attending at Guguftu health center, South Wollo, Northeast Ethiopia. *PLoS ONE*, *14*(7), 1–13. doi: 10.1371/journal.pone.0218961

Ghosh, J. K. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology* (Vol. 77) (No. 2). doi: 10.1111/j.1751-5823.2009.00085_26.x

Giorgi, E., Diggle, P. J., Snow, R. W., & Noor, A. M. (2018). Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys arXiv : 1802 . 06359v1 [stat . ME] 18 Feb 2018. *Package Prev Map*, 1–35.

Hadavand-Siri, M., & Deutsch, C. V. (2012). Some Thoughts on Understanding Correlation Matrices. *Centre for Computational Geostatistics Report*, *14*(4), 408.

Harding, K. L., Aguayo, V. M., Namirembe, G., & Webb, P. (2018). Determinants of anemia among women and children in Nepal and Pakistan: An analysis of recent national survey data. *Maternal and Child Nutrition*, *14*(May), 1–13. doi: 10.1111/mcn.12478

Kawo, K. N., Asfaw, Z. G., & Yohannes, N. (2018). Multilevel Analysis of Determinants of Anemia Prevalence among Children Aged 6 – 59 Months in Ethiopia : Classical and Bayesian Approaches. *Hindawi Anemia*, *2018*.

Khulu, C. (2019). Determinants of anemia in children under five years in Angola , Malawi and Senegal : A comparative study. *Research square*.

Khulu, C., & Ramroop, S. (2020). Key Determinants of Anemia among Youngsters under Five Years in Senegal , Malawi , and Angola. *Environmental Research and Public Health*.

Komiyama, T. (2008). Usability evaluation based on international standards for software quality evaluation. *NEC Technical Journal*, *3*(2), 27–32.

Macdonald, C., Alison, M., Mike, N., Rose, N., & Miriam, Y. (2010). Anemia - can its widespread prevalence among women in developing countries be impacted ? 1 . 0 Anemia in Women : A Global Health Priority. *A Global Health Priority*(October 2002), 1–42.

Ministry of Health. (2022). *National Multi-Sector Nutrition Policy* (No. April 2018). Retrieved from `https://extranet.who.int/nutrition/gina/en/`

node/39435

Molla, A., Egata, G., Mesfin, F., Arega, M., & Getacher, L. (2020). Prevalence of Anemia and Associated Factors among Infants and Young Children Aged 6-23 Months in Debre Berhan Town, North Shewa, Ethiopia. *Journal of Nutrition and Metabolism*, *2020*. doi: 10.1155/2020/2956129

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*, 1–9.

National Statistical Office. (2015). Malawi Demographic and Health Survey 2015-16. *National Statistics Office The DHS Program*, 1–658. Retrieved from http://dhsprogram.com/pubs/pdf/FR319/FR319.pdf doi: http://www.measuredhs.com/

National Statistical Office. (2017). Malawi Micronutrient Survey 2015-16: Key Indicators Report. *National Statistical Office and Centres for Disease Control and Prevention*(February), 1–36. Retrieved from http://www.nsomalawi.mw/images/stories/data_on_line/demography/mns2015_16/MalawiMicronutrientSurveyKeyIndicatorsReport.pdf

National Statistical Office. (2019). *2018 Malawi Population and Housing Main Report* (Tech. Rep. No. May). Retrieved from http://www.nsomalawi.mw/images/stories/data_on_line/demography/census_2018/2018MalawiPopulationandHousingCensusMainReport.pdf

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*, 370–384.

Ngwira, A., & Kazembe, L. N. (2015). Bayesian random effects modelling with application to childhood anaemia in Malawi. *BMC Public Health*, 1–11. doi: 10.1186/s12889-015-1494-y

Ngwira, A., & Kazembe, L. N. (2016). Analysis of severity of childhood anemia in Malawi : a Bayesian ordered categories model. *Open Access Medical Statistics*, 9–20.

Noor, A. M. (2011). The First Model-Based Geostatistical Map of Anaemia. *PloS Medicine*, *8*(6), 1–3. doi: 10.1038/ncomms110.PLoS

Ocas-co, S., Tapia, V., & Gonzales, G. F. (2018). Hemoglobin Concentration in Children at Different Altitudes in Peru : to Diagnose Anemia and Polycythemia. *HIGH ALTITUDE MEDICINE & BIOLOGY*, *19*(4), 398–403. doi: 10.1089/ham.2018.0032

Ogunsakin, R. E., Babalola, B. T., & Akinyemi, O. (2020). Statistical Modeling of Determinants of Anemia Prevalence among Children Aged 6-59 Months in Nigeria: A Cross-Sectional Study. *Anemia*, *2020*. doi: 10.1155/2020/4891965

Parbey, P. A., Tarkang, E., Manu, E., Amu, H., Ayanore, M. A., Aku, F. Y., … Kweku, M. (2019). Risk Factors of Anaemia among Children under Five Years in the Hohoe Municipality, Ghana: A Case Control Study. *Anemia*, *2019*. doi: 10.1155/2019/2139717

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rahman, M. S., Mushfiquee, M., Masud, M. S., & Howlader, T. (2019). Association between malnutrition and anemia in under-five children and women of reproductive age: Evidence from Bangladesh demographic and Health Survey 2011. *PLoS*

ONE, *14*(7), 1–18. doi: 10.1371/journal.pone.0219170

Rakanita, Y., Sinuraya, R. K., Suradji, E. W., Suwantika, A. A., & Syamsunarno, M. R. A. A. (2020). The Challenges in Eradication of Iron Deficiency Anemia in Developing Countries. *Eradication of Iron Deficiency Anemia*, *11*(5), 383–401. doi: 10.31838/srp.2020.5.55

Roberts, D. J., Matthews, G., Snow, R. W., Zewotir, T., & Sartorius, B. (2020). Investigating the spatial variation and risk factors of childhood anaemia in four sub-Saharan African countries. *BMC Public Health*, 1–10.

Roberts, D. J., & Zewotir, T. (2020). Copula geoadditive modelling of anaemia and malaria in young children in. *Journal ofHealth, Population and Nutrition*, *8*, 1–14.

Semedo, R. M., Santos, M. M., Baião, M. R., Luiz, R. R., & Da Veiga, G. V. (2014). Prevalence of Anaemia and Associated Factors among Children below Five Years of Age in Cape Verde, West Africa. *Journal of Health, Population and Nutrition*, *32*(4), 646–657. doi: 10.3329/jhpn.v32i4.3021

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42. doi: 10.12691/ajams-8-2-1

Simo, L. P., Agbor, V. N., Agborndip, E., Ekaney, D. S. M., Mbeng, E. N., Etombi, C. L., . . . Mbanya, D. (2020). Prevalence and determinants of anaemia in children aged 6-59 months in Africa: A protocol for systematic review and meta-analysis. *BMJ Open*, *10*(4), 1–4. doi: 10.1136/bmjopen-2019-032042

Sorsa, A., Habtamu, A., & Kaso, M. (2021). Prevalence and Predictors of Anemia Among Children Aged 6–23 Months in Dodota District, Southeast Ethiopia: A

Community-Based Cross-Sectional Study. *Pediatric Health, Medicine and Therapeutics*, *Volume 12*, 177–187. doi: 10.2147/phmt.s293261

StataCorp. (2015). Stata 14. *Stata Statistical Software*.

Tony, R., Ramroop, S., & Habyarimana, F. (2021). Heliyon Joint modelling of malaria and anaemia in children less than fi ve years of age in Malawi. *Heliyon*, *7*(March), e06899. Retrieved from https://doi.org/10.1016/j.heliyon.2021.e06899 doi: 10.1016/j.heliyon.2021.e06899

Wang, K., & Chen, Z. (2018). Stepwise Regression and All Possible Subsets Regression in Education. *Electronic International Journal of Education*(March).

WHO. (2001). Iron Deficiency Anemia, Assessment, Prevention, and Control: A Guide for Program Managers. *Iron Deficiency Anemia*.

WHO. (2012). Anaemia Policy Brief. *Global Nutrition Targets 2025*(6), 1–7. Retrieved from http://www.who.int//iris/bitstream/10665/148556/1/WHO_NMH_NHD_14.4_eng.pdf

World Health Organization. (2017). *Nutritional Anaemias : Tools for Effective Prevention*. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/259425/9789241513067-eng.pdf?sequence=1

Xue, J., Luo, Y., & Liang, F. (2017). Average (E)BIC-like Criteria for Bayesian Model Selection. *Working Paper*(Mcmc). Retrieved from https://people.clas.ufl.edu/yeluo/files/ave4.pdf

Yankson, R., Anto, E. A., & Chipeta, M. G. (2019). Geostatistical analysis and mapping of malaria risk in children under 5 using point-referenced prevalence data in Ghana. *Malaria Journal*, *18*(1), 1–12. Retrieved from https://doi.org/10.1186/s12936-019-2709-y doi: 10.1186/s12936-019-2709-y

APPENDIX

Appendix 1: Socio-demographic determinants of child anaemia

Table 8: Socio-demographic determinants of child anaemia

Covariates	Total	No	Yes	P.Value
	n = 4601	n = 1724	n = 2877	
No of HH Members				0.595
	5.55 (2.09)	5.58 (2.13)	5.54 (2.06)	
Region				0.238
North	819 (100%)	326 (39.8%)	493 (60.2%)	
Central	1599 (100%)	602 (37.6%)	997 (62.4%)	
South	2183 (100%)	796 (36.5%)	1387 (63.5%)	
Residence				$< .001$
Urban	736 (100%)	319 (43.3%)	417 (56.7%)	
Rural	3865 (100%)	1405 (36.4%)	2460 (63.6%)	
Water source				0.299
Unimproved	3293 (100%)	1218 (37%)	2075 (63%)	
Improved	1308 (100%)	506 (38.7%)	802 (61.3%)	
Toilet				0.024
Unimproved	367 (100%)	117 (31.9%)	250 (68.1%)	
Improved	4234 (100%)	1607 (38%)	2627 (62%)	
HH sex				0.502
Male	3370 (100%)	1273 (37.8%)	2097 (62.2%)	
Female	1231 (100%)	451 (36.6%)	780 (63.4%)	
HH age				0.058
	36.11 (11.99)	36.55 (11.78)	35.85 (12.11)	
Toilet share				0.043
No	2956 (100%)	1140 (38.6%)	1816 (61.4%)	
Yes	1645 (100%)	584 (35.5%)	1061 (64.5%)	
Wealth index				$< .001$
Poorest	994 (100%)	305 (30.7%)	689 (69.3%)	
Poorer	1037 (100%)	358 (34.5%)	679 (65.5%)	
Middle	900 (100%)	371 (41.2%)	529 (58.8%)	

Richer	851 (100%)	330 (38.8%)	521 (61.2%)	
Richest	819 (100%)	360 (44%)	459 (56%)	
U5 children				$<$ .001
	1.64 (0.73)	1.59 (0.74)	1.67 (0.72)	
breast feeding				0.933
No	77 (100%)	28 (36.4%)	49 (63.6%)	
Yes	4524 (100%)	1696 (37.5%)	2828 (62.5%)	
child stunt				$<$ .001
No	2917 (100%)	1158 (39.7%)	1759 (60.3%)	
Yes	1684 (100%)	566 (33.6%)	1118 (66.4%)	
child wasting				0.693
No	4463 (100%)	1675 (37.5%)	2788 (62.5%)	
Yes	138 (100%)	49 (35.5%)	89 (64.5%)	
child underweight				0.243
No	4024 (100%)	1521 (37.8%)	2503 (62.2%)	
Yes	577 (100%)	203 (35.2%)	374 (64.8%)	
Child sex				0.215
Male	2244 (100%)	820 (36.5%)	1424 (63.5%)	
Female	2357 (100%)	904 (38.4%)	1453 (61.6%)	
Child age (month)				$<$ .001
	36.71 (15.04)	39.48 (13.60)	35.05 (15.61)	
Maternal education				$<$ .001
None	532 (100%)	163 (30.6%)	369 (69.4%)	
Primary	3041 (100%)	1123 (36.9%)	1918 (63.1%)	
Secondary	984 (100%)	414 (42.1%)	570 (57.9%)	
Higher	44 (100%)	24 (54.5%)	20 (45.5%)	
Child fever				$<$ .001
No	3298 (100%)	1302 (39.5%)	1996 (60.5%)	
Yes	1303 (100%)	422 (32.4%)	881 (67.6%)	
Mother's age				$<$ .001
	28.06 (6.75)	28.59 (6.70)	27.74 (6.77)	
Birth weight				0.048
	3901.59 (2059.48)	3824.16 (1985.96)	3947.98 (2101.28)	
Birth order				0.071
	2.97 (2.04)	3.04 (2.06)	2.92 (2.03)	
Child diarrhea				$<$ .001
No	3804 (100%)	1470 (38.6%)	2334 (61.4%)	
Yes	797 (100%)	254 (31.9%)	543 (68.1%)	
Child cough				0.931

No	3522 (100%)	1318 (37.4%)	2204 (62.6%)
Yes	1079 (100%)	406 (37.6%)	673 (62.4%)

Appendix 2: Child anaemia prevalence across districts

Table 9: Child anaemia prevalence across districts

District	Prevalence
Nsanje	78.5
Chikwawa	77.6
Salima	74.9
Blantyre	74.0
Likoma	73.5
Machinga	73.1
Neno	72.4
Nkhotakota	71.3
Mwanza	71.3
Balaka	70.7
Mangochi	70.0
Nkhatabay	68.8
Zomba	68.4
Karonga	67.8
Ntcheu	66.6
Dowa	66.1
Ntchisi	65.8
Zomba City	65.4
Lilongwe City	65.0
Phalombe	64.6
Blantyre City	64.2
Lilongwe	64.1
Kasungu	64.0
Mzuzu City	63.4
Mchinji	62.7
Dedza	62.4
Thyolo	62.2
Mzimba	62.1
Mulanje	60.7
Chiradzulu	58.2
Rumphi	56.3
Chitipa	54.3

Appendix 3: Anaemia prevalence distribution by district

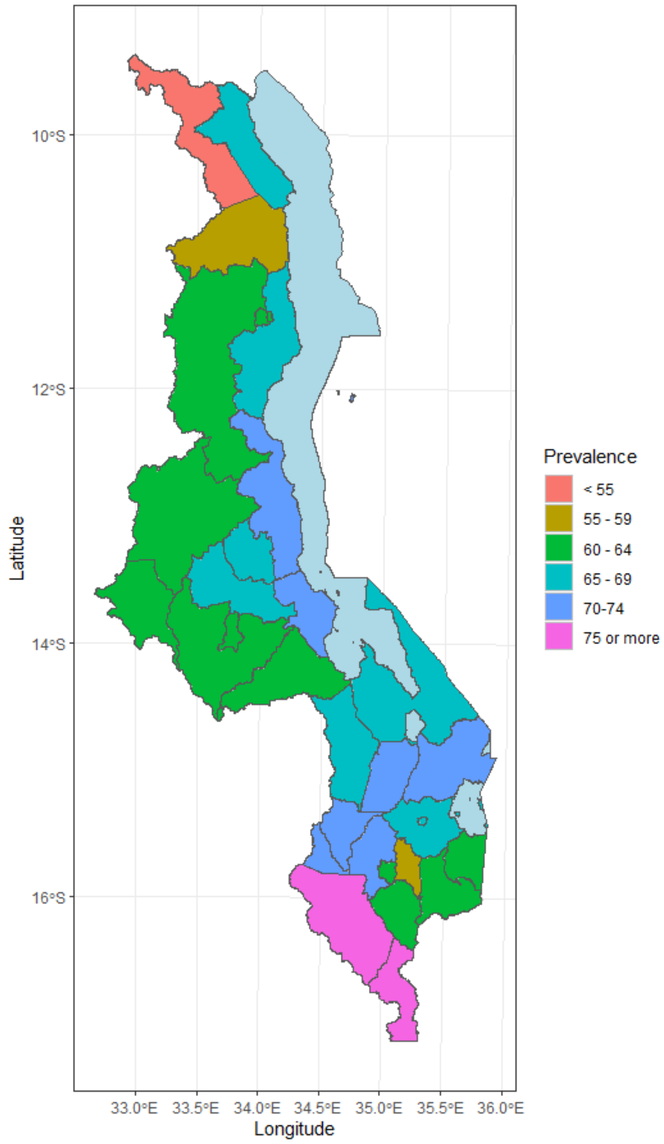


Figure 9: Anaemia prevalence predictions among children aged under 5 years in Malawi