COMBINING MULTIPLE DATA SOURCES TO MODEL UNDER-FIVE MORTALITY AT THE SUB-DISTRICT LEVEL IN MALAWI

MSc(BIOSTATISTICS) THESIS

RUTH VELLEMU

UNIVERSITY OF MALAWI

June, 2022

COMBINING MULTIPLE DATA SOURCES TO MODEL UNDER-FIVE MORTALITY AT THE SUB-DISTRICT LEVEL IN MALAWI

MSc(Biostatistics) Thesis

By

RUTH VELLEMU

Bsc(Statistics) - University of Malawi

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

June, 2022

DECLARATION

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

_____

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

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

Signature:  Date: 20th June, 2022

James Chirombo, PhD

Supervisor

Signature: _____ Date: _____

Patrick Sawerengera, MSc (Lecturer)

Programme Coordinator

DEDICATION

To my family, for all the love and support

Kezra and Amelia, I love you!!

ACKNOWLEDGEMENT

This work was supported by the DELTAS Africa Initiative, Sub-Saharan Africa Consortium for Advanced Biostatistics Training (SSACAB). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust and the UK Government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD, Wellcome Trust or the UK government.

I wholeheartedly respect and thank my supervisor James Chirombo, PhD, for his untiring support, constructive ideas and guidance throughout this project.

I thank God for guiding me through this work and special thanks to my loving husband Dieckens for his endless support and love.

To Olive and all my friends, thanks for the encouragement and all timely support that you rendered to me throughout this project work.

ABSTRACT

As countries continue making gains towards the attainment of sustainable development goals, surveillance of health outcomes at the sub-district level will become important as district level indicators may mask areas where progress is slow. To achieve this high level of surveillance, it may be necessary to pool data from multiple data sources with different spatial resolutions. The aim of this study was to estimate and model under-five mortality risk at the sub-district level in Malawi by combining multiple data sources. We used Bayesian hierarchical models to combine the Demographic and Health Survey (DHS) data with Census data in a principled framework. A binomial generalized linear geostatistical model was fitted to estimate the risk of under-five mortality in the presence of the various covariates. Results showed that mother's age and weight of child at birth were associated with under-five mortality.However, the posterior odds showed no significant differences in dying for children from mothers across different ages. In addition, the results showed that the risk of under-five mortality is higher in the northern region and along lakeshows as well as districts in the lower Shire. The study provided a means for performing small area estimation of population parameters of interest. In addition, using survey findings along with risk maps is essential for disease monitoring and surveillance purposes as well as for strengthening survey findings. More importantly, the project has improved our understanding of methods used in combining information from different sources.

TABLE OF CONTENTS

Figures

List of Tables

Abbreviations and acronyms

DHS	Demographic and Health Survey
GLGM	Generalized Linear Geostatistical Model
INLA	Integrated Nested Laplace Approximations
ITN	Insecticide Treated Nets
MCMC	Markov Chain Monte Carlo
MDHS	Malawi Demographic and Health Survey
MICS	Multiple Indicator Cluster Survey
NGO	Non-Governmental organisation
NHIS	National Health Interview Survey
NMS	National Maternity Surveys
PHC	Population and Housing Census
SAE	Small Area Estimation
SEA	Standard Enumeration Area

CHAPTER ONE
INTRODUCTION

This chapter presents a brief background of the study, the knowledge gap that was identified, the study objectives and the significance of the study.

1.1 Background Information

Collecting data that gives accurate and timely estimates of population quantities of interest is a challenge in most situations. Lohr and Raghunathan (2017) point out that probability sampling provides a means of collecting information efficiently as well as methods for assessing the suitability of the estimates obtained and has long been a foundation for producing national statistics for many countries (National Academies of Sciences, Engineering and Medicine, 2017). By definition, probability sampling means that "every item in the population has a positive chance of being included in the sample" (Taherdoost, 2018, p. 20). This sampling process uses some form of random selection and each unit is drawn with a known probability or has a nonzero chance of being in the sample. Probability sampling is more useful and precise when generalizing the findings from the sample to the whole population.

Many sampling techniques exist which fall under probability sampling including simple random sampling, systematic sampling, cluster sampling and stratified sampling. Different probability sampling methods are used depending on the nature of studies as well as how convenient and suitable the technique is. For example, a study conducted in Kampala District of Uganda used probability systematic sampling from the police register to determine the burden of alcohol use among the Uganda Police (Ovuga & Madrama, 2006). Surveys such as

Demographic and Health Survey (DHS), Multiple indicator Cluster surveys (MICs) use a combination of cluster and stratified sampling to select samples.

The probability sampling methods, however, face some challenges such as decreased response rates and in some occasions no response at all. In the United States of America, for instance, as reported in the National Center for Health Statistics, (2016), the US National Health Interview Survey (NHIS) which is a high-quality face-to-face survey, the response rate had declined from 92% in 1997 to 70% in 2015 and there were also issues of nonresponse among individuals within sampled households. Five National Maternity Surveys (NMS) conducted in England at varying intervals between 1995 and 2018 also showed a decline in response rate from 67% in 1995 to 29% in 2018 (Harrison et al., 2020).

A study to evaluate the relative importance of the factors associated with the decline of fertility in sub-Saharan Africa is another example showing declining response rates (Westoff et al., 2013). From 24 sub-Saharan African countries that were included in the study, the response rate declined to an average of about 7% (ranging from 0.2% to 20%) in the year 2013 from the average of 9% (ranging from 0.4% to 25%) response rate recorded between 2009 to 2011. The 2004 MDHS also showed a decrease in response rate compared with the 2000 MDHS. Specifically the response rates declined from 98% to 96% for women and from 97% to 95% for men (National Statistical Office (NSO) [Malawi] and ICF, 2017).

As a result of nonresponse, researchers have been increasingly inclined to implement data collection strategies to combat this trend, including longer field periods, increased numbers of call attempts, sending advance letters, offering incentives and attempting refusal conversions (Holbrook et al., 2007). Issues of nonresponse in surveys generally produce biased estimates of various population parameters.

Another challenge faced by probability sampling methods is the issue of misreporting. Issues of misreporting are usually experienced in surveys covering sensitive topics as a chosen respondent who agrees to participate in the survey fails to answer sensitive questions honestly and thereby creating measurement error (McNeeley, 2012). Survey respondents tend to misreport for various reasons such as avoiding embarrassment or stigmatization, avoiding potential repercussions, and trying to present themselves to the researcher in a positive manner (Pridemore et al., 2005). For example, in the Malawian setting, surveys about sexual activity are subject to misreporting due to social undesirability of such behaviour among different cultures within the country (Poulin, 2010).

Data collection methods based on probability sampling are also expensive and time-consuming (Lohr & Raghunathan, 2017). This, for instance, is reflected when socially disadvantaged groups like the homeless, chronically mentally ill and prostitutes are to be sampled (Bonevski et al., 2014). Under normal circumstances, such groups of people are hard to be interviewed and require more time in strategizing how to approach and engage them in health and medical research surveys. In some circumstances, incentives may be required to get such groups to participate in the survey which might mean additional costs.

Another challenge faced by surveys is to provide useful estimates for small sub-populations at higher spatial resolution widely known as small area estimation. By definition, small area estimation (SAE) is "any statistical technique involving the estimation of parameters for small sub-populations" (Rao & Molina, 2015, p. 3). This is generally used when the sub-population of interest is included in the larger survey and the term "small area" usually refers to a small geographical area. The demand for small area statistics has greatly increased worldwide due to their growing use in formulating policies and programs, in the allocation of government funds and in regional planning (Rao & Molina, 2015). Furthermore,

legislative acts by national governments have increasingly created a need for small area statistics. In the Malawian setting for instance, SAE is important so that limited resources for various interventions can be delivered to target populations at the right time.

SAE techniques were used to produce reliable, stable, representative and high precision small area estimates of poverty incidence at the district level in the State of Bihar in India (Islam et al., 2018). This was done by linking data from the existing Household Consumer Expenditure Survey data and the population census. The results would be necessary for effective planning, implementation and monitoring of various government schemes in Bihar such as focused and target-oriented intervention programs. In South Africa, SAE techniques were used to prove reliable district-level HIV prevalence estimates from national HIV prevalence survey (Gutreuter et al., 2019). A small area analysis has also been done pooling together national DHS surveys to provide estimates of under-five mortality rates (Zehang et al., 2019).

Even though small area estimation is challenging, it is an area of interest for researchers since it provides reliable estimates of population health indicators essential for monitoring trends and inequalities over time (Alexander & Alkema, 2018). Alexander and Alkema (2018) also recognized that there might be substantial differences that can occur across regions within a country and hence the need to measure and monitor trends at different smaller area levels to fully understand a country's progress and possible interventions. Moreover, according to Rose (2015), geographic variation in population parameters exist not only at a country level but also extends into sub-national and local areas and thus knowledge of such variations is necessary for decision making about resource allocation (Islam et al., 2018).

Further to this, small area estimates of the prevalence of risk factors play a crucial role in decision and policy-making and as such, quality of these estimates

must always be taken into account (Manzi et al., 2011). For example, when addressing area-specific health issues or lifestyle behaviours, researchers need to put into consideration the fact that some people live in deprived areas with limited access to screening programmes or preventive healthcare campaigns, or they may have a higher level of certain risk factors (Manzi et al., 2011). Knowledge of the prevalence of risk factors in small areas is essential to make health promotion strategies more effective.

To overcome the challenges faced by sample surveys such as high costs of data collection, declining response rates, misreporting and small area estimation discussed above, there is need to combine information from different data types and sources. Combining information from different data sources with varying spatial resolutions has many advantages. First and foremost, combining information from several sources provides a means for improving estimates of population parameters (Kim et al., 2018). For instance, "combining information from multiple data sources can enhance estimates of health-related measures by using one source to supply information that is lacking in another, assuming the former has accurate and complete data" (He et al., 2014, p. 1). Further, if the two data sources have common variables, then the produced estimates may have improved precision due to increased sample size for the common survey items, (Merkouris, 2010).

Since different data sources have different limitations such as nonresponse, noncoverage and measurement or response bias, combining information for the same set of variables reported from multiple sources might alleviate these errors and produce improved estimates of these variables (He et al., 2014). For instance, combining survey information with non-survey information such as census provides comprehensive and precise estimates of useful health indicators at a very fine spatial scale which improves decision making (Fung et al., 2010).

Reliable estimates of health status and many other population parameters may in turn improve global health (Finucane et al., 2014).

Also, a combined dataset can address analytic problems beyond the scope of a single survey (Lohr & Raghunathan, 2017) and can derive information on multiple sections of the population unlike when a single data source is used. A scenario may happen whereby two surveys conducted independently on the same population can have one or more variables in common and other variables that are not common for both surveys.

Because data is usually scarce and inadequate at small domain levels, combining information provides a solution to small area estimation, (Islam & Chandra, 2019; Gutreuter et al., 2019; Zehang et al., 2019) and it addresses the best prediction problem for small areas (Kim et al., 2018). Furthermore, the small area estimates produced by combining information from two surveys are more efficient than those produced from a single small survey (Islam & Chandra, 2019). This is made possible with the growing availability of data from several different surveys such as DHSs' as well as MICS and other auxiliary information outside samples used in surveys. The auxiliary information usually comes from large administrative record datasets like census and remote sensing data derived from satellite images (Rose, 2015).

Lastly, combining information can produce datasets without missing information in them. Usually, other data sources may contain information on variables that are not measured in a survey (National Academies of Sciences, Engineering and Medicine, 2017). Therefore, models developed on one data source may be used to impute missing variables in other sources thus making statistical inference beyond the scope of a single study possible (Lohr & Raghunathan, 2017).

1.2 Problem Statement

The declining response rates for national surveys worldwide affect the reliability of estimates. Declining response rates have contributed to higher costs for data collection and even if reliable estimates for subpopulations of interest may be calculated, they may require multiple years of data which when they are produced may be outdated (Lohr & Raghunathan, 2017). In resource limited settings like Malawi, it is difficult to obtain small area statistically representative estimates like at sub-district level. This is because national surveys such as the DHS, which is a nationally representative survey, is powered to provide district-level estimates and provides little or no information on the small area population characteristics that probably vary across Malawi's geographic space.

To obtain estimates for subpopulations of interest, it may be necessary to combine the information from surveys with other rich data sources such as population and housing census (PHC) or other surveys that have a high spatial resolution. However, methods on how to statistically combine such information to derive better estimates for small area population parameters are not easily accessible. For this reason, researchers usually make inferences about small areas based on the district level estimates which are readily available from the DHS. This research will investigate statistical methods for integrating DHS data with much higher spatial resolution and precision datasets (census-based data) to estimate under-five mortality at a fine spatial scale in Malawi.

1.3 Objectives of the study

1.3.1 Main objective

- To develop a principled statistical framework for combining data from census and survey sources with an application to Malawi DHS and Malawi Population Census.

1.3.2 Specific objectives

- To investigate different methods for combining data from multiple sources.
- To model under-five mortality at a sub-district level spatial resolution in Malawi.
- To spatially predict under-five mortality indicator values at unsampled locations.

1.4 Significance of the study

This study will help come up with estimates of important indicators in areas that were not sampled and will deepen the understanding of the methods used in combining information from different sources. This is important in resource constrained countries where there are several disjointed data sources. Furthermore, the study will contribute to the available work that has been done in the field of data fusion in order to come up with reliable small population estimates. In many cases, fine-scale data is not readily available, leading to the use of coarse data in monitoring and disease surveillance. By leveraging the multiple data sources available, small area estimates can be useful for

monitoring and evaluation purposes of different health outcomes at a fine spatial resolution.

CHAPTER TWO
LITERATURE REVIEW

The following chapter reviews statistical methods that are used in combining data from different sources. It further expounds on the concept of Bayesian Hierarchical Models and Generalised Linear Geostatistical Models.

2.1 Introduction

This chapter presents an overview of various statistical methods used for linking data from different sources. It also introduces the methods that have been used to link DHS data and census data for this study.

2.2 Methods for combining data

2.2.1 Statistical matching

Statistical matching is a model-based approach for providing joint statistical information based on variables and indicators collected through two or more sources, (Leulescu & Agafitei, 2013). Statistical matching requires that the two files containing information on a set of units be completely disjoint, (Christen, 2012; D'Orazio et al., 2006; Kiesl & Rässler, 2006). This means that data comes from two independent probability samples, with a few or no units in common, (Scanu, 2014). In scenarios where there are a few units common to both data sets, then these common variables are not able to identify the units. The statistical matches are then made based on similar characteristics and not a unique identifier, (Herzog et al., 2007; Radner, 1980; Winkler, 2014).

For example, if there are two data sources whereby one data source has information on the education level of persons, their gender, age and municipality and another data source has information on the occupation of (other) persons, their gender, age and municipality, then the information on gender, age and municipality can be used to statistically match similar units in the data sources with each other, (Waal, 2015).

The basic idea in statistical matching is that there are two data sources A and B sharing a set of variables X while the variable Y is available only in A and the variable Z is observed just in B, D'Orazio (2011). The X variables are common to both the data sources, while the variables Y and Z are not jointly observed. Statistical matching, therefore, investigates the relationship between Y and Z at "micro" or "macro" level.

In the micro case, statistical matching aims at creating a synthetic data source in which all the variables, X, Y and Z, are available. This synthetic data source is comprised of data from individual units in the different data sources (D'Orazio, 2011). The information from one data source is used to estimate target values in the other data set, (Waal, 2015, p. 4).

On the other hand, in the macro-level case, the data sources are integrated in order to derive an estimate of the parameter of interest, (D'Orazio, 2011; D'Orazio et al., 2006; Waal, 2015). Based on all the data, a parametric model such as a multivariate model is built and thereafter one estimates the parameters of the model which are then used to estimate population parameters of interest, (Waal, 2015, p. 5).

Statistical matching may be applied in different situations, such as in matching of two non-overlapping surveys with common background variables, matching of Big Data to survey or administrative data, and finding imputation values when for certain groups of units a number of variables are missing by design,

(Waal, 2015). For example, Wolff 1977 and Ruggles and Ruggles 1974 carried out statistical matches in the US using the 1969 Internal Revenue Service Tax Model and the 1970 Decennial Census Public Use Sample 15 percent file whose aim was to estimate and analyse the size distribution of household wealth, (Radner, 1980).

The major advantage of statistical matching is that it can enhance the complementary use and analysis of existing data sources such as cross-cutting statistical information that encompasses a broad range of socio-economic aspects without further increasing costs and response burden, (Leulescu & Agafitei, 2013). Besides, if statistical matching is done accurately, the linked data sets may provide more information than would be provided by each different data source and this opens up opportunities for performing multivariate relationships among the extra variables, (Christen, 2012; Leulescu & Agafitei, 2013). For example, matching patient addresses with spatial data can lead to the discovery of correlations between environmental factors and local hot-spots of disease cases, (Christen, 2012).

However, statistical matching is a complex operation which requires specific technical expertise and raises several methodological issues.

2.2.2 Imputation

Another method for combining information from different sources is by imputation. Imputation refers to the process of replacing missing data with substituted values. Imputation fills in responses for items not completed by the respondent, (Brick, 2011) and thus multiple imputation is a tool for handling nonresponse in sample surveys, (Gelman et al., 1998). After imputation, analysis is then done as if there were no missing values at all, (Zarnoch et al., 2010).

In the imputation process, variables that have missing values or variables missing from a data source are filled in using various techniques from information available from surveys or other data sources. When imputing missing data from several samples one can either impute missing data from each survey or can combine data from all the surveys and impute the missing data in a combined data matrix, (Gelman et al., 1998). When imputing data from each survey, models usually built from a certain data source are used to impute the missing variables from another data source, (Zarnoch et al., 2010). However, this method becomes challenging if each individual survey has a lot of missing information. Also, such methods are valuable when the survey with the missing variable is much larger than the survey with the observed variable, and/or where the available common variables are highly predictive of the outcome, (Elliott et al., 2018).

On the other hand, when imputing data in a combined data matrix, multivariate models or a sequence of regression models are often used. This method, however, does not take into consideration the differences between surveys, (Zarnoch et al., 2010). For instance, it does not take into account the differences in times of conducting the surveys, survey methodologies and/or organisations conducting the surveys, (Gelman et al., 1998).

A study by Gelman et al. (1998) used multiple imputation by adding a hierarchical regression model to existing methods of imputation designed for single surveys. This hierarchical model allowed covariates at both individual and survey levels and linked parameters in the different surveys. Imputations of item nonresponse were determined by data from that survey and imputations for questions not asked in a survey were determined by data from other surveys. This study was motivated by a study of pre-election public opinion polls in which not all the questions of interest were asked in all the surveys.

Results were a compromise between the approaches of no pooling and complete pooling of surveys which is one of the properties of the Bayesian approach. The estimated between-survey variation yielded wide posterior intervals for questions not asked in the survey. Including more variables (including those not asked in all surveys) in the imputation model, offered a flexible way to account for item nonresponse.

He et al. (2014) conducted a study on hospice-use by late stage cancer patients in which data was available from patients' abstracted medical records but there were issues of underreporting. The data was therefore supplemented with the patients' medicare claims that contained information on hospice use even though these data also had some missing information. A multiple imputation approach was applied using information from both sources whilst borrowing strenth from each other. This method yielded sensible results since it was able to account for misclassification of the hospice use from both data sources in an appropriate way. Clearly, this approach provided an effective means to synthesizing information from the two sources.

Imputation is advantageous because it can augment the amount of information available for analysis and to produce data sets without holes in them, (Lohr & Raghunathan, 2017; Soley-Bori, 2013; Zarnoch et al., 2010). In addition, it provides a means for inferring beyond the scope of each study. Multiple imputation allows for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them, (Sterne et al., 2009).

Despite these advantages, there are several challenges associated with combining information from multiple survey data sources via the multiple imputation approach. For example, surveys usually involve stratification, clustering and weighting for selection and nonresponse, (Lohr & Raghunathan, 2017). Though each survey may represent the same or a similar population, the complex

survey design differences have to be taken into consideration in deriving the combined estimates. Estimates based on combining information from multiple data sources are subject to errors due to incomparability as well as issues in modelling of those errors, (Soley-Bori, 2013). Lastly, many multiple imputation procedures assume that data are normally distributed, so including non-normally distributed variables may introduce bias, (Sterne et al., 2009).

2.2.3 Multiple frame sampling

Another method for combining data from different data sources is multiple frame sampling. A multiple frame survey is defined as "a set of several (single frame) surveys whose samples are combined to provide parameter estimates for the union of frames",(Biemer, 1984, p. 1). The objective of the dual-frame approach is to draw subpopulation samples from different sampling frames that, when combined, provide full coverage of the target population, (Baffour et al., 2016). The general principle in multiple frame sampling is that probability samples are selected independently from say Q sampling frames available, (Lohr & Raghunathan, 2017; Rao & Lohr, 2006). If $Q = 2$, then the survey is called a dual-frame survey and if $Q > 2$ then it is a multiple frame survey. Information is collected for every unit in each frame sample which is then used to classify each frame-specific sample data into disjoint domains, (Mecatti & Singh, 2014). For example, for a simple dual-frame case, with frames A and B, four frame-specific domain samples might be classified, that is, samples $a(A)$ and $ab(A)$ from frame A and $b(B)$ and $ab(B)$ from frame B. The collection of data from the Q frames is then used for estimation of the population parameter of interest and the union of all the frames represents the target population, (Mecatti & Singh, 2014).

Estimation occurs in different ways. First is a combined frame approach which is also known as a single frame estimation. In the combined frame approach,

all the samples drawn from different frames are combined into a single sample with appropriate weights included and then population parameter estimates are computed directly, (Lu et al., 2013). This method is simple and unbiased but does not use all the relevant information, (Lu et al., 2013).

The other approach is known as a separate frame approach and involves computing separate estimates of each domain using each sample that falls in that domain, (Mecatti & Singh, 2014). Afterwards, the domain estimates are aggregated over all the domains within and between frames in order to obtain an estimate of the population parameter.

Sampling units from multiple frames increases coverage and/or efficiency than when only a single sampling frame is used, (Brick, 2011). Multiple frame methodology can be used to improve survey coverage by complementing the strengths and limitations of one another and/or to reduce cost while maintaining broader coverage, (Chromy & Wilson, 2013; McMillen et al., 2015). For example, in a survey of businesses, where one frame is an incomplete list of businesses but easily accessible, and the other frame a list sample of businesses indicating their geographic areas, information from these two frames may be combined to provide better coverage and lead to efficiency, (Brick, 2011).

In the same vein, multiple frame surveys greatly decrease sampling costs due to the use of already available administrative records, (Rao & Lohr, 2006). Hartley (1962, 1974) showed that dual-frame surveys can cost far less than a single-frame survey that achieves the same precision. His applications concentrated on the situation where one frame is complete but expensive to sample; other frames are inexpensive to sample but incomplete. In many agricultural surveys, an area frame consists of segments of land; enumerators visit a probability sample of the segments. A list frame consists of the names and addresses of agricultural operators. The area frame is complete and insensitive to changes in farm ownership and activity, but very expensive to sample because of the

in-person visits. The list frames are usually less costly to sample, particularly if the commodity of interest is concentrated in the operators on the list, but the lists may not include all producers of the commodity.

It should be noted that the different frames from which information is derived from usually include different subsets of the population, (Lohr & Raghunathan, 2017) and that these methods are ideal when combining information from sources that are measuring same quantities.

2.2.4 Record linkage

Record linkage refers to "a process of pairing records from two files and trying to select the pairs that belong to the same entity", (Winglee et al., 2005, p. 4) and is a key technological tool that is used to exploit the wealth of information from different data sources, (Shlomo, 2019). Record linkage is conducted between two distinct data sources or within a single data-set to identify multiple entries for one person or record unit. This is done by matching and then merging records for a particular entity from a survey with other data sources believed to belong to the same entity using record identifiers such as name, date of birth and address, (Shlomo, 2019; Winglee et al., 2005).

There are two types of record linkage and these are exact or deterministic record linkage and probabilistic record linkage. In exact record linkage, records that have been linked from two different sources are declared to belong to the same entity if and only if they agree exactly on every character of every matching variable, (National Academies of Sciences, Engineering and Medicine, 2017; Shlomo, 2019). For example, "when comparing two records on first and last name, age and street number, the records are deemed to be a link if and only if the names agree on all characters, the ages are equal and the street numbers are identical", (Herzog et al., 2007, p. 82).

On the other hand, probabilistic record linkage is used when there is no unique identifier across data sources or when personal identifiers used in reporting or transcription do not differ, (Brown, 2017; Kabudula et al., 2014). By definition, probabilistic record linkage refers to "the process of determining which records in two databases correspond to the same underlying entity without a unique identifier", (McVeigh et al., 2019, p. 1). To determine whether a pair of records belong to the same entity, probabilities are used, (Machado, 2004). Under the probabilistic type of record linkage, a similarity score of likely matches is calculated using a pattern of agreements, disagreements, and near-agreements among the variables used in linking, (Lohr & Raghunathan, 2017). Threshold value is determined before-hand such that if the similarity score exceeds the threshold then a record from source A is linked with a record from source B.

All in all, data linkage should be conducted with optimal validity and reliability, and minimal risk to privacy and confidentiality, (Dusetzina et al., 2014).

2.2.5 Bayesian Hierarchical methods

Bayesian hierarchical models are "multi-level stochastic models in which a probability is decomposed into a series of levels linked by simple probability rules", (Arab et al., 2007, p. 2). The development of hierachircal models was a primary result of a shift in collaboration of statistics with other disciplines and inclusion of complex processes and recognition that prior knowledge in experiments plays a crucial role in statistical inference, (Gelfand, 2012). Hierarchical models offer a flexible framework for accommodating complex relationships between data and the process models while taking into account different sources of uncertainty in the model as well as priori scientific knowledge while retaining many advantages of a strict likelihood approach, (Arab et al., 2007). In hierarchical modelling, the joint distribution of a collection of random variables can be

decomposed into a series of conditional models. For example, if a, b and c are random variables then basic probability allows the factorization: $[a, b, c] = [a|b, c][b|c][c]$

whereby [.] specifies a probability distribution and the joint distribution describes the behaviour of the process at all spatial locations and possibly at all times.

There are three basic stages when modeling a complicated process in the presence of data. As motivated by Berliner (1996), a data model is the first stage and is an observational process which specifies the data distribution given the fundamental process of interest and parameters describing the data model, (Gelfand, 2012). The second stage is a process model which describes the process and is conditional on other process parameters. The final stage is a parameter model which models uncertainty in the parameters, from both the data and process stages. Mathematically, the three stages are written as follows:

$$\text{Stage 1: Data model: } [data|process, data\ parameters]$$

$$\text{Stage 2: Process model: } [process|process\ parameters]$$

$$\text{Stage 3: Parameter model: } [data|process, data\ parameters]$$

It has to be noted that each of these three stages can have many sub-stages.

Unlike the methods discussed above where paramaters are fixed and unknown, in Bayesian hierarchical models the paramaters are regarded as random variables and statements about these parameters are interpreted as the degree of belief based on some prior knowledge, (Held & Sabanés Bové, 2014). The beliefs about these parameters are then revised and summarized in a posterior distribution after getting the data, (Filippi & Holmes, 2017).

Advantages of Bayesian hierarchical models

Firstly, when modeling via the Bayesian approach, many different sources of uncertainty can be incorporated and projected through time. For instance, the prior distributions reflect uncertainty in parameters while the likelihood model captures stochasticity of the process and some sampling and non-sampling errors. As a result, one framework can handle data from multiple sources while taking into account the different quality of each source at the same time, (Alexander & Alkema, 2018).

Another advantage of Bayesian modeling is that since it incorporates prior information in the model, then the resulting posterior parameter estimates are influenced by the observed data. This is particularly useful in situations where data is limited in such a way that the little information available can be combined robustly through the Bayesian approach and hence allowing information to be pooled across different dimensions such as time and age. For example, trends observed in some areas can be used to inform trends in other similar areas with limited data, (Alexander & Alkema, 2018).

A study involving Bayesian hierarchical models for estimating agricultural yield from multiple repeated surveys was formulated by (Wang et al., 2012). In this study, prior distributions on model parameters were specified and details on model inference were presented via Markov chain Monte Carlo (MCMC) methods. In their model, information from multiple monthly surveys measured on different temporal supports was combined and the different levels of the hierarchy incorporated dependence between monthly surveys as well as serial dependence of annual yield. Results showed that the Bayesian model produced superior yield forecasts/estimates, while quantifying different sources of uncertainty. This study shows that hierarchical models are flexible in accommodating multiple data sources and different serial correlation structures. More importantly the model was able to produce root mean square error reduced by between 7.5% and 15.5% over other yield estimators. Finally, due to the model's ability to include

auxiliary information in its levels, the study was able to directly measure the effect of environmental conditions on end of season corn yield.

Finucane et al. (2014) conducted a study whose aim was to estimate population-level trends in measures of health status. In their study, Finucane et al. (2014) presented a Bayesian model that systematically combined disparate data to make country-, region- and global-level estimates of time trends in important health indicators, (Finucane et al., 2014). A total of 199 countries and territories from 1980 to 2008 were included to estimate trends in mean systolic blood pressure (SBP) for adults aged 25 years and older. The 199 countries were grouped into 21 subregions which were further grouped into seven merged regions. In this study, a hierarchical model was necessary to accommodate missingness when the data was being aggregated to regional and global levels and to provide inference for all country-year-age triplets. To borrow strength over time, countries and age prior distributions provided by the hierarchy were used while constraining plausible parameters. Finucane et al. (2014) fitted a Bayesian hierarchical model using MCMC approach which enabled inference in high dimensional constrained parameter space while providing posteriors important for statistical inference. Results showed that there had been a transition in risk of HBP for cardiovascular disease with decreasing blood pressure in high-income regions and increasing levels in many lower income regions.

A Bayesian framework was also applied in propagating large database of malaria field survey to evaluate trends in malaria infection prevalence across Sub-Saharan Africa between the years 2000 to 2015, (Bhatt et al., 2015). Data from 27,573 geo-referenced population clusters from sub-Saharan countries were combined in a spatio-temporal Bayesian geostatistical model to model malaria infection prevalence in children aged between 2 to 10 years. Results showed that the prevalence of malaria infection in these children declined from 33% in 2000

to 16% in 2015. Community-based surveys data across sub-Saharan Africa were also combined to determine malaria transmission cycles, (Snow et al., 2017). In this study, data from 50,424 surveys at 36,966 geocoded locations were combined in a Bayesian hierarchical binomial model in order to estimate stable spatial and temporal structured patterns of malaria prevalence in children aged 2 to 10 years between the years 1900 to 2015. Similarly, results showed a long-term decline in malaria prevalence from 40% between 1900-1929 to 24% between 2010-2015. Evidently, the Bayesian approach provided a framework for conveniently combining such different data sources.

Geostatistical modelling

Geostatistical data methods are a form of hierarchical specification which naturally lead to the adoption of a Bayesian framework methodology for inference and modeling purposes, (Gelfand & Banerjee, 2017). Geostatistics refers to "the sub-branch of spatial statistics in which the data consist of a finite sample of measured values relating to an underlying spatially continuous phenomenon", (Diggle & Ribeiro, 2007, p. preface).

The general theory in geostatistics is that measurements taken at locations close together are usually more alike than measurements taken at locations farther apart, (Gotway & Hartford, 1996). Geostatistics, therefore, provides methods for quantifying this spatial correlation and for incorporating it in statistical estimation and inference.

The other main objective of geostatistics lies in prediction. In geostatistics, prediction refers to "inference about the realization of the unobserved signal process $S(x)$", (Diggle & Ribeiro, 2007, p. 24) and this is called kriging.

CHAPTER THREE
MATERIALS AND METHODS

The following chapter expands in detail the materials and methods that were used in the study. It expounds on the statistical formulae for the analysis methods that were used in the study.

3.1 Data sources

3.1.1 MDHS data

The 2015-2016 MDHS was a cross-sectional survey and took place between October 2015 and February 2016 and it provides a comprehensive overview of population, maternal and child health issues in Malawi. In this survey, data that allows the calculation of key demographic indicators such as fertility, under-five and adult mortality rates were collected. In addition, the data allows the exploration of direct and indirect factors that determine the levels and trends of fertility and child mortality. The 2015-2016 MDHS used a two-stage selection process as follows: 850 standard enumeration areas (SEAs) were selected in 173 urban and 677 rural areas (stratum or SEAs) using a probability proportional to the SEA. In the second selection stage, a household list was used as a sampling frame for selecting households in the selected SEAs. A total of 30 households per urban SEA and 33 households per rural SEA were selected for interviews. The survey interviewed all women aged 15 to 49 who were either permanent residents of the selected household or those who slept in the household the night before the survey. The survey selected 27,516 households and identified 25,146

eligible women. Consequently, a total of 24,562 women were interviewed out of the 26,361 households that were occupied.

3.1.2 Census and population level covariates data

In addition to the point level covariate data from the MDHS referenced at each cluster, we also obtained population level geospatial datasets covering the whole country from Humanitarian Data Exchange (https://data.humdata.org/). In particular, we used vulnerability score to capture the social-economic structure of the country. Additionally, being a malaria endemic country and its threat among children, we included malaria risk as a potential predictor of mortality. Model-based geostatistics are used to measure infection prevalence for malaria and average proximity risk score for vulnerability by creating risk surfaces based on thousands of geolocated cross-sectional surveys. The inclusion of these covariates was influenced by studies that have shown that low socio-economic status and malaria are among the factors contributing to child mortality, (Johnson et al., 2010). Lastly, population data were extracted from the WorldPop data sets. All these population level data were in raster format at 100m spatial resolution. These data provided estimates at a high resolution thus allowing the calculation of important health indicators at a finer scale than would be possible if only district-level covariates were used.

Raster data allows for an easier way of integrating two types of data such as discrete and continuous data. Also, with raster data, analysis of the data is easy and quick to perform and do not require storage of geographic coordinates since the geographic location of each cell is implied by its position in the cell matrix.

3.2 Data management

Child and women data files from DHS were merged by matching their household numbers. Variables that were not necessary for the study were dropped and only kept the variables relevant to the under-five mortality problem. Data cleaning was carried out to remove duplicates and observations with missing data. One aspect of the data cleaning process was to reduce the number of variables in the merged dataset. Thus, from the 24,562 women aged between 15 to 49 that were interviewed, a total of 11680 women with children aged five years and below were included in the survey of which 286(2.4%) women had their children dead a year prior to the interview date while 11394(97.6%) women had their children still alive.

3.3 Data analysis

Descriptive analyses were first done to summarize the data. In particular, cross tabulations were done between the response variable and the explanatory variables. Chi-square test of association was performed to find the factors associated with under-five mortality. An exploratory spatial mapping was performed to produce maps detailing the spatial distribution of clusters within the 28 districts of Malawi and the underlying population densities. Spatial mapping of crude under-five mortality at district level was also done.

To investigate the factors affecting under-five mortality, all potential risk factors were then put into geostatistical models. During modeling the categories no education and primary education for mother's education were categorised as one and this was asigned as the reference group. Similarly, for wealth index, lowest and second weathy categories were categorised into a single category and this was the reference category. This was done because the first categories had few observations and could not be used as reference groups. After fitting the model, we did predictive mapping of under-five mortality across the whole

Malawi including at unsampled locations. All the covariates were used in the prediction model and the predicted values were posterior means realised from the posterior predictive distribution. In addition, approximate standard errors were also mapped. Both the predictive and standard errors maps were produced by overlaying their rasters on the Malawi districts map to estimate under-five mortality in Malawi.

All data analyses were carried out in the R environment for statistical computing using geostatsp package (R Core Team, 2020) which uses the Integrated Nested Laplace Approximation (INLA) on the backend. Data visualization was carried out in R.

Description of key study variables

Table 3.1 shows a description of variables used in the study.

Table 3.1: Variables used in the study

Variable	Description	Source
Alive(Response variable)	Child alive(0= Dead, 1=Alive)	DHS
Age	Mother's age in years	DHS
Sex	Sex of child(1=Male, 2=Female)	DHS
Education	Mother's level of education (0=None, 1=Primary, 2=Secondary, 3=Higher)	DHS
Wealth	Index showing a household well-being (1=Lowest, 2=Second, 3=Middle, 4=Fourth, 5=Highest)	DHS
Residence	Area of residence(1=Rural, 2=Urban)	DHS
Birth weight	Weight of child at birth in kilograms (1=Low birthweight($< 2.5Kgs$), 2=Normal birthweight($\geq 2.5Kgs$)	DHS
Malaria risk	Average malaria risk	Census
Vulnerability	Proportion of vulnerable individuals	Census

3.4 Study setting

Malawi is a small landlocked country in Southern Eastern Africa, sharing boundaries with Zambia, Tanzania and Mozambique. The country is divided into 3 administrative regions and further into 28 districts. The public health system comprises 4 central hospitalas, district hospitals and health centres. At the community level, health suiveillance assistants (HSAs) provide basic care including treatment of common illnesses in children. In addition, the faith-based health facilities primarily under the Christian Health Association of Malawi (CHAM) also provides a wide network of health facilities. Figure 3.1 provides more information about the districts and the distribution of public health facilities.

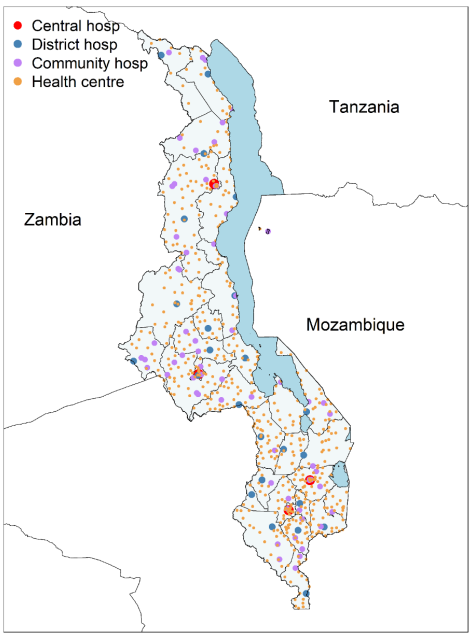

Figure 3.1: District boundaries, location and distribution of health care facilities in Malawi. Here, government and CHAM facilities are shown

3.5 Modeling framework

In this section, the Bayesian hierarchical method for data integration is discussed. Given that data is available, Bayesian methods fit in to estimate and predict distribution of the process as well as parameters in hierarchical settings. The parameters are considered to be random and not fixed. In Bayesian statistics, prior knowledge is used along with available observed data in order to come up with new estimates. These new estimates are derived from posterior distributions of the data through Markov Chain Monte Carlo (MCMC) or INLA approaches and are then used for statistical inference. Bayes' Theorem provides a mechanism for finding the posterior and the theorem is presented in the equation below where y represents a random variable and θ is a parameter of interest:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \tag{3.1}$$

In equation 3.1 above, $f(\theta|y)$ is the posterior distribution and is the conditional distribution of a specified parameters given the data, $f(y|\theta)$ is the likelihood function which is given by $f(y|\theta) = \int(f(y|\theta)f(\theta))d\theta$ and $f(\theta)$ is the prior distribution. Usually, the Bayes' theorem is written as the posterior distribution being proportional to the likelihood times the prior distribution, that is:

$$f(\theta|y) \propto f(y|\theta)f(\theta). \tag{3.2}$$

In this study, focus is only on the geostatistical modeling approach which is a special case of the Bayesian Hierarchical modeling that is used to combine different datasets.

3.6 Integrated Nested Laplace Approximation

This section describes how new estimates are derived through INLA approaches. Given the posterior distribution:

$$f(\theta_i|y)\int f(\theta_i,\psi|y)d\psi = \int f(\theta_i|\psi,y)f(\psi|y)d\psi \tag{3.3}$$

Interest is on obtaining posterior marginals $f(\theta_i|y)$ for each parameter in the vector and the estimates of the hyperparameters given by,

$$f(\psi_k|y)\int f(\psi_k|y)d\psi_{-k} \tag{3.4}$$

The following steps are followed in the INLA approximation. Firstly, the posterior marginals of the hyperparameters are approximated as given in the equation below:

$$\begin{aligned} f(\psi|y) = \frac{f(\theta,\psi|y)}{f(\theta|\psi,y)} &\propto \frac{f(\psi)f(\theta|\psi)f(y|\theta)}{f(\theta|\psi,y)}, \\ &\approx \frac{f(\psi)f(\theta|\psi)f(y|\theta)}{f(\theta|\psi,y)} \mid_{\theta_i*(\psi)} \tilde{f}(\psi|y), \end{aligned}$$

where $\tilde{f}(\theta|y)$ is a Gaussian approximation for $f(\theta|y)$ and θ^* is the mode. Secondly, the parameter vector is partitioned in such a way that $\theta = (\theta_i, \theta_{-i})$ and are again approximated using the Laplace procedure to obtain:

$$f(\theta_i|\psi,y) = \frac{f(\theta_i,\theta_{-i}|\psi,y)}{f(\theta_{-i}|\theta_i,\psi,y)} \approx \frac{f(\theta,\psi|y)}{\tilde{f}(\theta_1|\theta_i,\psi,y)} \mid_{\theta_{-i}=\theta^*_i(\theta_i,\psi)} \tilde{f}(\theta_i|\psi,y). \tag{3.5}$$

INLA bypasses the computational complexity of computing $\tilde{f}(\theta_i|\psi,y)$ by exploring the marginal joint posterior for the hyperparameters $f(\psi|y)$ in a grid search to select the important points ψ_k jointly with a corresponding set of weights Δ_k to give approximates to the posterior to the hyperparameters. Each marginal $\tilde{f}(\psi_k|y)\forall k$ can be obtained using log-spline interpolation bases on selected ψ_k

and Δ_k. For each k, the conditional posterior $\tilde{f}(\theta_i|\theta_i|\psi, y)$ is computed and a numerical integration:

$$\tilde{f}(\theta_i|y) \approx \sum_{k=1}^{K} \tilde{f}(\theta_i|\psi_k, y)\tilde{f}(\psi_k|y). \tag{3.6}$$

is, then, used to obtain $\tilde{f}(\theta_i|\psi, y)$.

3.7 Geostatistics

Geostatistics is a branch of spatial statistics concerned with the analysis of statistically discrete data that relates to an unobserved continuous phenomenon. In a geostatistical model, the data is represented as

$$(y_i, x_i) : i = 1, ..., n.$$

Here, the y_i are the realized values of the random variable Y_i associated with spatial locations $x_i \in A \subset \mathbb{R}^2$. In our application, the locations x_i are the specific locations (clusters) that were sampled during the DHS. Interest is on estimating the underlying mortality across the continuous spatial region. It is further assumed that the $Y_i's$ are dependent on an unobserved stochastic process $S = S(x) : x \in \mathbb{R}^2$ which is expressed as follows;

$$[S, Y] = [S][Y|S] \tag{3.7}$$

Let $p(x)$ be the prevalence of under-five mortality at location x. The resulting model is then binomial in nature yielding a generalized linear geostatistical model (GLGM). A standard GLGM is presented as

$$\log\left[\frac{p(x_i)}{1 - p(x_i)}\right] = d(x_i)'\beta \tag{3.8}$$

$d(x_i)'$ is a vector of explanatory variables associated with location x_i. With smoothing and random effects terms, the model becomes;

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = d(x_i)'\beta + S(x_i) + Z_i \tag{3.9}$$

where $\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean 0 and variance σ^2. The correlation function is provided by $\rho(x, x') = \text{Corr}\{S(x), S(x')\}$. It is assumed that the spatial process $\mathcal{S}$ is stationary and isotropic. Therefore, $\text{Corr}\{S(x), S(x')\} = \rho||x - x'||$ where $||.||$ is the Euclidean distance.

A Matern correlation function is used in this application

$$\rho(u; \phi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-}1\left(||x - x'||/\phi\right)^{\kappa}\mathcal{K}_{\kappa}\left(||x - x'||/\phi\right) \tag{3.10}$$

In this study, we draw data from different sources with different resolutions effectively splitting the covariate term, $d(x_i)'$ into different components. Therefore, the model 3.9 above becomes

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = d(x_i)'\beta + W(x)'\delta + Q(x)'\gamma + S(x_i) + Z_i \tag{3.11}$$

In this formulation, $d(x_i)'$ is the vector of covariates at the sampled locations as before, while $W(x)'$ and $Q(x)'$ are covariate vectors over the areal unit and not necessarily at individual locations x_i. Therefore, the model captures data at different spatial resolutions. In particular, the $W(x)$ and $Q(x)$ capture aggregate data for the entire spatial unit.

3.8 Spatial prediction for Generalised Linear Geostatistical Models

Spatial prediction is concerned with estimating unknown values of a stochastic process at locations where there was no data based on data available from

nearby location points. The first step in spatial prediction is to define the predictive target; let T^* be the target and is a property of the realisation of a spatial component of the set of values of $d(x)^T\beta + S(x)$ for all values of x in the region of interest, A. In our application where the geostatistical model is binomial in nature, the predictive target is the prevalence surface over the region of interest A and is shown by the equation below:

$$T^* = \{p(x) = exp\{T(x)\}/(1 + exp\{T(x)\}) : x \in A\}, \tag{3.12}$$

where $T(x) = d(x)^T\beta + S(x)$ and the prediction takes the form of a map.

Secondly, a number of random samples , say, B are drawn from the predictive distribution of the complete spatial surface $\{S(x) : x \in A\}$. Thereafter, the value of the specific target from each sample $T_1^*, ..., T_B^*$ is calculated and suitable summaries of the resulting empirical distribution of the T_i^* are reported.

During the actual prediction process, the region of interest A is approximated using a grid $\chi = \{x_1^*, ..., x_q^*\}$ consisting of q prediction locations in region A. To make inference on T^*, we obtain samples from its predictive distribution, $[T^*|y]$. It has to be noted that T^* can be calculated directly from the fitted model parameters and the spatial surface $S(x)$, hence samples of T^* can be obtained from the predictive distribution of $S^* = \{S(x) : x \in \chi\}$. Finally, the predictive samples s_h^* for $h = 1, ..., B$ can then be transformed into corresponding samples t_h^* from the predictive distribution of T^* by direct estimation. This can then be used to obtain any summary of the predictive distribution of interest at any or all of the q prediction locations x_j^*.

3.9 Geostatistical model for Malawi

The covariate vector $d(x_i)'$ captures the following covariates from DHS; mother's age, birth weight, residence, mother's education and wealth index. On the other

hand, the vector $W(x)'$ captures census covariates and these include malaria risk and vulnerability both of which are in raster format.

The first step was to find the variables that were associated with a child's death. Bivariate analysis was carried out to identify these variables which were later put into bayesian models for further analysis.

Default priors were used.

After fitting the GLGM, the posterior estimates were overlayed on a raster file for Malawi to generate predictions in under-five mortality in Malawi.

CHAPTER FOUR
RESULTS

In this chapter, results of the analysis are presented beginning with exploratory data analysis results and thereafter statistical inference based on the models built. Maps of under-five mortality risk prediction are also presented together with their associated errors.

4.1 Exploratory analysis

4.1.1 Cluster locations and population density

Figure 4.1 shows cluster locations within districts from which respondents from the survey were obtained and the underlying population density per 100,000. From the figure, it can be observed that more people are concentrated in cities. Lakeshore areas also have relatively more people. In addition, the Central and Southern regions have higher population densities compared to the Northern region.

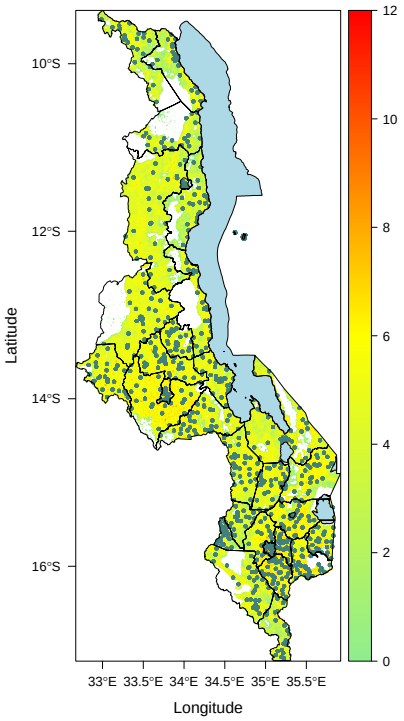

Figure 4.1: 2015-16 MDHS cluster locations and underlying population densities per 100,000. Water areas are shown in blue and white represent uninhabited and protected places e.g. national parks.

4.1.2 Malaria risk

Figure 4.2 A shows how the risk of malaria is distributed across Malawi. As shown in the map, it is observed that the risk is lower in the northern region and in the cities and higher in the southern region. Lakeshore areas also have a higher risk of malaria regardless of region. These risks are simply observed and there are likely to be underlying reasons for the disparities in malaria risk across the country. This therefore, is more likely to have an influence on under-five mortality as well as mortality patterns within the districts.

4.1.3 Vulnerability

Figure 4.2 B shows disparities in how populations are vulnerable to certain conditions. We observe that the uppermost Northern Malawi is relatively more vulnerable as compared to the central and southern part of the northern region. The Central region and the cities of Malawi are generally less vulnerable and most of the Southern region and lakeshore populations are the most vulnerable.

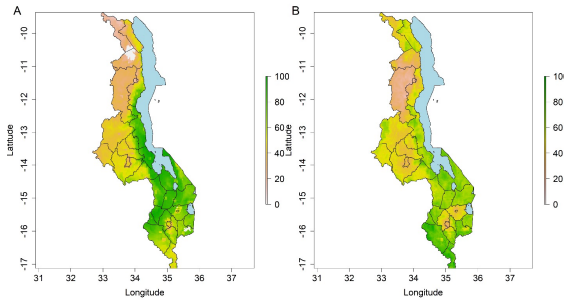

Figure 4.2: High spatial resolution population level covariate data (A) Average malaria risk (B) Proportion of vulnerable individuals

4.2 Under-five mortality rates

Figure 4.3 shows crude under-five mortality rates at district level obtained by aggregating number of deaths reported within each district and taking the average. It is observed that the northern part of Malawi has higher mortality rates followed by the southern region and finally the central region which has lower rates. However, all three cities namely Mzuzu, Lilongwe and Blantyre from northern, central and southern regions respectively have the lowest rates of under-five mortality, (< 0.02). As observed, the district level estimates of mortality mask the possible heterogeneities that are likely to be present at the sub district level due to differences in risk factors.

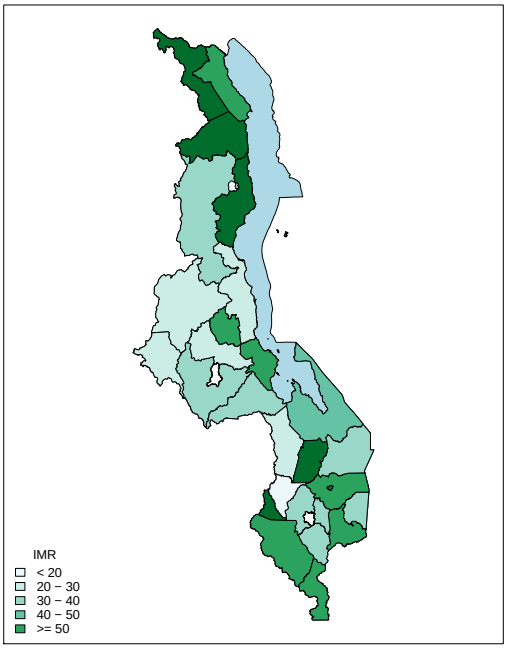

Figure 4.3: Crude under-five mortality rate per district in Malawi.

4.3 Association between under-five mortality and covariates

Table 4.1 presents baseline characteristics for the whole country. The mean age of mothers included in the study was 28.03 years and the mean birth weight of the children included in the survey was around 3.2kgs. These statistics were done before grouping the variables into discrete categories. A higher proportion of deaths was observed in rural areas (2.5%) compared to urban areas, (2.2%). There was an observed linear relationship between the level of education and proportion of children that had died with the highest proportion of mortality among mothers with primary education followed by mothers with secondary education then by those with higher education, (2.5%, 2.4% and 2.1% respectively). However, a slightly lower proportion of deaths was observed among mothers with no education at all (2.2%) compared to those with primary and secondary education.

A Chi-Squared test showed that mother's age and weight of a child at birth were associated with under-five mortality, ($p < 0.01$).

Table 4.1: Baseline characteristics of mothers with under-five children in Malawi as of 2015

	Outcome			
	Alive,N(%)	Dead,N(%)	Total,N(%)	p-value
N	11394(97.6)	286(2.4)	11680(100)	
Mother's age(yrs)				< 0.001
15-19years	926 (96.3)	36 (3.7)	962 (100)	
20-24years	3281 (98.1)	62 (1.9)	3343 (100)	
25-29years	2749 (97.9)	58 (2.1)	2807 (100)	
30-34years	2278 (97.9)	50 (2.1)	2328 (100)	
35-39years	1398 (96.9)	45 (3.1)	1443 (100)	
40-44years	565 (96.4)	21 (3.6)	586 (100)	
45-49years	197 (93.4)	14 (6.6)	211 (100)	
Birthweight(kilograms)				< 0.001
Low birthweight	1260 (96.1)	51 (3.9)	1311 (100)	
Normal birthweight	10134 (97.7)	235 (2.3)	10369 (100)	
Sex				0.127
Male	5620 (97.8)	128 (2.2)	5748 (100)	
Female	5774 (97.3)	158 (2.7)	5932 (100)	
Residence				0.356
Rural	9277 (97.5)	239 (2.5)	9516 (100)	
Urban	2117 (97.8)	47 (2.2)	2164 (100)	
Mother's education				< 0.873
None	1180 (97.8)	26 (2.2)	1206 (100)	
Primary	7340 (97.5)	189 (2.5)	7529 (100)	
Secondary	2637 (97.6)	66 (2.4)	2703 (100)	
Higher	237 (97.9)	5 (2.1)	242 (100)	
Wealth index				0.698
Lowest	2285 (97.7)	54 (2.3)	2339 (100))	
Second	2309 (97.3)	63 (2.7)	2372 (100)	
Middle	2182 (97.9)	47 (2.1)	2229 (100)	
Fourth	2218 (97.3)	61 (2.7)	2279 (100))	
Highest	2400 (97.5)	61 (2.5)	2461 (100)	

4.4 Model Results

Table 4.2 presents posterior means (given as odds ratios:OR) of a model with fixed effects from both DHS and census covariates and the corresponding 95% credible intervals (CI) for the GLGM. The odds of dying are not different for children from mothers with different ages, (OR=1.02). A unit increase in risk of malaria across a particular region.decreases the odds of children dying in such areas by one percent (OR=0.99).

The odds of dying for children born with normal weight at birth are 42% lower compared to children with a low birthweight, (OR=0.58) and the odds of female children dying are 15% lower compared to male children (OR=0.85). Children residing in urban areas have 46% lower odds of dying compared to their counteroarts in rural areas, (OR= 0.64). There is almost no difference in odds of children dying among mothers with secondary education and mothers with lower education (OR=1.01).

Children from highly educated mothers have 17% lower odds of dying compared to children from mothers with lower education. Children from richest households have 1.17 times higher odds of dying compared to children from poor households.and the odds of dying for children from middle to do households are 15% lower compared to children from poor households.

Table 4.2: Posterior estimates of model with fixed effects from both DHS and census data sets

Variable	OR	2.5% Quantile	97.5% Quintile
Intercept	0	0	0
Mother's age	1.02	1.01	1.04
Birth weight			
Normal weight	0.58	0.43	0.79
Sex of child			
Female	0.85	0.67	1.07
Residence			
Urban	0.64	0.43	0.96
Mother's education			
Secondary	1.01	0.74	1.39
Higher	0.83	0.32	2.11
Wealth index			
Middle	0.85	0.61	1.20
Fourth	1.06	0.76	1.48
Highest	1.17	0.76	1.75
Vulnerability	1.03	1.01	1.04
Malaria risk	0.99	0.98	1.00

4.4.1 Predicted risk of under-five mortality

Figure 4.4 shows the predicted mortality risk map. The Northern areas in general have a higher risk of under-five mortality. This is in agreement with the observed crude mortality map which showed high rates in the Northern areas. Lower risks are observed in Central and Southern areas. However, Nsanje, Neno, Chikwawa and Mwanza districts from the South Western region have higher risks of under-five mortality especially in areas close to borders. Within each district, the risk of under-five mortality is indeed varying as it was asserted at the beginning of the study.For instance, in Chikhwawa district, areas close

to the western border have a higher risk of under-five mortality compared to areas within the same dictrict but close to Blantyre and Thyolo districts.

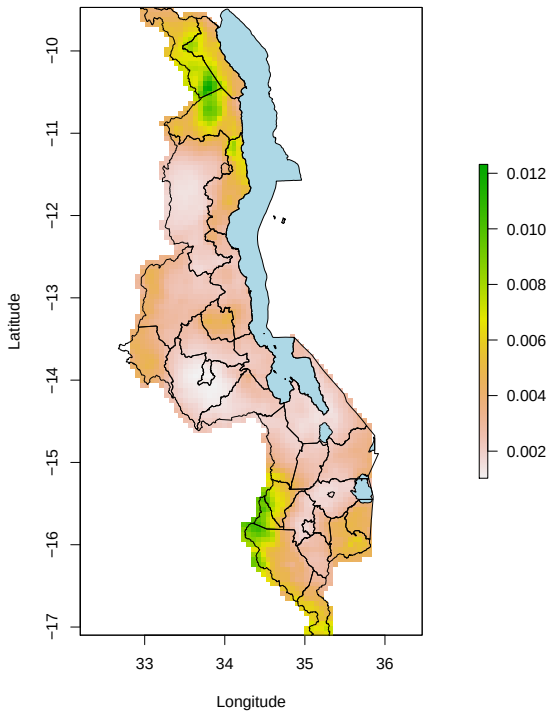

Figure 4.4: Map showing predicted risk of under-five mortality in Malawi.

Figure 4.5 shows there are relatively higher error values in the Northern part of Malawi as well as in border districts shown by greener colours as compared to the central and southern areas. This observation coincides with the sampling density in these areas. The general observation is that areas closest to the sampled locations have lower values of standard errors than those areas that are far. Areas around Lilongwe and Blantyre fall within the same standard error owing to the relatively large number of data points available for model estimation.

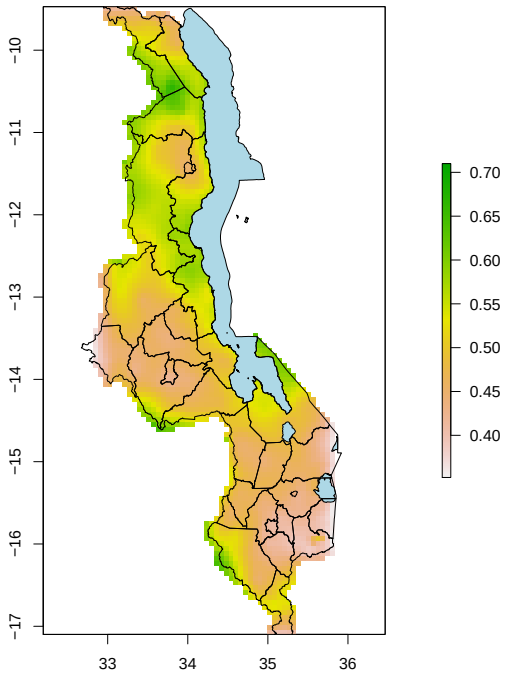

Figure 4.5: Map showing standard errors which are useful to quantify map precision.

CHAPTER FIVE
DISCUSSION AND RECOMMENDATIONS

5.1 Discussion

This chapter presents a discussion of the major findings of this study.

The study set out to combine multiple sources of data to model under-five mortality at the sub-district level in Malawi.

The results showed that the geospatial techniques employed identified hotspots of relatively high under-five mortality in Malawi. This was made possible with the combination of multiple data sources with different spatial resolutions. These results showed that combining datasets yields robust estimates at high spatial resolution and reveals heteregeneities within the districts. The predicted risk map is useful for focussed interventions, for example, it can advise areas to be targeted such as areas with higher risk of under-five mortality. In addition, since the estimates are at a local level, the estimates may provide a base against which intervention programmes may be assessed through follow-up surveys.

The bayesian hierarchical modelling approach performs better when using different sources of information by borrowing information, covariates within and between different time periods and it allows modelling of survey estimates, underfive mortality in our case, and underlying processes at different levels where data are sparse. In addition, the approach has enabled combining data from different sources that vary in their level of bias and timing and thus has informed more accurate, local area burden maps and this is necessary for improved risk stratification of high burden areas and identification of hot spots

The results showed that under-five mortality was strongly associated with mother's age and child's birth weight. These findings are in line with Ntenda's 2016 study which also found that mother's age and child's birth weight are positively associated with infant mortality, (Ntenda et al., 2014).

It was found that the odds of under-five children dying were higher for rural residents as compared to urban residents. This could be as a result of lack and not following proper disease prevention strategies attributed to poor settings such as not going to hospitals when a child is sick, not sleeping under bednets to prevent malaria, malnourishment resulting from shortage of food and balanced meals. This finding is similar to Ntenda's who found that the risk of children dying was higher for rural respondents compared to their counterparts in Malawi, (Ntenda et al., 2014). It is believed the urban/rural mortality differentials are attributed to various socioeconomic differences that exist within the country. In addition, factors such as better education, more public infrastructure that provides sanitation services, safer water supply, better systems to handle household waste and excreta removal, and easier access to healthcare services that are more favorable in urban than in rural areas can also explain this relationship.(Titaley et al., 2008; Hosseinpoor et al., 2005; Mekonnen et al., 2013).

Mortality is also likely to be driven by malaria as it one of the leading killers of children. Malaria risk was found to significantly contribute to underfive mortality with a one percent higher odds. This is true as it is evident in several studies that malaria is one of the leading causes of under-five mortality in Malawi, (Chilanga et al., 2020). Results showed that the risk of malaria was high in the Southern region and Lakeshore areas. This is the case probably because these areas are hot and coupled with high mosquito prevalence. The lower risk of malaria in the Northern region could be as a result of lower populations and therefore an advantage towards malaria initiatives. On the other hand, the lower risk in Chikwawa and Nsanje despite being hot areas

and providing a conducive environment for mosquitoes, could be attributed to the fact that these areas are hot spots for a number of NGOs involved in distribution of ITN which target pregnant mothers and children.

Similarly, the odds of dying are higher for vulnerable children as compared to less vulnerable children. For example, households experiencing economic instability are more vulnerable to spread of diseases and collapse of their health care systems as well as poor health conditions, (Kalipeni, 2000).

High vulnerability in most parts of Malawi is likely due to the fact that many Malawians are clustered close to the poverty line and due to shocks such as droughts, floods and fluctuations in food prices, (Devereux et al., 2006). Another reason could be due to a large proportion of Malawians relying on agriculture and thus erratic rainfall, landholding inequalities, constrained access to farm inputs and limited diversification and weak markets causing an increase in agricultural vulnerability, (Devereux et al., 2006).

5.2 Recommendations

This study recommends the integration of data from different sources with different resolutions as this helps in obtaining precise estimates of different population parameters. In addition, this helps to obtain estimates even in areas where data was not collected through prediction. Another recommendation is that findings from surveys such as the DHS should be used along with maps of risk prediction as this will enable effective allocation of resources. Also, the risk maps should be used in several studies as they help in strengthening survey findings. Furthermore, the risk maps should be updated regularly when new data become accessible.

One of the main major limitations of this study was the use of census and DHS data collected at different time points. We propose projection of DHS data to match the years of which census collects their data. In addition, the 2015/16

DHS data set is not very recent. However, the comprehensive nature of DH surveys make them quite useful for the period between successive surveys. The high-resolution population estimates from Worldpop are modelled estimates and not observed population values.

References

Alexander, M., & Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research, 38* (15), 335–372. https://doi.org/10.4054/DemRes.2018.38 .15

Arab, A., Hooten, M., & Wikle, C. (2007). Hierarchical spatial models. *Encyclopedia of Geographical Information Science*, 1–10. https:// doi.org/10.1007/978-0-387-35973-1 564

Baffour, B., Haynes, M., Western, M., Pennay, D., Misson, S., & Martinez, A. (2016). Weighting strategies for combining data from dual-frame telephone surveys: Emerging evidence from Australia. *Journal of Official Statistics, 32* (3), 549–578. https://doi.org/10.1515/JOS-2016 -0029

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Eckhoff, P. A., Wenger, E. A., Brie, O., Griffin, J. T., Fergus, C. A., . . . Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, *526*(7572), 207–211. https://doi.org/10.1038/nature15535

Biemer, P. P. (1984). *Methodology for optimal dual frame sample design* (Working Paper No. CENSUSI/SRD/RR-84/07). Washington D.C.: U.S. Bureau of the Census.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I., & Hughes, C. (2014). Reaching the hard-to-reach: A systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology, 14* (1). https://doi.org/10.1186/1471-2288-14-42

Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly*, *75* (5), 872–888. https://doi.org/10.1093/poq/nfr045

Brown, G. (2017). Methodology. In C. B. of Statistics (Ed.), *Statistical data warehouse design manual* (chap. 4). Netherlands: Excellence Center: On Micro Data Linking and Data Warehousing.

Chilanga, E., Collin-Vézina, D., MacIntosh, H., Mitchell, C., & Cherney, K. (2020). Prevalence and determinants of malaria infection among children of local farmers in Central Malawi. *Malaria Journal*, *19*(1), 1–10. https://doi.org/10.1186/s12936-020-03382-7

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution and duplicate detection*. Springer Science+Business Media. https://doi.org/10.1007/978-3-642-31164-2

Chromy, J. R., & Wilson, D. (2013). *Multiple Frame Approaches to Identify and Survey Victims of Rape and Sexual Assault*. Washington DC: RTI International. https://doi.org/10.1017/CBO9781107415324.004

Devereux, S., Baulch, B., Macauslan, I., Phiri, A., & Sabates-Wheeler, R. (2006). Vulnerability and social protection in Malawi. *IDS DiscussionPaper* (387).

Diggle, P. J., & Ribeiro, P. J. J. (2007). *Model-based Geostatistics*. SpringerScience+Business Media,LLC. https://doi.org/10.1007/978-0-387-98135-2

D'Orazio, M. (2011). Statistical matching and imputation of survey data with StatMatch. *Italian National Institute of Statistics*, 1–40.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. https://doi.org/10.1002/0470023554

Dusetzina, S., Tyree, S., Meyer, A., Meyer, A., Green, L., & Carpenter, W. (2014). *Linking data for health services: A framework and instructional guide [Internet]* (No. 14-EHC033-EF). Rockville (MD): Agency for Healthcare Research and Quality (US).

Elliott, M. R., Raghunathan, T. E., & Schenker, N. (2018). Combining estimates from multiple surveys. *Wiley StatsRef: Statistics Reference Online*, 1–10. https://doi.org/10.1002/9781118445112.stat08079

Filippi, S., & Holmes, C. C. (2017, December). A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Analysis, 12* (4), 919-938. https://doi.org/10.1214/16-BA1027

Finucane, M. M., Paciorek, C. J., Danaei, G., & Ezzati, M. (2014). Bayesian estimation of population-level trends in measures of health status. *Statistical Science*, *29* (1), 18–25. https://doi.org/10.1214/13-STS427

Fung, B., Wang, K., Fu, A., & Yu, P. (2010). *Introduction to privacy-preserving data publishing: Concepts and techniques*. https://doi.org/10.1201/ 9781420091502

Gelfand, A. E. (2012). Hierarchical modelling for spatial data problems. *Spatial Statistics,1*(4), 30–39. https://doi.org/10.1016/j.spasta.2012.02.005

Gelfand, A. E., & Banerjee, S. (2017). Bayesian modelling and analysis of geostatistical data. *Annual Review of Statistics and Its Application*,*4* (1), 245–266. https://doi.org/10.1146/annurev-statistics-060116-054155

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, *93*(443), 846–857. http://www.jstor.org/ stable/2669819

Gotway, C. A., & Hartford, A. H. (1996). Geostatistical methods for incorporating auxiliary information in the prediction of spatial variables. *Journal of Agricultural, Biological, and Environmental Statistics*, *1* (1), 17–39. https://doi.org/10.2307/1400558

Gutreuter, S., Igumbor, E., Wabiri, N., Desai, M., & Durand, L. (2019). Improving estimates of district HIV prevalence and burden in South Africa using small area estimation techniques. *PLOS ONE*, *14*(2), 1–14. https://doi.org/10.1371/journal.pone.0212445

Harrison, S., Alderdice, F., Henderson, J., Redshaw, M., & Quigley, M. A. (2020). Trends in response rates and respondent characteristics in five National Maternity Surveys in England during 1995 – 2018. *BMC Public Health*, *78*(1), 1–11. https://doi.org/10.1186/s13690-020-00427-

He, Y., Landrum, M. B., & Zaslavsky, A. M. (2014). Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: A multiple imputation approach. *Statistics in Medicine*, *33*(21), 3710–3724. https://doi.org/10.1002/ sim.6173

Held, L., & Sabanés-Bové, D. (2014). *Applied statistical inference: Likelihood and bayes*. https://doi.org/10.1007/978-3-642-37887-4

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. https://doi.org/10.1007/0-387-69505-2

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2007). *Advances in telephone survey methodology*. https://doi.org/10.1002/9780470173404.ch23

Hosseinpoor, A. R., Mohammad, K., Majdzadeh, R., Naghavi, M., Abolhassani, F., Sousa, A., Speybroeck, N., Jamshidi, H. R., & Vega, J. (2005). Socioeconomic inequality in infant mortality in Iran and across its provinces. *Bulletin of the World Health Organization*, *83* (11), 837–844. https://doi.org/10.1590/ S0042-96862005001100013

Islam, S., & Chandra, H. (2019). Small domain inference combining data from two independent surveys. *Indian Society of Agricultural Statistics*, *73*(1), 59–69.

Islam, S., Chandra, H., Aditya, K., & Lal, S. B. (2018). Small area estimation under a spatial model using data from two surveys. *International Journal of Agricultural and Statistical Sciences, 14* (1), 231–237.

Johnson, H. L., Liu, L., Fischer-Walker, C., & Black, R. E. (2010). Estimating the distribution of causes of death among children age 1–59 months in high-mortality countries with incomplete death certification. *International Journal of Epidemiology, 39*(4), 1103–1114. https://doi.org/10.1093/ije/dyq074

Kabudula, C. W., Clark, B. D., Gómez-olivé, F. X., Tollman, S., Menken, J., & Reniers, G. (2014). The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot studyfrom South Africa. *BMC Medical Research Methodology*, 1–10.

Kalipeni, E. (2000). Health and disease in southern Africa: A comparative and vulnerability perspective. *Social Science and Medicine, 50* (8), 965–983. https://doi.org/10.1016/S0277-9536(99)00348-2

Kiesl, H., & Rässler, S. (2006). *How valid can data fusion be?* Institute for Employment Research of the Federal Employment Services. https://doku.iab.de/discussionpapers/2006/dp1506.pdf

Kim, J. K., Wang, Z., Zhu, Z., & Cruze, N. B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level Model. *Journal of Agricultural, Biological, and Environmental Statistics, 23* (2), 175–189. https://doi.org/10.1007/s13253-018-0320-2

Leulescu, A., & Agafitei, M. (2013). *Statistical matching: a model based approach for data integration* (Tech. Rep.). Luxembourg: Publications Office of the European Union.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science, 32* (2), 293–312. https://doi.org/10.1214/16-STS584

Lu, B., Sahr, T. R., Weston, D., Iachan, R., & Duffy, T. P. (2013). Practical considerations in design and analysis of dual-frame telephone surveys for health policy research. *World Medical & Health Policy*. https:// doi.org/10.1002/wmh3.53

Machado, C. J. (2004). A literature review of record linkage procedures focusing on infant health outcomes. *Cadernos de saude publica, 20* (2), 362–371. https://doi.org/10.1590/S0102-311X2004000200003

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., & Thompson, S. G.(2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society, 174* (1), 31–50. https://doi.org/10.1111/j.1467-985X.2010.00648.x

McMillen, R. C., Winickoff, J. P., Wilson, K., Tanski, S., & Klein, J. D. (2015). A dual-frame sampling methodology to address landline replacement in tobacco control research. *BMC, 24*(1), 7–10. https://doi.org/10.1136/ tobaccocontrol-2012-050727

McNeeley, S. (2012). Sensitive issues in surveys: Reducing refusals while increasing reliability and quality of responses to sensitive survey items. In *Handbook of survey methodology for the social sciences* (p. 377-396). Springer. https://doi.org/10.1007/978-1-4614-3876-2 22

McVeigh, B. S., Spahn, B. T., & Murray, J. S. (2019). *Scaling bayesian probabilistic record linkage with post-hoc blocking: An application to the california great registers.* arXiv:Cornell University. https://arxiv .org/abs/1905.05337 10.48550/ARXIV.1905.05337

Mecatti, F., & Singh, A. C. (2014). Estimation in multiple frame surveys: A simplified and unified review using the multiplicity approach. *Journal de la Société Française de Statistique & revue de statistique appliquée, 155* (4), 51–69. http://www.sfds.asso.fr/journal

Mekonnen, Y., Tensou, B., Telake, D. S., Degefie, T., & Bekele, A. (2013). Neonatal mortality in Ethiopia: trends and determinants. *BMC Public Health, 13* (483). https://doi.org/10.1186/1471-2458-13-483

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society, 72* (1), 27–48. https://www.jstor.org/stable/40541573

National Academies of Sciences, Engineering and Medicine. (2017). *Federal statistics, multiple data sources, and privacy protection: Next steps*. Washington DC: The National Academies Press. https://doi.org/10.17226/24893

National Statistical Office (NSO) [Malawi] and ICF. (2017). *Malawi Demographic and Health Survey 2015-16.* Zomba, Malawi, and Rockville, Maryland, USA. NSO and ICF. http://dhsprogram.com/pubs/pdf/FR319/FR319.pdf

Ntenda, P. A. M., Chuang, K. Y., Tiruneh, F. N., & Chuang, Y. C. (2014). Factors associated with infant mortality in Malawi. *Journal of Experimental and Clinical Medicine, 6*(4), 125–131. https://doi.org/10.1016/j.jecm.2014.06.005

Ovuga, E., & Madrama, C. (2006). Burden of alcohol use in the Uganda Police in Kampala District. *African Health Sciences, 6*(1), 14–20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1831968

Poulin, M. (2010). Reporting on first sexual experience: The importance of interviewer-respondent interaction. *Demographic Research, 22*(11), 237–288. https://doi.org/10.4054/DemRes.2010.22.11

Pridemore, W. A., Damphousse, K. R., & Moore, R. K. (2005). Obtaining sensitive information from a wary population: A comparison of telephone and face-to-face surveys of welfare recipients in the United States. *Social Science & Medicine, 61*(5), 976–984. https://doi.org/10.1016/j.socscimed.2005.01.006

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. http://www.r-project.org/index.html

Radner, D. B. (1980). *An example of the use of statistical matching in the estimation and analysis of the size distribution of income* (No. 18). Division of Economic Research. https://www.ssa.gov/policy/docs/workingpapers/wp18.pdf

Rao, J., & Lohr, S. L. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association, 101*(475), 1019–1030. https:// doi.org/10.1198/016214506000000195

Rao, J., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118735855

Rose, A. N. (2015). *Data fusion methods for improved demographic resolution of population distribution datasets.* University of Tennessee. https://trace.tennessee.edu/utk_graddiss/3358

Scanu, M. (2014). Data integration: differences between record linkage and statistical matching. *Statistical Science, 4* (2), 14-19.

Shlomo, N. (2019). Overview of data linkage methods for policy design and evaluation: How access to microdata is transforming policy design. In C. Nuno & P. Paolo (Eds.), *Data-driven policy impact evaluation* (pp.47–65). Springer.

Snow, R. W., Sartorius, B., Kyalo, D., Maina, J., Amratia, P., Mundia, C. W., Bejon, P., & Noor, A. M. (2017). The prevalence of Plasmodium falciparum in sub Saharan Africa since 1900. *Nature, 550*(7677), 515–518. https:// doi.org/10.1038/nature24059

Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis* (Tech. Rep. No. 4). Boston University, School of Public Health. https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ, 338* (b2393). https://doi.org/10.1136/bmj.b2393

Taherdoost, H. (2018). Sampling methods in research methodology; How to choose a sampling technique for research. *International Journal of Academic Research in Management*, *5*(2), 18–27. https://doi.org/ 10.2139/ssrn.3205035

Titaley, C. R., Dibley, M. J., Agho, K., Roberts, C. L., & Hall, J. (2008). Determinants of neonatal mortality in Indonesia. *BMC Public Health*, *8* (232), 1–15. https://doi.org/10.1186/1471-2458-8-232

Waal, T. D. (2015). *Statistical matching: Experimental results and future research questions* (Tech. Rep.). Statistics Netherlands. https://doi: 10.13140/RG.2.1.1969.4161

Wang, J. C., Holan, S. H., Nandram, B., Barboza, W., Toto, C., & Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, *17* (1), 84–106. https://doi.org/10.1007/ s13253-011-0067-5

Westoff, C. F., Bietsch, K., & Koffman, D. (2013). *Indicators of trends in fertility in sub-Saharan Africa:DHS Analytical Surveys No. 34* .United States Agency for International Development. https://www .dhsprogram.com/pubs/pdf/AS34/AS34.pdf

Winglee, M., Valliant, R., & Scheuren, F. (2005). A case study in record linkage. *Survey Methodology*, *31*(1), 3–11. https://www.researchgate.net/ publication/265595065 A Case Study in Record Linkage

Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6* (5), 1–29. https://doi.org/10 .1002/wics.1317

Zarnoch, S. J., Cordell, H. K., Betz, C. J., & Bergstrom, J. C. (2010). Multiple imputation: An application to income nonresponse in the national survey on recreation and the environment. *United States Department of Agriculture, Forest Service*, *49*(3), 1–15. http://www.srs.fs.usda.gov/trends/{%}0Ahttps://www.fs.usda.gov/treesearch/pubs/35141

Zehang, L., Yuan, H., Godwin, J., Martin, B. D., Wakefield, J., & Clark, S. J. (2019). Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLOS ONE*, *14* (1), 1–17. https://doi.org/10.1371/journal.pone.0210645

Appendix

R Code

```
library(haven)
library(tidyverse)
library(sf)
library(raster)
library(rasterVis)
library(rgdal)
library(maptools)
library(RColorBrewer)
library(rgeos)
library(geostatsp)
library(geoR)
library(arsenal)

mwdistricts <- read.csv("data/mwdistricts.csv",
                        header = T,stringsAsFactors = F)
mergedatanew <- read.csv("data/mergeddata1.csv",
header = TRUE, stringsAsFactors = FALSE)

dat <- read.csv("data/merged_data.csv",header=TRUE,
          stringsAsFactors = FALSE)
pop <- raster("data/MWI_ppp_2015_adj_v2.tif")
clusters <- readOGR("data/Geographical data","MWGE7AFL")
mw.districts <- readOGR("data/Malawi districts","District")
lakes <- readOGR("data/Malawi districts","lake")
```

```
census <- read.csv("data/Censusdata.csv",header = T],
              stringsAsFactors = F)

# manage and change projections
mw.districts <- spTransform(mw.districts,
                            CRS("+init=epsg:4326")
)# set projection for lakes
proj4string(lakes) <- CRS("+init=epsg:32736")

lakesSP <- spTransform(lakes,CRS("+init=epsg:4326"))

# convert spatialpointdataframe to usual dataframe
clusterdf <- as.data.frame(clusters)

# reduce DHS cluster data
clusterdf <- clusterdf[,c("DHSCLUST","ADM1NAME",
                  "LATNUM","LONGNUM", "ALT_GPS")]
                  %>%
rename(clusterID="DHSCLUST")

# subset the data
dat <- dat[,c("v001","v012","v025","v106","v137",
          "v155","v190","b4_01","b5_01","v201","v208",
          "m19_1","m15_1","hv111_01","hv201","hv113_01",
       "hv115_01")]

# rename variables

dat <- rename(dat,clusterID="v001",mothAge="v012",
```

```
                residence="v025", educ="v106",totaChildren="v137",
                literacy="v155",wealthIndex="v190,sex="b4_01",
                alive="b5_01",totalChildrenBorn="v201",totalBirths
                ="v208",birthWeight="m19_1",placeDelivery=
                "m15_1",motherAlive="hv111_01",waterSource=
                "hv201",fatherAlive="hv113_01",maritalStatus=
                "hv115_01")


# merge the two datasets
dat <- left_join(dat,clusterdf,by="clusterID")

# aggregate the population
pop <- aggregate(pop,fact=13,fun=sum,na.rm=TRUE) # converting to
       roughly 1km resolution
mypop <-  aggregate(pop,fact=20,fun=sum,na.rm=T)
# extract the populations
dat$clusterpop <-  ceiling(raster::
extract(pop,dat$LONGNUM,dat$LATNUM))

#  determine  district
plot(mw.districts)
plot(lakesSP,add=TRUE,col="lightblue
")
points(dat$LONGNUM,dat$LATNUM,pch=19,col=2,cex=0.5)

# check with cluster coordinates fall into which district
clusterCoords <- dat[,c("LONGNUM","LATNUM")]
# remove all NA coords
clusterCoords <- clusterCoords[complete.cases(clusterCoords),]
```

```
colnames(clusterCoords) <-  c("Longitude","Latitude")

# convert  points  to  spatial  points
clusterCoordsSP <- SpatialPoints(clusterCoords,proj4string
                    =CRS("+init=epsg:4326"
                    ))

U <- over(clusterCoordsSP,mw.districts) # which points fall  into
     which  district?

j <-
which(!is.na(U$OBJECTID))
finaldata <- dat[j,]
finaldata$district <-
U$DISTRICT

# merge  DHS  with  Census  data
completedata <-  left_join(finaldata,census,by="district")

# summarize  of  deaths  by  district
# add mid year populations to the data from the census datanumDeaths
<-  completedata  %>%
group_by(district) %>%
summarize(deaths=sum(totaChildren,na.rm  =  TRUE),
pop = sum(clusterpop,na.rm = TRUE))

# exploratory maps
mydata <- completedata
xx  =  as.character(mw.districts@data$DISTRICT)
mergedata <- left_join(numDeaths,mydata,by="district")
```

```
## exploratory  graphics
# plot raster to show population density at very fine resolution# overlay the gridded polygon over the Guangdong raster
range(dat$clusterpop,na.rm = T) # to determine the range
breaks <- seq(0,12, by=0.01)
cols <- colorRampPalette(rev(c("red", "yellow",
          "lightgreen")))(length(breaks))
pdf("images/mw_pop.pdf", width = (15*0.39), height = (20*0.39))
par(mfrow=c(1,1))
levelplot(log(pop), maxpixels = ncell(pop),
col.regions = cols,
at =  breaks,
panel = panel.levelplot.raster,
interpolate = TRUE,
colorkey  =  list(space="right"),
margin = FALSE) +layer(sp.points(clusterCoordsSP,col=hsv(0.5,0.5,0.5,
alpha=0.5),pch=16,cex=0.5)) + layer(sp.polygons(lakes, fill =
"lightbluelayer(sp.polygons(mw.districts))
dev.off()


# some data  managementG
<- numDeaths
G$DISTRICT <-
G$districtH <-
mw.districts
V <- left_join(H@data,G,by="DISTRICT")

# number of births
```

```
births <- c(10152,14199,11078,8713,36415,367,38349,37304,
13841,34889,19168,66478,27723,31931,24502,43979,
25929,27702,12717,16123,26235,23765,15964, 22748,
 13195,17321,7330,41056,10233,31977,5883,4815)

# plot estimates of the child mortality rate
mw.districts$infantDeaths <- V$deaths
mw.districts$births <- births # births extracted from NSO projections
mw.districts$IMR <- mw.districts$infantDeaths/mw.districts$births*1000brks
<- c(10,20,30,40,50,60)
cols <- brewer.pal(6,"BuGn")
pdf("images/IMR.pdf",width = (20*0.39),height = (25*0.39))
par(mfrow=c(1,1),mar=c(2,2,2,2))
plot(mw.districts)
plot(lakes,add=TRUE,col="lightblue")
plot(mw.districts,col=cols[findInterval(mw.districts$IMR,
brks)],add=TRUE)
legend("bottomleft",
legend =
leglabs(brks,"<",">="),fill =
cols,
bty = "n",
cex = 1.1,
y.intersp = 0.9,
title = "IMR")
box(lwd=1,bty="o"
)dev.off()

# plot estimates of the child mortality rate
```

```
mw.districts$U5Deaths <- V$deaths
mw.districts$births <- births # births extracted from
NSO projections
mw.districts$UMR <- mw.districts$U5Deaths/mw.districts$births*1000
brks <- c(10,20,30,40,50,60)
cols <- brewer.pal(6,"BuGn")
pdf("images/UMR.pdf",width = (20*0.39),height = (25*0.39))
par(mfrow=c(1,1),mar=c(2,2,2,2))
plot(mw.districts)
plot(lakes,add=TRUE,col="lightblue")
plot(mw.districts,col=cols[findInterval(mw.districts$UMR,
brks)],add=TRUE)
legend("bottomleft",
legend =
leglabs(brks,"<",">="),fill =
cols,
bty = "n",
cex = 1.1,
y.intersp = 0.9,
title = "UMR")
box(lwd=1,bty="o"
)dev.off()

# exploratory models fitting
# exporatory GLMs
# geostatistical models

fit <- glm(alive ~ mothAge + factor(residence) + factor(maritalStatus),
family = binomial(link="logit"),data = mergedata)
```

```
#reading raster files
vulnerability <- raster("data/malawi_national_vulnerability_index.tif")
malrisk <- raster("data/malaria_risk.tif")

#Descriptives_DHS
dataview(completedata)


covariatelist <- list(vulnerabilitymw = vulnerability,
malriskmw = malrisk)

mergedatamw <- read.csv("data/mergeddata1.csv",header = TRUE,
      stringsAsFactors = FALSE)
mergedatamw <- read.csv("data/mergedatamw.csv",header=TRUE,
      stringsAsFactors = FALSE)

mergedatamwspdf <- SpatialPointsDataFrame(mergedatamw, coords=mergedatam
proj4string =CRS("+init=epsg:4326"))

mergeddatanewspdf <- SpatialPointsDataFrame(mergedatanew, coords=mergeda
proj4string= CRS("+init=epsg:4326"))


mwnewgeofit <- glgm(formula = IM ~ mothAge + birthWeight + sex +
  residence + educ + wealthIndex + vulnerabilitymw + malriskmw,
```

```
data = mergeddatanewspdf,
grid = 50,
covariates = covariatelist,
family = "binomial",
Ntrials = mergeddatanewspdf$Nwomen,verbose
= TRUE,
shape = 1,
buffer = 0,
priorCI = list(sd = c(0.1,3), range = c(0.5, 2)),
control.inla=list(strategy = "gaussian"))

View(mwnewgeofit[["inla"]][["summary.fixed"]])
plot(mwnewgeofit$raster[["predict.invlogit"]])
plot(mwnewgeofit$raster[["random.mean"]])
write.csv(mwnewgeofit[["inla"]][["summary.fixed"]],
"C:\\Users\\HP User\\Desktop\\MWnewgeofit.csv", row.names = TRUE)
```