Statistical Modelling of Clustered and Incomplete Data with Applications in Population Health Studies in Developing Countries

Oyelola Abdulwasiu Adegboye

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Statistics in the Department of Statistics and Population Studies, Faculty of Natural Sciences, University of the Western Cape

Supervised by Prof. Danelle Kotze

February 20, 2014

”Acquire knowledge; it enables the possessor to distinguish right from wrong; it guides us to happiness; it sustains us in misery; it is an ornament among friends, and an armour against enemies *The Prophet Muhammad (pbuh)*.”

To:

Dupe,

a faithful wife and friend;

Faiza, Fahd, Fahima and Farida,

apples of my eyes

Abstract

The United Nations (UN) Millennium Development Goals (MDGs) drafted eight goals to be achieved by the year 2015, namely: eradicating extreme poverty and hunger, achieving universal primary education, promoting gender equality and women empowerment, reducing child mortality, improving maternal health, combating HIV/AIDS, malaria and other diseases, ensuring environmental sustainability and lastly developing a global partnership for development. Many public health studies often result in complicated and complex data sets, the nature of these data sets could be clustered, multivariate, longitudinal, hierarchical, spatial, temporal or spatio-temporal. This often results in what is called correlated data, because the assumption of independence among observations may not be appropriate. The shared genetic traits in the studies of illness or shared household characteristics among family members in the studies of poverty are examples of correlated data. In cross-sectional studies, individuals may be nested within sub-clusters (e.g., families) that are nested within clusters (e.g., environment), thus causing correlation within clusters. Ignoring the structure of the data may result in asymptotically biased parameter estimates. Clustered data may also be a result of geographical location or time (spatial and temporal). A crucial step in modelling correlated data is the specification of the dependency by choosing the covariance/correlation function. However, often the choice for a particular application is unclear and diagnostic tests will have to be carried out, following fitting of a model. This study's view of developing countries investigates the prospects of achieving MDGs through the development of flexible predictor statistical

models.

The first objective of this study is to explore the existing methods for modelling correlated data sets (hierarchical, multilevel and spatial) and then apply the methods in a novel way to several data sets addressing the underlying MDGs.

One of the most challenging issue in spatial or spatio-temporal analysis is the choice of a valid and yet flexible correlation (covariance) structure. In cases of high dimensionality of the data, where the number of spatial locations or time points that produced the observations is large, the analysis of such data presents great computational challenges. It is debatable whether some of the classical correlation structures adequately reflect the dependency in the data.

The second objective is to propose a new flexible technique for handling spatial, temporal and spatio-temporal correlations. The goal of this study is to resolve the dependencies problems by proposing a more robust method for modelling spatial correlation. The techniques are used for different correlation structures and then combined to form the resulting estimating equations using the platform of the Generalized Method of Moments. The proposed model will therefore be built on a foundation of the Generalized Estimating Equations; this has the advantage of producing consistent regression parameter estimates under mild conditions due to separation of the processes of estimating the regression parameters from the modelling of the correlation. These estimates of the regression parameters are consistent under mild conditions.

Thirdly, to account for spatio-temporal correlation in data sets, a method that decouples the two sources of correlations is proposed. Specifically, the spatial and temporal effects were modelled separately and then combined optimally. The approach circumvents the need of inverting the full covariance matrix and simplifies the modelling of complex relationships such as anisotropy, which is known to be extremely difficult or impossible to model in analyzing large spatio-temporal data.

Lastly, large public health data sets consist of a high degree of zero counts where it is very difficult to distinguish between ”true zeros” and ”imputed” zeros. This can be due to the reporting mechanism as a result of insecurity, technical and logistics issues. The focus is therefore on the implementation of a technique that is capable of handling such a problem. The study will make the assumption that ”imputed” zeros are a random event and consider the option of discarding the zeros, and then model a conditional Poisson model, conditioning on all cases greater than 0.

Declaration

I hereby declare that ”Statistical Modelling of Clustered and Incomplete Data with Applications in Population Health Studies in Developing Countries” is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Signed: Oyelola A Adegboye February 20, 2014

Acknowledgements

Enlightenment is man's emergence from his self-imposed immaturity. Immaturity is the inability to use one's understanding without guidance from another. This immaturity is self-imposed when its cause lies not in lack of understanding, but in lack of resolve and courage to use it without guidance from another (*Immanuel Kant, 1784*).

The completion of this thesis could not have been possible without the help and support of a number of people. First and foremost, I would like to express my sincere appreciation to my supervisor, Prof. Danelle Kotze. Her thorough guidance, scientific insight, unwavering support and extreme patience over the last four years is astounding. She has been a great mentor and I have truly enjoyed collaborating with her.

I have also been fortunate to have met Prof. Denis H-Y Leung; he has been very generous with several discussions on generalized equations and R codes. I would also like to extend my sincere appreciation to other collaborators and co-authors: Prof. You-Gan Wang, Dr. Olasunkanni Rasheed Adegboye, as well as all other colleagues who have contributed to my PhD pursuit in some way or the other. Special thanks also goes to my friend, Dr. Timor Saffary for all his words of encouragement and his insightful discussions; to Siaka and Philomene for making each stay at UWC less boring and for their constant logistical support. I am grateful for generous access to various sources of data, in particular; Measure Demographic and Health Surveys, National Oceanic and Atmospheric Administration, and World Organization for Animal Health. I would like to extend my profound gratitude to all lecturers and non-teaching staff in the department, most especially Mr. Leslie Selbourne for his efforts, assistance on all administrative and procedural matters.

Finally to the most important people in my life; my wife Rashidat, thank you for your love and support, and for running the show while I was away all these years; to my kids, the fantastic four (Faizat, Fahd, Faridat and Fahimat) for all those long years without my presence; special appreciation and thank you to my parents, Prof. and Mrs. A. O. Adegboye, for their confidence, encouragement and source of inspiration for my doctoral studies.

The majority of the parts in this thesis is based on original publication in various journals as listed below:

Chapter 3: Adegboye, O.A. (2010). Under-five Mortality in Nigeria: Spatial Exploration and Spatial Scan Statistics for Cluster Detection. *International Journal of Statistics and Systems.* 5(2), 203-214.

Chapter 4: Adegboye, O.A. and Kotze, D. (2012). Disease Mapping of Leishmaniasis Outbreak in Afghanistan: Spatial Hierarchical Bayesian Analysis. *Asian Pacific Journal of Tropical Disease*, 2(4), 253-259. `http://dx.doi.org/10.1016/S2222-1808(12)60056-5`

Chapter 5: Adegboye, O.A. and Kotze, D. (2013). An Exploratory Look at Associated Factors of Poverty on Educational Attainment in Africa and In-depth Multi-level Modelling for Namibia. *Journal of Studies in Economics and Econometrics*, 37(1).

Chapter 6: Adegboye, O.A., Kotze, D., and Adegboye, O.A. (2013). Multi-year Trend Analysis of Childhood Immunization Uptake and Coverage in Nigeria. *Journal of Biosocial Sciences*, 45(4). doi:10.1017/S0021932013000254

Chapter 7: Adegboye, O.A. and Kotze, D. (2013). Epidemiological Analysis of Spatially Misaligned Data: A Case of Highly Pathogenic Avian Influenza Virus Outbreak in Nigeria. *Epidemiology and Infection.* doi:10.1017/S0950268813002136

Chapter 8: Adegboye, O.A. (2013). Causes and Patterns of Morbidity and Mortality in Afghanistan: Joint Estimation of Multiple Causes in the Neonatal period. *Canadian Studies in Population.*

Chapter 9: Adegboye, O.A., Kotze, D., Leung, D.H.Y. and Wang, Y.G. (2013) A Robust Method for Modelling Spatial Correlation: Analysis of Malaria Incidence in Afghanistan. *Submitted for publication*

Short versions of the chapters were also presented at international conferences as listed below:

1. Adegboye, O.A. and Kotze, D. (2013) Epidemiological Analysis of Causes and Patterns of Morbidity and Mortality in Afghanistan. *Proceedings of the 2013 Annual Meeting of the Population Association of America, New Orleans, LA. 11-13 April, 2013*

2. Adegboye, O.A., Kotze, D., Leung, D.H.Y. and Wang, Y.G. (2012) Spatial Analysis Of Malaria Incidence In Afghanistan Using Combined Estimating Equations. *Proceedings of the 12 Islamic Countries Conference on Statistical Science, Doha, Qatar. 19-22 December, 2012.*

3. Adegboye, O.A., Kotze, D., Leung, D.H.Y. and Wang, Y.G. (2012) Spatial Analysis of Malaria Incidence in Africa: An Efficient Parameter Estimation in the Premise of Generalized Estimating Equation. *Proceedings of 26th International Biometric Conference, Kobe, Japan. 26-31 August, 2012.*

4. Adegboye, O.A. and Kotze, D. (2012) Disease Mapping of Malaria Incidence in Sub-Sahara Africa. *Proceedings of the 26th International Biometric Conference, Kobe, Japan. 26-31 August, 2012.*

5. Adegboye, O.A. and Kotze, D. (2012) Evidence-based Evaluation of Immunization in Nigeria: Multi-year Trend Analysis. *Proceedings of the 1st Asia Pacific Clinical Epidemiology and Evidence-Based Conference, Kuala Lumpur, Malaysia. 6-8 July, 2012.*

6. Adegboye, O.A. (2011). Bayesian Spatial Analysis and Disease Mapping of Leishmaniasis Outbreak in Afghanistan, 2003-2009. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland. 21-24 August, 2011.*

7. Adegboye, O.A. and Kotze, D. (2011). Is Poverty a Determinant of Educational Attainment? Perspectives from Africa. *Proceedings of the Joint Statistical Meeting, Miami, USA. 30 July-4 August, 2011, pp. 3834-3848.*

8. Adegboye, O.A. (2011). Is Poverty a Determinant of Educational Attainment? Perspectives from Africa. *Presented at the Post-Graduate Seminar*, Department of Statistics, University of the Western Cape, South Africa. 20 July, 2011.

9. Adegboye, O.A. (2010). Spatio-temporal Analysis of Transmission of Avian Flu Outbreak in Nigeria. *Pproceedings of the 25th International Biometric Conference, UFSC, Brazil. 5-10 December, 2010.*

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
HIV	Human Immunodeficiency Virus
DHS	Measure Demographic Health Surveys
AIS	AIDS/HIV indicator Survey
MIS	Malaria Indicator Survey
MDG	Millennium Development Goals
UN	United Nations
UNESCO	United Nations Educational Scientific and Cultural Organization
UNAIDS	Joint United Nations Programme on HIV/AIDS
UNICEF	United Nations Children's Fund
BCG	Bacillus Calmette-Guerin
DPT	United Nations Children's Fund
NPC	(National Population Commission
WHO	World Health Organization
HPAI	Highly Pathogenic Avian Influenza
PAN	Poultry Association of Nigeria
SSA	Sub Sahara Africa
HMIS	Heath Management Information Systems
AMS	Afghanistan Mortality Survey
GEE	Generalized Estimating Equations
OIE	World Organization for Animal Health
ACL	Anthroponotic Cutaneous Leishmaniasis
ZCL	Zoonotic Cutaneous Leishmaniasis
SIR	Standard Incidence Rate
GLMM	Generalized Linear Mixed Models

DIC	Deviance Information Criterion
EPI	Expanded Program on Immunization
NPI	National Programme on Immunization
OPV	Oral Polio Vaccine
GLM	Generalized Linear Models
ALR	Alternating Logistic Regression
USAID	U.S. Agency for International Development
ICD	International Classification of Disease
CSO	Central Statistics Organization
APHI	Afghan Public Health Institute
MoPH	Ministry of Public Health
VA	Verbal Autopsies
AIC	Akaike Information Criteria
HSA	Habitat Suitability Analysis
HSI	Habitat Suitability Index
ENFA	Ecologic Niche Factor Analysis
H5N1	Influenza A Virus Subtype
SMR	Standard Mortality Rate
ITN	Insecticide-Treated mosquito Net
NOAA	National Oceanic and Atmospheric Administration
LISA	Local Indicators of Spatial Association
GMM	Generalized Method of Moments

Chapter 1

General Introduction

1.1 Population Health Studies

Acquisition and subsequent analysis of health outcomes and health indicators involve a large pool of information from the target population. Health data may be based on population-based health problems from health care providers, health agencies or individuals conducting research in this area. This may include studies to improve social, economic and environmental conditions, investigating the efficacy of a new vaccine or medication, or to understand the transmission of a new disease. In order to conduct applied and policy-relevant research on population health problems, adequate and large data sets must be available. Many population health research institutes and centres around the world provide harmonized data on key health problems. Measure Demographic Health Surveys (DHSs) collects a wide range of data on demographic, AIDS/HIV indicator Surveys (AISs), Malaria Indicator Surveys (MISs) and other key indicators around the globe. Population health data can also be obtained from a country's ministry of health, public health institutes, international organizations and the United Nations affiliated agencies. The United Nations (UN) Millennium Development Goals (MDGs) drafted eight goals to be achieved by the year 2015, namely: eradicating of extreme poverty and hunger, achieving universal primary education, promoting gender equality and women empowerment, reducing child mortality, improving maternal health, combating HIV/AIDS, malaria and other diseases, ensuring environmental sustainability and lastly, developing a global partnership for development. This research study draws on issues surrounding the statistical modelling of public health indicators in developing countries as well as the methodological aspects.

1.1.1 Correlated Data

In recent times, most population health studies involve complex surveys that result in complex data structures such as clustered, multivariate, longitudinal, hierarchical or spatial. Therefore, the assumption of independence among observations may not be appropriate; for instance, because of the shared genetic traits in studies of illness or shared household characteristics among family members in studies of poverty. In cross-sectional studies, individuals may be nested within sub-clusters (e.g., families) that are nested within clusters (e.g., environment), thus causing correlation within clusters. Clustered data may also be a result of geographical location or time (spatial and temporal). For example, links between climate and disease transmission or/and incidence sometimes lead to the mechanism underlying the disease epidemic to be spatial or temporal or both (spatio-temporal). Analyzing such data must be done with caution as observations are now correlated (clusters) and nested (hierarchical), hence ordinary statistical methods assuming the independence of observations are no longer valid. One plausible unit of analysis here may be the individual, and the correlation of individuals within a family must be taken into account. Ignoring the structure of the data may result in asymptotically biased parameter estimates. Recognizing the structure of the survey means using appropriate statistical modelling techniques that model individuals nested within households.

1.1.2 Incomplete Data

Missing data is a common phenomenon in any complex data sets, especially in surveys. Missing data can occur because of non-response, such as when there is no information provided for some items, the whole unit or non-coverage (when the target population is not included in the survey's sampling frame). Non-response or incompleteness may be a result of the interviewer being unable to locate the sampled person or data lost during data entry, or perhaps as a result of sampled persons refusal to participate in the survey. A high degree of missing data may jeopardize statistical analysis. The quality of the data is sometimes measured by the level of missingness, and incomplete data sets are known to sometimes have huge effects on results, thus, their interpretation. Quite often, statistical analysis is hampered due to the presence of missing data, a term referring to the problem that arises when the intended measurements are not obtained or not present in the data set. Missingness reduces the amount of observations in the data set, because only a fraction of the intended observations is available for some individuals (Fitzmaurice *et al.*, 2008). In order to obtain a valid inference for the data set, it is important to first indicate the nature of the missingness. Most real life population health studies result in a combination of incomplete and correlated data.

In this study, missing data are characterized by ”non-true zeros” or ”imputed zeros”. Data acquisition and data collection in most developing countries is worrisome; the reliability of such data sets are questionable, especially in the cases of a high percentage of zero disease counts. In spatial data, where data sets are collected at spatial location (e.g. province), most of the locations may report no cases of the disease and at times this claim cannot be verified. It is very difficult to distinguish between ”true” and ”imputed” zeros, because of the disease reporting mechanism in some countries as a result of insecurity, technical or logistics issues.

1.2 Conceptual Framework and Objectives of the Research

As most of the research in the subsequent chapters has appeared in peer-reviewed journals (Chapters 3 to 7 have been published, Chapter 8 has been accepted for publication and is currently in press, Chapter 9 submitted for publication and Chapter 10 recently completed), this section has been included to combine Chapters 3 to 10 in a coherent whole and place the thesis into context.

This study view of developing countries investigates their prospects of achieving MDGs through the development of flexible predictor statistical models. The author combines flexible statistical models for correlated and incomplete data on the following themes; poverty and education, HIV/AIDS, malaria, maternal and child health (including vaccination) and some other diseases. The data sets for this research are from Demographic Health Surveys (DHSs), Malaria Indictor Surveys (MISs), Heath Management Information Systems (HMIS), the Afghanistan Mortality Survey (AMS) and the World Organization for Animal Health (OIE).

The general framework of the project is divided into two parts. The idea was to start with simple and known techniques, then build up to new ideas of handling different types of correlation structures, in particular; multilevel, spatial, temporal and spatio-temporal. The first part consists of providing flexible models using existing statistical methods for correlated data from population health studies in developing countries (see Section 1.1.1). The first two chapters present the general idea and introduction of the study, the conceptual framework and the data sources, and descriptive summaries of the motivating examples. The next two chapters (3 and 4), illustrate the application of spatial exploration of under-five mortality in Nigeria and spatial Bayesian models to study the outbreak of Leishmaniasis in Afghanistan, respectively. Due to problems associated with maximum likelihood models resulting in complexity in their formulation, Bayesian methods provide an interesting alternative (Bernardinelli and Songini, 1995). Spatial Hierarchical Bayesian (SHB) models provide a good model for the over-dispersion of the relative risk of diseases and take into account the risk dependence between close areas. The dynamics of Leishmaniasis disease in Afghanistan was investigated using the Bayesian spatial model. In Chapter 5, the influence of poverty on educational attainment and enrollment, and the effects of socio-economic realities of poor households was analyzed, while Chapter 6 is dedicated to multi-year trend analysis of childhood immunization uptake and coverage in Nigeria. In both chapters, it is expected that measurements within a family are more alike than measurements from a different family. The challenge is to seek statistical models that correctly account for different sources of variability or heterogeneity. Particularly, the research study explores different plausible candidates of models that correctly account for the variability in the data set. Possible models range from full-likelihood based estimation methods (Willaim, 1975; Molenberghs and Ryan, 1999), pseudo-likelihood (Arnold and Strauss, 1991) to Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) and Alternating Logistic Regression (ALR) (Carey *et al.*, 1993). Chapter 7 presents the epidemiological analysis of the Highly Pathogenic Avian Influenza (HPAI) H5N1 virus in Nigeria using misaligned data. The chapter discusses the difficulties that may arise from a mismatch between data measured at different resolutions resulting in spatial misalignment. Spatial misalignment of the exposure and response variables can bias the estimation of health risk (Peng and Bell, 2010). The weather stations were realigned with the locations of the infected farms and possible association between H5N1 outbreaks and environmental factors such as altitude, temperature, wind and dew were investigated. In Chapter 8, the use of the spatial bivariate probit to investigate the leading cause(s) of death and the preventable factors in Afghanistan using data from verbal autopsies was introduced. In some cases, death is a result of cumulative effects of different causes; some ailments may not directly contribute to the death process and therefore, are not included as the underlying cause of death. The main idea was to investigate the associations between diseases and possibly, the cause of death.

The second part of the dissertation is dedicated to methodological contributions to spatial (Chapter 9), spatio-temporal and incomplete data (Chapter 10) (see also Section 1.1.2). One of the most challenging issues in modelling spatial and spatio-temporal data is the choice of a valid and yet flexible correlation (covariance) structure. Some examples of correlation structures can be found in Cressie and Huang (1999); Gneiting (2002); Stein (2005) and Porcu *et al.* (2007), among others. Unfortunately, most of these correlation structures are either extremely complicated or infeasible to manipulate due to their high dimensions. Moreover, in some cases, a single correlation structure may not give an explanation for the correlation in the data. Hence, in Chapter 9, the data was analyzed using Generalized Estimating Equation (GEE, Liang and Zeger, 1986) based on different correlation structures and then the resulting estimating equations were combined using the platform of Generalized Method of Moments (GMM, Hansen, 1982).

Chapter 10 deals with spatio-temporal and incomplete data. In spatio-temporal analysis, the correlation structures fall into one of two types: separable, in which case it is assumed that the space-time correlation can be written as a product of a correlation for the space dimension, and one for the time dimension, or non-separable, where the space-time correlation is modelled as a single entity. Incomplete data sets are known to sometimes have significant effects on the estimations, and thus also on their interpretations. Here, incomplete data sets refer to data that include a high number of zeros in which it is very difficult to distinguish between ”true zeros” and ”imputed” zeros. This can be due to the reporting mechanism as a result of insecurity, technical and logistics issues. The focus is therefore on the implementation of techniques for handling these problems. One example would be to consider the option of discarding the zeros, and then model a conditional Poisson model, conditional on all cases greater than 0.

Chapter 2

Motivating Examples on the Themes of the Research

2.1 Demographic and Health Surveys (DHS)

Demographic and Health Surveys (DHSs) have been conducted in many countries through their national statistics offices for more than two decades. Starting with DHS phase I in 1984-1989 to DHS phase VI 2008-2013, the surveys are based on national representative data and provide information on the population and the health situation of a country. These are available for downloading on the website of ICF Macro International (`www.measuredhs.com`). The main interest is specifically to first collect information at individual and household level, then aggregate to state, country or regional level. Such information also includes: maternal and child health, childhood mortality, wealth index, ownership of basic facilities, use of bed nets, awareness and behavior of HIV/AIDS and HIV testing, to mention a few of the MDGs indicators. Data sets from standard demographic and health surveys (DHSs), malaria indicator surveys (MISs) and AIDS/HIV indicator surveys (AISs) were used in this thesis. The background to motivating examples considered in Part I on correlated data (1.1.1) is captured in 2.1.1 to 2.1.4 and the background to Part II on incomplete data (1.1.2) is presented in 2.2 under other data sets.

2.1.1 Under-five Mortality

Under-five mortality in Africa is occuring at an alarming rate, with about 40% of the global deaths occurring in Sub-Sahara Africa. As a leading indicator of level of child health, the under-five mortality rate has also been incorporated as a United Nations Millennium Development Goals indicator, with the goal of reducing the under-five mortality rate by two-thirds by 2015. Reducing the under-five mortality has been said to be a key health outcome in developing countries (Korenromp *et al.*, 2004). It was reported that in 2007, 9.2 million children born alive across the world died before their fifth birthday. Most of these children live in developing countries (UNICEF, 2010). Reports from UNICEF have shown a decline in under-five mortality of more than 50 percent in some parts of the world, including Latin America, the Caribbean, Central and Eastern Europe, the Commonwealth of Independent States, East Asia and the Pacific. However, in 2007, UNICEF country-by-country statistics showed Nigeria as 8th in the world in under-five mortality. Although the Nigerian mortality rate decreased from 213 per 1000 in 1990 to 143 in 2010, Nigeria's under-five mortality was ranked 12th world wide in 2010 (Adegboye, 2010a; UNICEF, 2010).

Several factors have been identified to have a significant impact on under-five mortality: pneumonia, diarrhea, poverty, safe water and sanitation, mother's education and mother's age. The data for this study was extracted from the DHS mentioned in Section 2.1 above. The survey includes information on fertility, family planning, child health, nutrition, women's health, women's characteristics and status. The data type used in this study is count data which arises when the cases of events are aggregated over a region (i.e. the number of disease events at a location). These data sets were

Table 2.1: Responses to survey questions about problems around time of birth in Nigeria DHS 2003 (Percentage)

Risk factors	Residence			
	Urban		Rural	
	Yes	No	Yes	No
Long labour more than 12 hours	6.99	18.23	25.97	48.81
Excessive bleeding	4.49	20.72	18.85	55.93
Fever/Bad smelling vaginal discharge	3.12	22.10	11.74	63.05
Convulsions not caused by a fever	1.38	23.78	3.75	71.09

obtained at the household level: the number of cases of under-five deaths, number of births in each household and later aggregated to regional level. Information collected includes age at first marriage, birth interval, and place where a child was given birth to: home, public hospital, private medical sector. Table 2.1 presents the response summary to a sample question from the DHS data. The table indicates the proportions for responses to questions on problems around the time of birth by a female respondent who lost her child(ren).

The data on under-five deaths was used in Chapter 3, where spatial exploratory tools were applied to study their suitability for modelling the 2003 Nigeria Demographic and Health Survey (NDHS) data set.

2.1.2 Poverty and Education in Africa

Access to education, particularly in the developing countries, has been discouraging. The United Nations (UN) Universal Declaration of Human Rights Article 26 states that everyone has the right to education (United Nations, 1948). The Jomtien 1990 declaration of the "World Conference on Education For All" stipulates that every person (child, youth and adult) shall be able to benefit from educational opportunities designed to meet his/her basic needs. Education has also been described as a tool for economic development

Table 2.2: Summary of the DHS data sets on poverty and education

Survey Phase	Observation	Male	Female	Number of Households
DHS 2	261205	50.8	49.2	51788
DHS 3	373865	50.99	49.01	81820
DHS 4	596607	50.65	49.35	119237
DHS 5	485268	50.72	49.28	129006

and the eradication of poverty. Schooling improves productivity, health and reduces negative features of life such as child labour, as well as bringing empowerment (EFA Global Monitoring Report, 2002). Bledsoe *et al.* (1999) found an association between schooling and fertility in less developed countries. The mother's educational attainment plays an important role in the household and has a significant effect on her bargaining power and, thus, her drive for education. The United Nations Educational Scientific and Cultural Organization (UNESCO) reported that Sub-Sahara Africa had an increase in its average enrolment from 54% to 70% between 1999 and 2006. The Africa Recovery July 2000 report indicated that Tanzania has been more successful than many developing countries in achieving gender equality in education, with girls making up to 49.6% of all enrolled primary school students in 1997.

A total of 1,716,945 observations from 381,851 households and from a total of 36 African countries were collected from 1988 (DHS 2) to 2008 (DHS 5). Table 2.2 and 2.3 summarize the data characteristics. Only students between the ages of 5 years to 30 years were considered, because in most African countries this is referred to as the school age (especially primary to tertiary). This data provides information on individual children and household characteristics. Household members shared some information collected at the household level, thus dependencies or correlation among these responses we would be expected (i.e. within subject dependency).

The poverty and education data sets were used in Chapter 5, where the focus is on nested data structures. In the data, individuals are nested within households and

Table 2.3: Characteristics of the variables (poverty and education)

Variable	Description
Child	Relationship (son/daughter, adopted/foster/stepchild or other relation)
Single parent	Both parents are dead or at least one is alive
Medulevel	Mother's educational level-no education, primary, secondary
Fedulevel	Father's educational level-no education, primary, secondary
Fage	Father's age
Mage	Mother's age
Wealthindex	Relative scale (from poorest 20% to richest 20%)
Weight	Sample weight :probability of sample
Nohh	Total number of family members
Underfive	Number of under five year old children
Residence	Type of residence : (urban or rural)
Gender	Gender of the individual
Eduattain	Educational attainment :no education or at least primary education
Age	Age in years
Source	Source of drinking water: safe water or unsafe
Toilet	Type of toilet facility: good sanitation or bad sanitation
Rooms	Number of rooms used for sleeping in the household
Sexofhead	Gender of household head
Ageofhead	Age of household head in years
Rank	Position of child in household
Literacy gap	Gender difference in the percentage of male and female with at least primary education

information were collected for every eligible member at a household level. Multi-level models that allow for information to be pulled together from multiple levels and enable interrelationships between the different levels to be explored and facilitated for overall interpretation, were used to address multiple complex data issues.

2.1.3 Vaccination in Nigeria

Vaccination is an effective way of eradicating the spread of a large number of preventable diseases that account for the high proportion of under-five deaths. Wammanda *et al.* (2011) reviewed that only 22% of children receive their Bacillus Calmette-Guerin (BCG)

Table 2.4: Summary of vaccination uptake in Nigeria from 1990 to 2008

Characteristic	Year of the survey				**Total**
	1990	1999	2003	2008	
Number of communities(Clusters)	298	393	361	886	1938
Number of households	3995	2837	2161	11014	17380
Number of observations	7902	3552	6029	28647	46130
Percentage fully vaccinated	58	55	62	60	59

within the first 3 days of life and 36.2% within the first 7 days of life. It is expected that a fully vaccinated child should have received a BCG, three doses of Diphtheria, Pertussis and Tetanus (DPT), at least three doses of Polio and one dose of Measles vaccines by their first birthday (National Population Commission (NPC) [Nigeria] and ICF Macro, 2009).

The data from this study was collected from the Nigeria Demographic and Health Surveys (NGHSs) implemented by National Bureau of Statistics with the technical support of ICF macro from 1990 to 2008. Standard NDHSs were conducted in 1990, 1999, 2003 and 2008, and they are categorized by the phase of the survey as phase II, phase III, phase IV, and phase V, respectively. The percentage of children who received complete antigens (BCG, DPT1, DPT2, DPT3, OPV1, OPV2, OPV3, Measles) in 1990, 1999, 2003 and 2008 in this study are 58%, 55%, 62% and 60%, respectively (Table 2.4). In the NDHS 2008, mothers were asked regarding the place of all live births in the five years preceding the survey; 62% responded that they had their birth delivered at home, 13% at private health facilities and 20% at a public health facility (National Population Commission (NPC) [Nigeria] and ICF Macro, 2009).

These data sets were used in Chapter 6, where the study evaluates the uptake of childhood immunization in Nigeria from 1990 to 2008. A flexible model that simultaneously regresses the response on explanatory variables as well as modelling the association among responses in terms of pair-wise odds ratios was used to address data complexities.

2.1.4 Causes and Patterns of Morbidity and Mortality in Afghanistan

For about three decades, Afghanistan has been devastated by armed conflicts, starting with the Soviet invasion in 1979, the civil war between 1989 and 1994, and the rise of the Taliban to power in 1994. Over 5 million people have been displaced to Pakistan and Iran during this time period (Bhutta, 2002). Humanitarian relief efforts since the fall of the Taliban in Afghanistan have seen more than $2 billion spent on the health sector (Reithinger and Coleman, 2007; Toby *et al.*, 2006). The health care system is barely functioning; the life expectancy for males and females is 47 years and 50 years, respectively. Mortality statistics provided by the World Health Organization (WHO, 2012) show the under-five mortality rate to be 199 per 1000, and the probability of dying between the age of 15 and 60 to be 440 and 352 per 1000 for males and females, respectively.

The Afghanistan Mortality Survey was conducted in 2010 by the Afghan Public Health Institute (APHI) of the Ministry of Public Health (MoPH) and the Central Statistics Organization (CSO) with technical and logistic support from ICF Macro and the Indian Institute for Health Management Research (IIHMR) (APHI/MoPH, CSO, ICF Macro, IIHMR and WHO/EMRO, 2011). This is the first nationwide survey of its kind in Afghanistan that covers about 87% of the country and 34 provinces of Afghanistan. It is a special survey that focuses on mortality, causes of death and maternal mortality. Specific information collected include maternal and child health, childhood mortality, wealth index, ownership of basic facilities, availability of health facility, and verbal autopsies of deaths to mention a few at household and/or individual level. This study will focus on data from verbal autopsies of deaths within a household up to seven years before the survey (2003-2010) in the national representative Afghanistan Mortality Survey 2010.

The data include verbal autopsies from 1105 neonatal (0 - 28 days), 997 perinatal and children (29 days - 11 years), and 1831 adults (above 12 years). The list as well as the summary statistics of variables used in this study is provided in Table 2.5.

The data presents special challenges when analysing correlated binary response variables. The data sets from verbal autopsies were used in Chapter 7 to model the joint probability of causes of death using multivariate techniques that can accommodate dependency that arise from spatial sources.

2.2 Other Data Sets

Other data sets used in this study to highlight the analysis of intricate data structures include: Data on the incidence of Leishmaniasis in Afghanistan reported to the Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) in Afghanistan from 2003 to 2009. HPAI H5N1 outbreaks data reported to the World Organization of Animal Health (OIE) in 2006. Climatic data for the weather stations in Nigeria and Afghanistan were extracted from the National Climatic Data Center of the National Oceanic and Atmospheric Administration, US Department of Commerce.

2.2.1 Highly Pathogenic Avian Influenza (H5NI) in Nigeria

The World Health Organization (WHO) describes avian influenza as an infectious disease of animals (usually birds and less commonly pigs) caused by type A strains of the influenza virus. The disease spreads extensively in poultry and some transmissions to humans and other mammals have been documented. The Highly Pathogenic Avian Influenza (H5N1) virus was first reported in Nigeria in 2006; the disease spread to several states within months (Ojo, 2008). The epidemic of H5N1 has huge social-economic consequences with

Table 2.5: List of explanatory variables and their summary statistics (AMS, 2010)

Variables	Neonatal		Perinatal and Children		Adults	
	N	%	N	%	N	%
Disease classification						
Infectious	220	19.91	874	87.66	989	54.01
Complication	713	64.52	140	14.04	111	6.06
Non-infectious	40	3.62	280	28.08	1798	98.20
Symptoms & Signs	290	26.24	821	82.35	1433	78.26
Injury	65	5.88	393	39.42	711	38.83
Gender						
Male	624	56.62	548	55.47	1045	57.26
Female	478	43.38	440	44.53	780	42.74
Mean age		2.5 months		5.14 years		54.26 years
Mother's age (years)						
$\leq$ 15	55	4.98				
16-20	241	21.81				
21-25	294	26.61				
26-30	206	18.64				
31-35	157	14.21				
$\geq$ 36	152	13.75				
Size at birth (birth weight)						
Smaller than normal	431	43.98				
Normal	493	50.31				
Larger than normal	56	5.71				
Place of birth						
Home	680	61.87				
Hospital	419	38.13				
Antenatal						
Yes	522	48.47				
No	555	51.53				
Single of multiple birth						
Yes	1025	93.01				
No	77	6.99				
Duration of pregnancy						
1-3 months	12	1.08				
4-6 months	122	11.04				
6-8 months	304	27.51				
9 months	611	56.25				
10 months	41	3.76				
Pregnant at time of death						
Yes					415	15.89
No					217	84.11

Table 2.6: The most affected states of H5N1 virus in 2006

States	Cases(%)	Duration(Days)
Plateau State	43 (29.7%)	143
Kaduna State	18 (12.4%)	346
Kano State	17 (11.7%)	346
Bauchi State	17 (11.7%)	322
Kastina State	9 (6.2%)	21

respect to animal welfare, international trade and cost. The federal government of Nigeria set aside the sum of 1.5 billion Naira (11.5 million USD) for compensation for the birds that were killed throughout the nation on suspicion of carrying the virus, in order to contain any potential spread. An economic impact assessment made by the Poultry Association of Nigeria (PAN) put the loss to farmers in the first four weeks of the outbreak to 14.4 billion Naira, an estimate which does not include those very small-scale farmers whose stock constitutes over 60% of the total national poultry population (Duru, 2006).

HPAI H5N1 virus outbreaks were reported to the World Organization of Animal Health (OIE) in 2006. The information collected include; date of infection, species and the location of the infected farms. Figure 2.1 shows the location of the infected farms and location of the weather stations masked with altitude. Climatic data for the available weather stations in Nigeria in 2006 was extracted from the National Climatic Data Center of the National Oceanic and Atmospheric Administration, US Department of Commerce. A total of 145 outbreaks were recorded; the first outbreak occurred on the 10th of January 2006 in Kaduna State, in the North-Western part of Nigeria. Most of the outbreaks occurred in Plateau State (29.7%) followed by Kaduna State (12.4%), then Kano (11.7%) and Bauchi (11.7%) respectively (Table 2.6).

In this study, the locations of the infected farms are not linked with the climatic data measured at different weather stations. The first task was to realign the weather stations with the locations of the infected farms, then the Bayesian Kriging method was used

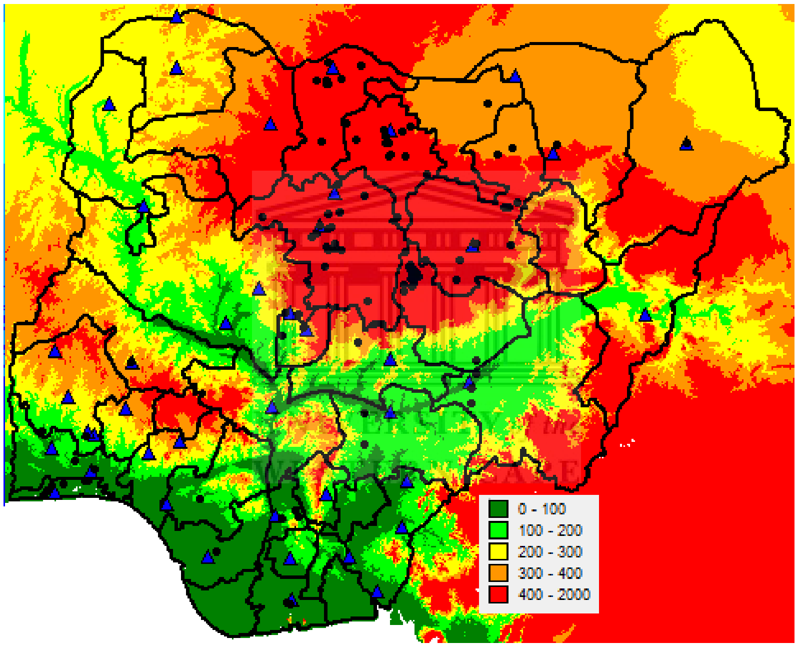


Figure 2.1: Map showing locations of H5NI infected farms (black dots) and weather stations (blue triangle): Altitude in the background

to estimate the mean weather variables for the infected farms. In Chapter 8, a flexible predictor model was used to fit the complex intensity of H5N1 and possible association between H5N1 outbreaks and environmental factors such as altitude, temperature, wind and dew were investigated.

2.2.2 Malaria Incidence

Malaria is caused by the Plasmodium parasite transmitted by the female Anopheles mosquito. It is a preventable and curable disease. The disease accounted for nearly one million deaths in 2008, mostly among children living in Africa. Furthermore, malaria is a leading cause of under-five deaths in SSA, where a child dies every 45 seconds of malaria. Malaria epidemics often occur in areas with non-immune populations living in highlands characterized by arid and desert-fringe zones. Malaria has been reported to decrease the gross domestic product by as much as 1.3% in countries with high disease rates. This disease is synonymous with poverty, loss of life, low productivity due to illness, premature deaths, low school attendance and attainment.

An estimated 10.5 million number of malaria cases, including 37,000 deaths of under-five children, were recorded in the Eastern Mediterranean Region (World Health Organization, 2006). About 248 million people in the region are at risk of malaria transmission from both Plasmodium *falciparum* and Plasmodium *vivax*. Afghanistan is a malaria-endemic country and has the second highest burden of malaria in the region (Safi *et al.*, 2009), with Plasmodium *vivax* accounting for nearly 90% of malaria cases. P. *vivax* malaria is associated with rice growing areas and transmitted by the endophilic and exophilic rice-field breeders Anopheles *pulcherrimus* and A. *hyrcanus* (Faulde *et al.*, 2007). Monthly cases of malaria incidence were reported to the Health Management Information System of the Ministry of Public Health in the 34 provinces of Afghanistan in 2009. Of a total of 521,817 malaria slides examined, 242,127 were positive and

Table 2.7: Description of the variables in the Malaria data

Variables	**Description**
Pv	Plasmodium vivax
Pf	Plasmodium falciparum
Mal	Number of positive and clinical malaria cases
Temp	Average monthly temperature (Celsius)
Wind	Average monthly wind speed (Knots)
Precip	Average monthly precipitation (Inches)
Alt	Altitude of the provincial headquarters (Metres)
Pop	Population size
Wealth	Percentage in the middle and higher wealth quintile in the province
Residence	Place of residence: Rural or Urban
Remoteness	Percentage of the population living in the most and second most remote areas
Sanitation	Percentage in the province with bad sanitation (eg. open latrine, no toilet at all, bucket etc)
Water	Percentage in the province with access to clean and safe water (eg. piped borne, bottle water, spring water etc)
HHmember	Average number of household members
Lat	Latitude of the centroid
Long	Longitude of the centriod

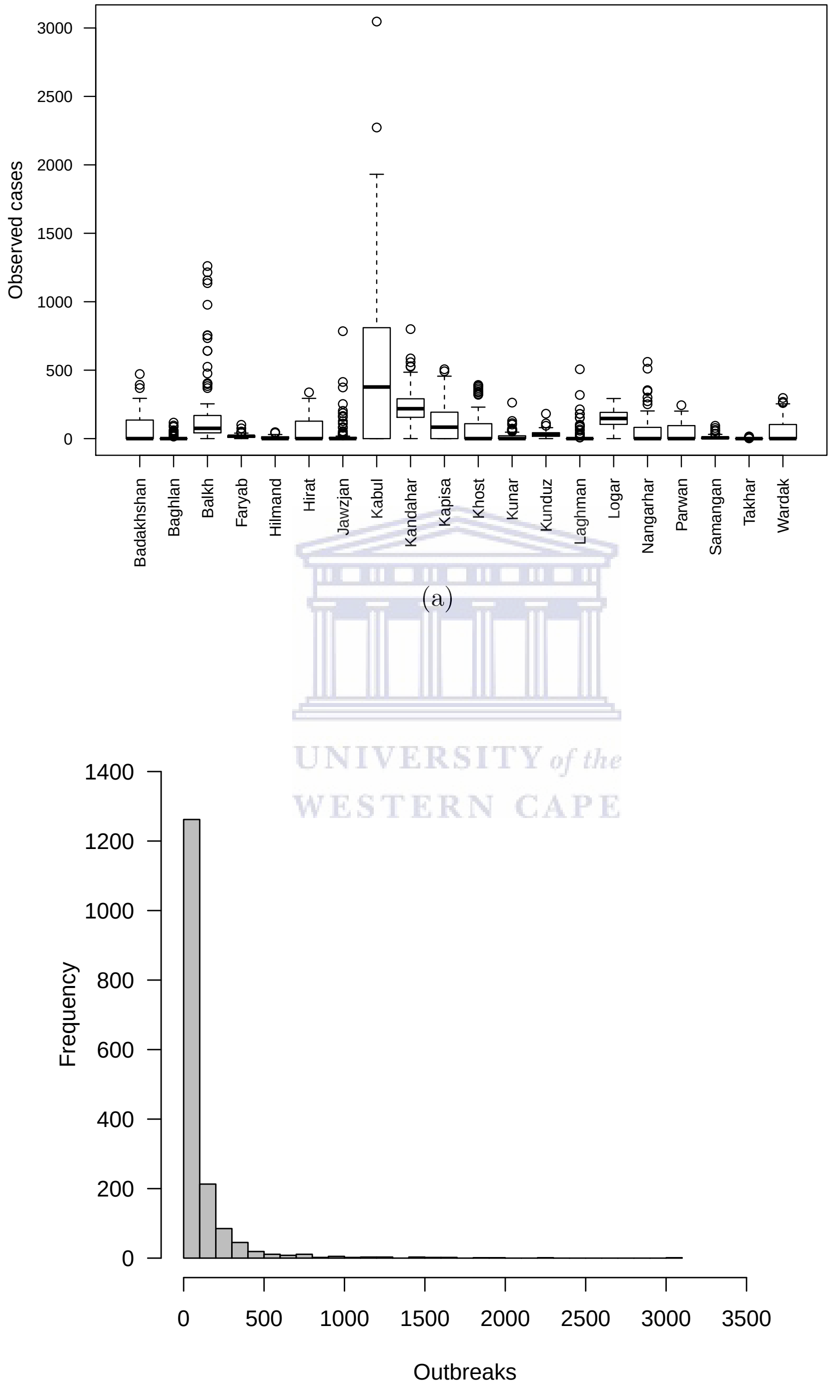
clinical malaria cases, 94% were of P. *vivica*, 6% P. *falciparum* and 148,602 cases of malaria hemophagocytic syndrome. Data on the population of Afghanistan in 2009 was obtained from the Central Statistics Organization of Afghanistan. Additional province level information was extracted from the Afghanistan Mortality Survey (AMS) that was conducted in 2010. The variables extracted from AMS include the remoteness index of the city, migration of the member of the household, the wealth status of the household, number of rooms used for sleeping, number of household members. An overview of the variables used in the study is presented in Table 2.7.

In Chapter 9, the spatial aspect of the transmission is considered, since spatial locations encompass a number of the factors simultaneously. One of the most challenging issues in spatial analysis is the choice of a valid and yet flexible correlation (covariance) structure. We are of the opinion that classical spatial structures may not singly adequately reflect the spatial dependency in the data. The chapter presents a more robust method for modelling spatial correlation to address the multifaceted data issues.

2.2.3 Leishmaniasis in Afghanistan

It was reported that in 2000 there were an estimated 1.5 million annual cases of Leishmaniasis worldwide and Afghanistan, Algeria, Saudi Arabia, Brazil, Iran, Iraq, Peru and Syria accounted for over 90% of the cases (Michael *et al.*, 2008). There are about 250,000 estimated new cases of cutaneous Leishmaniasis incidence in Afghanistan and 67,000 cases in Kabul, thus making it the city with the highest incidence worldwide (Reithinger *et al.*, 2003). Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection. It is contracted through bites from sand flies, which are themselves not poisonous, but the parasitic Leishmania in its saliva can result in chronic and non-healing sores. This mostly occurs on exposed skin and can lead to itchy skin irritation, and disfiguring and painful ulcers. The burden of the disease is overwhelming and the psychological effect can be disturbing. In some societies, women infected with this disease are stigmatized and deemed unsuitable for marriage and motherhood (Reithinger *et al.*, 2005). The impact of environmental influences on Leishmaniasis cannot be ruled out and human activities play a significant role in the dispersion of the vectors, thereby changing the geographical distribution of the disease. The study consists of cases of Leishmaniasis reported to the Afghanistan Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) between 2003 and 2009. A total of 148,945 new cases of Leishmaniasis came from 20 provinces in Afghanistan between 2003 and 2009 (of these, 17,425 occurred in Kabul in 2009). Table 2.8 shows the summary of the data used in this study. The data set is characterized with a high number of zero incidence at many locations (Figure 2.2).

These data sets were used in Chapter 4 to study the quantification and prediction of



(a)

(b)

Figure 2.2: (a) Distribution of total cases of Leishmaniasis at provincial level in Afghanistan (2003-2009)(b) Histogram of the number of times a Leishmaniasis case occurred.

Table 2.8: Summary of the Leishmaniasis data

Year	Gender		% Age group			**Total**
	% Male	% Female	0-4	5-14	$\geq$ 15	
2003	51.60	48.40	23.5	38.5	38.0	10944
2004	53.27	46.75	22.7	38.2	39.1	9203
2005	52.98	47.02	22.3	40.2	37.5	12951
2006	50.19	49.81	18.3	56.2	25.5	19689
2007	48.07	51.93	18.2	43.4	38.4	30273
2008	48.41	51.59	21.7	37.6	40.7	24813
2009	50.03	49.97	23.5	37.2	39.3	41072

Leishmaniasis disease incidence rates across provinces in Afghanistan. Also, spatial scan statistics were used to identify the disease clusters. In Chapter 10, issues surrounding modelling of spatio-temporal correlation/covariance in the data and the high number of zero incidences in many locations, complicating data analysis, were dealt with.

Part I: Statistical Methods for Correlated Data

Chapter 3

Under-five mortality in Nigeria: Spatial Exploration and Spatial Scan Statistics for Cluster Detection

This chapter has been published as:

Adegboye, O.A. (2010). Under-five mortality in Nigeria: Spatial exploration and spatial scan statistics for cluster detection. *International Journal of Statistics and Systems.* 5(2), 203-214.

Summary

Reducing under-five mortality has been said to be a key health outcome in developing countries. Several factors have been identified to have a significant impact on under-five mortality: pneumonia, diarrhea, poverty, safe water and sanitation, mother's education, mothers age. Disease mapping techniques were applied to data on 4911 recorded cases of under-five mortality in the 2003 Nigeria Demographic and Health Survey. The Northwestern region accounts for about 37% of under-five mortality and has the highest proportion of (32%) out the 77% home delivery. The average age at first marriage was 18 years for the Southwest and 14 years for the Northwest. 33% of the respondents who had lost a child indicated that they experienced long labour, 23% excessive bleeding, 15% fever and/or bad vaginal smelling and 25% experienced convulsions not caused by fever around the time of giving birth. SaTScan confirms cluster detection in the Northwest region with a p-value of 0.001. This study suggests possible geographical variations in under-five mortality in Nigeria.

3.1 Introduction

Reducing the under-five mortality has been said to be a key health outcome in a developing countries (Korenromp *et al.*, 2004). It was reported that in 2007, 9.2 million children born alive across the world died before their fifth birthday and most of these children live in a developing country (UNICEF, 2010). As a leading indicator of level of child health, the under-five mortality rate has also been incorporated by most United Nations (UN) member states as a UN Millennium Development Goal (MDG) indicator, with the goal of reducing the under-five mortality rate by two-thirds between 1990 and 2015.

Reports from UNICEF have shown a decline in under-five mortality of more than

50% in some parts of the world, including Latin America, the Caribbean, Central and Eastern Europe, the Commonwealth of independent states, East Asia and the Pacific. However, in 2007 UNICEF country-by-country statistics showed Nigeria as 8^{th} in under-five mortality in the world, even though the mortality rate decreased from 230/1000 in 1990 to 189/1000 in 2007. Several factors have been identified as having a significant impact on under-five mortality; pneumonia, diarrhea, poverty, safe water and sanitary, mothers' education, and mothers' age (Yohanne *et al.*, 1992; Woldernicael, 2001; Kandala *et al.*, 2008). It has also been reported that malnutrition contributes to over a third of these deaths and that improvement of public health services are essential.

In this study geographical location was used instead of environmental factors, to provide insight into the question of how geographical variations correlates with rates of under- five mortality in Nigeria. Spatial exploratory analysis was employed to analyze the 2003 Nigeria Demographic and Health Survey (NDHS) data set.

3.2 Data and Methods

3.2.1 Data

The data for this study were collected from the 2003 NDHS. The survey includes information on fertility, family planning, child health, nutrition, women's health, women's characteristics and status, HIV/AIDS and other STIs at the household level. The data types used in this study are count data which arise when the cases of disease/event are aggregated over a region (i.e. the number of disease events at a location). These data sets were obtained at the state level; the number of cases of under-five death, number of births in each state and the coordinate of the state capital.

3.3 Methods

3.3.1 Disease Mapping

Different disease mapping tools were presented for aggregated data at state level ($i = 1, ..., 37$), specifically crude count, crude rates, probability maps, and the mapping of standard mortality rate (SMR_i). The estimation of SMR_i is subject to its own right (Clayton and Bernardinelli, 1992), thus in this study SMR was calculated such that for each state i, the expected counts E_i, were defined as

$$E_i = n_i \left(\frac{\sum_{i=1}^{37} Y_i}{\sum_{i=1}^{37} n_i} \right) \tag{3.1}$$

and

$$SMR_i = \frac{Y_i}{E_i}$$

where Y_i is the total number of observed under-five deaths in the i^{th} state and n_i is the total number of births in each state. The SMR_i for the under-five is the ratio of observed number of cases to the expected number of cases in the i^{th} state, and these quantities were plotted as a choropleth map. Furthermore, a Poisson probability map was used to display the data owing to the problems of small area count of SMR.

3.3.2 Non-parametric Regression

Mapping SMR_i is very common and often subject to noise. Therefore, the non-parametric regression smoothing method was used to remove the background noise so that the underlying spatial pattern could be seen. The idea is to maximize the weighted log-likelihood of the data given by

$$\sum_{j=1}^{N} w_{ij}\mathrm{log}f(y_i|\theta) \tag{3.2}$$

Assuming $Y_i \sim Poisson(n_i r_i)$ the smoothed rate is given as:

$$\hat{r} = \frac{\sum\limits_{j \in N_j} w_{ij} Y_j}{\sum\limits_{j \in N_j} w_{ij} n_j} \tag{3.3}$$

In this study the kernel functions was used for the weights, i.e.

$$w_{ij} = \mathrm{kern}\left(\frac{s_i - s_j}{b}\right) \tag{3.4}$$

where s_i and s_j are the spatial locations of the capital of state i and j, b is the bandwidth and controls the amount of smoothing.

3.3.3 Test For Randomness

Several tests are available for spatial randomness that adjusts for an uneven background population. Such test statistics are used to test whether or not the geographical distribution of disease is random (Song and Kulldorff, 2005). Previous studies have shown that the spatial scan statistic has good power in detecting hot spot clusters, and Tango's MEET has good power in detecting global clustering (Song and Kulldorff, 2005). In this study, Moran's I, Tango's MEET and spatial scan statistics were used to test for spatial randomness.

Denote c_i as the number of cases in location i, n_i as the population size of location i, C as the total number of cases in the data set, N as the total population, and d_{ij} as the distance between the centroids of locations i and j.

3.3.4 Moran's I

Moran's I (Allepuz *et al.*, 2007) was originally proposed to analyze continuous attribute data such as weight or income. Subsequently, this statistic has often been used to analyze count data, as well, such as cancer incidence. Where L is the total number of locations, the test statistic is defined as:

$$I = \frac{\sum_{i} \sum_{j>1} w_{ij} \left(\frac{c_i}{n_j} - \bar{r} \right) \left(\frac{c_j}{n_j} - \bar{r} \right)}{\sum_{i} \left(\frac{c_i}{n_j} - \bar{r} \right)} \tag{3.5}$$

$$\bar{r} = \frac{1}{L} \sum_{i} \frac{L_i}{n_i} \tag{3.6}$$

$$w_{ij} \begin{cases} 1 & \text{if j is a neighbour of i} \quad (3.7a) \\ 0 & \text{otherwise} \quad (3.7b) \end{cases}$$

The concept of neighbour can be defined in different ways. In this article, j is either defined as a neighbour of i if location j is one of location $i's$ nearest neighbours in terms the distances d_{ij} between them, or, j is defined as a neighbour of i if $d_{ij} < D$ for some fixed distance D. The null hypothesis is rejected when I is large (Moran, 1950).

3.3.5 Tango's Maximized Excess Events Test

Tango's Maximized Excess Events Test (MEET) (Clayton and Bernardinelli, 1992) has been shown to have very good statistical power in detecting global disease clustering. A nice feature of this test is that it considers a range of spatial scale parameters, adjusting for the multiple testing. This means that it has good power to detect a wide range of clustering processes (Song and Kulldorff, 2005).

For a given parameter λ_{ij}, Tango's Excess Events Test is defined as

$$EET(\lambda) = \sum_i \sum_j \exp^{\frac{-4d^2}{\lambda^2}} \left(c_i - n_i \frac{C}{N} \right) \left(c_j - n_j \frac{C}{N} \right) \tag{3.8}$$

This is simply a weighted sum of the excess number of events (observed minus expected) in location i times the excess number of events in location j, where the weight is higher when locations i and j are close. Choosing a large λ makes the test sensitive to geographically large clusters, whereas a small λ makes it more sensitive to small clusters.

To be able to detect clustering irrespectively of its geographical scale, Tango proposed the Maximized Excess Events Test (MEET) (Tango, 1995)

$$MEET = \min_{0<\lambda<U} P(EET(\lambda) > eet(\lambda)|H_0, \lambda) \tag{3.9}$$

where $eet(\lambda)$ is the observed value of the Excess Events Test statistic conditioning on λ, and U is an upper limit on λ specified by the user. The null hypothesis is rejected when MEET is small.

3.3.6 Cluster Detection Tests

Cluster detection tests are concerned with local clusters, and are used when there is simultaneous interest in detecting their location and testing their statistical significance. The circular spatial scan statistics were obtained by using the SaTScan as described by Kulldorff (1997) . Suppose that a map contains p points and consider all circles $C_{i,r}$, where $i = 1, ..., p$ indicates which point is the center of the circle, and the radius, r, ranges from 0 to some maximum radius. When the number of cases is assumed to follow a Poisson model, then for circle $C_{i,r}$, the likelihood ratio is

$$L_{i,r} = \left(\frac{n_{i,r}}{\mu_{i,r}} \right)^{n_{i,r}} \left(\frac{N - n_{i,r}}{N - \mu_{i,r}} \right)^{N - n_{i,r}} \tag{3.10}$$

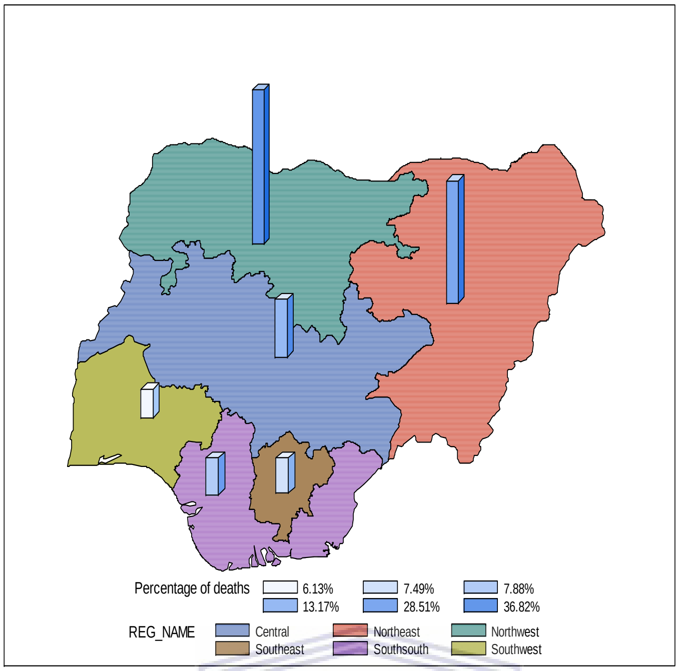


Figure 3.1: Map showing percentage of under-five mortality in Nigeria

where $n_{i,r}$ is the number of cases inside $C_{i,r}$, N is the total number of cases, and $\mu_{i,r}$ is the expected number of cases inside $C_{i,r}$. Note here that the expected number of cases inside the circle is the total number of cases, N, times the proportion of the population inside the circle. In practice, only values of r for which $C_{i,r}$ contains a distinct set of points are considered, since 2 circles with different radii, but containing the same set of points, will have the same likelihood ratio. The likelihood ratio is calculated for each circle; the scan statistic is the maximum likelihood ratio over all distinct circles and corresponds to the most likely cluster.

3.4 Results

A total of 4911 under-five mortality cases were recorded for the 37 states, including the Federal Capital territory of Nigeria. The North-Western region of the country recorded the highest proportion under-five mortality cases (36.82%) while the South-Western region

Table 3.1: Responses to survey questions about problems around time of birth (Percentage)

	Residence			
	Urban		Rural	
	Yes	No	Yes	No
Long labour more than 12 hours	6.99	18.23	25.97	48.81
Excessive bleeding	4.49	20.72	18.85	55.93
Fever/Bad smelling vaginal discharge	3.12	22.10	11.74	63.05
Convulsions not caused by a fever	1.38	23.78	3.75	71.09

recorded the lowest proportion (6.13%), Figure 3.1. About 56% of the deaths were boys with an average age of 16.5 months, in contrast to 17.5 months for girls. Table 3.1 shows the proportions for responses to questions on problems around the time of birth by female respondents who lost her child(ren). When asked if they experienced any problems around the birth, the vast majority who replied ”YES” to questions about long labour, excessive bleeding, fevers and convulsions were from rural areas (Table 3.1). Few people answered ”YES” in urban areas. Also, correlations between these responses were statistically significant for all questions ranging from 0.2225 for long labour and convulsion to 0.3978 for long labour and bleeding.

Figure 3.2 displays the average age at first marriage for the respondents at the six geo-political zones in Nigeria. As it is shown on the map, the South-Western part of the country had the highest value of 18 years, while the North-West had the lowest (14 years). The data indicated that about 77% of the under-five mortality recorded was given birth to at ”HOME”, 14% at ”PUBLIC HOSPITAL” and 9% at ”PRIVATE MEDICAL SECTOR”. As it is shown in Figure 3.3, the North-East and North-West account for 26% and 32% of the total of home deliveries respectively.

The correlations between average age at first marriage, age at first intercourse, age at death of a child and age at first birth indicate a statistically significant correlation among all pairs of variables. There exists a strong positive correlation between age at

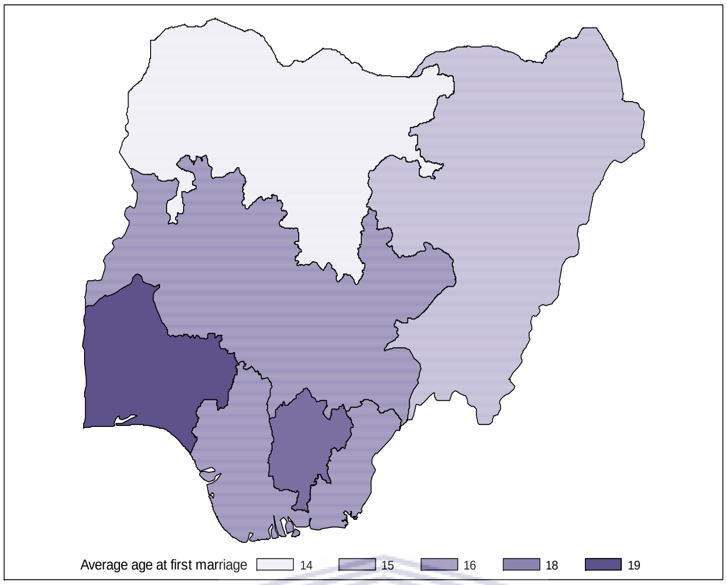


Figure 3.2: Map showing average age at first marriage in Nigeria

first marriage and age at first birth of respondents and weak negative correlation between age at child's death and age of respondent at first marriage, first intercourse and first birth, respectively.

Crude rates for under-five mortality were obtained for each state by finding the ratio of the observed count to population; these estimates, together with the SMR, were plotted on the maps as displayed in Figure 3.4 and 3.5 respectively. The two maps suggest higher values for states in the North-Western and North-Eastern parts of the country. Due to problems of small area counts, a Poisson probability map was used to explore the data, as displayed in Figure 3.6. While there is some indication that there seems to be segregation in the mortality rates in the North-Eastern and North-Western parts of Nigeria, further analyses are required to make inferences about this.

To get a clearer picture of the rates, non-parametric regression was used to reduce the noise in the maps. The smoothing techniques were used with several bandwidths,

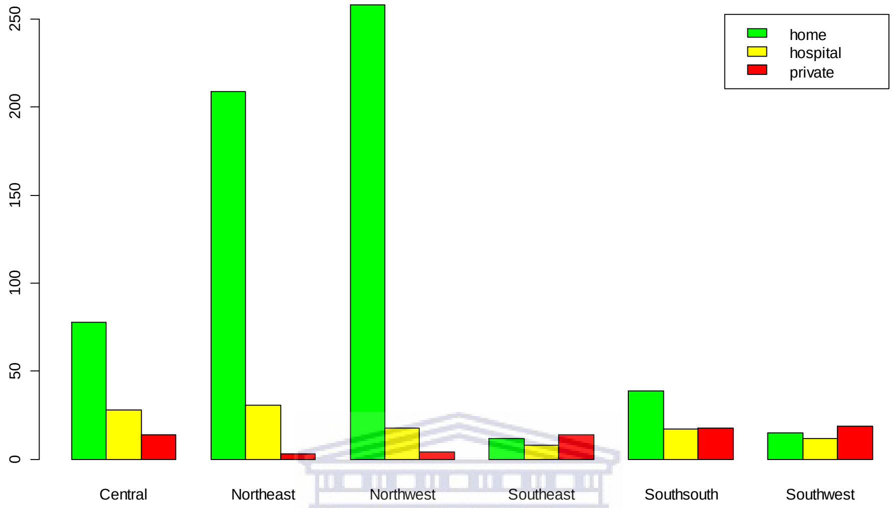


Figure 3.3: Plot showing place of delivery by region in Nigeria

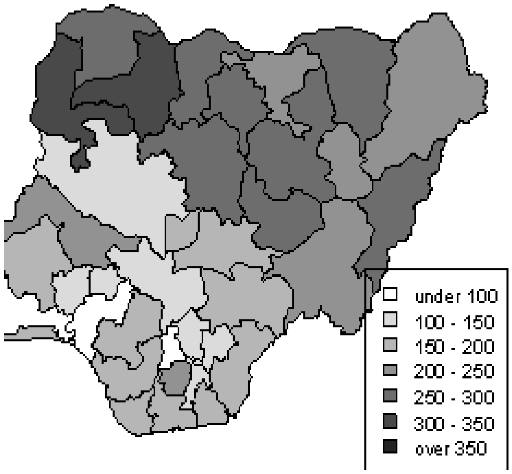


Figure 3.4: Map showing the crude rate of under-five mortality in Nigeria

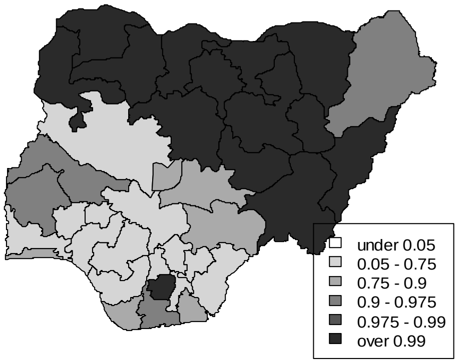


Figure 3.5: Map showing standard mortality rate of under-five mortality in Nigeria

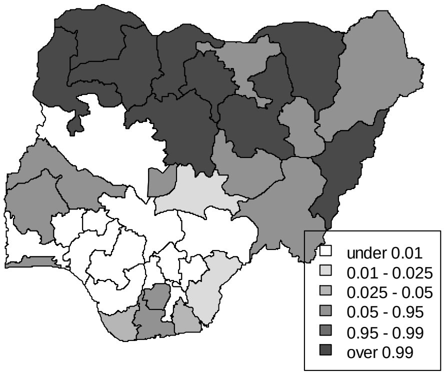


Figure 3.6: Poisson probability map of under-five mortality rate in Nigeria

(0.5, 1, 2, 3), the larger the bandwidth the more variability/noise is taken out. As is shown in Figure 3.7a, b, c and d, one can observe great geographical differences in the rates around the North-Eastern and North-Western regions of Nigeria; these differences disappear with an increase in bandwidth size.

Moran's I is a measure of spatial autocorrelation in the data. It indicates how the values of a variable are related, based on the locations where they were measured. Using the functions in the `ape` library in R software, Moran's I statistics was calculated as 0.1260 with a standard deviation of 0.0293 and a p-value of < 0.000. This statistic indicates the existence of spatial autocorrelation in the case of under-five mortality in Nigeria.

Tango's Maximized Excess Events Test (MEET) has been shown to have very good statistical power in detecting global clustering (Song *et al.*, 2006). Results from this analysis showed a Tango's MEET statistic of 0.00177 and p-value of 0.01, also rejecting the null hypothesis. Moran's I and Tango's MEET both suggest the possibility of clustering in the data. Further cluster detection analysis will be carried out using spatial scan statistics by Kulldurf (Kulldorff, 1997).

So far, different spatial clustering statistics have been explored and they all attested to the possibility of clusters in the data. The SaTScan is a software program written to implement the scan statistics, it can be used to find clusters in space and/or time (Kulldorff, 1997). The scan statistics was used for cluster detection, and the mostly likely clusters were detected in the North-Western region of Nigeria with a p-value of 0.001 (see Figure 3.8).

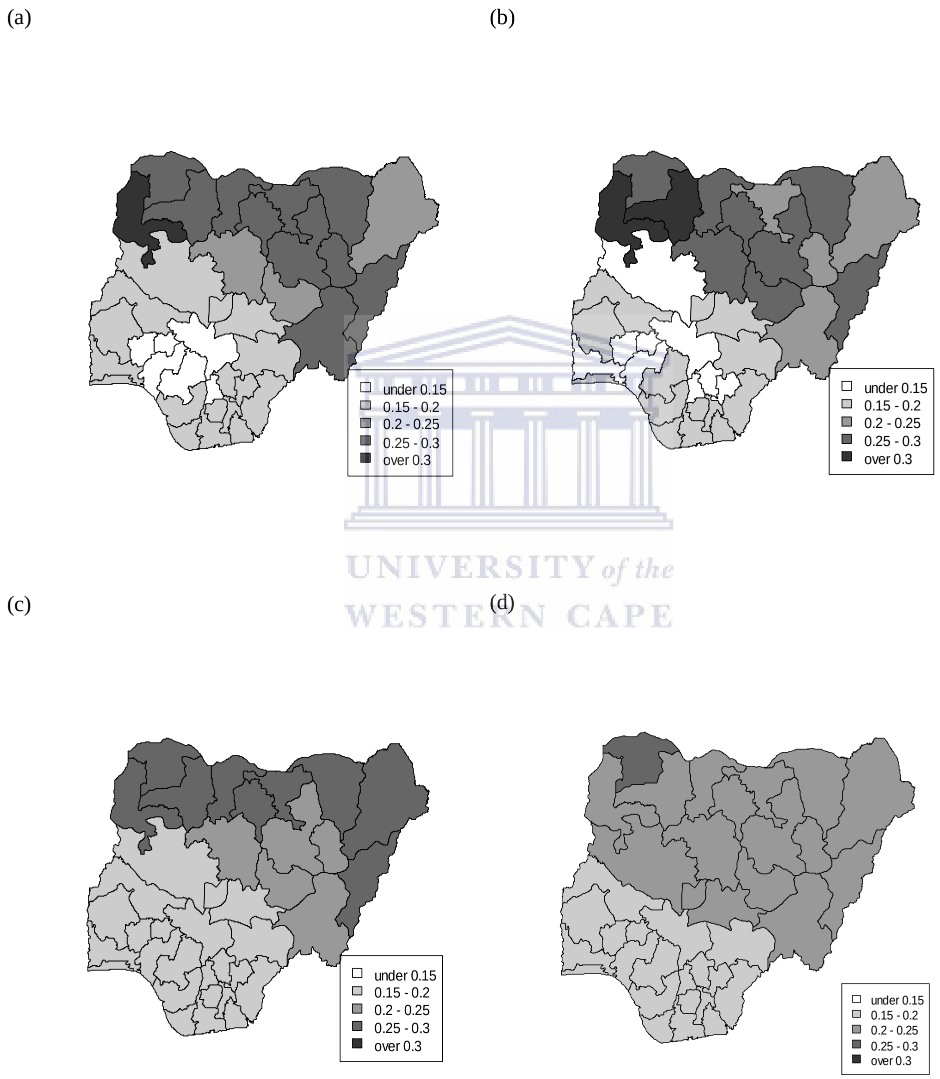


Figure 3.7: Nonparametric regression smoothed map using (a) bandwidth = 0.5, (b) bandwidth = 1, (c) bandwidth = 2, (d) bandwidth = 3

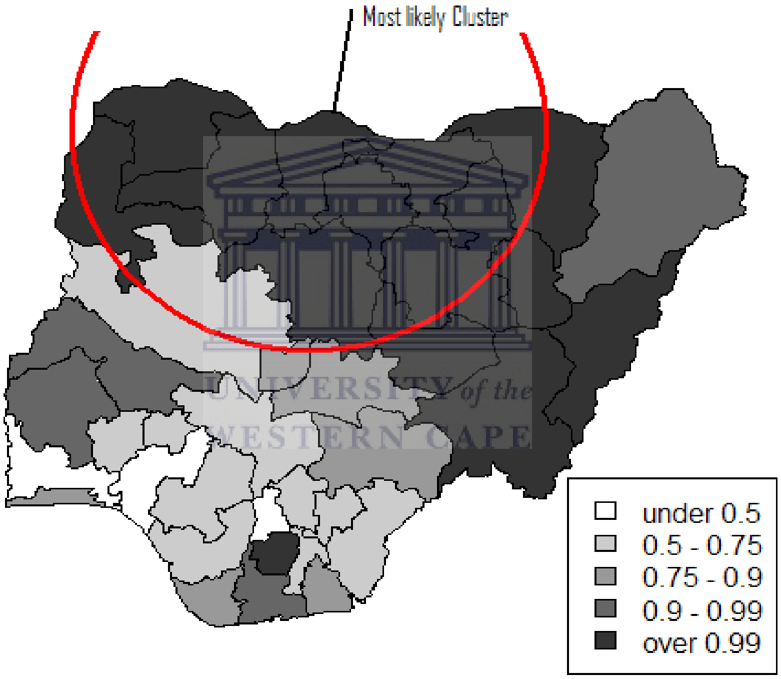


Figure 3.8: Map indicating mostly likely clusters of under-five mortality detected by spatial scan statistics

3.5 Conclusion

In this paper, data on under-five mortality from the Nigeria Demographic and Health Survey, 2003, was explored spatially. Possible social-demographic variables were mapped; these maps showed geographical variations in all the estimates. In addition, clustering was confirmed by spatial scan statistics (SaTScan) which showed a significant cluster with high relative risk in the North-Western part of the country.

The research identifies that a high percentage of respondents experience one type of complication at birth in rural areas, which could be due to accessibility of pre-natal care or cultural beliefs. Early marriage is another variable that was shown to be a possible factor for the regional disparities; the North-Western region has a mean age at first marriage of 14 years. Also, about 77% of the under-five mortality cases had given birth at home and the Northern region accounts for more than half of the cases of home delivery.

In conclusion, reorientation plays a large part in changing people's perception about early marriage, prenatal care and adequate medical care during delivery. There is a great need to integrate people's ideology, beliefs and cultural values, both in the rural areas and the Northern part of Nigeria, into health policy formation. Early marriage has been said to contribute to early stoppage in education for girls in many developing countries, thereby hindering young girls from having a good education. This research identifies a high mortality rate in the Northern region, especially the North-Western part. Therefore, additional resources, education and policy attention directed to the region are recommended.

Chapter 4

Disease Mapping of Leishmaniasis Outbreak in Afghanistan: Spatial Hierarchical Bayesian Analysis

This chapter has been published as:

Adegboye, O.A. and Kotze, D. (2012). Disease Mapping of Leishmaniasis Outbreak in Afghanistan: Spatial Hierarchical Bayesian Analysis. *Asian Pacific Journal of Tropical Disease*, 2(4); 253-259. `http://dx.doi.org/10.1016/S2222-1808(12)60056-5`

Summary

In this paper we utilize provincial level geo-referenced data to analyze the spatial pattern of Leishmaniasis disease in Afghanistan. The disease is contracted through bites from sand flies and is the third most common vector-borne disease. Leishmaniasis is a serious health concern in Afghanistan with about 250,000 estimated new cases of cutaneous infection nationwide and 67,000 cases in Kabul. This makes Kabul the city with the largest incidence of the disease worldwide. A Bayesian hierarchical Poisson model was used to estimate the influence of hypothesized risk factors on the relative risk of the disease. A random components was used to take into account the lack of independence of the risk between adjacent areas. Statistical inference is carried out using Markov Chain Monte Carlo simulation. The final model specification includes altitude, two random components (intercept and slope) and utilizes a conditional autoregressive prior with a deviance information criterion of 247.761. Spatial scan statistics confirm disease clusters in the North-Eastern and South-Eastern regions of Afghanistan with a p-value of $<$ 0.0001. The study confirms disease clusters in the North-Eastern and South-Eastern regions of Afghanistan. Our findings are robust with respect to the specification of the prior distribution and give important insights into the spatial dynamics of Leishmaniasis in Afghanistan.

4.1 Introduction

Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection. It is contracted through bites from sand flies and can result in chronic and non-healing sores. This mostly occurs on exposed skin and can be disfiguring and painful. The burden of the disease is overwhelming and the psychological effect can be disturbing. In some societies, women infected with this disease are stigmatized and deemed unsuitable for marriage and motherhood (Reithinger *et al.*, 2005).The World Health Organization (WHO) in 2000 reported that there are an estimated 1.5 million annual cases of Leishmaniasis worldwide and Afghanistan, Algeria, Saudi Arabia, Brazil, Iran, Iraq, Peru and Syria account for over 90% of the cases (Michael *et al.*, 2008).

There are about 250,000 estimated new cases of cutaneous Leishmaniasis incidence in Afghanistan and 67,000 cases in Kabul, thus making it the city with the largest incidence worldwide (Reithinger *et al.*, 2003). Humanitarian relief efforts since the fall of the Taliban in Afghanistan have seen more than US billion spent on the health sector according to the most recent statistics. However, despite this huge investment, health indicators in Afghanistan have shown very little improvement (Toby *et al.*, 2006; Reithinger and Coleman, 2007).

Several different methods from epidemiology, geostatistics and small area modelling have been used to analyze disease incidence rates. The simplest model assumes a Poisson log-linear relationship between disease rates and other covariates with random effects used to capture extra variation in the Poisson model. The simple model ignores the spatial pattern and may be inadequate to explain the variation in the occurrence of the disease. Bayesian modelling has the advantage of allowing the exact analysis of random effects and coefficient models. The impact of environmental factors on the transmission of Leishmaniasis cannot be ruled out and human activity is likely to play a significant role in the dispersion of the vectors thereby changing the geographical distribution of the disease.

The purpose of this paper is to model the transmission dynamics of Leishmaniasis, (the quantification and prediction of the disease incidence rates) across provinces in Afghanistan. We estimate the incidence of Leishmaniasis at the provincial level and explore the effect of altitude on the outbreak of Leishmaniasis. We use a spatial hierarchical Bayesian model to model the over-dispersion of the relative risk of the disease.

This specification allows the risk dependence between close areas to be taken into account. By introducing random components into the model specification, the lack of independence of the risk between areas are taken into account.

Most literature on Leishmaniasis in Afghanistan ranges from the economics of the disease burden to epidemiological evaluation. An investigation into the association of household-level characteristics with the incidence of Anthroponotic Cutaneous Leishmaniasis (ACL) in Kabul, identified that household construction material, design, density (in terms of household members per room) and presence of disease in other households are significant risk factors for the incidence of ACL (Reithinger *et al.*, 2010). An epidemiological evaluation of Zoonotic Cutaneous Leishmaniasis (ZCL) outbreak conducted around Mazar-e Sharif revealed the role played by high rodent infestations as the ZCL natural host in the outbreak (Michael *et al.*, 2008). The results further showed that seasonality in the occurrence of ZCL in humans can be attributed to seasonal activity of the ZCL vector (sand fly). Other studies include the cost-effectiveness of treating Cutaneous Leishmaniasis in Afghanistan (Reithinger *et al.*, 2003).

The contribution of this paper is the first application of spatial Bayesian models to study the outbreak of Leishmaniasis in Afghanistan. The rest of this paper is organized as follows. In the next section we discuss the methods, data sources and provide descriptive statistics of the dynamics of Leishmaniasis disease in Afghanistan. This will be followed by results from Bayesian spatial modelling and finally, the discussion and conclusion will follow.

4.2 Methods

4.2.1 Data

In this paper data are analyzed on cases of Leishmaniasis incidence reported to the Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) in Afghanistan. The data were collected and aggregated at the provincial level and includes a total of 148,564 new cases of Leishmaniasis observed annually in Afghanistan over the period 2003 to 2009. Population sizes for this period were obtained from Central Statistics Organization (CSO) of Afghanistan and the latitude and longitude of the central district was supplied by Afghanistan Information Management Services (AIMS).

Out of the 34 provinces in Afghanistan cases of Leishmaniasis were not available for the following provinces: Badghis, Bamyam, Frah, Ghazni, Ghor, Nimroz, Nuristan, Paktika, Sari Pul, Uruzgan and Zabul (Figure 4.1, bottom). The data indicate that the incidence of Leishmaniasis disease in Afghanistan has been on the rise especially in Kabul (Figure 4.2). While the number of cases of the disease reported across the provinces has not been consistent, Kabul province has recorded a steady increase since 2003 and accounts for about 30% of the total cases (Figure 4.2).

4.3 Estimation Results

Exploratory data analysis on incidents of Leishmaniasis in Afghanistan reveals geographical disparity in the occurrence of the disease (Figure 4.1). The standard incidence rate (SIR) for each of the provinces in Afghanistan ($i = 1, , 34$) was calculated and then mapped in Figure 4.3. The map shows areas with high and low risk; dark regions indicating high risk of Leishmaniasis and the light regions indicating low risk. The SIR

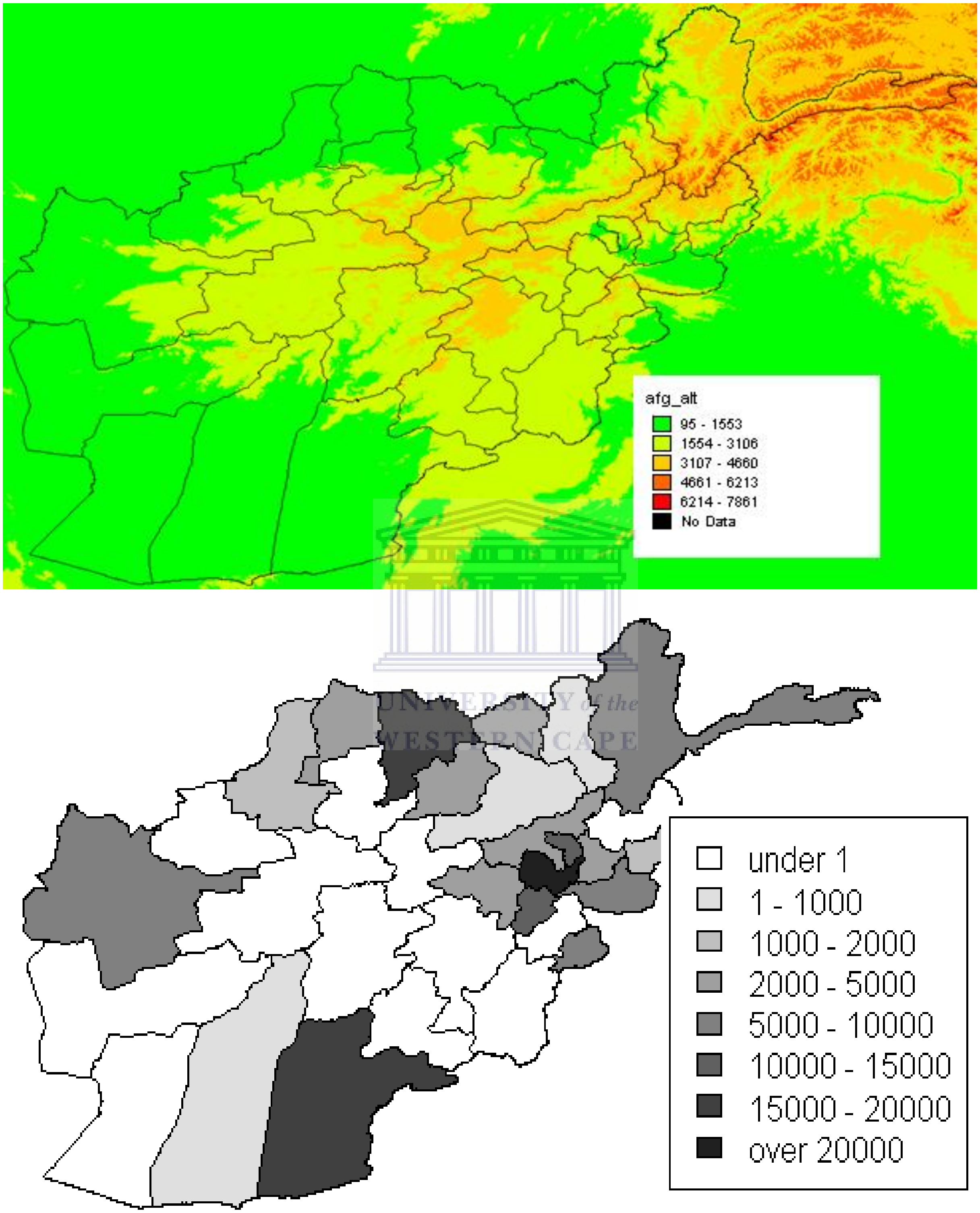


Figure 4.1: Top: SRTM30 1km digital mapping of continuous elevation surfaces data of Afghanistan aggregated to 30 seconds. Bottom: Distribution of cases of Leishmaniasis incidence in Afghanistan 2003-2009.

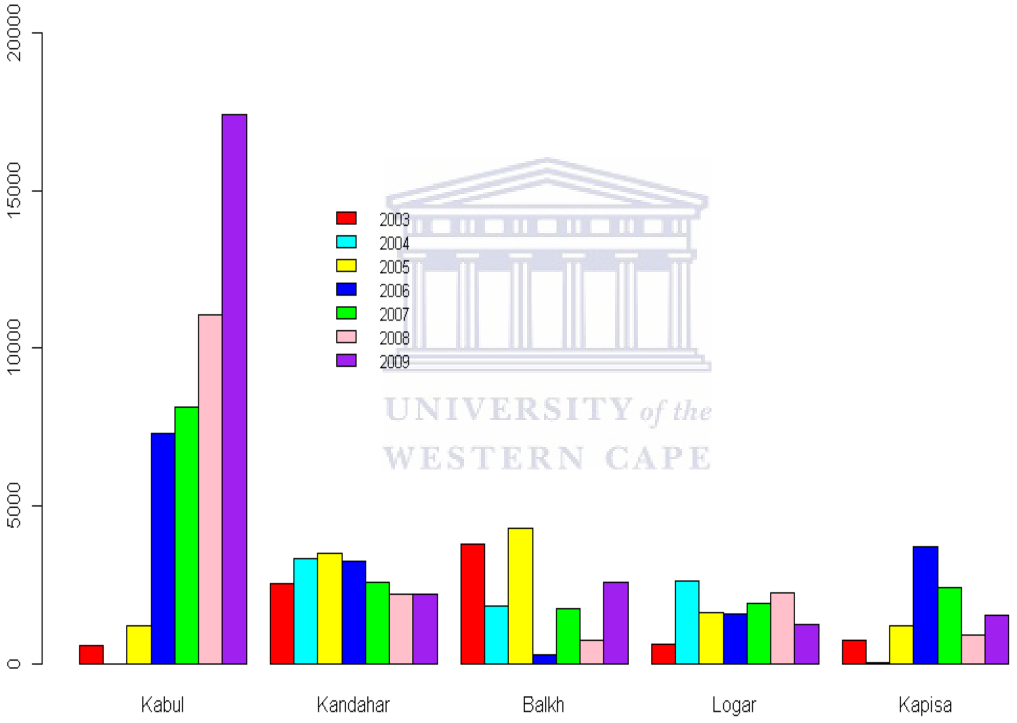


Figure 4.2: Outbreak of Leishmaniasis disease in Afghanistan provinces with the highest incidence (2003-2009)

can be defined as:

$$SIR = \frac{Y_i}{E_i} = \frac{Y_i}{n_i\left(\frac{\sum_i^{34} Y_i}{\sum_i^{34} n_i}\right)} \tag{4.1}$$

where Y_i is the observed count of cases at provincial level and n_i is the number of individuals at risk. The expected number of Leishmaniasis cases, E_i is calculated as indicated by the denominator of equation (4.1). Estimation of standard incidence rates are deficient because of small area disease count where extreme rates occur. The populations that are smallest and geographically close areas tend to have similar disease rates (Odoi *et al.*, 2003). Leishmaniasis disease is a non-contagious vector borne disease and the observed cases at provincial level (Y_i) are assumed to occur independently and follow a Poisson distribution. To overcome the limitations of SIR a spatial hierarchical Bayesian (SHB) model was implemented. This model makes it possible to combine the specific provincial rate with the influence of the spatial neighbourhood. The altitude of the provincial capital is included as a covariate in the SHB. The reason for implementing the latter is that the SIR_i is a crude estimate of relative risk and covariate adjustment can improve this estimate by providing an estimate of logarithm of the relative risk, $log(\theta_i)$. Another approach is the use of random effects and random coefficients in generalized linear mixed models (GLMMs) to model extra variation in the Poisson model (Lowe *et al.*, 2010). A simple model constructed for this scenario assumes a Poisson log-linear relationship between numbers of cases of Leishmaniasis (Y_i) , with mean μ_i and independently distributed as

$$\begin{aligned} y_i &\sim Poisson(\mu_i) \\ E(y_i) &= \mu_i = e_i\theta_i \end{aligned} \tag{4.2}$$

where e_i is the expected rate of Leishmaniasis at province i and θ_i is the relative

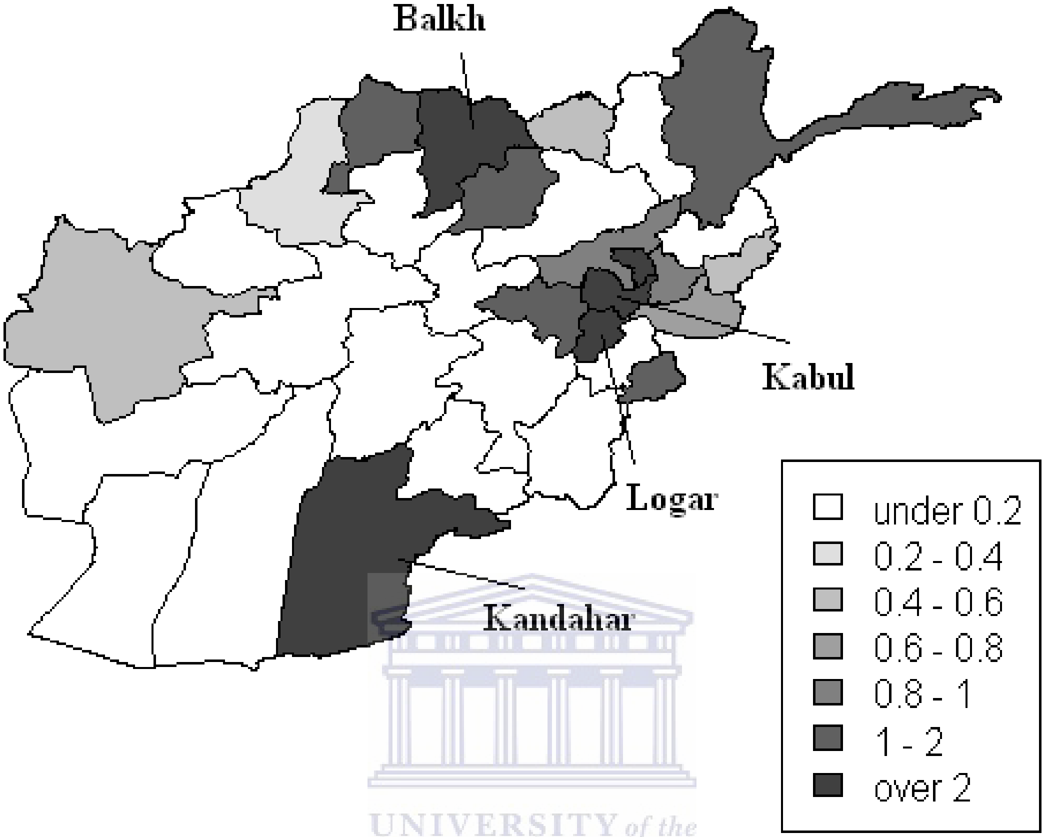


Figure 4.3: Map showing standardized incidence rate of Leishmaniasis in Afghanistan during the period 2003 to 2009

risk for the i^th province.

4.3.1 Spatial Hierarchical Bayesian Modelling

When the observed data are sparse, maximum likelihood (ML) estimation may lead to unstable and largely uninformative estimates of the area-specific linear trends due to Poisson sampling variation (Bernardinelli and Songini, 1995). Bayesian modelling has the advantage of allowing the exact analysis of random effects and coefficient models (Lawson and Zhou, 2007). Several authors have used the spatial hierarchical Bayesian approach to model disease epidemics (Allepuz *et al.*, 2007; Lawson and Zhou, 2007; Stevenson *et al.*, 2005; Durr *et al.*, 2005). In this paper a SHB Poisson model was used to capture over-dispersion of the relative risk and take into account the risk dependence between spatially close areas. The SHB Poisson model was used to quantify the influence of the hypothesized risk factors on the provincial level relative risk of Leishmaniasis disease. To take into account the lack of independence of the risk between provinces, a random component is used.

In the Bayesian context, the likelihood of the data is defined as $L(y|\theta)$ where y is the vector of counts of the disease occurrence in the small areas and θ is a parameter vector describing underlying disease rate. The parameters, θ's, have prior distributions that define the investigator's beliefs about the extra or unobserved random variation; the reader is referred to Lawson and Zhou (2007) for more information on choice of the prior distribution. The joint prior distribution of θ is denoted by $p(\theta)$. The analysis seeks to examine the posterior distribution of θ given the data, denoted by

$$p(\theta|y)\alpha L(y|\theta)p(\theta). \tag{4.3}$$

All models were implemented in WINBUGS software using Gibbs sampling, this allows the iterative exploration of the posterior surface (Lunn *et al.*, 2000) and leads to a set of parameter values rather than a single value which is typical of ML methods (Lawson and Zhou, 2007).

4.3.2 Prior Distribution

When modelling using a Bayesian framework, one needs to specify a prior distribution for the observed data. Several prior distributions for this study were explored, namely flat distributions thus providing a non-informative prior, a gamma distribution and a conditional autoregressive (CAR) distribution. The prior in the spatial model is similar to that proposed by Lunn *et al.* (2000), where alpha0 is assumed to follow a flat distribution, and the unstructured variability parameter (u_i) is assumed to follow a normal distribution with mean 0 and a precision variable; the structured variability term (v_i) was allowed to depend on the neighbours. This is sometimes called the convolution Gaussian distribution or intrinsic Gaussian CAR (Mariella and Tarantino, 2010).

4.3.3 Model Selection and Assessment

For this study different classes of Bayesian Poisson hierarchical models of increasing complexity were formulated. The models closely follow the approaches of Lawson and Zhou (2007) and Stevenson *et al.* (2005). These models include a random component with and without spatial structure. Several adjustments were made to the model to give rise to what was called spatially smoothed and non-spatially smoothed models and were explored with varying prior distributions for the random effect. The model is defined as

$$log(\mu_i) = X'\beta + Z'b \tag{4.4}$$

where X' and Z' are vectors of explanatory variables - the location of the disease and altitude, β's is a parameter vector, the b's are random effects. For all models three chains were ran to help assess convergence and was visualized with time series plots and Gelman-Rubin statistics. The deviance information criterion (DIC) was used to compare all models and the model with the smallest DIC is said to be the "best fit".

4.3.4 Areas of High Risk

Several tests are available for spatial randomness that enables adjustment for unevenness in the background population. The latter tests produce statistics to test whether or not the geographical distribution of the disease is random. Previous studies have shown that

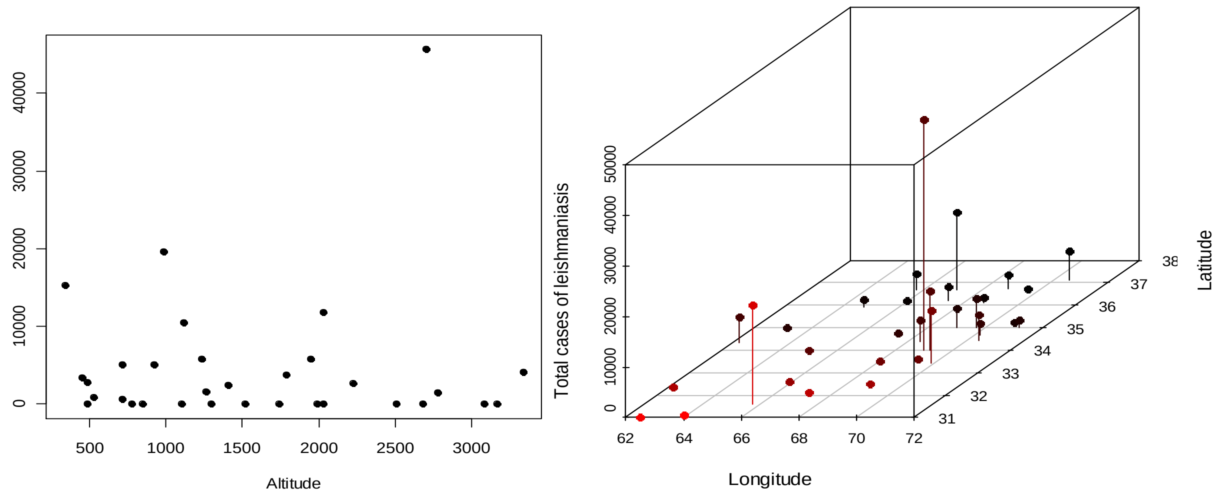


Figure 4.4: Left: Scatter plots of total cases of Leishmaniasis against altitude. Right: 3D scatter plot of total cases of Leishmaniasis against the latitude and longitude of the Centrum of the province

the spatial scan statistic has good power in detecting hot spot clusters (Kulldorff *et al.*, 2009). SaTScan is a software program written to implement the scan statistic; it can be used to find clusters in space and/or time (Kulldorff *et al.*, 2009).

4.4 Results

The data consist of cases of Leishmaniasis incidence from 34 provinces in Afghanistan for the period 2003 to 2009 collected by a health provider and reported to HMIS. A total of 148,564 new cases were reported to HMIS and MoPH during this period with Kabul recording the highest number (30%), followed by Kandahar (13%) and Balkh (10% of all new cases). As indicated in Figure 4.4 (the graph on the left), there seems to be an association between Leishmaniasis and altitude.

The North Eastern region of Afghanistan recorded the highest incidence of the

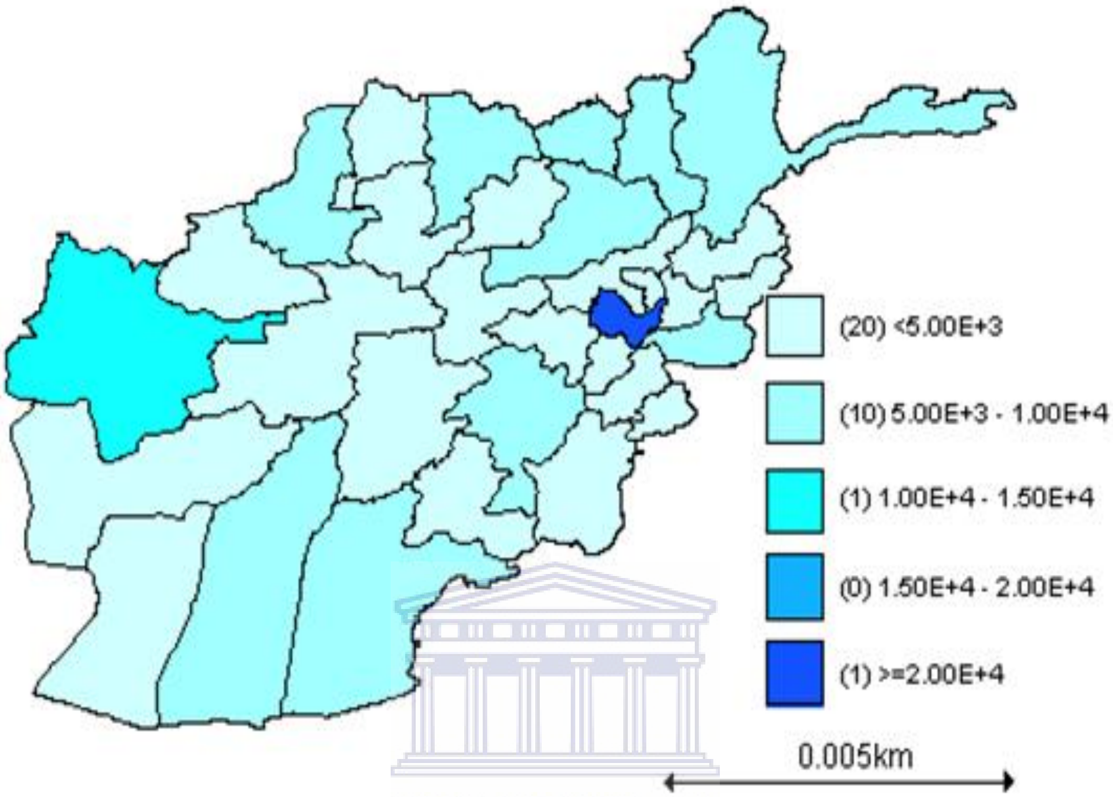


Figure 4.5: Expected incidence of Leishmaniasis in Afghanistan during the period 2003-2009.

disease, while some provinces in South Western region recorded no cases of the disease (Figure 4.1). The expected number of cases was estimated for each province and displayed as maps in Figure 4.5. The standard incidence rate provides an assessment of excess risk expected in a province. The map of the standard incidence rate also indicates geographical disparities in the risk of the disease. The North Eastern region has high risk of the disease, more than the rest of the country.

These crude rates (SIR) must be interpreted carefully and may be misleading, because they are influenced by the population size of the regions and neighbouring provinces.

As mentioned before, the SIR has drawbacks because it assumed provinces are

Table 4.1: Summary of results of hierarchical Bayesian model from WinBUGs with different complexities.

Model	Description	Dbar	Dhat	pD	DIC
Non-spatial random effects					
Model 1	Non-structured random intercept	245.3570	201.0600	44.2970	289.6530
Model 2	Altitude with non-structured random intercept	254.1260	201.0250	53.1010	307.2270
Spatial random effects					
Model 3	Altitude with non-structured & spatial random intercept	234.1620	200.8080	33.3530	267.5150
Model 4	Spatial random intercept	53654.2000	206.5380	53447.6000	107102.0000
Model 5	Non-structured & spatial random intercept	261.2680	200.9060	60.3620	321.6310
Model 6	Altitude with spatial random intercept	71671	4662	67009	138680
Model 7	Altitude with non-structured & spatial (random intercept & slopes)	224.6120	201.4630	23.1490	247.7610

independent. However, from a spatial point of view, this is not the case and more interest lies in the more global, spatial distribution of the number of cases of Leishmaniasis. A spatial hierarchical Bayesian analysis with random components to take into account the lack of independence of the risk between provinces was formulated. Several models (Models 1-7, Table 4.1) with different complexities were explored that included a random component with non-spatially structured heterogeneity and spatial structured heterogeneity. The models follow that of Mariella and Tarantino (2010); the unstructured variability parameter (u_i) was assumed to follow a normal distribution, and the structured variability (v_i) term followed the multivariate normal conditional autoregressive (CAR) distribution. The non-spatial smoothing adjusted the relative risk estimates for province with low numbers towards the overall mean, while including the spatially structured heterogeneity term was to condition the smoothing on neighbouring provinces. The model is presented below

$$log(\mu_i) = log(e_i) + \alpha_0 + (\beta_1 + b_0) \times Altitude + u_i + v_i \tag{4.5}$$

Table 4.2: Posterior summary of results of hierarchical Bayesian model from WinBUGs: Non-spatial regression models, spatial regression model with random intercept only and spatial regression with both random intercept and slope models

Model	Mean	Standard error	Mc error	Credible interval	
				2.5%	97.5%
Non-spatial random effects:					
Intercept (non-structured)					
Intercept	-0.1060	0.1471	0.0101	-0.4595	0.1120
Variance of random intercept (non-spatial)	2.3490	165.1000	1.3540	0.0072	0.0316
Spatial random effects:					
Intercept (non-structured & spatial)					
Intercept	-0.4203	0.1345	0.0092	-0.6200	-0.1696
Altitude	480.2533	679.0460	6.3778	0.4797	2400.7896
Variance of random intercept (non-spatial)	3.5801	255.1184	2.0947	0.0112	
Variance of random intercept (spatial)	0.0680	0.0483	0.0028	0.0043	0.1861
Spatial random effects:					
Intercept (non-structured & spatial) & slopes					
Intercept	-0.1426	0.1825	0.0125	-0.4883	0.2016
Altitude	0.0001	0.0001	0.0000	-0.0001	0.0002
Variance of random intercept (non-spatial)	19770.0000	7476.0000	107.6000	8349.0000	36120.0000
Variance of random intercept (spatial)	208.0000	489.1000	23.7200	1.7770	1389.0000
Variance of random slope	1176.0000	1629.0000	74.8300	21.4100	5641.0000

Table 4.3: Areas with high risk of Leishmaniasis cases for 2003-2009 in provinces of Afghanistan: From SaTScan purely spatial analysis

Clusters	Province	Observed cases	Expected cases	Relative risk	P-value
Primary	Logar	11765	2271.4820	2.9500	< 0.0001
	Paktya	0	3153.8420	2.9500	< 0.0001
	Kabul	45631	22154.1700	2.9500	< 0.0001
	Parwan	3759	4032.4480	2.9500	< 0.0001
	Khost	5836	3178.1880	2.9500	< 0.0001
	Paktika	0	2528.3280	2.9500	< 0.0001
	Kapisa	10546	2567.5910	2.9500	< 0.0001
	Wardak	4132	3428.2020	2.9500	< 0.0001
Secondary	Kandahar	19568	6795.9690	3.1800	< 0.0001
	Balkh	15246	7321.0160	2.0800	< 0.0001

All the models were run in WINBUGS via Gibbs sampling and the posterior distributions were estimated. Three chains of 15,000 iterations were run and the convergence was checked by trace plot and visualized by Gelman and Rubin plot. Model selection was done by looking at the Deviance Information Criterion (DIC) to assess the goodness-of-fit and model complexity (Farnsworth and Ward, 2009). Table 4.1 summarizes the models with their level of complexity and value of DIC (the smaller the better). The results for the three models with the smallest DIC were selected for presentation in Table 4.2.

The models with smaller DIC values are those with altitude as covariate and with random components (Table 4.1). The final model (Model 7 in Table 4.1) selected as the best include altitude as covariate and two random components that is, random intercept (unstructured and structured) and random slope. The chosen model has a DIC value of 247.761 and allow for over-dispersion and spatial correlation through the use of the conditional autoregressive prior. The unstructured heterogeneity term (u_i) followed a normal distribution with a mean of 0 and variance σ, while the structured heterogeneity term (v_i), estimated using a normal distribution with a provincial dependent mean and variance weighted by adjacent provinces. In this model the SIR is smoothed locally

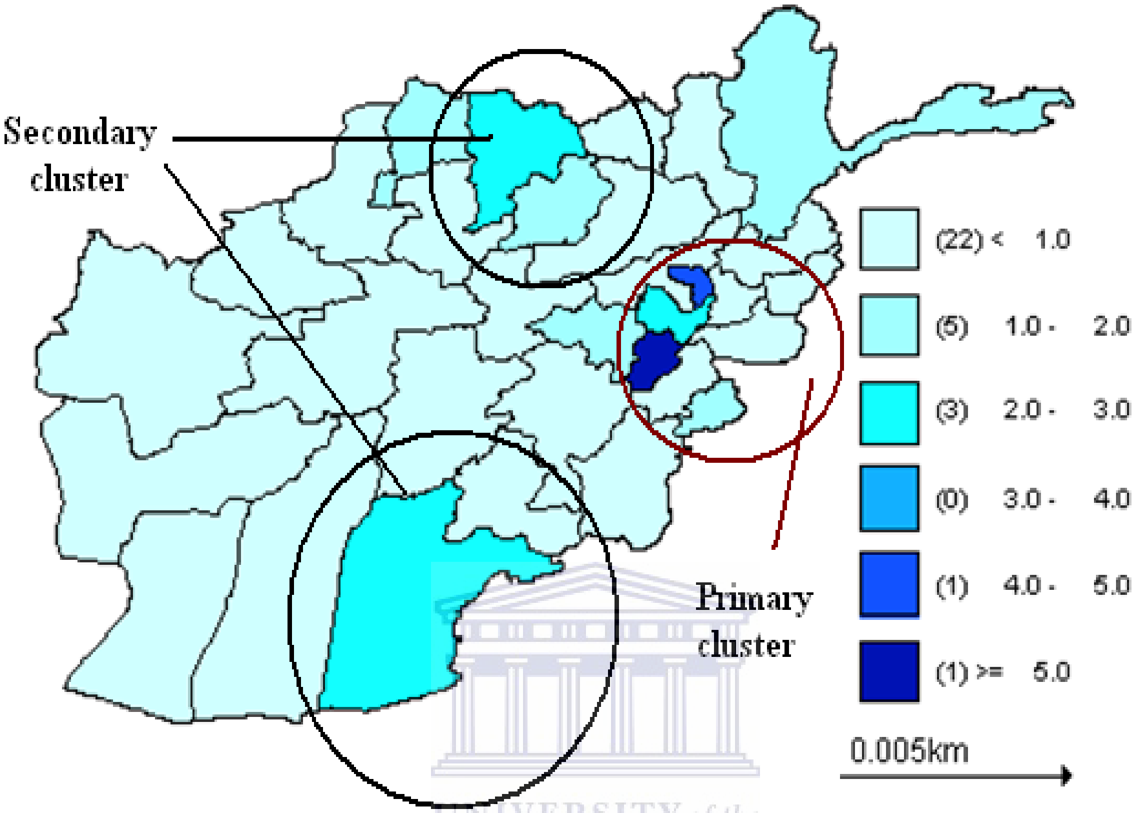


Figure 4.6: Relative risk estimated by hierarchical Bayesian model for Leishmaniais cases in Afghanistan from 2003 to 2009 with non-structured and spatial random intercept and random slope controlling for altitude and population.

towards the mean risk in the set of neighbouring areas (Lorenzo-Luaces Alvarez *et al.*, 2009).

Table 4.2 shows the summary statistics for the precision terms and posterior summaries of the final model. The relative risk (RR) from the final model, together with the clustered regions are displayed in Figure 4.6. Table 4.4 shows the relative risks and the credibility interval, higher RR was observed in North-Eastern, central and South-Eastern (Figure 4.6).

Table 4.4: Relative risks (with credibility interval) for the spatial Bayesian hierarchical CAR model per province.

Province	Relative risk	Credible interval		Random intercept	Random slope
		Lower	Upper		
Badakhshan	1.0620	1.0350	1.0890	0.0542	0.0137
Takhar	0.0160	0.0128	0.0195	-0.0033	0.0177
Jawzjan	1.0880	1.0510	1.1260	-0.0306	0.0071
Balkh	2.0820	2.0500	2.1160	-0.1017	0.0239
Kunduz	0.4788	0.4612	0.4965	0.0172	0.0187
Faryab	0.2715	0.2584	0.2851	-0.0003	-0.0221
Samangan	1.2170	1.1720	1.2630	-0.0275	0.0073
Baghlan	0.1783	0.1668	0.1899	-0.0213	0.0195
Sari Pul	0.0001	0.0000	0.0006	-0.0151	-0.0021
Nuristan	0.0001	0.0000	0.0010	-0.0032	0.0236
Badghis	0.0001	0.0000	0.0004	-0.0056	-0.0283
Parwan	0.9322	0.9033	0.9620	0.0516	0.0183
Hirat	0.4919	0.4788	0.5055	0.0528	-0.0355
Kunar	0.5935	0.5638	0.6239	0.0239	0.0402
Bamyan	0.0000	0.0000	0.0002	-0.0046	0.0006
Ghor	0.0000	0.0000	0.0002	-0.0033	-0.0230
Laghman	0.9375	0.9007	0.9751	0.0449	0.0224
Kapisa	4.1080	4.0280	4.1860	0.0348	0.0236
Kabul	2.0600	2.0410	2.0790	0.0833	0.0195
Wardak	1.2050	1.1690	1.2430	0.0140	0.0049
Nangarhar	0.6010	0.5846	0.6176	0.0278	0.0339
Uruzgan	0.0001	0.0000	0.0006	-0.0087	-0.0204
Logar	5.1800	5.0860	5.2730	0.0857	0.0125
Paktya	0.0000	0.0000	0.0002	-0.0022	0.0105
Ghazni	0.0000	0.0000	0.0001	-0.0030	-0.0031
Khost	1.8360	1.7900	1.8840	0.0056	0.0009
Farah	0.0001	0.0000	0.0005	-0.0095	-0.0386
Paktika	0.0000	0.0000	0.0003	-0.0023	-0.0043
Hilmand	0.1117	0.1030	0.1208	-0.0264	-0.0392
Zabul	0.0001	0.0000	0.0005	-0.0054	-0.0176
Kandahar	2.8790	2.8390	2.9200	-0.0306	-0.0378
Nimroz	0.0007	0.0000	0.0031	-0.0331	-0.0471

4.5 Discussion

The spatial scan statistic of Kulldorf for cluster detection and test of local clusters were used. The results confirm earlier findings using the Bayesian hierarchical analysis. The summary results from the spatial scan statistics identified eight provinces as primary cluster and another two as a temporary cluster. These provinces are mostly located in the North-Eastern and South-Eastern part of the country and are termed regions with high risk of the disease with a statistically significant p-value of < 0.0001.

The environment appears to play an important part in the transmission and occurrence of vector borne diseases, with areas in close proximity to each other having similar risk. The use of spatial statistics and geographical information systems in the study of geographical heterogeneity in hypothesized risk of Leishmaniasis in Afghanistan is crucial.

The model in this paper suggests the presence of excess risk of Leishmaniasis in the North-Eastern and South-Eastern regions of Afghanistan. The model includes a non-spatially structured component, an unstructured heterogeneity term with a normal prior distribution and gamma hyper-priors for the precision terms. The model also includes a spatially structured term with CAR (allowing for spatial dependencies in the estimation of relative risks) priors plus a random slope. Further epidemiological analysis that includes additional demographic and environmental variables could be explored to obtain a more consistent explanation.

The results confirm geographical heterogeneity of Leishmaniasis. This will enable governmental and non-governmental organization to better target health interventions and choose areas to implement control measures against Leishmaniasis in a more efficient way. This research is limited by the availability of data. Afghanistan is emerging from decades of war so the quality of data sets are questionable, under-reporting of the disease is likely and the sample sizes are limited. Further epidemiological research that incorporates additional demographic and environmental variables (temperature and wind) is called for to shed more light on the dynamics of this disease. An investigation using Poisson Kriging techniques will be conducted to explore other risk structures.

Chapter 5

An Exploratory look at Associated Factors of Poverty on Educational Attainment in Africa and In-Depth Multi-Level Modelling for Namibia

This chapter has been published as:

Adegboye, O.A. and Kotze, D. (2013). An exploratory look at Associated Factors of Poverty on Educational Attainment in Africa and In-depth Multi-level Modelling for Namibia. *Journal for Studies in Economics and Econometrics*, 37(1).

Summary

This study examines several indicator variables related to education and poverty in Africa from the Demographic and Health Surveys (DHS). Many have described income and education as one of the fundamental determinants of health and as one of the indicators for socio-economic status. Firstly, data from thirty-six African countries were explored, geographical heterogeneity of the countries were discussed. Secondly, an in-depth multilevel analyses was carried out on data for 72,230 respondents and from 5,436 households in the Namibia DHS (1992-2006). Results from statistical analyses indicate that age of household head, socio-economic status of household, parent's level of education, family size and the position of a child in the family play a significant role in the educational attainment of household members. It was found that these household level characteristics are important predictors of educational attainment. Thus, government policy aimed at reducing household level poverty should be implemented to alleviate the economic power at household level thereby increasing educational attainment.

5.1 Introduction

Access to education particularly in the developing countries has been discouraging. The United Nations (UN) Universal Declaration of Human Rights Article 26 states that everyone has the right to education (United Nations, 1948). The Jomtien 1990 declaration of the "World Conference on Education For All" stipulates that every person (child, youth and adult) shall be able to benefit from educational opportunities designed to meet his/her basic needs. Education has also been described as a tool for economic development and eradication of poverty. Schooling improves productivity, health and reduces negative features of life such as child labour as well as bringing empowerment (EFA Global Monitoring Report, 2002). Education paves the way to empower people to obtain access to jobs and higher wages. This in turn allows individuals to acquire resources (economic power) to access basic health facilities and, thus, improve the health of the population. It was reported that a country with a higher percentage of its youth in schools considerably reduces its risk of conflict (Collier, 2007). Many have described the link between income and education as one of the fundamental determinants of health and one of the indicators for socio-economic status.

A study based on cross sectional data from nine countries showed that an earnings inequality increases with educational inequality (Chiswick, 71). In developing countries, the male child is favoured to go to school rather than the girls. The number of siblings may affect the continuation of school for some other family member as this poses an alternative cost. Bledsoe *et al.* (1999) found an association between schooling and fertility in less developed countries. The mother's educational attainment plays an important role in the household and has a significant effect on her bargaining power and thus her drive for education.

The United Nations Educational Scientific and Cultural Organization (UNESCO) reported that Sub-Sahara Africa had an increase in its average enrolment from 54 per cent to 70 per cent between 1999 and 2006. In Namibia, although the "primary education net enrolment keeps improving at levels of 92,3% in 2007 to 98,3% in 2009, there is a worrying trend of not retaining the number of enrolled primary school learners in secondary education" (The Ombudsman, 2010, p. 2). The primary school completion rate in Namibia was about 80% in 2006 while the net enrollment rate in grades 1 to 7 had reached 92,3% in 2007 (Van der Berg and Moses, 2011). The Africa Recovery July 2000 report indicated that Tanzania has been more successful than many developing countries in achieving gender equality in education, with girls making up to 49,6 per cent of all enrolled primary school students in 1997. Further, the report noted that early marriage tends to cut short a girl's education at the upper primary and secondary levels in Tanzania. Other details from this report showed that 76% of children in Nigeria had access to primary school education; the southwest region recorded the highest percentage while the southeast recorded the lowest percentage. Accessibility to basic schooling and a region dummy could explain the 99% variation in income inequality in Nigeria and suggested that income redistribution in favour of the northern region will reduce income inequality in Nigeria (Alabi, 2008). The war in southern Sudan was associated with educational inequality in that country according to Deng (2003). Aluede (2008) discussed the variations in educational development in Nigeria; he noted that this disparity could be traced to historical educational development in Nigeria. Alabi and Abu (2008), in Alabi (2008) particularly attributed the persistent crisis in the Niger-Delta to low educational attainment of the people in that region. In 2004 Namibia was the country with the greatest income inequality in the world with a Gini coefficient of 0,7 (The Ombudsman, 2010; Levine and Roberts, 2008). The Central Bureau of Statistics of Namibia in 1996 reported that about 38% of the people were poor and 9% were severely poor (Levine and Roberts, 2008).

For the past five decades there has been a clamour for an increase in educational enrolment and attainment in developing countries. The first two goals of the UN Millennium Development Goals target the increase in school enrolment to be observed by 2015, an eradication of extreme poverty and hunger by the same date, and that all children must have access to and be able to complete primary school by this time.

The current study is structured as follows: Firstly, the performance of African countries in achieving these goals was descriptively examined, in space and time and to gain a better understanding of factors that need improvement. These are the factors that influence educational enrolment and, thus, educational attainment using empirical evidence to explain the relationships. It is important to understand how poverty affects the educational attainment and enrolment and to explore the socio-economic realities of poor households. It becomes necessary to gain an idea of education distributions, wealth distributions and their inequalities. Moreover, it is of interest to document trends in educational attainment and poverty using data from Demographic Health Surveys. During this investigation, other household characteristics: parent education, socio-economic status, family size, living conditions and location will be explored. Secondly, the exploratory analyses of the wider group of African countries' educational attainment will be followed by a representative in-depth analysis of associated factors of poverty on educational attainment in Namibia using data from Namibia Demographic and Health Surveys (1992-2006).

5.2 Data and Variables

5.2.1 Data

Demographic and Health Surveys (DHSs) have been conducted in more than 85 countries worldwide since 1984, including in Africa. The DHSs are based on national representative data and provide information on the population and health situation. The data are available for download on the website of ICF Macro International (`www.measuredhs.com`). The main interest is specifically to collect information at household and individual levels, and the information collected includes fertility, family planning, maternal and child health, educational characteristics, wealth index, ownership of basic facilities and location.

The data sets were restricted to data drawn from Phase II (1988-1993) to Phase V (2003-2008), due to the fact that Phase I surveys are outdated and sometimes do not match with the current questionnaire used in other phases of the survey. Data are not available for all African countries; therefore, only data sets on countries where available and for different phases were used.

Table 5.1: Distribution of the survey data used in this study

Survey Phase	Observation	Male	Female	Number of Households
DHS 2	261205	50.80	49.20	51788
DHS 3	373865	50.99	49.01	81820
DHS 4	596607	50.65	49.35	119237
DHS 5	485268	50.72	49.28	129006

Similar variables were extracted from all surveys at household and individual level. Individuals in the same household shared the same characteristics such as age of parents, parent's education, number of rooms, wealth index, household head etc. The data sets were aggregated at three levels: household, national and continent, and using common survey indicators across countries and DHS phases.

A total of 1,716,945 observations from 381,851 households and from a total of 36 African countries were collected from 1988 (DHS II) to 2008 (DHS V). Table 5.2 and (Figure A.1 and Table A.1) in the Appendix summarize the data characteristics. Only respondents between the ages of 5 years to 30 years are considered , because in most African countries this is referred to as the school age (especially primary to tertiary). This data provide information on individual children and household characteristics.

5.2.2 Variables Definitions

Educational Attainment

Several definitions of educational attainment have been discussed in the literature (Thomas *et al.*, 2001; Gardner, 1998; Psacharopoulos and Arriagada, 1986; Barro and Lee, 2010). These centred on using different indicators, like years of schooling completed, level of education completed, ability to read (literacy) and so on.

Several problems are associated with obtaining the educational attainment indicator.

In addition to problems associated with the definition of educational attainment, most African countries have different systems of education, as summarized in Figure A.2 in the Appendix. The age of entrance to primary school varies across different countries, usually 6-7 years but can be as low as 4 years in Morocco. Moreover, the length of primary education also ranges from 6 years in many countries to 9 years in Mali, Morocco and Egypt. In Namibia, 84% of children ages 6-12 attends primary school (83% of boys and 85% of girls) (Education Policy and Data Center, 2012).

The years of compulsory schooling differ in a number of the countries under study; it ranges from as low as 4 years in Senegal to 10 years in Comoros and Ethiopia. Egypt, Mali and Morocco have the lowest number of years of secondary education (3 years) with some countries having up to 7 years of secondary education (such as Benin, Burkina Faso, Cameroon, Central Africa, Chad, Comoros, Congo Brazzaville, Niger, Namibia, Senegal and Togo). In order to minimize problems associated with indicators not reported for some countries and the differences in education systems across African countries, only the levels of education completed as an indicator for educational attainment were considered. That is, the highest level of education completed by an individual is referred to as his/her educational attainment indicator variable and was divided into no education, primary, secondary education and above secondary education. These may be further sub-divided as need arises for appropriate statistical analysis, such as proportion of the population with no education.

Education Gini

The Education Gini was used as a measure of inequality in educational attainment while the standard deviation of schooling measures absolute dispersion. The methods of Thomas *et al.* (2001) using both direct formulae were adapted as follows:

Gini: $E_L = \frac{1}{n}\sum_{i=2}^{1-1}\sum_{j=1}^{1-1} P_i \mid y_i - y_j \mid P_j$

Educational attainment (mean education levels): $\mu = \sum_{i=n}^{n} P_i y_i$

Standard deviation: $\sigma = \sqrt{\sum_{i=1}^{n} P_i (y_i - \mu)^2}$

where E_L is the education Gini based on educational attainment distribution, large populations, μ is the average educational attainment for the concerned population, P_i and P_j are the proportions of the population with certain levels of educational attainment, while y_i and y_j are the different educational attainment levels and n is the number of levels/categories in attainment data.

The Gini value can be interpreted as follows: a Gini value of zero implies a perfect equality while a Gini value of one implies a perfect inequality. The severity of the inequality depends on how close the Gini coefficient is to 1.

Poverty Indicators

Providing a single, generally accepted definition of poverty and its measurement is a very difficult task. Townsend (1979) defined poverty as ”the absence or inadequacy of those diets, amenities, standards, services and activities which are common or customary in society”. Many authors have proposed ways of estimating poverty: Gibson (2002) suggested monetary or non-monetary measures when using poverty-focused household surveys, Ravallion (1994) suggested the ”well-being” for poverty and further argued for welfarist and the non-welfarist approaches.

Measuring poverty indicators over time from household surveys have setbacks as data may not be collected from the same household over time presenting a repeated cross-sectional survey. In order to make comparison analyses across countries, any poverty indicator must be available in all surveys. DHS does not provide information on household income or expenditures; however, a poverty profile can be constructed using information on household assets.

DHS provide wealth quintiles as a measure of economic status based on all household assets and utility services, including country-specific items and sometimes ownership of agricultural land and domestic servants while excluding family size and age structure.

The wealth quintiles were calculated using principal component analysis (PCA); this procedure first standardizes the indicator variables (calculating z-scores), then, the factor coefficient scores (factor loadings) are calculated, and finally, for each household, the indicator values are multiplied by the loadings and summed to produce the household's index value. In this process, only the first of the factors produced is used to represent the wealth index. The resulting sum is itself a standardized score with a mean of zero and a standard deviation of one (Rutstein and Kiersten, 1994). The idea of PCA is to find an orthogonal transformation of the original variables (vector of assets correspondent to every household) to a new set of uncorrelated variables called principal components, which are ranked in decreasing order of importance (Chatfield and Collins, 1980).

For this study, the DHS wealth quintile was used. These poverty indicators (the wealth quintiles or wealth index score) were estimated relatively, that is, the proportion of the current status (for example, poorest 20 per cent of the population in each year). The advantage of using DHS wealth quintiles is that it allows a potential comparison across countries and over time, and can be linked to other indicators like education and health. In addition to the wealth index, family size and household structure (family structure) were included in the poverty indicator. Other explanatory variables considered are as follows: number of wives, age, sampling weight, gender, country, household characteristics, gender of household head, own child, type of residence, literacy gap, parent educational attainment, sanitation and access to clean water.

5.3 Descriptive Analyses

Various data exploration techniques were applied: summary statistics were obtained in order to gain insight into the data. Plots for various covariates were used to investigate any trend or pattern in the indicator variables (See Appendix A.1 for the list of variables). Significant association exists between poverty and health outcomes; wealthy people have limitless access to good health facilities and good education and will most likely be better educated and live longer. An investigation of the education distribution can explain aspects about the poverty level in a household, region and country at large. Furthermore, the wealth distribution and education distribution may also play an important role in health distribution.

Education has been described as one of the indicators for national socio-economic development and that the proportion of literate population is a good indicator of development (Gardner, 1998). As it is shown in Appendix A.3, the proportion of literate population in Africa has risen over time (survey phase), while population with no education has decreased. Educational indicators will be used to make clear the distribution and inequality of educational attainment across countries.

The plots of educational attainment in Appendix A.4 showed that in DHS II (1988-1993), Burkina Faso, Morocco, Niger and Senegal had more people with no education than primary education. Over the years, many countries have seen an increase in primary enrolment and have seen a rise in the proportion of the population with at least primary education, but countries like Cote D'Ivoire, Liberia, Mali and Niger, however, still have more people with no education. Although, Namibia has seen an increase in the enrolment rates over the years, the proportion of the population with no education was not as low (6%) as it was in DHS II (1992): 8% in DHS IV (2000) and 9% in DHS V (2006) (not shown).

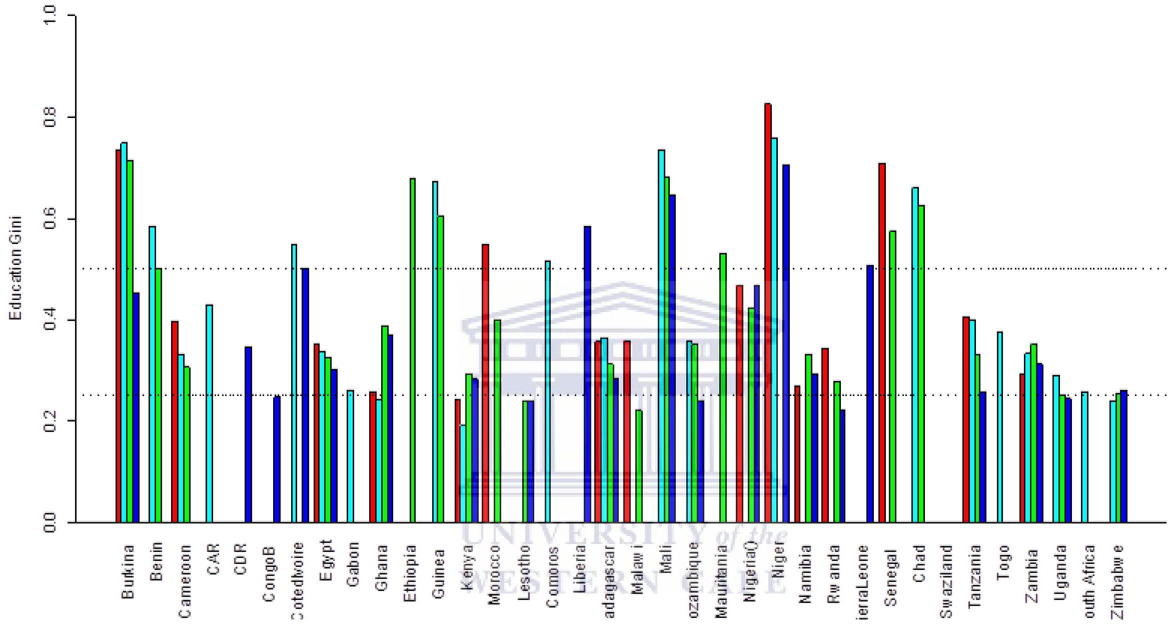


Figure 5.1: Education Gini for countries under study across different survey phases
Source: Authors' calculation (See Appendix 1)

The Education Gini coefficient was calculated for each country to allow for country wide comparison and to assess countries with greater or lower educational inequality and to check the dynamics over time. Figure 5.1 gives a graphical display of the Gini coefficients across the 36 countries under study over different survey periods (DHS II to DHS V); this allows for comparison and assessment of education inequality among the countries (see Appendix A.5). Most of the countries under study have been experiencing a steady decrease in educational inequality; some still record a high level of inequality. For example, Burkina Faso, Chad, Ethiopia, Guinea, Mali, Niger and Senegal have an

Education Gini of at least 0,60 (Figure 5.1). The educational inequality in Namibia is one of the lowest in Africa with a Gini coefficient of 0,26 in DHS II (1992), 0,33 in DHS IV (2000) and 0,29 in DHS V (2006) (Author's calculation, Figure 5.1 and Appendix A.5). This can be attributed to the changes in enrolment between 1992 and 2006. Looking at the Lorenz curve (not shown), Namibia, South Africa, Zimbabwe and Lesotho have the smallest area in DHS II, III, IV and V respectively. Niger has the highest inequality in DHS II, DHS III and DHS V with a Gini value of 0,83, 0,76 and 0,70 respectively, although the Gini value is decreasing.

Parent perception of education plays an important role in children's education; there exists a strong correlation between household head with no education and children with no education (Figure 5.2). Most countries clustered in the third quadrant of the graph are countries with less than 50% of the population with no education. In the extreme case, this is displayed in the first quadrant, indicating countries with over 50% population with no education (household head and children); Burkina Faso, Mali and Niger (Figure 5.2).

Information about the distribution of the wealth index quintile by country over time is displayed in Appendix A.6. The proportion of the poorest 40% of the population increased from less than 30% in DHS II in Burkina Faso to about 40% in DHS V. This is lower than the proportion of the richest 20%. The proportion of the poorest 40% generally has a share of above 25% of the population for the countries under study at all survey periods except Cote D'Ivore in DHS III, Mozambique and Nigeria in DHS V (Appendix A.6). With high income inequality, the proportion of households in the poorest 40% in Namibia was about 42% in 1992 (DHS II) but dropped to about 33% in 2000 (DHS IV) and increased in 2006 (DHS V) to about 38% (Authors calculation: Appendix A.5)

Gender differences in educational attainment (literacy gap) can be measured by the ratio of the proportion of male population with at least primary education to that of their

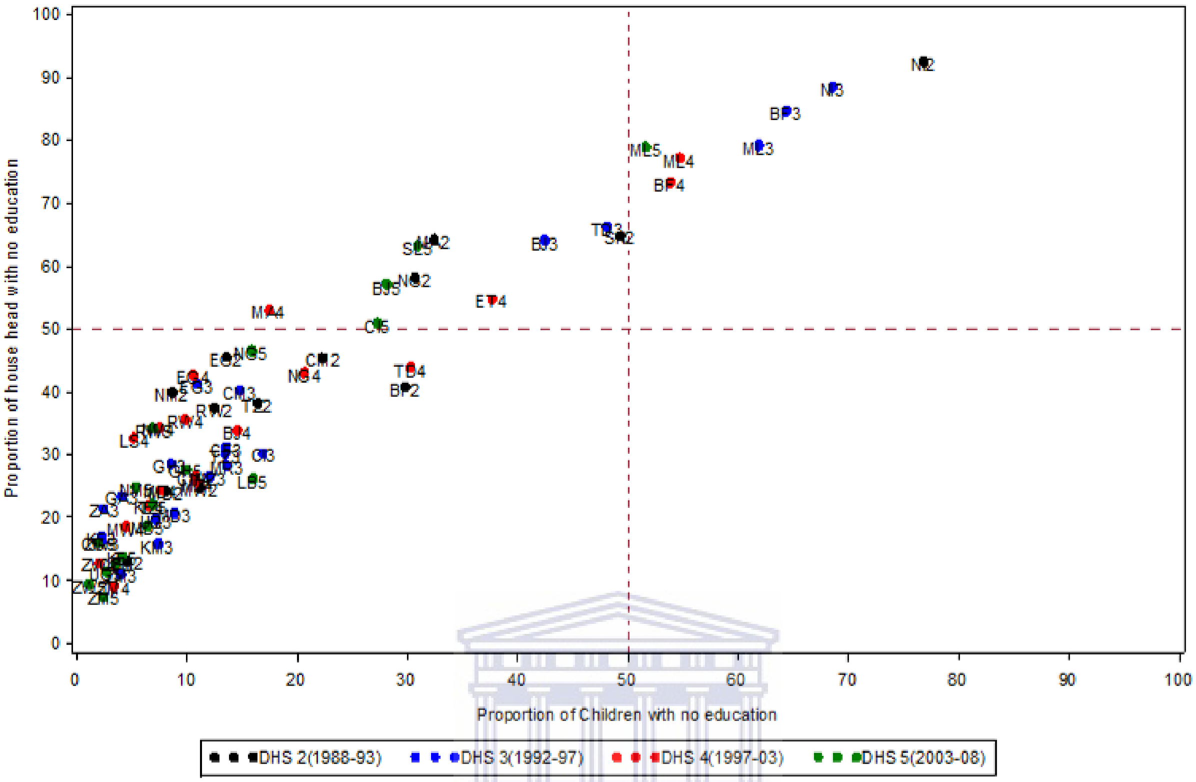


Figure 5.2: Relationship between the proportion of household members (children) with no education and the proportion of household heads with no education (country ID indicated on the dot for different survey phase)

female counterpart. Appendix A.2 illustrates the result from these estimates; in DHSII, Niger recorded the highest difference in male to female ratio of educational attainment of about 25%.

The proportion of males with at least primary education was about 35% more than females in Nigeria for DHS V. For most of the countries the percentage of males are more than that of females except Namibia (DHS II), South Africa (DHS III), Lesotho, Namibia and Rwanda (DHS IV) and Congo Brazzaville, Lesotho, Namibia and Rwanda (DHS V). These countries have seen significant changes in gender equality in educational attainment over time. Although there are fluctuations, it is safe to say that most of the countries under study are bridging the gap in the gender differences in educational attainment.

These fluctuations may be attributed to changes in educational policies and systems over the years.

Relationship between gender differences in the literacy gap and the educational attainment of a country cannot be overemphasized. There is a positive relationship between literacy gap and population with no education. Countries with higher literacy gap (above 20%), like Niger, Burkina Faso, Mali and Benin, with the exception of Nigeria, have about 65% or higher percentage with no education. Similarly, these countries also have a lower proportion of the population with at least primary education. Namibia is one of the few countries in Africa that has achieved gender parity in education with fairly more females with at least primary education than males: 4% in 1992, 2,2 % in 2000 and 2% in 2006.

The lack of access to and availability of clean, safe drinking water and clean sanitation is another major cause of poverty in Africa. It is difficult to go to school when time is spent on a daily basis finding and transporting water. Many countries still cannot provide these basic amenities to their people. Moreover, several diseases are attributed to unhygienic living conditions and dirty water which, in turn, lead to poor health and poor productivity; poverty is inevitable in this situation. However, the percentage of population without access to clean and safe water is decreasing over time in many of the countries under study. A country like Nigeria has experienced a drastic reduction in the percentage of population without access to clean and safe water from above 95% in DHS II to below 20% in DHS V. Egypt has maintained a low percentage over time and has the lowest percentage of people without access to clean water (below 10%) in DHS V, while Kenya, Liberia, Madagascar, Mali, Sierra Leon, Niger and Uganda still have around 70% of the population without access to clean water (source: authors' calculation). In 1992 about 50% of households in Namibia did not have access to clean and safe water, this reduced to less than 30% in 2006. The average household size also reduced from about

10 in 1992 to 7 household members in 2006.

5.4 A Case Study of Namibia DHS II, DHS IV and DHS V.

5.4.1 Methodology

Marginal models are often a better choice when dealing with dependencies in the data set, without the need for complex and unattainable assumptions, as found in some other methods and can be used to answer research questions directly at the intended marginal level. In this study, individuals are nested within households; data were collected for every eligible member of the study at household level. Ignoring the structure of the data may result in parameter estimates to be asymptotically biased. In recognizing the structure of the survey data, multi-level modelling of the individuals nested within households was used. Multi-level analysis allows for information to be pulled together from multiple levels and enable interrelationships between the different levels to be explored and facilitated for overall interpretation. Household members shared some information collected at household level, thus, it would expected that, there exist dependencies or correlation among these responses (i.e. within subject dependency). These dependencies or correlations must be accounted for by methods appropriate to the data (Diggle *et al.*, 1994). Statistical methods that take the dependencies in the data into account should be used. Several models have been proposed for the analysis of such data. Most of these are extensions of the well-known logistic regression, a particular case of generalized linear models with logit or probit link functions (McCullagh and Nelder, 1989). They are usually classified into marginal or random-effects models.

A simple model for discrete data may assume a Poisson log-linear relationship between rates and other explanatory variables. Different approaches are available for implementing a multi-level analysis for cluster models or correlated data, that is, methods that simultaneously model all the outcomes elicited from an individual.

Let Y_{ij} denote a binary outcome corresponding to the j^{th} household ($j = 1ton_i$)of the i^{th} individual. Let also X_{ij} be a design matrix of covariates ($1p$) vector, with the first element being 1 for the intercept. The marginal model, also called the population-averaged model (Zeger *et al.*, 1988), estimates the model, thus:

$logit(E(Y_{ij} \mid X_{ij})) = logit(P(Y_{ij} = 1 \mid X_{ij})) = X_{ij}\beta$ and under the marginal model the Odds Ratio $= exp(\beta)$.

The marginal model supposes that the relationship between the outcome Y and the covariate X is the same for all subjects. Moreover, dependencies between observations within the same household are handled by fitting the vector of parameters, β, using the Generalized Estimating Equations (GEE) (Liang and Zeger, 1986), wherein the covariance matrix is structured by using a working correlation matrix $R(\alpha)$, fully specified by the vector of parameters β. This working correlation matrix is assumed to be the same for all the subjects, reflecting an average dependence among the observations for all subjects. In the marginal model, several specific choices of the structure of the working correlation matrix $R(\alpha)$ are possible (seeLiang and Zeger (1986)). An advantage of the marginal model, as demonstrated by Liang and Zeger (1986), is that β and their robust variance are consistent (the estimator converges towards the parameter being estimated as the sample size increases) even when the correlation structure is misspecified. However, choosing the working correlation structure closest to the true structure increases the statistical efficiency of the parameter estimator. Consequently, it is recommended to specify the working correlation as accurately as possible, based on the knowledge of the process (Albert, 1999).

5.5 Results

This section discusses statistical analysis of a preselected data set, by looking at the effect of some indicators on educational attainment. In Section 5.2.2, some indicator variables were explored and countries with little or no improvement in these variables were identified. In order to provide an in depth analysis of these variables and assess their significant influence, the statistical analysis using Generalized Estimating Equations (Liang and Zeger, 1986) will be carried out on a sample of the data. The study will require a country with at least three survey data sets as an example, hence, the choice of Namibia. Namibia has data available for DHS II (1992), DHS IV (2000) and DHS V (2006). There is a total of 20,173 observations in DHS II, 22,332 in DHS IV and 29,725 in DHS V from 1410, 1687, and 2339 households respectively. The effect of wealth, parents' educational characteristics, household head characteristic, family characteristics and individual characteristics will be investigated.

Table 5.2 presents the results from the analyses of DHS data from Namibia in 1992 (DHS II), 2000 (DHS IV) and 2006 (DHS V) respectively. Synonymous with the saying ”two heads are better than one”, children in the household with a single parent have a lower probability of attaining at least primary education compared with children in the household with both parents. Children in a household with both parents have a higher chance of having at least primary education, and these probabilities increase over time. The implication is that the contribution of having both parents on educational attainment in Namibia cannot be overemphasized. The odds increased from about 1 in 1992 to 21 in 2006.

The position of a child in the family plays a very important role in who goes to school first and who does the domestic work. The results here showed that the lower ”ranking” a child has in a family, the less chance he/she has to attain at least primary education.

That implied the first child has more priority than others. Also, these significant results indicate an increase in its importance over time.

Interestingly, the gender of the household head does not play a significant role in educational attainment in Namibia. The age of household head and number of household members were both positive and significantly associated with educational attainment. These results must be carefully interpreted, as the definition of household plays a very important role here.

The number of wives in the household, which is correlated with number of household members, has a significant influence on educational attainment of the children. The influence of number of wives was not significant in 1992 and 2000, but was negatively significant in 2006. Therefore, the more the number of wives in a household increases, the less are the chances of their children attaining at least primary education.

Gender has been a very important variable, especially in Africa. Interestingly, this may be a prejudice in the case of Namibia; in 1992 the effect gender was significant and the chance of a male child was about 1.2 times that of female child of attaining at least primary education. By 2006, the chances of a male child have dropped and gender is no longer a significant variable. Looking at the type of residence, the likelihood of people living in rural areas attaining at least primary education is less than that of those living in urban areas.

Table 5.2: Results from the modelling of educational attainment for Namibia: DHS II-DHS V (No education against at least primary education)

Parameter	DHS II (1992)			DHS IV (2000)			DHS V (2006/2007)		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
Intercept	-0.0372	0.8473	0.9649	-1.3180	1.3190	0.3178	3.2441	2.1362	0.1289
Single Parent (YES)	-0.2100	0.1641	0.2006	-0.2040	0.2217	0.3583	-0.1985	0.2415	0.4110
Rank of Child	-0.2884	0.0233	< .0001	-0.4830	0.0449	< .0001	-0.6620	0.0314	< .0001
Sex of head (MALE)	0.4273	0.2390	0.0738	0.4019	0.5089	0.4297	-1.8242	1.0446	0.0808
Age of head	0.0450	0.0037	< .0001	0.0635	0.0044	< .0001	0.0799	0.0047	< .0001
Number of household member	0.1662	0.0203	< .0001	0.3284	0.0381	< .0001	0.4551	0.0241	< .0001
Number of wives	-0.1732	0.1216	0.1544	-0.1310	0.4824	0.7857	-2.2922	1.0419	0.0278
Mother alive (NO)	-0.8234	0.5325	0.1221	-1.0830	0.4686	0.0209	-2.1744	0.4260	< .0001
Father alive (NO)	-1.3950	0.4519	0.0020	-1.0720	0.3997	0.0073	-0.9289	0.4285	0.0302
Gender (MALE)	0.1502	0.0623	0.0158	0.0749	0.0769	0.3298	0.0815	0.0552	0.1399
Type of residence (RURAL)	-0.2329	0.1311	0.0757	-0.4150	0.1153	0.0003	-0.1916	0.0822	0.0197
Wealth Index	0.0076	0.0406	0.0415	0.0536	0.0393	0.0032	0.0694	0.0332	0.0367
Father's highest education	0.0465	0.0654	0.4771	0.2824	0.0448	< .0001	0.2293	0.0574	< .0001
Mother's highest education	0.2476	0.0647	0.0001	0.2244	0.0553	< .0001	0.2703	0.0604	< .0001

5.6 Conclusion

Against the descriptive background of educational attainment in African countries, an in-depth analysis of Namibia revealed its status with respect to poverty factors. Results from the analysis of correlated DHS survey data from Namibia indicated that the following indicator variables had a significant effect on educational attainment: rank of the child, age of household head, number of household members, father alive, gender, parent's education and wealth index. The wealth index quintile also has a positively significant influence on educational attainment. As one would expect, the richer the household is, the more they are likely to have at least primary education.

In conclusion, the case study of Namibia from 1992-2006 indicated that socio-economic indicators play a significant role in educational attainment. Families with higher socio-economic resources tend to send their children to school more often than poorer families, and the parent's level of education may also play an important role in their perception of the value for schooling. In developing countries, parents are bestowed with the sole responsibility of sending their children to school; therefore they weigh the immediate cost carefully (Jannie and S., 2009). Families with both parents gainfully employed may have higher bargaining power than a family with one parent providing for the entire family. Further, family size plays a very important role and impacts on the wealth distribution in the household. The position of a child in the family may affect the continuation of schooling for other members of the family.

Finally, the idea of this study was to provide an empirical analysis of geographical heterogeneity of African countries in educational attainment and poverty with special reference to Namibia as case study. Further studies will include investigating statistical issues as related to parameter estimation techniques and missing data.

Chapter 6

Multi-year Trend Analysis of Childhood Immunization Uptake and Coverage in Nigeria

This chapter has been published as:

Adegboye, O.A., Kotze, D., and Adegboye, O.A. (2013). Multi-year Trend Analysis of Childhood Immunization Uptake and Coverage in Nigeria. *Journal of Biosocial Sciences*, 45(4). doi:10.1017/S0021932013000254

Summary

As a leading indicator of child health, under-five mortality was incorporated in the United Nations Millennium Development Goals with the aim of reducing the rate by two-thirds between 1990 and 2015. Under-five mortality in Nigeria is alarmingly high, and many of the diseases that result in mortality are vaccine preventable. This study evaluates the uptake of childhood immunization in Nigeria from 1990 to 2008. A multi-year trend analysis was carried out using alternating logistic regression on 46,130 children nested within 17,380 mothers in 1938 communities from the Nigerian Demographic and Health Surveys from 1990 to 2008. The findings reveal that the mother-level and community-level variability were significantly associated with immunization uptake in Nigeria. The model also indicates that children delivered at private hospitals had a higher chance of being immunized than children who were delivered at home. Children from the poorest families (who are more likely to be delivered at home) have a lower chance of being immunized than those from the richest families (OR=0.712; 95% CI, 0.641-0.792). Similarly, the chance of children with a mother with no education being immunized is decreased by 17% compared with children whose mother has at least a primary education. In the same way, the mother's occupation (those that are gainfully employed) and age are statistically significantly associated with higher odds of a child being immunized. The effect of household head on child immunization differs significantly between households with female- and male-headed households. The statistical significance of the community-survey year interaction term suggests an increase in the odds of a child being immunized over the years and spread over communities. Evidence-based policy should lay more emphasis on mother- and community-level risk factors in order to increase immunization coverage among Nigerian children.

6.1 Introduction

Under-five mortality in Africa is at an alarming level, with about 40% of global deaths occurring in sub-Sahara Africa. As a leading indicator of level of child health, the under-five mortality rate has been incorporated by most United Nations (UN) member states as a UN Millennium Development Goal (MDG) indicator (Goal 4), with the goal of reducing the under-five mortality rate by two-thirds by 2015 (United Nations, n.d.). Although the developmental agenda emboldened in the MDGs addresses all countries of the world, there can be no doubt that sub-Saharan African countries stand to benefit most from the promotion of its principles as compared with other regions of the world. In comparison with the rest of the world, sub-Saharan African countries have the highest rates of poverty and illiteracy as well as the highest rates of child and maternal mortality, and hence the importance of MDG-4 to this region.

Despite the decrease in the under-five mortality rate in Nigeria from 213 per 1000 live births in 1990 to 143 per 1000 live births in 2010, Nigerian under-five mortality was ranked 12th worldwide in 2010 (Adegboye, 2010b; UNICEF, 2010). The total fertility rate was 4.82 (average number of children per woman in 2010) and life expectancy at birth was 51 years (UNICEF, 2010). Several factors have been identified to have significant impact on under-five mortality: pneumonia, diarrhoea, poverty, lack of safe water and poor sanitation, mother's education and mother's age (Adegboye, 2010b). Others have found poor socioeconomic development, a weak health care system and low socio-cultural barriers to care utilization as associated risk factors for the high rate of maternal and child mortality in Nigeria (Ogunjimi *et al.*, 2012); people's beliefs, attitudes and behavioural practices (Ogunjuyigbe, 2004) add to the burden. Bosch-Capblanch *et al.* (2012) found caregiver's and partner's education, and caregiver's tetanus toxoid status to be strongly associated with being unvaccinated. Also, Kayode *et al.* (2012) reported that the mother's age at first marriage plays an important role in reducing under-five mortality in Nigeria;

other favourable practices include health-seeking behaviour, breast-feeding children for more than 18 months, use of contraception, small family size, having one wife, low birth order, normal birth weight, child spacing, living in urban areas and good sanitation.

Immunization remains the most important and cost-effective public health intervention, protecting individuals, families and communities from vaccine-preventable diseases and conferring herd immunity thereby breaking cycles of disease transmission. It also serves as a response to outbreaks of diseases as well as an in-road for other primary health care services. However, routine immunization coverage in Nigeria has continued to fall below average. Preventable infectious diseases such as tuberculosis, poliomyelitis, diphtheria, tetanus and measles are the main causes of morbidity and mortality in children, especially in developing countries like Nigeria. Vaccination is a very effective way of reducing (and where possible eradicating) the spread of these preventable diseases. The Nigeria Expanded Program on Immunization (EPI) introduced in 1979, was restructured and renamed the National Programme on Immunization (NPI) in 1997 and later merged with the National Primary Health Care Development Agency (NPHCDA) in 2007. The main goal of the immunization programme is to develop, promote and sustain the immunization programme towards reducing childhood morbidity and mortality through adequate immunization coverage (90% by the year 2020) of at-risk populations. It was designed to reach children from birth to five years of age, all pregnant women and at-risk populations. It is expected that a fully vaccinated child should have received a Bacillus Calmette-Gurin (BCG), three doses of Diphtheria Pertussis and Tetanus (DPT), at least three doses of poliomyelitis and one dose of measles vaccines by their first birthday (National Population Commission (NPC) [Nigeria] and ICF Macro, 2009) (see Table 6.1 for details of vaccines). In 2004, the country included hepatitis B and yellow fever vaccines in its immunization schedule. By 2012 the *Haemophilus influenzae* type b ’Hib’ vaccine was added to form a pentavalent vaccine consisting of diphtheria, pertussis, tetanus, hepatitis B and *Haemophilus influenzae* type b.

Table 6.1: Schedule of antigens for children under the age 12 months

Contact No.	Antigens	Schedule	Dosage
First	Bacillus Calmette-Guerin (BCG)	At birth	0.05ml
Second	Diphtheria, pertussis and tetanus (DPT) I	6 weeks	0.5ml
Second	Oral polio vaccine (OPV) I	6 weeks	2 drops
Third	Diphtheria, pertussis and tetanus (DPT) II	10 weeks	0.5ml
Third	Oral polio vaccine (OPV) II	10 weeks	2 drops
Fourth	Diphtheria, pertussis and tetanus (DPT) III	14 weeks	0.5ml
Fourth	Oral polio vaccine (OPV) III	14 weeks	2 drops
Fifth	Measles	9 months	0.5ml

The use of poliomyelitis vaccines has eliminated the disease in many countries. After the World Health Assembly in 1988 resolved to eradicate poliomyelitis globally, the number of poliomyelitis-endemic countries reduced from more than 125 countries in 1988 to four countries (Afghanistan, India, Nigeria and Pakistan) in August 2008 (WHO, 2008). The new statistics from WHO showed that only Afghanistan, Nigeria and Pakistan remain poliomyelitis endemic (with Nigeria at the top of the list), while Chad and Niger are classified as non-endemic countries (WHO, 2012a). Although there has been a 71% global decrease in deaths due to measles from 542,000 in 2000 to 158,000 in 2011, while new cases dropped by 58% during the same period, this success story due to vaccination is dented by the high number of children that did not receive the first-dose measles vaccine in 2011 (20 million worldwide with 1.7 million alone in Nigeria) (WHO, 2012b). Nigeria recorded the third highest number of new cases of measles in 2011 (18,843 cases), preceded by India (29,339 cases) and DRC with 134,042 new cases.

Success towards achieving the target of having 80% or above of children fully immunized is still a problem. Coverage in many parts of Nigeria has fallen below 50% (Antai, 2009; Kunle-Olowu *et al.*, 2011; Abdulraheem *et al.*, 2011). The decline in the attainment of the universal child immunization target in Nigeria in 1990 can be attributed to a number of factors: from political will to poor service delivery, culture, funding, community involvement and beliefs (Federal Ministry of Health, Nigeria). Problems of immunization uptake in Nigeria have been associated with mothers' poor knowledge of immunization against targeted diseases, parents' concern about immunization safety, long waiting time at the health facility and long distance from the hospital (Maekawa *et al.*, 2007; Abdulraheem *et al.*, 2011). Apart from these problems, false believe in contraindications to immunization like catarrh and mild fever in the child at the time of immunization, failure to administer simultaneously all vaccines for which the child was eligible and lack of information on the vaccination regimen are reported causes of missed opportunities to immunize in Nigeria (Kabir, M. and Iliyasu, Z. and Abubakar, I. S. and Nwosuh, J. I., 2004; Adeiga *et al.*, 2005; Onyiriuka, 2005; Anah *et al.*, 2006).

The safety of pregnant women and their babies depends on the success of the programme. Wammanda *et al.* (2011) reported that only 22% of children receive their BCG within the first 3 days of life and 36.2% within the first 7 days of life. Adeiga *et al.* (2005) indicated that reasons for failure to immunize or complete the immunization of children included poor knowledge of immunization and belief about immunization in 47%, lack of information in 40.7% and lack of motivation in 11.6%, and only 11% of the children in their study were not vaccinated against measles in the Lagos metropolis.

The Demographic and Health Survey (DHS) is a national representative survey that provides information on the population and health situation of a country. The main interest of the survey is specifically to collect information on fertility, family planning, maternal and child health, immunization, educational characteristics, wealth index, ownership of basic facilities and HIV/AIDS at household and individual levels. For this study, similar variables were extracted from all four surveys conducted in 1990, 1999, 2003 and 2008 at the community- and mother-level. Clusters were introduced at two levels: first, children sharing the same mother level characteristics such as age of parents, parents' education, number of rooms, wealth index, household head and so on; second, children from the same community sharing the some community-level characteristics, such as availability of and distance to health care facilities and place of residence. This study's outcome variable is whether a child has been fully immunized or not, which is at the child's level but nested within mother and within a community, thus introducing dependency in the data.

For clustered binary outcomes, several approaches have been suggested: E.g. Generalized Estimating Equations (GEE) techniques developed by Liang and Zeger (1986) to extend the Generalized Linear Models (GLM) to accommodate correlated data. Dependencies between observations are handled by fitting the vector of parameters, using the GEE techniques, wherein the covariance matrix is structured by using a working correlation matrix fully specified by the vector of parameters. It is often preferred to talk in terms of association rather than correlation when dealing with a binary response and it is important to make inferences about the association between any two measurements within a cluster. Carey *et al.* (1993) proposed Alternating Logistic Regression (ALR) as an alternative to GEE. Alternating Logistic Regression simultaneously regresses the response on explanatory variables as well as modelling the association among responses in terms of pair-wise odds ratios. This method uses the odds ratio to capture association between categorical outcomes. The odds ratio is a particularly straightforward measure to capture association between categorical outcomes (Molenberghs and Lesaffre, 1994; Fitzmaurice *et al.*, 2004). These models estimate Pair-Wise Odds Ratios (POR) to capture the association structure between clustered binary data (Diggle *et al.*, 2002). (Molenberghs and Verbeke, 2006, p. 217) praised ALR as a tool to fit a marginal model based on odds ratios in a way that inferences can be made, not only about the marginal parameters, but about the pair-wise associations as well.

The goal of this study was to identify the associated risk factors and estimate the mother- and community-level clustering of the immunization status of children between 12 months and 59 months of age, in order to evaluate the impact of immunization programmes and assess their coverage in Nigeria.

6.2 Methods

6.2.1 Data Sources

The data for this study were obtained from the Nigerian Demographic and Health Surveys (NDHS) implemented by the National Population Commission with technical support from ICF Macro. The NDHS started in 1990 and follow-up surveys were done in 1999, 2003 and 2008. The information collected included key health indicators for women aged 15-49 years, men aged 15-59 years and children between the ages of 0 and 5 years (National Population Commission (NPC) [Nigeria] and ICF Macro, 2009). These are national representative data that are available for download with permission from the website of Measure DHS (http://www.measuredhs.com/).

The data sets consist of 46,130 children aged 12-59 months from 17,380 mothers in 1938 communities. The motivation for assessing the immunization status of children between the ages of 12 and 59 months is that: (i) generally all children should have completed the Nigerian immunization schedule by 12 months of age; (ii) most cases of vaccine-preventable death occur before the age of 5 years (under-five mortality rate is a well known child health index). Moreover, those children who are not immunized within 'the standard timeframe' may not necessarily get immunized at an older age because the parents may presume that they are out of the immunization age range based on the television and radio enlightenment jingles that emphasize the timeframe for immunization.

The outcome variable is whether a child has fully completed his/her immunization schedule by receiving all eight doses of antigens or not (Table 6.1). The other explanatory variables include the child-, mother- and community-specific variables. There are two levels of clustering used in the study: mother and community. Communities were identified using the Primary Sampling Unit (PSU) referred to as a cluster in the NDHS. Tables 6.2 and 6.3 summarize the demographic characteristics and list the variables.

Table 6.2: Variables and their definitions

Variables	**Categories**
Survey year	1990, 1999, 2003, 2010
Region	North-East, North-Central, North-West, South-East, South-West, South-South
Child's level	
Completed all immunization	Yes or No
Sex of child	Male or female
Place of delivery	Hospital (public or private) or home
Mother's level	
Wealth quintile	Poorest, poorer, middle, richer, richest
Religion	Catholic, Protestant, Other Christian, Islam, Traditionalist, Other
Mother's age	Age of mother in years
Mother's education	No education vs at least primary education
Mother's occupation	Currently employed or worked in the 12 months preceding the survey or not
Age at first birth	Mother's age at first birth
Mother's marital status	Never married, married, living together, widowed, divorced or not living together
Sex of household head	Male or female
Community level	
Wealth	Proportion in the lowest and second lowest wealth quintile in the community
Place of residence	Rural or urban
Time to source of water	Mean time to source of water
Sanitation	Proportion in the community with sanitation
Access to clean water	Proportion in the community with access to clean and safe water

6.2.2 Statistical Analysis

Multi-level analysis was applied to identify associated factors for partial or non-immunization of a child. Let y_{ijk} be the k^{th} binary response ($y_{ijk} = 1$, completed, or

$y_{ijk} = 0$, not completed his/her immunization) from a child of the j^{th} mother (subcluster) in the i^{th} community (cluster). The interest here is in fitting a marginal probability of a child not completing his/her immunization schedule, $\pi_{ijk} = Pr(y_{ijk} = 1)$, given as:

$$logit(\pi_{ijk}) = x'_{ijk}\beta$$

where x_{ijk} is the vector of covariates at each cluster level. The association model for the pair-wise odds ratios (POR) among observations with three-level cluster data is given as:

$$\log \text{POR}(Y_{ijk}, Y_{ij'k'}) = \alpha_0 + \alpha_1 z'_{ijk},$$

where z_{ijk} if the pair of observations is from the same subcluster ($j = j'$) and 0 otherwise, α_0 is the log odds ratio for association among observations from different subclusters, and $\alpha_0 + \alpha_1$ is the log odds ratio for within-subcluster association (Preisser *et al.*, 2003).

Four kinds of models were specified: Model 1, which is the null model, was fitted to examine whether the mother- or community-level ORs varied. The second model (Model 2) contains the survey year, community and the interaction term survey year and community, while the third model (Model 3) contains the survey year and other explanatory variables, and Model 4 includes all explanatory variables plus the interaction terms. All models were implemented in `SAS` procedures `PROC GENMOD` (SAS Institute Inc., 2008).

Table 6.3: Summary of vaccination uptake in Nigeria from 1990 to 2008

Characteristic	Survey year				**Total**
	1990	1999	2003	2008	
Number of communities(Clusters)	298	393	361	886	1938
Number of households	3995	2837	2161	11,014	17,380
Number of observations	7902	3552	6029	28,647	46,130
Percentage fully vaccinated	60.32%	63.70%	77.79%	70.68%	68.97%

6.3 Results

About 60%, 64%, 78% and 71% of the children in the study were fully vaccinated in 1990, 1999, 2003 and 2008, respectively (Table 6.3). In the NDHS 2008, 62% had their birth delivered at home: 13% at private health facilities and 20% at a public health facility (National Population Commission (NPC) [Nigeria] and ICF Macro, 2009). Figure 6.1 illustrates the distribution of vaccination by place of delivery. The proportions of children that were not fully immunized among those delivered at home were 37.5% in 1990, 36.7% in 1999, 31.6% in 2003 and 32.2% in 2008. As shown in Figure 6.2, the percentage and distribution of antigens taken varies; for example, in 2008 about 47% received BCG only, 67% received the first dose of polio vaccine only, 58% took the second dose of polio vaccine and 40% took the third dose of polio vaccine. Also, about 49% received only the first dose of DPT, 33% received DPT2, 20% received DPT3 and 43% received the measles vaccine only.

Table 6.4 presents the odds ratio estimates together with the 95% confidence interval from the ALR model (of different specifications) for the associated risk factors for a child not completing the set of antigens. Accounting for the clustering (mother and community) in the vaccination uptake was handled by pair-wise odds ratios (PORs). The PORs indicate the association between pairs of observations from children with the same mother-level characteristics or in the same community. Evidence from Model 1, without adjusting for any covariates, indicated a significant mother-level variability (OR=87.96, p<0.0001) and community-level variability (OR=33.22, p<0.0001). The result also showed that the mother-level clustering (POR=92.08, 157.71 and 156.04) and the community-level clustering (POR=36.97, 79.25 and 78.81), after adjusting risk factors in Model 2, Model 3 and Model 4, respectively, were statistically significant. A simpler model (Model 1) with only the survey year (not shown) shows a significant association between the survey year and the odds of being immunized: the estimated odds ratio of a child being immunized in 1999 was OR=1.189 (95% CI, 0.922-1.528), in 2003 it was OR=4.506 (95% CI, 3.171-6.403) and in 2008 it was OR=1.836 (95% CI, 1.439-2.342) compared with 1990.

The results from Model 2, with two-way interaction between survey year and community, showed a significant interaction term indicating an increase in the odds chance of a child being immunized over time and space by 0.2% and 0.3% in 1999 and 2008, respectively, compared with 2003. The three models (Model 2, Model 3 and Model 4) showed a significant increase in the chance of being immunized in 1999, 2003 and 2008 compared with 1990. The inclusion of additional risk factors also increased the within-community clustering POR from 36.97 in Model 2, to 79.25 in Model 3 and to 78.81 in Model 4. A similar pattern was observed with the within-mother clustering.

The results from the final model (Model 4), in which all the risk factors that were considered in this study were adjusted for, together with the interaction term (survey year and community), are displayed in Table 6.4. The final model also indicates that children delivered at hospitals have about a 42% higher chance (OR=1.415; 95% CI, 1.304-1.535) of being immunized than children delivered at home.

The estimated odds ratio shows that children from the poorest families have a greater chance than those from the richest families of being immunized (OR=0.712; 95% CI, 0.641-0.792). Mother's education is a significant risk factor for childhood immunization, with children of mothers with no education having a lower chance of being immunized than those of mothers with at least a primary education (OR 0.844). In the same regard, the proportion of fully immunized children was higher when the mother is gainfully employed (mother's occupation) than when the is not gainfully employed (OR=1.001; 95% CI, 1.000-1.001). The likelihood of a child being fully immunized increased with age. Age is statistically significantly associated with higher the odds of a child being immunized (OR=1.012; 95% CI, 1.006-1.017). Children from households with a female head are 0.926 times less likely to be immunized than those from a male-headed household (OR=0.926;

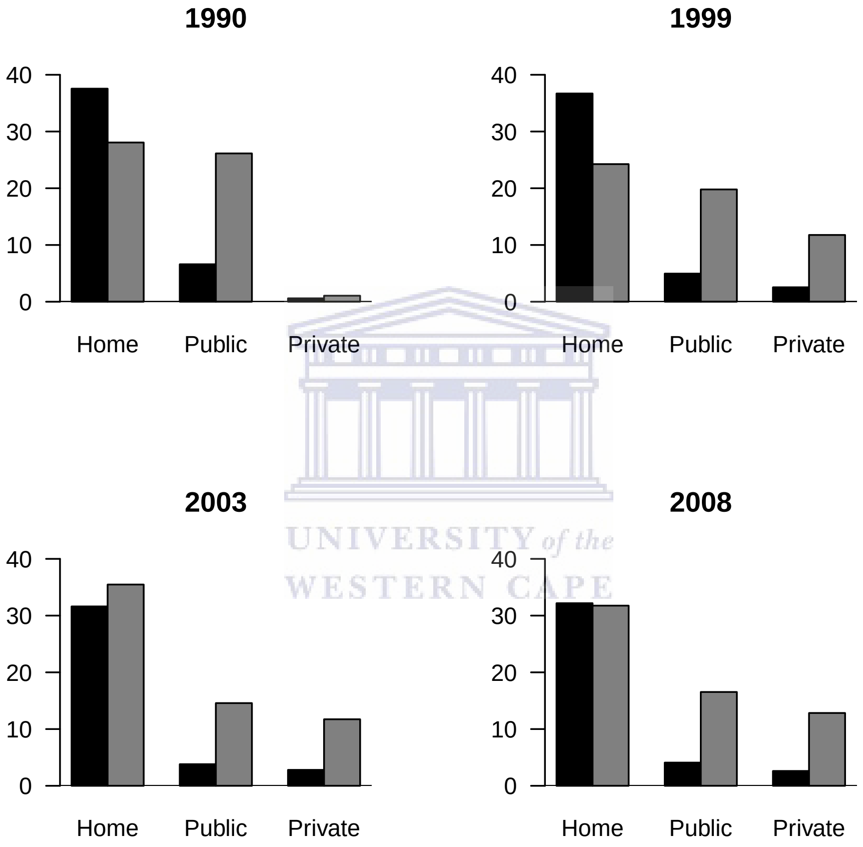


Figure 6.1: Proportion of children having received the full schedule immunization (Gray) or not (Dark) by place of delivery.

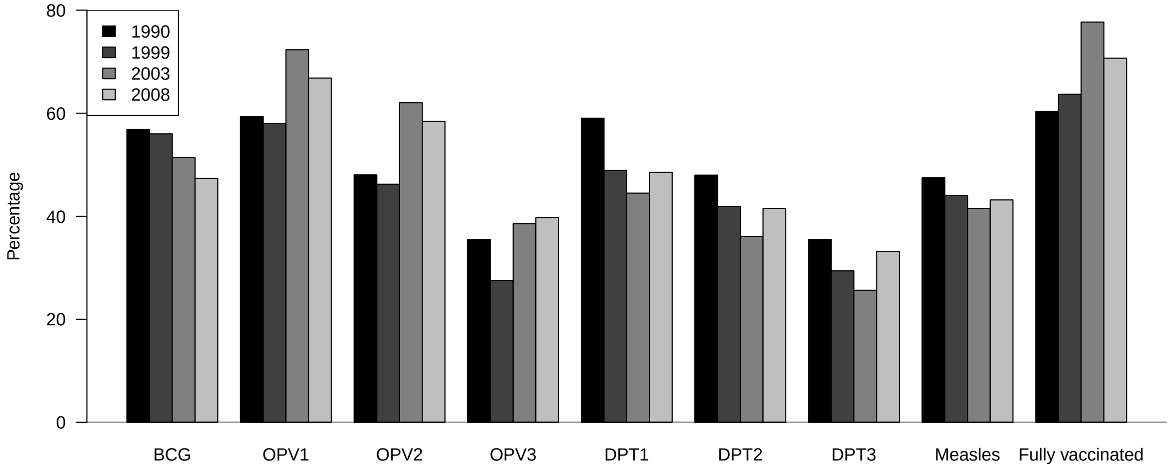


Figure 6.2: Percentage of children having received different sets of antigens in Nigeria (1990-2008)

95% CI, 0.861-0.996). An increase in the proportion of households with good sanitation in communities increases the chance of a child being immunized (OR=1.006; 95% CI, 1.004-1.009).

6.4 Discussion and Conclusion

Vaccination is a very effective way of eradicating the burden of a large number of preventable diseases that account for approximately 22% of child deaths in Nigeria, amounting to over 200,000 deaths per year (USAID, n.d.). The results of this study show the advantage of the use of alternating logistic regression, which allows estimation of the correlations using pair-wise odds ratios in the data set. The inclusion of within-cluster association improved the inference (Preisser *et al.*, 2003) on the risk factors for vaccination uptake in Nigeria. Regardless of the adjusted risk factors, the mother-level and community variability remain statistically significant. This study confirms the mother-

and community-level clustering, reaffirming that children with the same mother-level characteristics or in the same community exhibit a similar likelihood of been immunized (Diddy, 2009; Wiysonge *et al.*, 2012).

Table 6.4: Odds ratio estimates from ALR models for associated risk factors for unimmunized children in Nigeria, 1990-2008.

Risk factor	Model 2		Model 3		Model 4	
	OR	95% CI	OR	95% CI	OR	95% CI
Survey year						
1999 vs 1990	**2.060**	1.175, 3.611	**1.776**	1.147, 2.750	**1.788**	1.014, 3.155
2003 vs 1990	**2.071**	1.115, 3.849	**3.529**	2.528, 4.924	**3.191**	1.709, 5.958
2008 vs 1990	0.660	0.390, 1.119	**1.283**	1.013, 1.626	0.946	0.548, 1.633
Community	1.000	0.999, 1.001	**1.000**	1.000, 1.000	1.000	1.000, 1.000
Community × 1990	0.999	0.999, 1.000			1.000	1.000, 1.000
Community × 1999	**1.002**	1.000, 1.004			1.000	0.998, 1.003
Community × 2008	**1.003**	1.002, 1.003			**1.001**	1.000, 1.001
Region						
Central vs South-West			0.900	0.779, 1.039	0.938	0.809, 1.087
North-East vs South-West			**1.318**	1.110, 1.565	**1.444**	1.192, 1.749
North-West vs South-West			**0.786**	0.677, 0.913	0.870	0.736, 1.029
South-East vs South-West			**0.713**	0.615, 0.826	**0.691**	0.593, 0.804
South-West vs South-West			1.190	0.975, 1.452	1.077	0.863, 1.344
Child's sex			0.994	0.970, 1.018	0.993	0.969, 1.018
Place of delivery						
Hospital vs home			**1.409**	1.299, 1.529	**1.415**	1.304, 1.535
Residence: Rural vs urban			0.873	0.732, 1.042	0.857	0.716, 1.026
Wealth Index						
Poorest vs richest			**0.711**	0.639, 0.791	**0.712**	0.641, 0.792
Poorer vs richest			**0.813**	0.747, 0.886	**0.815**	0.749, 0.888
Middle vs richest			0.934	0.858, 1.017	0.935	0.859, 1.018
Richer vs richest			**1.201**	1.087, 1.327	**1.201**	1.087, 1.326
Age at fist marriage			0.999	0.988, 1.010	0.998	0.987, 1.009

Table 6.4 – Continued

Risk factor	Model 2		Model 3		Model 4	
	OR	95% CI	OR	95% CI	OR	95% CI
Mother's marital status						
Never married vs divorced or separated			1.056	0.914, 1.219	1.059	0.917, 1.223
Married vs divorced or separated			0.939	0.734, 1.201	0.943	0.738, 1.205
Living together vs divorced or separated			1.034	0.781, 1.369	1.033	0.780, 1.369
Widowed vs divorced or separated			0.832	0.594, 1.165	0.831	0.593, 1.165
Mother's education (none vs at least primary)			**0.846**	0.762, 0.940	**0.844**	0.760, 0.937
Mother's age			1.012	1.006, 1.017	**1.012**	1.006, 1.017
Mother's occupation (gainfully employed)			**1.001**	1.000, 1.001	**1.001**	1.000, 1.001
Household head (female)			**0.924**	0.860, 0.993	**0.926**	0.861, 0.996
Prop. of 40% poorest			0.997	0.994, 1.000	**0.998**	0.994, 1.001
Religion						
Catholic vs Other Christian			1.071	0.900, 1.275	1.058	0.887 ,1.261
Catholic vs Islam			0.788	0.658, 0.943	0.836	0.695, 1.005
Catholic vs Traditionalist			0.575	0.451, 0.733	**0.558**	0.440, 0.707
Catholic vs Other			0.655	0.412 ,1.042	0.663	0.415, 1.060
Average time to water source			0.999	0.998, 1.000	0.999	0.998, 1.001
Prop. with clean water			1.001	0.998, 1.004	1.001	0.998, 1.004
Prop. with good sanitation			**1.006**	1.004, 1.009	**1.006**	1.004, 1.009
Within-community POR		**36.97**		**79.25**		**78.81**
Within-mother POR		**92.08**		**157.71**		**156.04**

Bold:$p < 0.05$

The geographical heterogeneity in the coverage of vaccination can be attributed to variations between communities within the different regions in Nigeria. The South-South region in particular is characterized by extensive mangrove forests, lagoons and swamps stretching over hundreds of kilometres inland, as well as poverty, poor social infrastructure and conflicts that are exacerbated by environmental degradation from crude oil pollution (Diddy, 2009). Sorungbe (1989) suggested regular review of the programme and intensive training of personnel. The mother-level variability confirmed in this study attested to (Abdulraheem *et al.*, 2011), who claimed that long-distance trekking and the high cost of transportation are limiting factors for mothers completing immunization schedules for their children. Routine immunization in northern Nigeria is one the lowest in the world. The region is known to be hesitant about immunization uptake, and this has been attributed to cultural beliefs and interpretations, health worker malpractices, fear of injection, level of education and lack of adequate knowledge about immunization (Renne, 2010). Furthermore, some vaccinators and health care staff have been attacked and even murdered in recent times in Nigeria.

More children received different antigens in 2008 than in previous years, and it is safe to say that more children were immunized in 2003 and 2008 than in previous years (62% and 60% respectively). Multi-year analysis of immunization in Nigeria, using survey year as the time factor, suggests a significant association between survey year and the odds of being immunized. The chance of a child been fully vaccinated increased from 19% in 1999 to about 84% in 2008 compared with 1990.This indicates an increase in the propensity for full immunization of children over the years, except from 2003 to 2008, which showed a slight trough in the percentage of children that were fully vaccinated (from about 78% in 2003 to 71% in 2008). Coverage evaluation was also carried out by the incorporation of a two-way interaction term in the model; there was evidence of progressive coverage of immunization in communities over the years. The interaction term survey year (2008) and community was significant, suggesting an increase in the odds of a child being immunized over time and across communities.

The sex of a child is not associated with the chance of being immunized. The non-significance of sex suggests that the odds of a child receiving all doses of the antigens are not affected by the sex of the child. The chance of children from the poorest families being fully immunized decreases by 36% compared with children of the richest families. Although the proportion of home birth deliveries is very high, most 'completers' are those that had their birth delivery at a public or private health facility. There are more unimmunized children among the home birth delivery group in 1990 and 1999. The analysis indicates that children delivered at hospitals have about 1.5 times higher chance of being immunized than children delivered at home. This may be the case in a rural area where the nearest health facility is some kilometres away and where the trekking distance or cost of transportation may deter parents. Also, in some cases the mother may not be strong enough to embark on such trip and some cultural beliefs restrict the mother from going out for the first 40 days after delivery. In such cases, a home visit vaccination service would provide the necessary breakthrough.

Similarly, mother's education plays a significant role in her child's immunization status: a child whose mother is educated has a higher chance (17%) of completing his/her immunization than a child from a mother with no education. The proportion of children that are fully immunized was higher when the mother is gainfully employed than when the mother is not gainfully employed. Children of households with a female head are less likely to be immunized than those from male-headed households. An increase in the proportion of households with good sanitation in communities will increase the chance of a child being immunized. These analyses indicate that even after adjusting for known risk factors, some children have a greater propensity to be immunized than others.

This work is, to the best of the authors' knowledge, the first to provide a multi-year coverage evaluation of vaccination uptake in Nigeria using DHS data sets using a multi-level technique. Most of the methods used to analyse binary outcome variables are either generalized linear models or their extensions. Rammohan *et al.* (2012) used logistic regression analysis to explore the likelihood of a child under five years of age being vaccinated for measles separately for Indonesia, India, Pakistan, Nigeria, Democratic Republic of Congo and Ethiopia, while Wiysonge *et al.* (2012) used multi-level logistic regression to examine the risk factors associated with childhood immunization in 24 countries in sub-Saharan Africa at the child, community and country level. In his study, Antai (2009) employed the use of a three-level multi-level logistic regression model to assess the risk factors for childhood immunization in Nigeria in 2003 using DHS data at the child level nested within mothers who were in turn nested within communities. Bosch-Capblanch *et al.* (2012) carried out a logistic regression analysis on data from 241 nationally representative household surveys in 96 countries using the unique or most recent survey for each country. They dichotomized the vaccination status as children having not received any vaccination ('unvaccinated') vs children who had received at least one dose (partially vaccinated) of any vaccine and fully vaccinated children. The present study's method (alternating logistic regression) allows the simultaneous regressing of the outcome variable on the explanatory variable, as well as modelling of the association among the outcome variables by means of a pair-wise odds ratio instead of correlations. The odds ratio provides a better interpretation for binary outcome variables.

Some limitations of this study have been identified. Firstly, the use of survey year as the time factor in cross-sectional data may be seen as a limiting factor; the years of the surveys are not evenly spaced (i.e. 1990, 1999, 2003 and 2008). Secondly, it is difficult to say that the same households or primary sample units were sampled each year, as is the case in longitudinal studies. It would be interesting to be able to identify those households that were sampled over the years. Although the data may have these limitations, they do provide extensive information that is crucial for assessing the risk factors for full immunization over the years.

In conclusion, the significance of the mother- and community-level variability showed the important role played by the mother and community in immunization. This suggests the need for more emphasis on mother- and community-level characteristics. The community-survey year interaction term suggests an increase in the chance of a child being immunized over the years and spread over communities. Evidence-based policy should lay more emphasis on mother- and community-level risk factors in order to ensure that all children are immunized. It is evident in this study that the importance of hospital delivery cannot be overemphasized. Interventions aimed at increasing and encouraging hospital delivery should be put in place to increase the coverage and uptake of vaccination. Parental education is crucial in the uptake of immunization in Nigeria; this could be in the form of community enlightenment and full communication of medical risk to alleviate their fears. Also, it is essential to involve private and independent bodies to provide medical information to the local population because of the distrust between the government and the local population, especially in the Northern region. Adequate security should be provided to health care staff and vaccinators to curb the persistent attacks on them in this region.

There is a need to reach out to every district and ward through re-establishment of outreach services, community links with service delivery, monitoring and use of data for action planning and management. Routine immunization must be prioritized at every level of government, especially at the local level, in order to capture the local population who are mostly at risk of vaccine-preventable diseases in Nigeria. The national orientation agency has to take drastic action to provide adequate information about immunization to quell rumours about vaccines. Perhaps a vaccination card could be used as a prerequisite for pre-school enrolment in order to identify those children that are not vaccinated. More effort should be made to access hard-to-reach communities, perhaps through home visits. Efforts should be made to make immunization practice more 'patient friendly', for instance by reducing the number of doses required to complete a full course, hence reducing dropout. The present pentavalent vaccine consisting of diphtheria, pertussis, tetanus, hepatitis B vaccine and *Haemophilus influenzae* type b vaccines is a welcome innovation, but a single vaccine containing all the eight vaccines would be better.

Chapter 7

Causes and Patterns of Morbidity and Mortality in Afghanistan: Joint Estimation of Multiple Causes in the Neonatal Period

This chapter has been accepted for publication in Canadian Studies in Population:

Adegboye, O.A. and Kotze, D. (2013). Causes and Patterns of Morbidity and Mortality in Afghanistan: Joint Estimation of Multiple Causes in the Neonatal Period. *Canadian Studies in Population.*

Summary

This paper focuses on investigating the leading cause(s) of death as categorized by the International Classification of Diseases and the preventable factors in Afghanistan using data from verbal autopsies of infant deaths within households in the Afghanistan Mortality Survey (AMS) of 2010. The presence of a disease in a person may increase the risk of another disease that may contribute to the death process. Therefore, it is desirable to simultaneously model the correlations between the two diseases. The joint analysis of causes of death was analyzed using the multivariate spatial bivariate probit model to accommodate dependency that arises from spatial sources. The influence of individual- and community-level variables such as size at birth (birth weight), sanitation, remoteness and distance to health facilities on infant morbidity and mortality in Afghanistan were examined. About 65% of deaths in neonates occurred as a result of complications of pregnancy, childbirth, and the puerperium. Exploratory data analysis indicated a highly significant association between disease classified as complications and symptomatic conditions ($p < 0.0001$). The results from the multivariate analysis showed that the size at birth, remoteness index, residence, age of the child, wealth index, sanitation and duration of pregnancy are important predictors of complications at birth and the puerperium while age, size at birth, place of birth and duration of pregnancy are strongly associated with diseases classified as symptomatic conditions. The findings highlight the importance of joint analysis of causes of death. The results of this study suggest the existence of multiple causes of death in the AMS. In Afghanistan complications of pregnancy are clearly a problem and must be adequately improved. The results can be used to strengthen policies that will adequately address the identified risk factors.

7.1 Introduction

Most information on medical history are provided on the death certificate or in the vital registration system that may not be available in some developing countries. Measuring cause-specific mortality rates are very difficult in this situation especially where the vital registration systems are lacking for the majority of the population. Providing a good analysis and estimates for cause-specific mortality are essential for understanding the overall profile and burden of disease in a population (Bickel *et al.*, 2006). To resolve the issue of the death certificate and vital registration, many countries have set up mortality surveillance systems that provide community-based death reviews exploring the medical and non-medical causes of death. This kind of review identifies personal, family and community factors that may have contributed to the death process.

The verbal autopsy (VA) has been used to estimate cause-specific mortality in a variety of methodological settings, the most common being in the context of an epidemiological study (Quigley, 2005). There is a substantial and growing body of literature on the reliability and validity of VA; it provides accurate cause of death especially in the absence of a death certificate or vital registration (Bapat *et al.*, 2012; King and Lu, 2008; Lee *et al.*, 2008; Quigley, 2005; Sibai *et al.*, 2001; Philip *et al.*, 2005). The use of a questionnaire for VA has an advantage over ad hoc investigations because of its high specificity when the list of target diseases is extensive (Fauveau, 2005; Garenne and Fauveau, 2006). The World Health Organization has a long history of structured questionnaires for systematic recording of signs and symptoms for assessing causes of death (WHO, 1978, 1995).

Afghanistan has been devastated by armed conflicts for about three decades now and has seen over 5 million people been displaced during this period (Bhutta, 2002). This has resulted in a barely functioning health care system and in some areas non-availability of health care facilities. The population of Afghanistan consists of about 43% child population and the country has a physician density of about 21 physicians to 100,000, where about two-thirds of the physicians live and work around Kabul. There is limited literature on infant and child mortality in Afghanistan, most research were conducted at provincial or regional level due to security and logistic reasons. Moreover, most mortality statistics and data in Afghanistan were collected in the past decade. For instance, in their study, Kim *et al.* (2012) assess the caesarean section deliveries across secure sites in 31 out of the 34 provinces in Afghanistan, Amowitz *et al.* (2002) conducted a cross-sectional survey of 4886 Afghan women living in 7 districts in Afghanistan's Herat province while Mayhew *et al.* (2008) conducted a cross-sectional study in all 33 provinces in 2004. A retrospective cohort study of women of reproductive age (15-49 years) was done by Bartlett *et al.* (2005) in four districts of Kabul province, Laghman province, Kandahar province and Badakshan province. Gessner (1994) surveyed 312 displaced families and 300 resident families in Kabul to assess their mortality rates.

The high rates of maternal and perinatal mortality in Afghanistan were attributed to low rates of caesarean section (CS) and many cases of CS are either emergencies or referrals from another health care facility (Kim *et al.*, 2012). Bartlett *et al.* (2005) reported that only 13% of respondents in their study had used skilled birth attendants during delivery. Diarrhoea and acute respiratory infections are responsible for about 36% of the childhood deaths in Afghanistan (Gessner, 1994; Prasad, 2006). Bartlett *et al.* (2005) investigated the deaths among women of reproductive age through verbal-autopsy interviews of family members. Their study revealed that remoteness played a significant role in increasing the maternal mortality in Afghanistan. Other risk factors for high child and maternal mortality rates in the country include: home delivery (90% of total births are without medical assistance) (Neelon *et al.*, 2012); poor sanitation and lack of access to clean water have an indirect relation to high infant mortality in the Herat province, Afghanistan (Amowitz *et al.*, 2002). Gessner (1994) found injuries sustained directly as a result of war as the most common cause of death in the displaced and resident population in Kabul. Most of these studies are restricted to district level or provincial level and also neglected the possibility that mortality rates may be spatially dependent.

In some cases, death is a result of cumulative effects of different causes, some ailment may not directly contribute to the death process and therefore, not included as the underlying cause of death. For instance, the presence of a chronic disease in a person may increase the risk of developing another related disease that may contribute to the death process. Moreover, the causes of death are not listed in the AMS data, rather the diseases leading to the death of the household member were listed. Consequently, there may exist correlations between multiple causes of death; ignoring such vital information may bias the mortality estimates and thus, their simultaneous effects must be jointly investigated.

The main idea of this paper is to investigate the associations between diseases and possibly the cause of death. This will allow the possibility of identifying a range of illnesses that may have contributed to the death process thereby providing comprehensive information. The study will investigate the influence of some demographic variables and risk factors (e.g. remoteness, wealth index, residence, sanitation, access to safe and clean water) on morbidity and mortality. Risk factors for multiple causes of deaths will be simultaneously analyzed. As in most health outcomes, geographical variations cannot be ruled out and risk factors may vary geographically (Neelon *et al.*, 2012). Thus, this study addresses the issue of spatial dependency between provinces by extending the bivariate probit model to incorporate spatial information where the spatial dependency was handled through random effects to investigate if association among different causes of death, dependent on certain demographic variables, varied geographically. The study, to the best of the authors' knowledge, is the first empirical analysis on the use of VA in Afghanistan and the first to provide multiple causes of death analysis using AMS data sets.

The rest of the paper is organized as follows: The next section presents the data sources and methodology, describes the data, explains the disease classification of diseases and outline the proposed multivariate analysis model. Then general empirical application will be introduced and a summary of the results of an in depth analysis of the neonatal mortality data will be presented. The paper will be concluded with the discussion and conclusion from the study.

7.2 Methods

7.2.1 Data Sources

The data used in this research is part of the Afghanistan Mortality Survey (AMS 2010) APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt (2011) which is the first nationwide survey of its kind in Afghanistan. The survey was conducted in 2010 by the Afghan Public Health Institute (APHI) of the Ministry of Public Health (MoPH) and the Central Statistics Organization (CSO) with technical and logistic support from ICF Macro and the Indian Institute for Health Management Research (IIHMR) APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt (2011). The AMS is a special survey that focused on mortality, causes of death and maternal mortality. The survey covered about 87% of the country and the 34 provinces of Afghanistan.

The study centres on data from verbal autopsies (VA) of deaths within a household up to three years before the AMS (2007-2010). A team of 15 physicians participated in the AMS 2010; they underwent special training in the WHO International Classification of Death (ICD10) (WHO, 2011) to assign a cause of death. Three types of questionnaires were used for the verbal autopsies (VA); death of an infant age 0-28 days (VA1); death of a child aged 29 days - 11 years (VA2) ; death of an adult aged 12 years and above (VA3). The respondents were key informants that were present at the time of death, they were allowed to provide an open narrative section to describe verbatim the illness and events that led to the death. The VA questionnaires included a wide range of questions on signs and symptoms during the disease or injury preceding the death and were reviewed independently by 2 physicians. Also, the VA provides information on medications used, utilization of health services, and behavioural and environmental risk factors. In addition to basic information about the respondent, information on the deceased, and date/place of death and delivery history were collected in the verbal autopsies of deaths as well as data on maternal and child health, childhood mortality, the wealth index, ownership of basic facilities, the availability of health facility at household and/or individual level. The data sets consist of VA from 1105 neonatal (0 - 28 days), 997 perinatal and children (29 days - 11 years), and 1831 adults (above 12 years). The list, as well as the summary statistics of some of the variables used in this study, are provided in Table 7.1.

7.2.2 Classification of Diseases

The classification of the causes of death was done using the International Classification of Disease (ICD)(WHO, 2011) which is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. For easy comparison and interpretability of results, the classification of the disease was done in the following manner. All causes of death due to infections or parasitic infestations were classified as INFECTIOUS (INF). Disease that takes care of primary systemic diseases like neoplasm, diseases of the blood and blood-forming organs, disease of the nervous system, sense organs, circulatory system, respiratory system, digestive system, genitourinary system, skin and subcutaneous tissue, musculoskeletal system and connective tissue were

Table 7.1: Characteristics of deaths in the three age group and summary of the risk factors

Variables	Neonatal		Perinatal and Children		Adults	
	N	%	N	%	N	%
Disease classification						
Infectious	220	19.91	874	87.66	989	54.01
Complications	713	64.52	140	14.04	111	6.06
Non-infectious	40	3.62	280	28.08	1798	98.20
Symptoms & Signs	290	26.24	821	82.35	1433	78.26
Injury	65	5.88	393	39.42	711	38.83
Gender						
Male	624	56.62	548	55.47	1045	57.26
Female	478	43.38	440	44.53	780	42.74
Mean age		2.5 months		5.14 years		54.26 years
Mother's age (years)						
$\leq$ 15	55	4.98				
16-20	241	21.81				
21-25	294	26.61				
26-30	206	18.64				
31-35	157	14.21				
$\geq$ 36	152	13.75				
Size at birth (birth weight)						
Smaller than normal	431	43.98				
Normal	493	50.31				
Larger than normal	56	5.71				
Place of birth						
Home	680	61.87				
Hospital	419	38.13				
Antenatal						
Yes	522	48.47				
No	555	51.53				
Single of multiple birth						
Yes	1025	93.01				
No	77	6.99				
Duration of pregnancy						
1-3 months	12	1.08				
4-6 months	122	11.04				
6-8 months	304	27.51				
9 months	611	56.25				
10 months	41	3.76				
Pregnant at time of death						
Yes					415	15.89
No					217	84.11

Table 7.1 – Continued

Variables	Neonatal		Perinatal and Children		Adults	
	N	%	N	%	N	%
Smoke						
Yes					256	14.17
No					1551	85.83
Small at birth						
Yes			140	30.11		
No			325	69.89		
Malnutrition						
Yes			174	17.45	208	11.36
No			809	81.14	1594	87.04
Residence						
Urban	320	28.96	210	21.06	618	33.75
Rural	785	71.04	787	78.94	1213	66.25
Region						
North Eastern	166	15.02	176	17.65	263	14.36
Northern	161	14.57	158	15.85	335	18.30
Western	142	12.85	140	14.04	230	12.56
Central Highlands	32	2.90	21	2.11	40	2.18
Capital	188	17.01	142	14.24	361	19.72
Eastern	181	16.38	158	15.85	260	14.20
Southern	109	9.86	89	8.93 195	10.65	
South Eastern	126	11.40	113	11.33	147	8.03
Sanitation						
Bad	888	80.36	824	82.65	1394	76.13
Good	217	19.64	173	17.35	437	23.87
Source of water						
Unsafe	473	42.81	442	44.33	754	41.18
Safe	632	57.19	555	55.67	1077	58.82
Wealth index						
Poorest	245	22.17	232	23.27	345	18.84
Poorest	218	19.73	206	20.66	373	20.37
Middle	202	18.28	167	16.75	369	20.15
Richer	224	20.27	201	20.16	360	19.66
Richest	216	19.55	191	19.16	384	20.97
Remoteness						
Most remote	261	23.62	268	26.88	443	24.19
Least remote	130	11.76	112	11.23	251	13.71
Total number of deaths						
1	764	69.14	770	77.23	1461	79.79
≥ 2	341	30.86	227	22.77	370	20.21

classified as NON-INFECTIOUS (NNN). Also, all deaths related to pregnancy, childbirth, and the puerperium and conditions that have their origin in the perinatal period even though death or morbidity occurs later were classified as COMPLICATIONS (COMP). Other classifications are endocrine, nutritional and metabolic diseases as ENM; mental and behavioural disorders which were classified as MENTAL. Also symptoms, signs, and ill-defined conditions were classified as SYMPTOMS (SSID), while injury, poisoning and certain other consequences of external causes were classified as INJURY (INJ).

Although it is important to understand the specific cause of death independently, the interaction between causes offer valuable information that cannot be overemphasized. Lets assume that there are N main causes of death each with a YES/NO response. This will give 2^{N-1} possible combinations of multiple causes excluding the case where both responses are NO. These combinations will have different implications in terms of effect on demographic variables; it will provide a very important understanding of their occurrence in the population.

7.2.3 Data Analysis

Exploratory analysis was based on cross tabulation of statistics (frequencies/ percentages) from the mortality data. The analyses were carried out on the contingency tables using the Cochran-Mantel-Haenszel chi-square test. The dimensions of the tables were defined by the number of the causes of death. The analysis was used to investigate the association between causes of deaths.

The simultaneous analysis of causes of deaths was modelled by multivariate probit which is appropriate when correlated binary responses (causes of deaths) are regressed on explanatory variables (demographic variables and risk factors) (Lesaffre and Molenberghs, 1991). Consider the modelling of the joint probability in the bivariate case [1], the 2-equations for a multivariate probit model with three response variables are given as:

$$Y^*_{ijm}|\phi_{im} = \beta'_m X_{ijm} + \phi_{im} + \epsilon_{im}, \qquad m = 1, 2; \quad i = 1, ..., 34; \quad j = 1, .., n_i \tag{7.1}$$

$$Y_{ijm} = \begin{cases} 1, & \text{if } Y^*_{ijm} > 0 \\ 0, & Y^*_{ijm} < 0 \end{cases}$$

where the continuous Y^*_{ijm} is an unobservable latent variable -related to dichotomous response Y_{ijm} via a threshold concept. $Y_{ijm} = 1$ implies that the j-th individual in the i-th province die of m-th cause and $Y_{ijm} = 0$ otherwise, that is, an unobserved variable representing the latent variable which represents the probability of each disease classification m vs not, with the latent vector having a 2-dimensional normal distribution. X_{im} is a vector of covariates associated with disease classification m, β_m is the parameter to be estimated and ϕ_{im}, is the vector of spatially dependent random effects for the i-th province which is distributed as multivariate normal with a mean of zero and variance-covariance matrix Σ, ϵ is the error term where $\epsilon_{im} \sim$ i.i.d. N(0,Ω),

$$\Omega = \begin{pmatrix} \tau_1^2 & \\ \tau_{21} & \tau_2^2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

The two error terms can be written as:

$$\eta_{im} = \phi_{im} + \epsilon_{im} \tag{7.2}$$

From the above assumptions $Var(\eta_{im}) = \tau_m^2 + \sigma_m^2$; this implies that within units η_{im}'s will be correlated:

$$Corr(\eta_{im}, \eta_{im'}, m \neq m') \equiv \rho = \frac{\tau_{21}^2 + \sigma_{21}^2}{\{(\tau_m^2 + \sigma_m^2)(\tau_{m'}^2 + \sigma_{m'}^2)\}^{\frac{1}{2}}} \tag{7.3}$$

[1] Two major causes of death each with Yes/No responses will result in 3 possible combinations of multiple outcomes excluding the NO-NO combination

For simplicity, let us assume that the two disease classifications are the main causes of death. The joint probability can be written as:

$$Pr(Y_1 = 1, Y_2 = 1|\phi_{im}) = \int_{A_1} \int_{A_2} \psi(\beta'_m X_{ijm} + \phi_{im}, \Sigma) dy_1 dy_2, \tag{7.4}$$

where ψ is the density function of a multivariate normal distribution and A_m is the interval $(-\infty, \beta'_{im} X_{ijm})$.

7.3 Results

An explorative summary of 1105 deaths from neonatal, 997 from perinatal and child, and 1831 adult deaths in the AMS data was first carried out. Results from the combined data sets indicated that two-thirds of urban households fall in the richest quintile while overall 54% of the households have access to safe drinking water. Also, only one-fifth of the households have an improved toilet facility. The fertility rates are 5.2 and 4.7 children per women for rural areas and urban areas respectively. Death from pregnancy related causes is very high (higher in rural areas) with 1 in every 50 women dying of pregnancy related causes, 2 in 5 pregnancy related deaths occur during pregnancy and about 20% within two months after delivery (APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt , 2011). Injury related deaths in men were 20%, half of these deaths were war and violence related in 15+ years (APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt , 2011).

The number of times a specified disease occurs on the verbal autopsy form as well as some demographic variables are presented in Tables 7.1 and 7.2. Among the 1105 deaths in the neonatal data, about 65% of the deceased were reported to have had complications of pregnancy, childbirth, and the puerperium (Table 7.2). Because of these

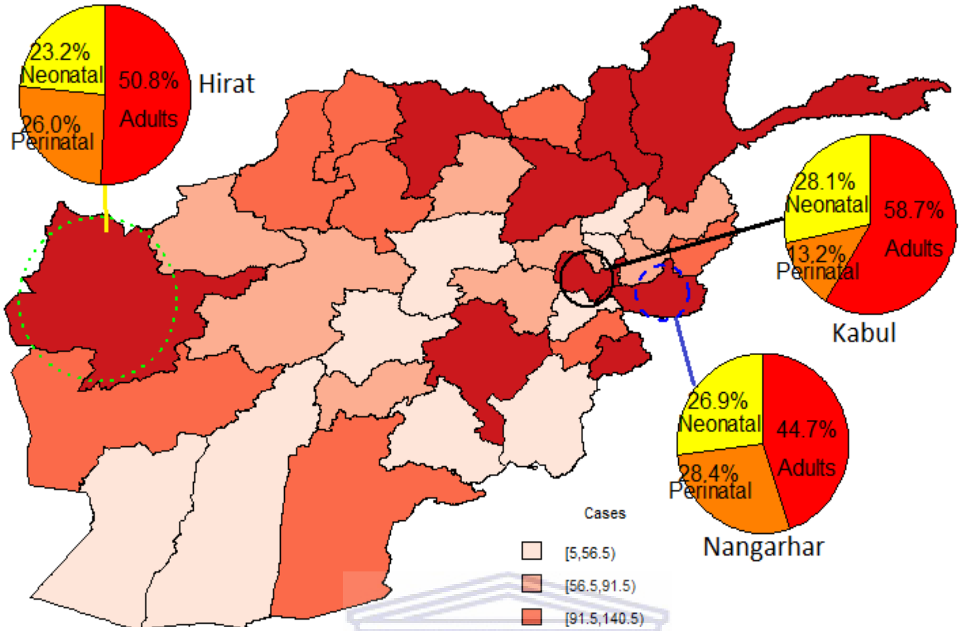


Figure 7.1: Map of Afghanistan showing the distribution of mortality. The pie chart indicates the breakdown of the proportion of deaths from neonatal, perinatal and child, and adults in the provinces with the top three mortality rates

complications, babies are born before term; some due to early end of pregnancy (39%), premature (40.6%), 44% cases of low birth weight and few cases of malformation, thus making diseases classified as complications (i.e. complications of pregnancy, childbirth, and the puerperium) the most common cause of death in neonatal. Table 7.3 reflects the number of deaths with the indicated combinations of causes of death. Complications alone occurred in about 54% of all deaths in neonatal while symptomatic diseases alone occurred in about 6% (Table 7.3). Among the 997 deaths in perinatal, the most common cause of death was diseases classified as infectious and parasitic diseases (e.g. diarrhoea, fever, cough and tuberculosis) which constitutes 88% of all deaths in this age group. Also, 82% of the deceased in perinatal were reported to have had diseases classified as symptomatic conditions (symptoms, signs, and ill-defined conditions) prior to their death. Infectious and parasitic diseases, and/or symptomatic conditions together constitute about 25% of all deaths in this category.

Table 7.2: Observed frequency of the disease profile by age group in the study[a]

Categories(ICD)/ Diseases	Neonatal 1105	Perinatal 997	Adults 1831
Symptoms and signs			
Vomited	83		452
Distension	70		
Breathe difficulties	166		
Headaches		155	726
Total	*290 (26.3%)**	*821 (82.3%)**	*1433(78.3%)**
Infectious			
Fever	186	650	778
Cough	67	360	83
Diarrhoea	36	316	234
Tuberculosis		4	131
Excessive bleeding			73
Total	*220(19.9%)**	*874(87.7%)**	*989(54.0%)*
Injury			
Any injury at birth	65		
Injury and accident		393	711
Total	*65(5.8%)**	*393(39.4%)**	*711(38.8%)**
Non-infectious			
Paralysis	15	40	
Rash	28		
Heart		83	
Bleeding from nose		874	
Asthma		122	579
Convulsion	142		
Epilepsy			71
High blood pressure			697
Total	*40(3.6%)**	*280(28.1%)**	*789(43.1%)**
Complications			
Malformation	51		
Premature	449	140	
End of pregnancy	431		
Total	*713(54.5%)**	*140(14.1%)**	
ENM			
Diabetes		194	194
Malnutrition		174	208
Total		*368(36.9%)**	*402(22%)**
MENTAL			
Mental			342
Total			*342(18.7%)**

[a] For the above total, we are counting the number of deaths in each categories only, cases where the disease appear more once will be counted once. *The percentages indicated the ratio of the number of deaths in each category to the total number of deaths in each age group.

Table 7.3: Percentage of overlapping disease classification for Neonatal data: Two-way classification[a]

	Complications	Non-infectious	Symptoms	Injury	Infectious
Complications	53.57†	2.3	7.3	2.9	3.9
Non-infectious		0.09†	2.5	0.54	1.7
Symptoms			5.79†	2.35	5.74
Injury				1.18†	1.27
Infectious					3.71†

[a] Higher dimension cross classifications are not included because of low percentage and complex structures. Also, percentages lower than 1% are excluded from the table. †Diagonal indicates the percentage share of each disease classification only

Out of the 1831 deaths in adults, the verbal autopsy revealed that 98% of them reported to have had diseases classified as non-infectious/non-communicable and neoplasm (NNN) diseases. About 28% of deaths from adults were reported to have been jointly caused by diseases in the three major classifications: non-infectious, infectious and symptomatic conditions. Deaths from adults generally exceed those from neonatal and perinatal, for instance, in Kabul 58.7% of the deaths occurred in adults (Figure 7.1). Also noted was the geographical heterogeneity in distribution of complications at birth and the puerperium in neonatal (Figure 7.2).

We conduct a separate in-depth analysis of the neonatal data. Exploratory data analysis of the contingency table was used to assess if single cause analysis will be appropriate or not. The Cochran-Mantel-Haenszel chi-square test statistics for independence of causes of death (binary responses) showed a strong association among the variables (p-value < 0.0001). From this result, the association between the two most common causes of death (Complications and Symptomatic conditions) was further explored. The percentage of deaths involving the complications which also include symptomatic diseases was calculated to be 7.3%. The percentage is smaller than would be expected if the two causes were occurring independently in the deaths population. The chi-square test for independence is statistically highly significant, indicating that the two

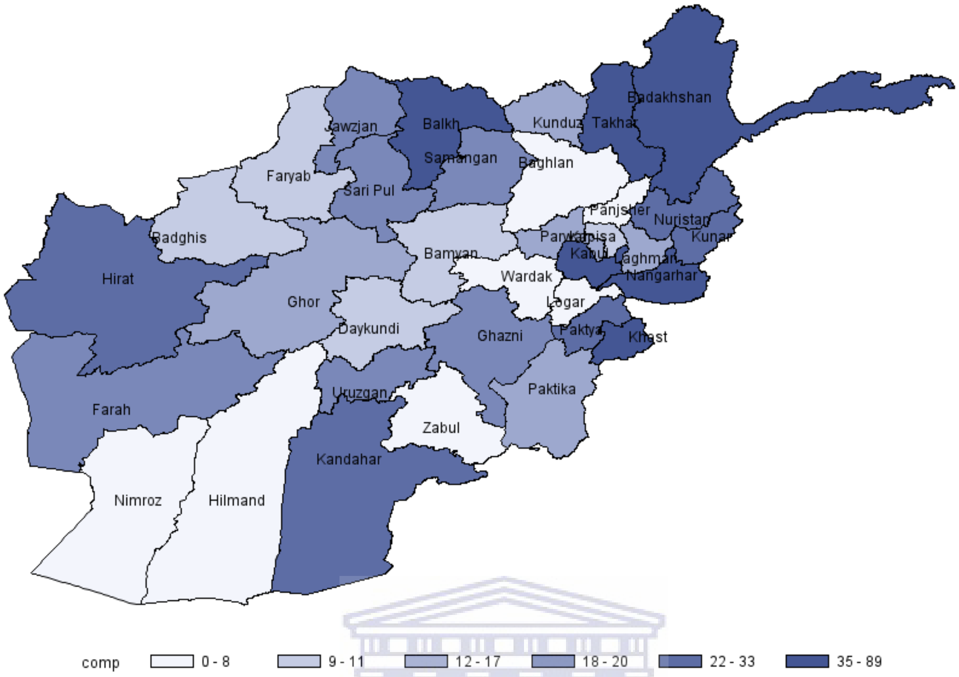


Figure 7.2: Spatial distribution of percentage of the population with complications in the neonatal data

causes are not independent. These results indicated a clear association between causes of death and the appropriateness of multiple cause analysis to capture the association between different causes of death and thus the use of the bivariate probit model.

The bivariate probit model was applied to the neonatal data; the results as illustrated in Table 7.4 suggest a number of highly significant risk factors for disease classified as complications and symptomatic conditions. The spatial bivariate model that accommodates the spatial random effects (Model 3) was compared with the other two models that are similar to that of Neelon *et al.* (2012). The models (Model 1 & 2) include an independent spatial bivariate probit model and a model with uncorrelated spatial random effects.

Joint probabilities of disease classified as complications and symptoms were computed considering the remoteness index, place of residence, age of the child, size at birth, place of birth, wealth index, sanitation and mother's age as risk factors. The Akaike Information Criteria (AIC) was used to assess the performance of the three Models. The results showed that the model which incorporates the spatial level dependency and other risk factors best, describe the joint probabilities of diseases classified as complications and symptoms based on smaller AIC. In Table 7.4, the efficiency of the parameter estimates increased after the inclusion of the correlated random effects. The independent model showed that remoteness index, residence, wealth index, place of birth and sanitation are not associated with the risk of symptoms and complications, but these risk factors were associated in the uncorrelated spatial and correlated spatial models respectively. The standard errors also reduce and thereby produce larger parameter estimates. The differences observed in the estimates among the Models 1-3 are due to the incorporation of spatial correlations in Model 3.

It was found that the correlations between the two major causes of deaths in neonatal (complications and symptoms are statistically significant, p-value <0.0001). The results also indicated an increase in the odds of complications at birth for home delivery (35% higher chance for home delivery). There exists a significant difference in proportions of complications between poorest families and the richest families and a 7% higher chance of complications at birth for families in the poorest quintile. Additionally the analysis of remoteness indicates a significant difference (p-value ¡0.0001) between families in the most remote areas compared with families in the least remote areas (2% higher chance for families in the most remote).

Evidence from the spatial correlated model as illustrated in Table 7.4 and Figure 7.3 indicates that disease classified as symptoms is prominent: In children delivered at home, there is a 25% higher chance than hospital delivery; in children that are small at birth there is a 51% higher chance than normal size; in families in the poorest quintile with a 8% higher chance compared to families in the richest quintile.

Table 7.4: Spatial bivariate model fitted to the neonatal mortality

Risk factors	Independent		Uncorrelated		Correlated	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
Symptoms						
Remoteness: Least remote	0.0299	0.5281	0.0323	0.4853	0.0053	0.9056
Residence: Urban	-0.0154	0.8996	0.0451	0.7408	-0.0467	0.7057
Age	-0.1264	<.0001	-0.1022	<.0001	-0.1381	<.0001
Size at birth: Small	-0.2876	0.0063	-0.2131	0.0404	-0.4119	<.0001
Place of birth: Home	0.1526	0.0577	0.1431	0.1489	0.2188	0.0033
Wealth Index: Poorest	-0.0919	0.0748	-0.0588	0.2811	-0.0718	0.0999
Duration of pregnancy	-0.9665	<.0001	-1.2412	<.0001	-0.7659	<.0001
Sanitation: Poor	0.1903	0.2061	0.1954	0.1965	0.2749	0.0766
Mother's age	-0.0222	0.5426	-0.0108	0.7669	-0.0072	0.8268
Complications						
Remoteness: Least remote	0.0102	0.8064	0.0101	0.805	0.0173	<.0001
Residence: Urban	0.1142	0.4088	0.1088	0.3879	0.1394	0.0146
Age	0.0504	<.0001	0.0514	<.0001	0.0571	<.0001
Size at birth: Small	0.0646	0.4018	0.0645	0.4734	0.1289	0.0203
Place of birth: Home	-0.0373	0.7149	-0.0396	0.6446	-0.0305	0.3521
Wealth Index: Poorest	0.0715	0.1540	0.0585	0.1453	0.0785	0.0147
Duration of pregnancy	0.2953	<.0001	0.2857	<.0001	0.2800	<.0001
Sanitation: Poor	-0.2681	0.0569	-0.2721	0.0571	-0.3562	0.015
Mother's age	-0.0416	0.2056	-0.0467	0.1692	-0.0558	0.0827
σ_1^2	0.1535	0.058	0.05045	0.8287	0.9368	<.0001
σ_2^2	-0.1548	0.1091	0.1012	0.3038	0.0267	<.0001
σ_{12}					-0.0251	<.0001
$\rho = corr(Y^*_{ij1}, Y^*_{ij2} \mid \phi)$					-0.5098	<.0001
ρ_ϕ					-0.0007	<.0001
AIC	1896.7		1813.8		1793.6	

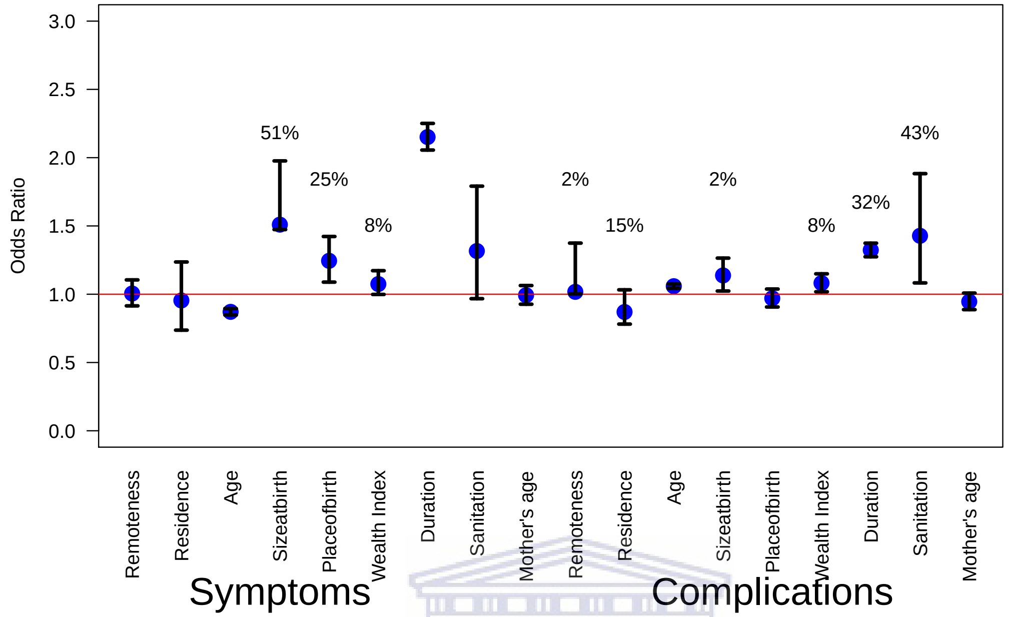


Figure 7.3: Odds ratio for various risk factors from the correlated spatial bivariate probit model fitted to the neonatal mortality and the percentage chances are shown

7.4 Discussion and Conclusion

An understanding of the dynamics of cause-specific mortality is crucial in assessing the health of a population. Several diseases may play different roles in the death process and thus, the concept of multiple analysis. The evaluation of multiple causes of deaths and risk factors can be done by the analysis of their joint probabilities. This study provides an important and crucial empirical study of multiple causes of deaths. In this study, the prevalence of multiple cause of death risk factors was investigated in the neonatal population in the AMS 2010. The AMS 2010 is the first national representative and comprehensive mortality survey in Afghanistan.

Complications at pregnancy, and the puerperium and symptomatic conditions are the top two major of causes death in neonatals. It is known that the contribution of different diseases can vary in the mortality process, some are viewed as a lethal sequel of other underlying conditions (Stallard, 2002). This study revealed significant information about the strength of association between causes of death.

The use of no trained or professional birth attendants is common in developing countries, thus the high rates of complications at birth. In Afghanistan, 1 in every 50 women die of pregnancy related causes, 2 in 5 pregnancy related deaths occur during pregnancy and about 20% within two months after delivery (APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt , 2011). Complications during pregnancy or at birth has serious consequences for the baby and the mother. In this study, 39% of babies are born before term due to early end of pregnancy, 40.6% are born premature and 44% are born with low birth weight. The high rate of complications at pregnancy and the puerperium in Afghanistan could have many causes, including lack of access to skilled birth attendants during delivery and high remoteness (Bartlett *et al.*, 2005), home delivery (Neelon *et al.*, 2012), poor sanitation and lack of access to clean and safe drinking water (Amowitz *et al.*, 2002).

The findings also indicated heterogeneity in the geographical distribution of the causes of death and was confirmed by significant correlation between random effects. Accounting for the within-subject association between causes of death improved the parameter estimates, this improvement was confirmed by smaller AIC. This is consistent with the findings in this study that mothers living in remote and rural areas are prone to complications at birth. The observed geographical differences in the mortality rates could be a result of the barely functional health care system in Afghanistan especially in the rural areas (high mortality rates are observed in the rural areas). The supply of trained health personnel is almost non-existent in the rural areas and about 75% of the physicians in Afghanistan are concentrated around Kabul (Prasad, 2006).

The following limitations are acknowledged in this study: Firstly, in collecting information through VA, the assessment is based on the respondents' ability to provide adequate account of the death process. The enquiries about the death may be upsetting especially in a very sensitive society, thereby raising the issue of the reliability of the data. In developing countries though, where death certificates are lacking for the majority of the population, VA is very crucial for investigations into cause of death. Secondly, although there is a strong correlation between COMPLICATIONS = 1 and SYMPTOMS = 1, this may not be true for other possible combinations, besides there may be other disease classifications that are acting as a confounding factor. Finally, although the use of the ICD is to reduce the number of diseases and complexity with the analysis, the classification may be too broad to pin point the specific disease(s).

In spite of the above mentioned limitations, the use of the spatial bivariate probit model for joint analysis of multiple causes of death provides an opportunity to investigate the influence of both the individual- and province- level risk factors on mortality. It also allows for accounting for dependency at both levels of the data (Neelon *et al.*, 2012). This model is superior to the univariate model when the outcomes are correlated. The model provides an opportunity for a flexible tool for joint analysis using random-effects models. The extension of the bivariate probit model to account for geographical heterogeneity through random effects provides an avenue to model the dependency in the data and allowed for the possibility to estimate the extent of correlation between the different causes of death. The approach not only indicates the dependence of the responses on the demographic variables and other risk factors but it also provides additional information about the relationship between complications and symptoms.

The effect of the social economic status of the family on mortality cannot be overemphasized, the struggle for food and other resources for child survival is overwhelming, especially in the rural areas. Disparities in socioeconomic groups in

Afghanistan is mind-boggling, two-thirds of urban households fall in the richest quintile (APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt , 2011). This study revealed that remoteness, place of residence and wealth index played a significant role in increasing neonatal deaths in Afghanistan.

About 46% of the households in this study do not have access to safe drinking and only one-fifth of the households have an improved toilet facility. Good hygiene is crucial and a simple way of combating many diseases, health seeking of mothers have been shown to have a positive impact on child survival (Kayode *et al.*, 2012). Previous studies have reported that poor sanitation and lack of access to clean water have indirect relation to high infant mortality (Amowitz *et al.*, 2002). The chance of complications of pregnancy or at birth is higher for households with bad sanitation. The same can said about households without access to safe drinking water.

In conclusion, this study has presented the analysis of multiple causes of death using three models. This study has shown that accounting for correlations between multiple causes of death provide better results and precise parameter estimates. The models were assessed by AIC which indicates that the model that analyzes two causes of death simultaneously (jointly), has the smallest AIC. The study has shown significant differences in the mortality rates across Afghanistan and the role played by the remoteness index, especially as related to complication during pregnancy and child birth. The study has revealed that the size at birth, remoteness index, residence, age of the child, wealth index, sanitation and duration of pregnancy are important predictors of diseases classified as complications while age, size at birth, place of birth and duration of pregnancy are strongly associated with diseases classified as symptoms. This study is quite novel because the dependency between two causes of death has not been exploited before in Afghanistan. The findings provide the possibility of simultaneously looking into issues of multiple causes of death in Afghanistan.

Chapter 8

Epidemiological Analysis of Spatially Misaligned Data: A Case of Highly Pathogenic Avian Influenza Virus Outbreak in Nigeria

This chapter has been published as:

Adegboye, O.A. and Kotze, D. (2013). Epidemiological Analysis of Spatially Misaligned Data: A Case of Highly Pathogenic Avian Influenza Virus Outbreak in Nigeria, *Epidemiology and Infection.*
doi:10.1017/S0950268813002136

Summary

This research is focused on the epidemiological analysis of the transmission of the highly pathogenic avian influenza (HPAI) H5N1 virus outbreak in Nigeria. The data included 145 outbreaks together with the locations of the infected farms and the date of confirmation of infection. In order to investigate the environmental conditions that favoured the transmission and spread of the virus, weather stations were realigned with the locations of the infected farms. The spatial Kolmogorov-Smirnov test for complete spatial randomness rejects the null hypothesis of constant intensity ($p < 0 : 0001$). Preliminary exploratory analysis showed an increase in the incidence of H5N1 virus at farms located at high altitude. Results from the Poisson log-linear conditional intensity function identified temperature (-0.9601) and wind speed (0.6239) as the ecological factors that influence the intensity of transmission of the H5N1 virus. The model also includes distance from the first outbreak (-0.9175) with an Akaike Information Criterion of -103.87. Findings in this study showed that geographical heterogeneity, seasonal effects, temperature, wind as well as proximity to the first outbreak are very important components of spread and transmission of HPAI H5N1.

8.1 Introduction

The World Health Organization (WHO) has described avian influenza as an infectious disease of animals (usually birds and less commonly pigs) caused by type A strains of the influenza virus. The main reservoir of the virus is wild birds and in the last two decades it has been seen in commercial and domestics poultry in Asia and Africa. Few cases of transmissions to humans and other mammals have also been documented by the United States Centres for Disease Control and Prevention, and World Health Organization. This virus was first reported in one state in Nigeria in 2006. The disease spread to 25 states and the Federal Capital Territory (FCT) within weeks (Ojo, 2008). The epidemic of H5N1 has huge social-economic consequences with respect to animal welfare, international trade and cost. The federal government set aside the sum of 1.5 billion Naira ($11.5 million) for compensation alone for suspected birds that were culled throughout the nation to contain the spread of the disease. An economic impact assessment made by the Poultry Association of Nigeria (PAN) put the loss to farmers in the first four weeks of the outbreak to 14.4 billion Naira, which does not include those very small-scale farmers whose stock constitutes over 60% of the total national poultry population (Duru, 2006). The consequence of the outbreak is devastating to the poultry owners as well as the consumer.

The initial boycott of chicken and eggs created high demand and pressure on the supply of fish and other meat products. Outbreaks of HPAI had been confirmed in some parts of Africa: Benin, 2008, 2007; Burkina Faso, 2006; Cameroon, 2006; Cote d'Ivoire, 2006; Djibouti, 2007, 2006; Egypt, 2008,2007, 2006; Ghana, 2007; Nigeria, 2009, 2008, 2006; Niger, 2006; South Africa, 2012, 2011, 2006; Sudan, 2007, 2006; Togo, 2009 , 2008, 2007 (World Organization for Animal Health , 2012). Despite the spread and incidence of H5N1, some farmers do not believe in its existence. This is very disturbing and may influence human infection (Fasina, 2008).

Many studies have attributed the spread of the H5N1 virus in Nigeria and Africa to international and inter-state poultry trade (Fasina, 2008; Bello *et al.*, 2008); as a result of regional influx of wild birds (Fasina, 2008); presence of visitors on farm premises; purchase of live poultry and poultry products by farmers, and farm workers living outside the farm premises (Fasina *et al.*, 2011). Many households practiced poultry farming at home (backyard) thereby increasing the risk of transmission of the avian virus to humans (Musa *et al.*, 2010). The management of backyard poultry farming has been identified as a challenge to the epidemic of H5N1 in Nigeria and improved biosecurity measures was advocated to prevent and control the spread of the disease (Musa *et al.*, 2010). In their study about one-third of the participants were aware of bird to man transmission of the avian virus and about two thirds were not aware of the existence of H5N1 infection at all. The knowledge and awareness of avian flu and risk of infection is low and hygiene is an important element to the disease control (Musa *et al.*, 2010). Three distinct waves of the H5N1 epidemic in Nigeria were identified with peaks in January to March in the North West, North Central and North East and July to September in the South West and South South as well as disease clusters in the North West, North East and South West, respectively (Ekong *et al.*, 2012). A previous analysis using the methods for spatial point processes with altitude as the only covariate in making inferences about H5N1 outbreaks revealed a significant effect of altitude in estimating the pattern of H5N1 virus (Adegboye, 2010b).

It is difficult to describe the mechanism underlying the H5N1 epidemic, its origin, the geographies, climatic and other factors establishing and influencing the spread of the disease. Such difficulties can be a result of mismatch between data measured at different resolutions resulting in spatial misalignment (Banerjee *et al.*, 2004). Spatial misalignment of the exposure and response variables can bias the estimation of health risk (Peng and Bell, 2010). The first major aim of this study was to realign the weather stations with the locations of the infected farms. The locations of the outbreaks of H5N1 were not linked with the climate data measured at different weather stations. The use of interpolation to estimate the weather data to provide a mean estimate for each infected farm was considered. Moreover an appropriate functional relationship between H5N1 outbreaks and possible association with environmental factors such as altitude, temperature, wind and dew were investigated. The geographical constraints that operate under the outbreak of the disease were explored to understand the effect of distances between farms.

8.2 Methods

8.2.1 Data Sources

The data for this study consists of the HPAI H5N1 outbreaks in 2006 that were reported to the World Organization for Animal Health (OIE) (World Organization for Animal Health , 2012). The information collected include the date of confirmation of infection (diagnosed), species, number of susceptible birds, number of infected birds and the locations of the infected farms (X,Y coordinate system). The data sets can be found on the website of the World Organization for Animal Health (World Organization for Animal Health , 2012). Poultry farms (subsequently referred to as farms) in Nigeria is categorized into backyard flocks which comprises of a couple of dozen birds and commercial poultry with up hundreds of thousands chickens. Figure 8.1 shows the locations of the infected farms which include cases from both commercial and backyard farms. The disease was either diagnosed using clinical signs or sent for a test at the national laboratory (National Veterinary Research Institute, Vom, Nigeria or the OIE's reference laboratory in Padova, Italy) (World Organization for Animal Health , 2012). A total of 145 diagnosed outbreaks were recorded; the first outbreak occurred on the 10th of January 2006 in Kaduna State, in the North-Western part of Nigeria. The virus spread to the central part of the country within weeks, and about 70% of the total reported outbreaks occurred between January and March, 2006 (Figure 8.2).

The climatic data for the available weather stations (Figure 8.1 shows the location of the weather stations) in Nigeria in 2006 were extracted from the National Climatic Data Centre of the National Oceanic and Atmospheric Administration, US Department of Commerce. Daily weather data were collected from 48 weather stations across Nigeria; these data included the temperature (degree Celsius), dew point (degree Celsius), wind speed (knots), pressure (millibars) and elevation measured in metres.

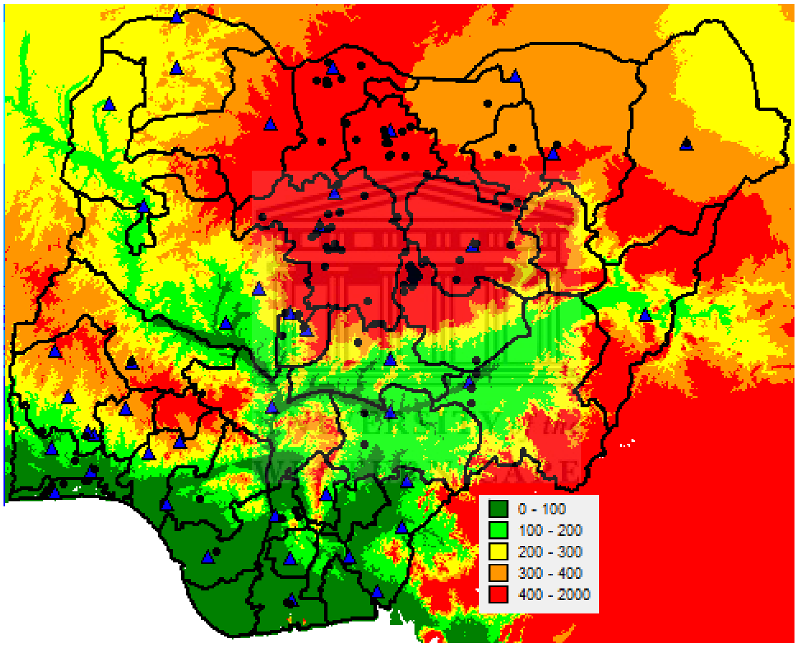


Figure 8.1: The locations of HPAI H5NI virus infected farms (black stars) and weather stations (red triangle): altitude in the background

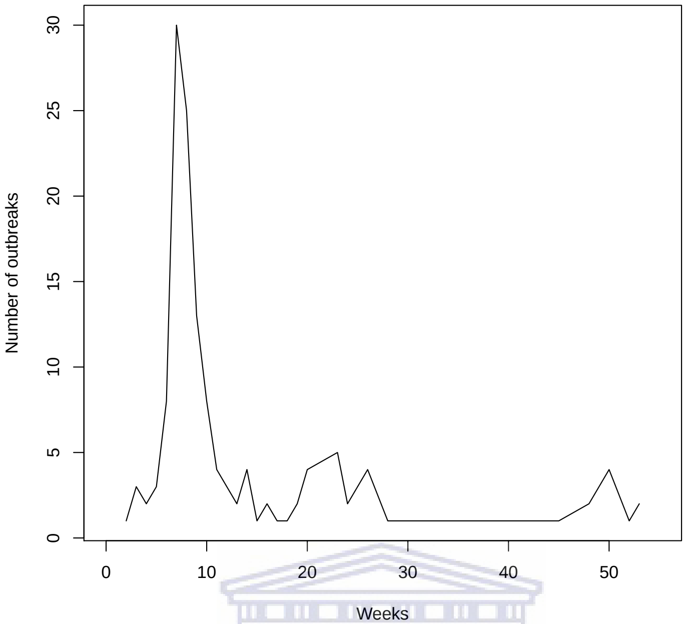


Figure 8.2: Weekly number of cases of HPAI H5N1 since the beginning of the outbreak in Nigeria, January 2006

8.2.2 Bayesian Kriging

Misalignment is often encountered in spatial analysis, this occurs when sampling at different spatial scales are not linked. In this study, the outbreaks of the H5N1 virus were not linked with the weather variables measured at different weather stations. One possible way to tackle this problem is through interpolation, a method of constructing new data points within the range of a discrete set of known data points.

Let $s_i, i = 1, 2, ..., N$ be the location of the infected farms and $Z(s_l), l = 1, 2, ..., L$ the climatic variables measured at the weather stations. Usually, it is desirable to measure these climatic variables $Z(s_l)$ at the location of the infected farms, which involves the use of interpolation (Lawson, 2008). The spatial Gaussian process was assumed for the measured predictor and a conditional prediction of the predictors at the set location were constructed. The Gaussian process model was defined for the sets of weather stations,

$\theta = (\psi, \tau)'$ as the parameter vector. Let $Z'_s = Z(s_1), ..., Z(s_l) : Z_s|\alpha, \theta \ N(\mu, \Gamma)$, where $\mu_{sl} = \mu(s_l, \alpha)$ is a predictor at the l^{th} farm and Γ is the spatial covariance matrix. Often, μ_s will consist of trend surface components and $\Gamma_{ll'} = \tau\rho(s_l - s_{l'}; \psi)$ (where τ is the variance and $\rho(.)$ is a correlation function for the Z's at separation distance $s_l - s_{l'}$. Using the conventional geo-statistics approach for interpolation, "kriging", the covariance structure is estimated first, then the estimated covariance is used for interpolation. The exponential form was used for $\rho(s_l - s_{l';\psi})$ given by:

$$\rho(s_l - s_{l';\psi}) = exp\{-\psi\|s_l - s_{l'}\|\}, \tag{8.1}$$

where $\|s_l - s_{l'}\|$ is the Euclidean distance between farm l and l'. The matern function was used to estimate $\gamma(h; \theta)$, $h = s_l - s_{l'}$ given by

$$\begin{cases} \gamma(h;\theta) = c_0 + c_k \left[1 - \dfrac{1}{2^{\alpha-1}\Gamma(\alpha)}(\dfrac{h}{a_k})^{\alpha} K_{\alpha}\dfrac{h}{a_k}\right], & \text{if } h > 0, \quad (8.2a) \\ 0, & \text{if } h = 0. \quad (8.2b) \end{cases}$$

where $\theta = (c_0, c_k, a_k, \alpha); c_0 \geq 0, c_k \geq 0, a_k \geq 0, \alpha \geq 0, K_\alpha(.)$ is the modified Bassel function of the second order of α. Here, c_0 measures the nugget effect and c_k is the partial sill (so $c_0 + c_k$ is the sill). As in the stable family, the behaviour near the origin is determined by α, and the parameter a_k controls the range. An advantage of this model is that the behaviour of the semivariogram near the origin can be estimated from the data rather than assumed to be a certain form (Waller and Gotway, 2004).

Bayesian Kriging analysis was adopted for predicting a new set of locations $Z(s_i), i = 1, 2, ..., N$. The posterior predictive distribution of $Z(s_i)$ given the observations $Z(s_l)$ is

$$f(Z(s_i|Z(s_l)) = \int f(Z(s_i|Z(s_l), \alpha, \theta)f(\alpha, \theta|Z(s_l))\partial\alpha\partial\theta \tag{8.3}$$

where $f(\alpha, \theta|Z(s_l))$ is the posterior of the model parameters.

The matern function 8.2a was implemented in `geoR` and the Bayesian analysis was implemented using the function `krige.bayes` (Ribeiro and Diggle, 2001).

8.2.3 Modelling H5N1 Intensity

The variations of the disease intensity were measured by inhomogeneuos K-function. In general, the intensity of a point process varies from place to place and the intensity function assumed that the expected number of points falls in a small area du around a location u_i that satisfies $E[N(X \cap B)] = \int_B \lambda(u)du$. The inhomogeneous K-function stipulates that if $\lambda(u)$ is the true intensity function of the point process X then each point x_i will be weighted by $w_i = \frac{1}{\lambda(u)}$:

$$K_{inhom(r)} = E[\frac{1}{\lambda(u)} \sum_{x_i \in X} \frac{1}{\lambda(x_j)} 10 < \|u - x_j\| \leq r | u \in X] \tag{8.4}$$

The space-time cluster interaction was measured at both spatial and temporal scale using

$$K = \frac{LR}{n_i n_j} \sum \sum \frac{I_{h,t}(d_{ij})}{w_{ij}} \tag{8.5}$$

where n_i and n_j are the numbers of observations within distance h and time interval t for the pair of events respectively .

The spatial point process model was fitted to the data in which the point pattern is dependent on spatial covariates such as geographical location, week of infection, elevation, temperature, dew point and wind. The conditional intensity, presence per unit area (Cressie, 1993), expressed as a log-linear function of the covariates is

$$\lambda(u) = exp(\beta' Z(u)) \tag{8.6}$$

where β is vector parameter and covariates at location u.

The Akaike Information Criterion (AIC) (Burnham and Anderson, 2002) was used to evaluate the "goodness of fit" of each model. All data analyses were performed and implemented in R (R Development Core Team, n.d.).

8.2.4 Habitat Suitability Analysis

An Habitat Suitability Analysis (HSA) provides a way to model the relationship between species and habitat characteristics. The Habitat Suitability Index (HSI) describes the suitability of a given habitat by combining the interactions of all key environmental variables on a species' population characteristics and ultimately, survival (Atlantic Ecology Division, 2013). Although the HSA is popular with species modelling, it is a growing area in spatial analysis of disease epidemiology that characterizes the distribution of the disease in a space defined by environmental parameters. The model can be constructed in several ways such as Ecologic Niche Analysis (ENA) which involves the use of geography and ecology of disease transmission (Peterson, 2006) with the aim to determine the distribution of the species using the location where the species has been found (Fischer *et al.*, 2011). The idea is to use the observed presence (and absence) data together with ecological variables at those sites to provide a reasonable likelihood for the species to be present (Fischer *et al.*, 2011). Hirzel *et al.* (2002) proposed the Ecological Niche Factor Analysis (ENFA), a multivariate approach to study the geography of species distribution which does not require an absence data. ENFA computes the suitability function by computing the species distribution in the eco-geography variables space with that of the whole set of cells (Arnese, 2007). In this study, ENFA was used to compute the suitability function and produce the habitat suitability map for H5N1 where the HSI values ranges from 0 to 100%. HSI values close to 100% indicates an area where a species will flourish and values close to 0 imply an area where the species will not do very well.

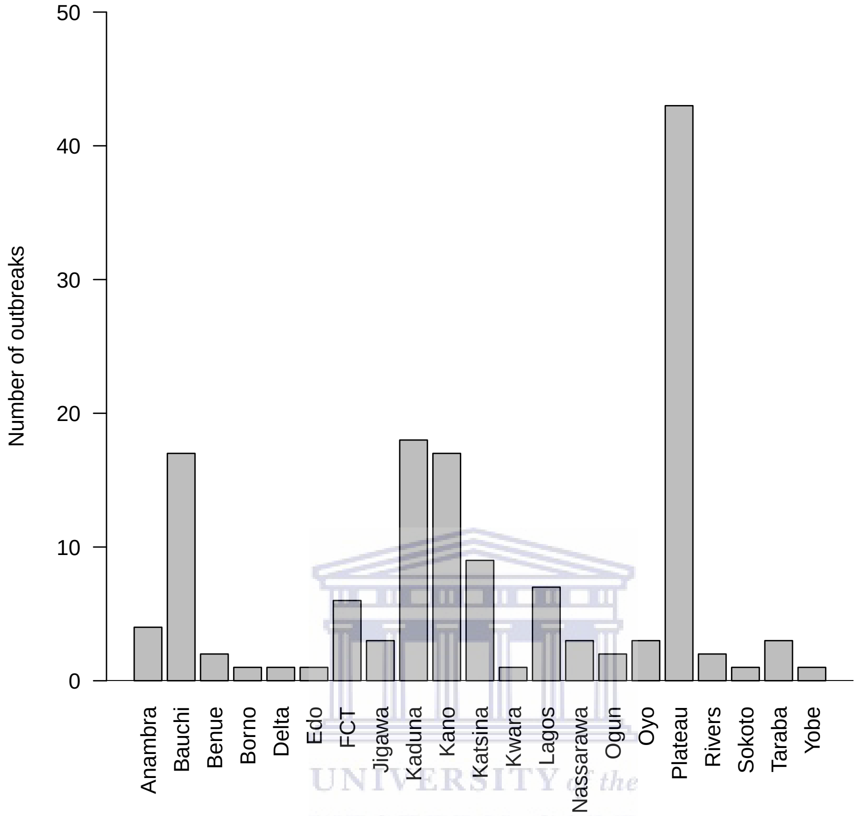


Figure 8.3: Cases of HPAI H5N1 virus outbreak in Nigeria in 2006

The ecological niche factor analysis was performed using the `enfa` function implemented in `R` (Basille *et al.*, 2008).

8.3 Results

A high percentage of the outbreaks occurred in the Plateau State (29.7%) followed by Kaduna State (12.4%), then Kano (11.7%) and Bauchi (11.7%) (Table 8.1 and Figure 8.3). Preliminary exploratory analysis showed an increase in the incidence of H5N1 virus at farms located at high altitude.

Table 8.1: The most affected states of H5N1 virus in 2006

States	†Cases(%)	‡Duration(Days)
Plateau State	43(29.7%)	143
Kaduna State	18(12.4%)	346
Kano State	17(11.7%)	346
Bauchi State	17(11.7%)	322
Kastina State	9(6.2%)	21

†Percentage of all cases

‡Number of days from the first reported cases and the last reported cases.

8.3.1 Estimating Weather Variables

A total of 48 weather stations across Nigeria were used for the interpolation of the weather data in this study. The location of the weather stations (black star) and the infected farms (red triangle) are depicted in Figure 8.1. The occurrence of H5N1 has been described to take place differently in waves (Ekong *et al.*, 2012), thereby suspecting seasonal effects on the outbreaks (Figure 8.2). This influence may be explained by some climatic variables and the disease epidemiology can be explored using ecological data. The weather stations are misaligned with the location of infected farms. The first task is to realign the weather stations. Interpolation was used to estimate the mean weather variables (temperature, wind, dew) for the infected farms (Figure 8.4). Daily data were not available for some weather stations and thus weekly average weather predictions were used.

Different parametric models were fitted to the empirical semivariogram estimated to the data and the best fitted model (matern) was chosen by varying the parameters. The estimated Bayesian variograms were checked against the empirical variogram by plotting the posterior distribution means, medians and modes. Histograms of the posterior distribution for the model parameters indicated a good fit of the Bayesian Kriging prediction of the weather variables for the locations of infected farms (not shown here).

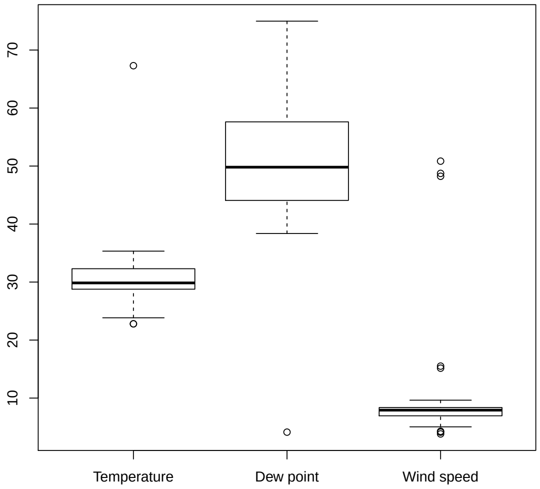


Figure 8.4: Box plots of the predicted weather variables

8.4 H5N1 Risk Factors

The spatial Kolmogorov-Smirnov test for complete spatial randomness rejects the null hypothesis of constant intensity with a $p-value < 0.0001$. This implies that the locations of outbreaks are dependent on each other and that propensity varies from location to location (inhomogeneous). The inhomogeneous K-function plot shows that estimates of K(r) using different techniques are roughly the same (Figure 8.5). This suggests that the outbreaks of H5N1 appear to be clustered after accounting for temperature, elevation, dew, wind and distance from the first outbreak. In order to test whether the point pattern depends on other covariates, an inhomogeneous Poisson process intensity was fitted as a function of the covariates. The full model consist of temperature, wind, elevation, dew point and distance. Using the method of backward elimination, the model with temperature, wind and distance was found to best fit the data and to be suitable for predicting the intensity of H5N1 outbreak with an Akaike Information Criterion (AIC) of -103.87.

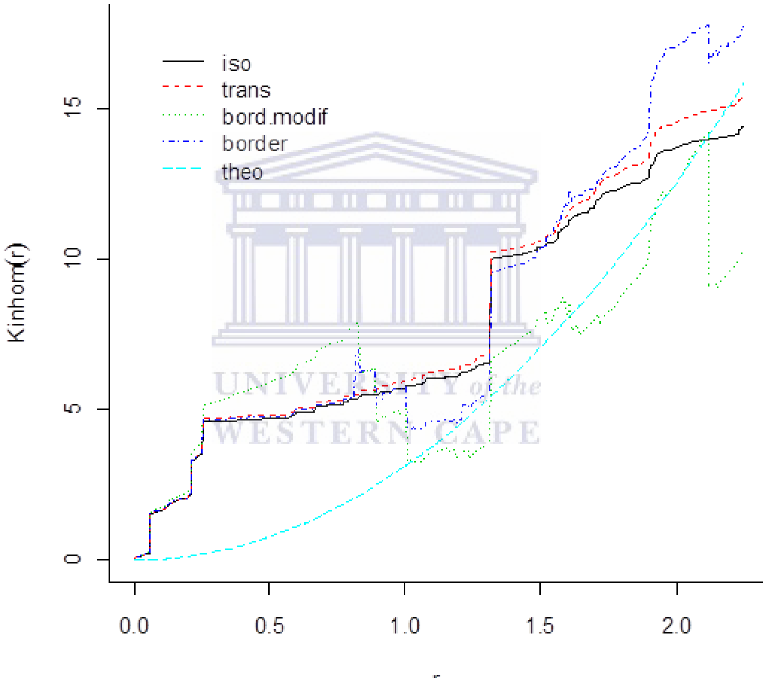
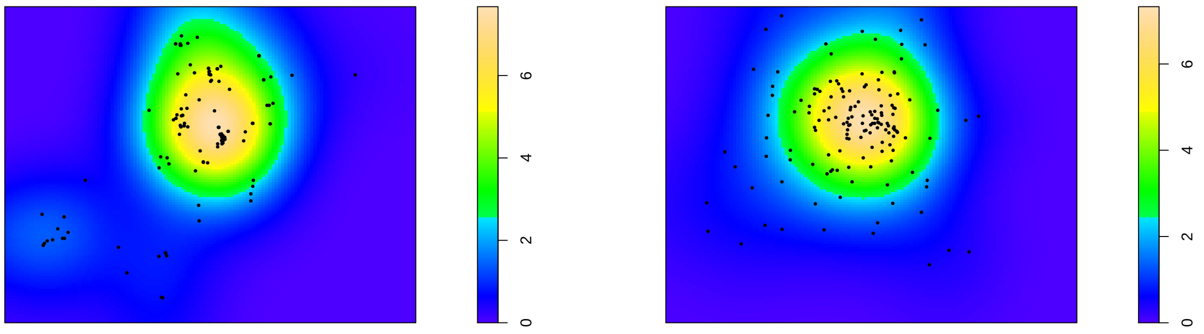


Figure 8.5: K-inhomogeneous functions using different edge corrections: theoretical Poisson, border-corrected estimate, translation-corrected estimate and Ripley's isotropic correction estimate



(a) Observed H5N1 (b) Predicted H5N1

Figure 8.6: HPAI H5N1 risk map from Poisson point process model with log linear intensity: The intensity of the transmission of the virus (i.e. the number of outbreaks per unit area) is indicated by different colours with values ranging from 0 to 8.

Results from the Poisson point process model with a 1 by 1km regular grid of quadrature points showed negative effect of temperature and distance. This suggest that for a unit increase in temperature and distance from the first outbreak, the intensity of the transmission (number of outbreaks per unit area) of H5N1 outbreak decreases by 0.9601 and 0.9175, respectively. The result shows that an increase in wind speed will increase the conditional intensity of the disease outbreak. Also, the negative effect of distance on the intensity of the transmission, implied that an increase in the distance result in a decrease in the intensity of the transmission of the virus. Figure 8.6 presents the intensity of the transmission of the H5N1 virus using Poission process model; the fitted model exhibit similar patterns in the transmission of the virus as the observed H5N1 outbreak. The intensity of the transmission (indicated by different colours in Figure 8.6) suggest that most of the infected farms are located in the North Central region of Nigeria and occurred within a short time frame. The habitat suitability map for the disease is presented in Figure 8.7; higher HSI were noticed around the Northern part of the country (60%-100%); this implied that the H5N1 virus is favoured by the climatic condition in the North Central region of Nigeria and areas in close proximity to the first outbreak.

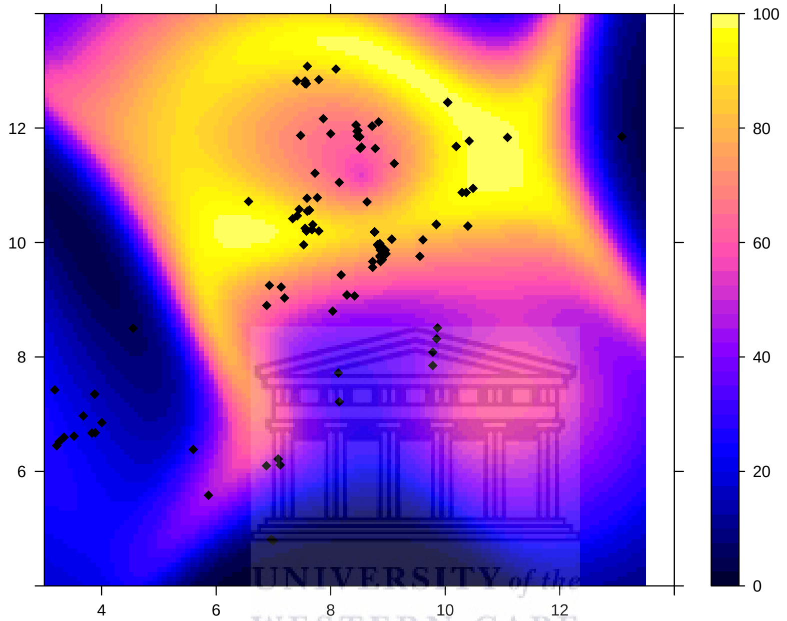


Figure 8.7: Habitat Suitability Index (0-100%) for HPAI H5N1 virus outbreak in Nigeria using the predicted weather data as environmental covariates

8.5 Discussion and Conclusion

It is difficult to describe the mechanism underlying the H5N1 epidemic, its origin, the geographies, the climatic, and other factors establishing and influencing the spread of disease. Information on the spread of infectious diseases is crucial for the control and surveillance initiative programs. The role played by weather variables in the transmission and spread of HPAI should not be overlooked. The H5N1 virus was found to be a result of regional influx of wild birds (Fasina, 2008); migratory waterfowl (Hanson *et al.*, 2005;

Krauss *et al.*, 2004) trade routes (Gilbert *et al.*, 2006) and also the effect of transportation of farm staff and poultry products (Fasina *et al.*, 2011; Oyana *et al.*, 2004). The results showed that the spread and transmission of HPAI H5N1 is favoured by some weather variables and areas in close proximity to the first outbreak.

The movement of people, equipment and animals within infected farms may contribute to the spread of the pathogen. Airborne transmission of the pathogen may also occur through dust and feathers. This method of transmission has been previously studied (Spekreijse *et al.*, 2011; Herfst *et al.*, 2012; Ssematimba *et al.*, 2012; Mikkelsen *et al.*, 2003); this attest to the findings in this study, suggesting that there is a relationship between the intensity of the disease in a given region and, wind speed and temperature. The contribution of wind in the transmission of the pathogens has been described to occur only in short distances and the extent may be relatively small (Ssematimba *et al.*, 2012; Mikkelsen *et al.*, 2003). Although the effect of biosecurity has been discussed in another paper (Musa *et al.*, 2010), temperature and wind may induce dryness of the environment which may in turn make it easier for the virus to be transmitted between birds. Moreover the effect of alternating temperature highs and lows may imply seasonality which could alter the transmission and survivability of the virus, which is consistent with the temporal patterns observed in their work (Ekong *et al.*, 2012). In their work, Ekong *et al.* (2012) suggested that the outbreak occurred in different waves. Brown and Rohani (2012) have looked into the effect of alteration between migratory shorebird and horseshoe crabs as a result of climate change. Vandegrift *et al.* (2010) also reviewed how anthropogenic change may alter the evolution and transmission of influenza viruses. The initial outbreak of the disease in Nigeria was attributed to movement of migratory birds (Fasina *et al.*, 2009); this supports the claim wind is a risk factor because the wind speed and direction may also affect the path of the migratory birds thereby altering the spread of the virus.

This study provides empirical analysis of how spatial models can be used to capture the intensity and mechanism of the spread of the H5N1 virus with misaligned weather data. The application of the Bayesian Kriging method to predict weather variables in a misaligned data setting was demonstrated. These techniques were applied to the HPAI H5N1 virus outbreak in Nigeria in 2006. The spatial Poisson process intensity fitted as a log-linear function has shown that intensity combined with temperature, wind and distance from the first outbreak provides a better estimate. A habitat suitability analysis was conducted using ecologic niche modelling on the point process and the predicted weather data, to explore the habitat suitability of H5N1 and to predict the geographical distribution of the disease outbreak.

The major limitation of this study lies in the data gathering and risk factors. The data on HPAI H5N1 virus were only extracted from the OIE database and cases of under reporting cannot be ruled out. Also, Nigeria is a very extensive country with very poor data acquisition and the lack of e-data sources may affect timely reporting of disease occurrence. It would be interesting to be able to have additional variables linked with the location of the events (for example, those that address biosecurity, farm characteristics and measure virus dispersion). This study only considered the ecological aspect of the disease transmission due to availability of data. Furthermore, it was not possible assess the extent (distance) of the contribution of the wind in the transmission of HPAI based on limited data. In spite of these limitations, OIE provide a very reliable and considerable data that yielded extensive information crucial for HPAI in Nigeria. The analysis based on this data offers the opportunity to detect whether environmental and ecological variables are potential risk factors used to map potential suitable habitats.

The analysis using point process modelling also showed that geographical heterogeneity and seasonal effects are very important components of spread and transmission of HPAI H5N1. The study calls for adequate surveillance and quick quarantine of infected farms in close proximity to the first outbreak. Further studies could incorporate and utilize the wind direction and dust dispersion (as proxy for airborne dispersion of the virus), farm population density and biosecurity in determining the pattern and mechanism of the spread of disease.

Part II: Robust Methods for Spatial, Temporal and Spatio-Temporal data

Chapter 9

A Robust Method for Modelling Spatial Correlation: Analysis of Malaria Incidence in Afghanistan

This chapter has been submitted for publication as:

Adegboye, O.A., Kotze, D., Leung, D.H.Y. and Wang, Y.G. (2013) A Robust Method for Modelling Spatial Correlation: Analysis of Malaria Incidence in Afghanistan. *Submitted for publication.*

Summary

Malaria epidemics often occur in areas with non-immune populations living in arid and desert-fringe zones. It is of high importance to properly identify the risk factors that are associated with the incidence of malaria. A crucial step in modelling spatial data is the specification of the spatial dependency via a spatial correlation function. However, often the choice for a particular application is unclear and diagnostic tests will have to be carried out following fitting of a model. To resolve this problem, a more robust method is proposed for modelling spatial correlation by combining a few candidates using the generalized method of moments. The method is applied to malaria data from Afghanistan. Results show that the use of a combination of candidate correlation structures is superior to those using any one of the candidate structures singly.

9.1 Introduction

Malaria accounted for nearly one million deaths in 2008 globally and about 500 million annual cases of malaria incidence, predominantly in tropical and sub-tropical countries. Malaria is caused by the protozoan parasite of the genus Plasmodium debilitates (Mbanefo *et al.*, 2009). Afghanistan is a malaria-endemic country. It has the second highest burden of malaria in the region (Safi *et al.*, 2009), with Plasmodium *vivax* accounting for nearly 90% of malaria cases. Plasmodium *vivax* malaria is associated with rice growing areas and is transmitted by the endophilic and exophilic rice-field breeders Anopheles *pulcherrimus* and Anopheles *hyrcanus* (Faulde *et al.*, 2007).

In Afghanistan, parts of the country that lie more than 2000 m above sea level are free of malaria but in lower elevation areas, malaria is prevalent between April and December (Centers for Disease Control and Prevention, 2012). Some cases of Plasmodium *falciparum* have been documented in areas of high altitude (> 2400 m) (AbdurRab *et al.*, 2003). The intensity of transmission may depend on a number of factors including, climate, geography, human population, socio-economic level and vector ecology as well as control strategies (Senn *et al.*, 2010).

In this study, the spatial aspect of the transmission since spatial locations encompass a number of the factors simultaneously was considered. It was assumed that malaria incidence across different spatial locations (provinces) may be correlated. The interests are to understanding the risk factors that are associated with the incidence of Malaria in these locations and to be able to make a prediction at unobserved locations. One of the most challenging issues in spatial analysis is the choice of a valid and yet flexible correlation (covariance) structure. In cases of high dimensionality of data, where the number of spatial locations that produced the observations is large, the spatial analysis of such data sets presents great computational challenges.

Spatial dependence in the data can be modelled via Generalized Estimating Equations (GEE, Liang and Zeger, 1986), wherein the covariance matrix is structured by using a working correlation matrix, fully specified by a vector of parameters. An advantage of GEE is that the parameter estimates and their robust variances are consistent even when the correlation structure is misspecified. However, choosing the working correlation structure closest to the true structure increases the statistical efficiency of the parameter estimates. Consequently, it is desirable to use a working correlation matrix that models closely the underlying correlation structure. The choice of a particular correlation is often unclear and diagnostic tests will have to be carried out following fitting of a model (McShane *et al.*, 1997). Cressie and Huang (1999); Gneiting (2002); Stein (2005); Porcu *et al.* (2007) discussed some of the popular choices of correlation structures; other covariance functions include that of Zimmerman (1989); Fan *et al.* (2007); Cressie and Johannesson (2008) to mention a few. Zimmerman (1989), suggested that the covariance matrix possesses some types of pattern that can be exploited to reduce the computational burden of analysis. The use of semiparametric models for the covariance function that imposes a parametric correlation structure while allowing a nonparametric variance function in longitudinal data was proposed by Fan *et al.* (2007). Kriging is a very popular method for spatial analysis, even with data sets with a high spatial dimension (see for example: Banerjee *et al.* (2008); Cressie and Johannesson (2008)). Banerjee *et al.* (2008) used a class of models that were motivated from the idea of kriging, Cressie and Johannesson (2008) proposed a method called fixed rank kriging for very large spatial data sets, where the covariance matrices were specially designed so that the matrix manipulations were of a fixed magnitude (Bai *et al.*, 2012). it is of opinion that some of these spatial structures may not singly adequately reflect the spatial dependency in the data.

The goal of this paper is to resolve this problem by proposing a more robust method for modelling spatial correlation. This approach is in the same spirit as the combining estimating equations of Leung *et al.* (2009) and Bai *et al.* (2012). Leung *et al.* (2009) proposed the use of a hybrid method that combines multiple GEEs based on different working correlation models, using the Empirical Likelihood (EL) method (Qin and Lawless, 1994) in a longitudinal setting. This method is computationally intensive and require a lot of programming for the EL algorithm. In their study, Bai *et al.* (2012) proposed the use of a joint composite estimating function to estimate the spatiotemporal covariance structures. This version uses different correlation structures and combines the resulting estimating equations using the platform of Generalized Method of Moments (GMM, Hansen, 1982).

9.2 Methodology

9.2.1 Model

Let $y_1, y_2, ...y_n$ be the counts of malaria incidence at the $n = 34$ provinces $s_1, s_2, ..., s_n$ of Afghanistan. Associated with the provinces are covariates $x_1, ..., x_n$ that measure the spatial location (provincial headquarters) as well as other covariate information. It is natural to model malaria incidence at each province as a Poisson count. However, preliminary investigation revealed possible overdispersion of malaria incidence across the different spatial locations. In order to model spatial correlation and overdispersion, a nonnegative weakly stationary must be assumed for the latent process $e_1, ..., e_n$ such that conditional on the e's, the y's are independent and follow a log-linear model given by

$$\mathrm{E}(y_i|e_i) = \exp(x_i^T\beta)e_i, \quad \text{and} \quad \mathrm{Var}(y_i|e_i) = \mathrm{E}(y_i|e_i), \tag{9.1}$$

where the $\beta's$ are a set of unknown parameters. Assume $\mathrm{E}(e_i)$ to be 1 so that $\exp(x_i^T\beta)$ represents the marginal mean of y_i. The latent process e_i is assumed to have a variance of σ^2 and the covariance between e_i, e_j is given by

$$\mathrm{Cov}(e_i, e_j) = \sigma^2\rho(z_i, z_j, \alpha) \tag{9.2}$$

where z_i, z_j are covariates from s_i, s_j that jointly induce spatial correlation and α are unknown parameters. This model was used in Zeger (1988) for handling correlation in time series data and McShane *et al.* (1997) for spatial counts. Under this model, it can be easily shown that

$$\begin{aligned}
\mathrm{E}(y_i) &= \exp(x_i^T\beta) = \mu_i(\beta), \\
\mathrm{Var}(y_i) &= \mu_i(\beta) + \mu_i(\beta)^2\sigma^2, \\
\mathrm{Cov}(y_i, y_j) &= \rho(z_i, z_j, \alpha)\{1 + (\sigma^2\mu_i(\beta))^{-1}\}\{1 + (\sigma^2\mu_j(\beta))^{-1}\}^{\frac{1}{2}}.
\end{aligned}$$

If $\rho(z_i, z_j, \alpha) = 0$, an overdispersion Poisson model will be formed; furthermore, if $\sigma^2 = 0$ then, this will result in a standard Poisson model at each spatial location.

A crucial step in modelling spatial data is the specification of the spatial correlation. Cressie (1993, pp 61-64) discussed some popular choices of the correlation matrix. However, often the choice for a particular application is unclear and diagnostic tests will have to be carried out following fitting of a model (McShane *et al.*, 1997). To resolve this problem in this article, a more robust method for modelling spatial correlation was adopted.

Following Liang and Zeger (1986), a Generalized Estimating Equations (GEE) type model can be used to estimate the parameters β, by solving the following estimating equation:

$$S(\beta, \alpha) \equiv D^T V^{-1}\{y - \mu\} = 0, \tag{9.3}$$

where $\mu = (\mu_1, ..., \mu_n)^T$, $D = \partial\mu/\partial\beta^T$, and V is a $n \times n$ variance-covariance matrix of $y = (y_1, ..., y_n)^T$. The matrix V can be expressed as $A + \sigma^2 AR(\alpha)A$, where $A = \text{diag}(\mu_1, ..., \mu_n)$ and $R(\alpha)$ is a $n \times n$ correlation matrix, with the i, j-th element equal to $\rho(z_i, z_j, \alpha)$, and σ^2 is the scale parameter used to model overdispersion.

Now consider different, linearly independent choices of $R(\alpha)$, say $R^j(\alpha), j = 1, ..., J$, and write $S^j(\beta, \alpha)$ for the estimating equation (9.3) using working correlation matrix $R^j(\alpha)$. Let $h(\beta, \alpha) \equiv (S^1(\beta, \alpha)^T, ..., S^j(\beta, \alpha)^T, ..., S^J(\beta, \alpha)^T)^T$ and note that $h \equiv h(\beta, \alpha)$ is a function of β, α only. Also note that

$$h(\beta, \alpha) = (S^1(\beta, \alpha)^T, ..., S^J(\beta, \alpha)^T)^T = \begin{pmatrix} D^T\{A + \sigma^2 AR^1(\alpha)A\}^{-1}\{y - \mu\} \\ \vdots \\ D^T\{A + \sigma^2 AR^J(\alpha)A\}^{-1}\{y - \mu\} \end{pmatrix}. \tag{9.4}$$

In general, the dimension of h may be higher than that of β. The Generalized Method of Moments (Hansen, 1982, GMM) can be used to combine the estimating equations

$S^1, ..., S^J$. The idea of the GMM is to find β, α that jointly minimize the following quantity:

$$h(\beta, \alpha)W(\beta, \alpha)h^T(\beta, \alpha), \tag{9.5}$$

where W is a weight matrix. Hansen (1982) showed that the optimal choice of W^{-1} is the variance covariance matrix of the components of h, in the sense that for the resulting estimate of β, α is semiparametrically efficient under the assumptions of the model. The solution to (9.5) can easily be obtained by algorithms such as the Newton-Raphson method.

The interest is if one of the $S^1, ..., S^J$ is the correct estimating equation, in the sense that it solves (9.3) with $A + \sigma^2 AR(\alpha)A = V^{-1}$. In that case, then the GMM estimate will be optimal. If none of them is correct, the GMM estimate is still consistent and combines optimally the information in S^1,...,S^J.

In practice, a few popular choices of $R(\alpha)$, for example, those discussed in Cressie (1993) may be used; other simpler forms of correlation matrices, for example, exchangeable, $AR(1)$ and $MA(1)$ will be explored. These simpler forms of correlation matrices may become valuable in practice if they can jointly approximate unknown and difficult forms of the correlation structures.

9.2.2 Correlation Structures

In spatial analysis of disease incidence the specification of spatial association must be chosen carefully. If the correlation structure is misspecified, the estimators of the fixed effects will be inefficient and moreover, standard errors of the estimate will be incorrectly estimated, resulting in erroneous inference. Few choices of the spatial correlation $\text{Corr}(Y_{ij}, Y_{ik})$ considered are listed below:

i. Compound symmetry:

$$\text{Corr}(Y_{ij}, Y_{ik}) \begin{cases} 1 & \text{j=k} \qquad (9.6a) \\ \alpha & \text{j} \neq \text{k} \qquad (9.6b) \end{cases}$$

where α is the correlation.

ii. Independent:

$$\text{Corr}(Y_{ij}, Y_{ik}) \begin{cases} 1 & \text{j=k} \qquad (9.7a) \\ 0 & \text{j} \neq \text{k} \qquad (9.7b) \end{cases}$$

iii. User defined 1:

$$\text{Corr}(Y_{ij}, Y_{i,j+1}) = \alpha^{\delta_{ij}} \tag{9.8}$$

where α is the correlation and δ_{ij} is the Euclidean distance.

iv. User defined 2:

$$\text{Corr}(Y_{ij}, Y_{i,j+1}) \begin{cases} \alpha^{\delta_{ij}} & \delta_{ij} \leq 1 \qquad (9.9a) \\ 0 & \delta_{ij} > 1 \qquad (9.9b) \end{cases}$$

where α is the correlation and δ_{ij} is the Euclidean distance.

Under these formulations, α parametrizes the level of correlation in malaria incidence between two provinces. If $\alpha = 0$, then there is no correlation between provinces. Compound symmetry means that all provinces have the same correlation to each other, irrespective of distance apart. The spatial correlation structures (iii) and (iv) are distant dependent, under (iv) it was assumed that provinces (centroid) that are more than one Euclidean distant apart to be independent.

9.3 Analysis of Malaria Incidence in Afghanistan in 2009

9.3.1 Data Sources

The data in this study are monthly cases of malaria in 2009 in 34 provinces of Afghanistan reported to the Health Management Information System of the Ministry of Public Health. Of a total of approximately 520,000 malaria slides examined, about 240,000 were positive and clinical malaria cases. Among these, 94% was Plasmodium *vivica*, 6% Plasmodium *falciparum* and about 150,000 were malaria hemophagocytic syndrome. Data on the population of Afghanistan in 2009 was obtained from the Central Statistics Organization of Afghanistan. Additional province level information was extracted from the 2010 Afghanistan Mortality Survey (AMS). The variables extracted from AMS include remoteness index of the household location, household wealth status, number of rooms used for sleeping, number of household members and migration records of the household.

Data on the daily weather was extracted from the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NOAA), US Department of Commerce. Since the weather stations are spread unevenly across Afghanistan with some provinces having more than one station while some have none. The average monthly weather data at the provincial level was estimated by interpolation and Bayesian kriging (Lawson, 2008). The method was implemented in `geoR` and the Bayesian analysis was carried out using the function `krige.bayes` (Ribeiro and Diggle, 2001). Detail information on the variables used in this study are displayed in Table 9.1.

Table 9.1: Description of the variables used in the study

Variables	**Description**
Pv	Plasmodium vivax
Pf	Plasmodium falciparum
Mal	Number of positive and clinical malaria cases
Temp	Average monthly temperature (Celsius)
Wind	Average monthly wind speed (Knots)
Precip	Average monthly precipitation (Inches)
Alt	Altitude of the provincial headquarters (Metres)
Pop	Population size
Wealth	Percentage in the middle and higher wealth quintile in the province
Residence	Place of residence: Rural or Urban
Remoteness	Percentage of the population living in the most and second most remote areas
Sanitation	Percentage in the province with bad sanitation (eg. open latrine, no toilet at all, bucket etc)
Water	Percentage in the province with access to clean and safe water (eg. Piped borne, bottle water, spring water etc)
HHmember	Average number of household members
Lat	Latitude of the centroid
Long	Longitude of the centriod

Table 9.2: Summary results of confirmed malaria incidence clusters in Afghanistan from SaTScan

Provinces	Cases	Relative Risk	P-value
Farah, Takhar	5339	5.33	<0.0000
Ghazni	16436	2.33	<0.0000
Badakhshan, Bamyan, Faryab, Kapisa, Hilmand, Kunduz, Zabul	76.173	1.32	<0.0000
Baghlan, Balkh, Wardak	4511	2.17	<0.0000

9.3.2 Illustration and Discussion

Returning to the analysis of malaria incidence in Afghanistan in 2009. The data consist of the number of blood samples examined, number tested positive for malaria, number of clinical malaria, number of Plasmodium *falciprum* and Plasmodium *vivax* malaria and malaria hemophagocytic syndrome (HPS) cases. Data were aggregated at the province level. The distribution of the crude rates across Afghanistan indicates that about 30% of the malaria incidence occurred in the Nangarhar province, 13% in Kunar and 7% in Badakhshan province (Figure 9.1). Figure 9.2 shows the 25th, 50th and 75th percentile of the malaria cases in the 34 provinces of Afghanistan in 2009. Although, few cases of P. *falciparum* have been documented in areas with high altitude (>2400 m) (AbdurRab *et al.*, 2003); malaria cases are more pronounced in areas that are less than 1000 m above sea level and accounted for about 45% of the total malaria cases (out of this, 95% are cases of P. *vivax*) (Figure 9.3). The presence of heterogeneity in the standard incidence rates (SIR) of malaria across the provinces is illustrated in Figure 9.4. The plot of SIR shows high incidence in the Central and North-Eastern region of Afghanistan. Using spatial scan statistics (SaTScan, Kulldorff, 1997), the Central area is identified as a primary cluster and the North-Eastern region as a secondary cluster. Table 9.2 gives summary statistics of clusters of malaria cases in Afghanistan in 2009.

The variation in the data was explored by the use of variogram. The variogram was

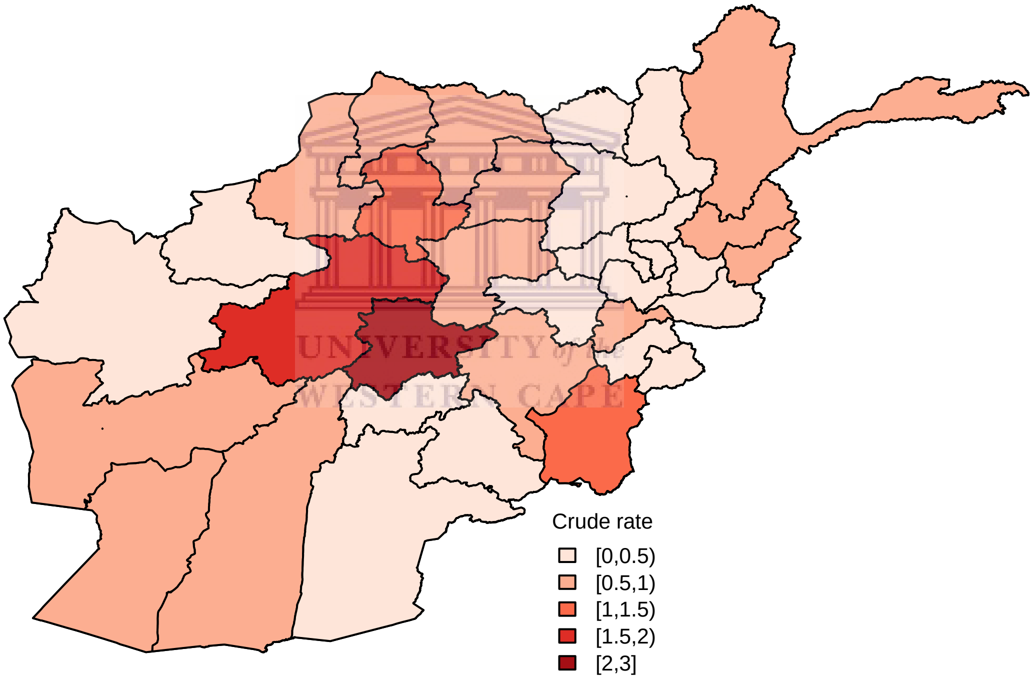


Figure 9.1: Map of Afghanistan indicating showing crude rates of malaria incidence

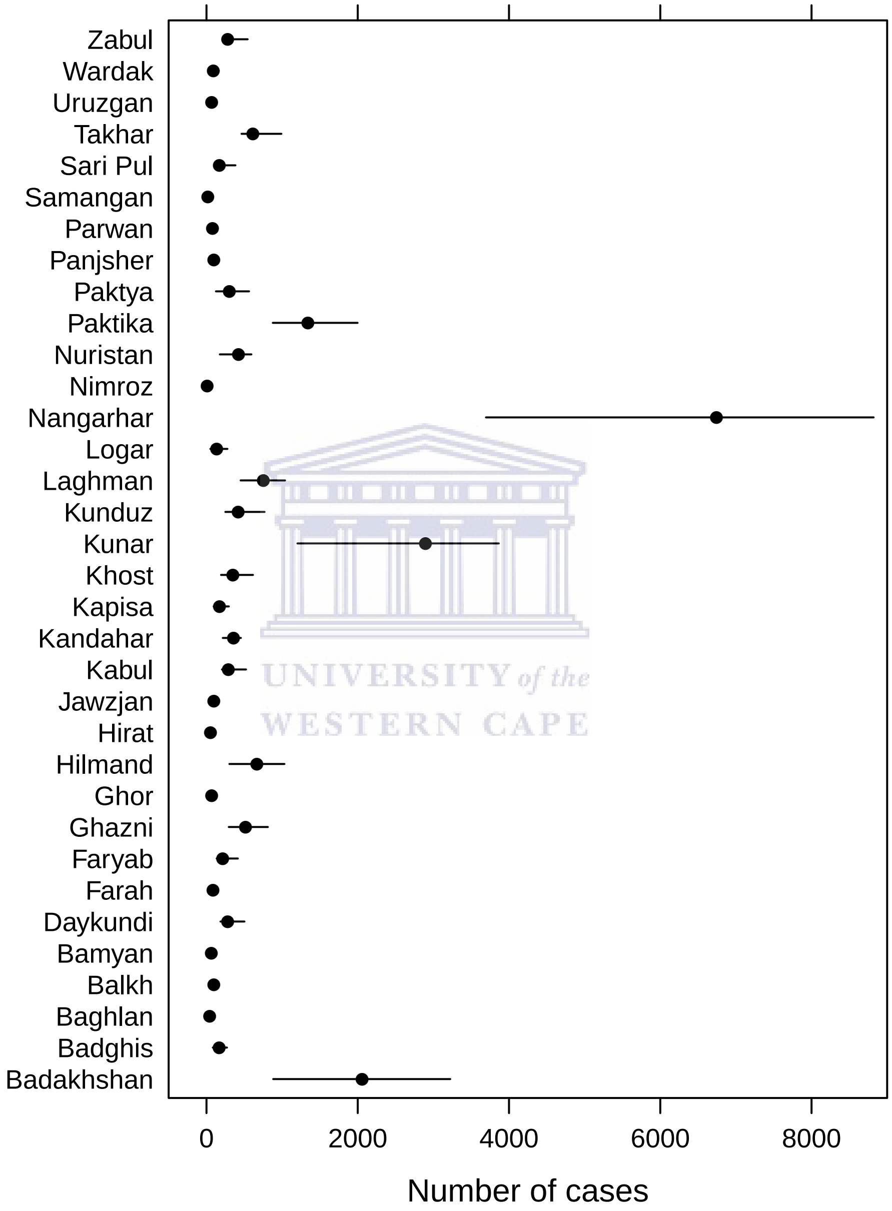


Figure 9.2: Distribution of total number of reported malaria cases; the lines run from the 25th percentile to the 75th percentile and the dot indicates the 50th percentile (median)

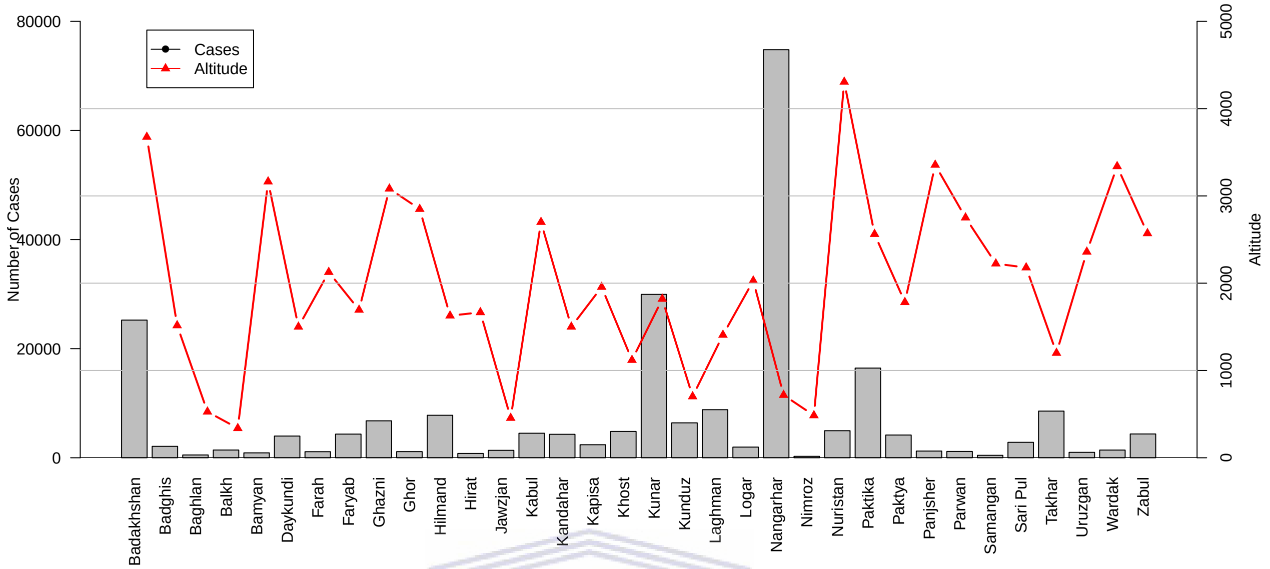


Figure 9.3: Bar plot showing Altitude and cases of malaria incidence

evaluated at distances covered by the data. Figure 9.5 shows the empirical semivariogram from smoothed malaria rates, robust semivariogram as well as three fitted models (wave, gaussian and exponential) for a robust semivariogram. The spatial dependence structure was modelled by a wave semivariogram fitted the with nugget=0.44, range=0.67 and sill=0.54.

The working spatial correlation between measurements taken at two locations (provinces) were those given the correlation structures in Section 2.3. Two methods of analysis were considered: (1) Using GEE based on each of these working correlations and (2) A GMM combining all these working correlations. Table 9.3 presents the results from the analysis implemented in R. These are point estimates together with their corresponding Akaike's Information Criteria (AIC) to assess the best model. The results assuming independent and distant dependent working correlation are similar while that of compound symmetry and the proposed method are similar. The merit of the models as assessed by AIC indicates that the proposed method has the lowest AIC implying its

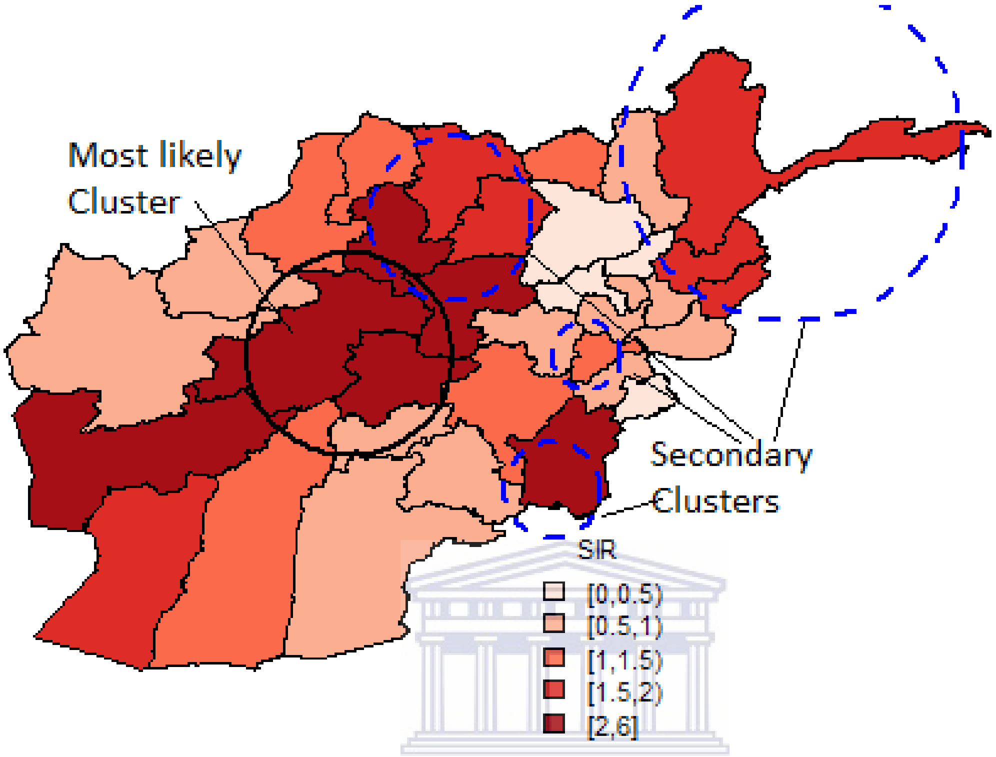


Figure 9.4: Map of Afghanistan showing the locations of the confirmed malaria incidence clusters in Afghanistan (2009) by SaTScan

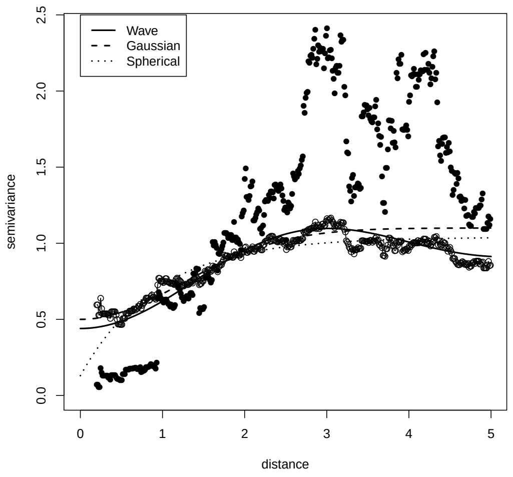


Figure 9.5: Plots of empirical semivariograms for smoothed malaria incidence and fitted (wave, gaussian and spherical models): The open dots were robust empirical semivariogram

Table 9.3: Parameter estimates using different correlation structures (and the proposed method) to analyze malaria incidence in Afghanistan

Risk factors	GEE_{inde}	GEE_{CS}	$GEE_{UserDefine}$	EE_{GMM}
Constant	9.0030	-0.5189	0.6213	-0.5262
Altitude	-0.00004	0.00004	-0.00016	0.00003
Population density	-0.00004	-0.00001	-0.00073	-0.00001
AIC	1326.714	626.701	701.246	623.512

superiority.

The interpretation of the best model using the GMM shows a positive effect of altitude and negative effect of population density on risk of malaria incidence. The positive effect of altitude is not surprising because about 21% of the total cases of malaria incidence occurred in highland areas (altitude > 2500) and Plasmodium *vivax* accounted for 92% of cases in this region. For example, Badakhshan province in the North-East which account for 7% of the total malaria cases in Afghanistan in 2009 is characterized by highlands and is associated with rice growing thereby providing a breeding environment for malaria vector Anopheles *pulcherrimus* and A. *hyrcanus* (Faulde *et al.*, 2007).

Also the rural areas are characterized by low population densities, vacant lands and farms that provide suitable habitat for malaria vectors, thereby increasing the transmission of malaria. Studies on the effect of population density on malaria transmission have indicated that areas with low population density may not provide enough people to aid the transmission (Snow *et al.*, 1999). High population density areas have reduced risk, probably due to urbanization which implies access to preventative and curative measures, and better health care facilities that make the urban population less biologically or economically vulnerable to malaria infection (Hay *et al.*, 2005; Tatem *et al.*, 2008).

In conclusion, this study has shown that the combined GEE method provides better results and precise parameter estimates, at least when compared with models with independent, compound symmetry, and user defined 1 working correlations separately. Apart from the efficient parameter estimation and inferences thereof, this method also optimally combines the information in $S^1, ..., S^j$. The merit of the proposed method was confirmed by AIC and can be implemented using the `geepack` package (Hjsgaard *et al.*, 2006; Yan and Fine, 2004; Yan, 2002) in `R` with some additional coding.

Chapter 10

Modelling of Spatio-temporal Correlation Structures in Disease Profile

Summary

This paper studies the spatio-temporal pattern in disease profile of Leishmaniasis incidence in Afghanistan in 2009. The data is characterized by a high percentage of zero disease counts, the zero counts may be the result of no disease incidence or lapse of data collection. To resolve this issue, a model truncated at zero was used. The proposed model was built on a foundation of the generalized estimating equations (Liang and Zeger, 1986). It has the advantage of producing consistent regression parameter estimates under mild conditions due to separation of the processes of estimating the regression parameters from the modelling of the correlation, and therefore, estimates of the regression parameters are consistent under mild conditions. To account for the spatio-temporal nature of the data, a method that decouples the two sources of correlations was proposed. Specifically, the spatial and temporal effects were modelled separately and then combined optimally. This approach circumvents the need of inverting the full covariance matrix and simplifies the modelling of complex relationships such as anisotropy, which is known to be extremely difficult or impossible to model in analyzing spatio-temporal data.

10.1 Introduction

One of the most challenging issues in modelling spatio-temporal data is the choice of a valid and yet flexible correlation (covariance) structure. Some examples of correlation structures can be found in Cressie and Huang (1999); Gneiting (2002); Stein (2005) and Porcu *et al.* (2007), among others. The correlation structures fall into one of two types: separable, in which case it is assumed that the space-time correlation can be written as a product of a correlation for the space dimension, and one for the time dimension or non-separable, where the space-time correlation is modelled as a single entity. Unfortunately, most of these correlation structures are either extremely complicated or infeasible to manipulate, due to their high dimensions. Furthermore, in most previous works, the space-time correlation is considered jointly, a step that it is believe to be unnecessary or unrealistic.

Analysis of correlations arising from spatial and temporal sources may be inherently distinct. This study proposes a method that decouples these two sources of correlations. This view motivated us to consider an approach that separates the modelling of the space- and time-correlations. There are at least two advantages in taking this approach. Firstly, it circumvents the need to invert a large correlation matrix, which is a commonly encountered problem in spatio-temporal analyses (*e.g.*, Yasui and Lele, 1997). Secondly, it simplifies the modelling of complex relationships such as anisotropy, which would have been extremely difficult or impossible if spatio-temporal correlations were simultaneously considered.

The proposed model was built on a foundation of the generalized estimating equations (GEE, Liang and Zeger, 1986). The standard GEE assumes longitudinal measurements within within cluster are correlated but independent between cluster. But in this situation, the observations are disease counts and covariates for the different spatial locations, and there may be spatial dependencies. An advantage of the generalized estimating equations is the separation of the processes of estimating the regression parameters from the modelling of correlation, and therefore, estimates of the regression parameters are consistent under mild conditions. However, efficiency and valid inference still depends on the correct specification of the covariance structure. Three challenges were identified in the modelling of a spatio-temporal process: (1) accommodation of covariances that arise from spatial and temporal sources; (2) choosing the correct covariance structure and (3) extending to situations where a covariance is not the natural measure of association. The method was illustrated using data from Leishmaniasis in Afghanistan.

It was reported that in 2000 there were an estimated 1.5 million annual cases of Leishmaniasis worldwide and Afghanistan, Algeria, Saudi Arabia, Brazil, Iran, Iraq, Peru and Syria account for over 90% of the cases (Michael *et al.*, 2008). There are about 250,000 estimated new cases of cutaneous Leishmaniasis incidence in Afghanistan and 67,000 cases in Kabul, thus making it the city with the highest incidence worldwide (Reithinger *et al.*, 2003). Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection. It is contracted through bites from sand flies, which are themselves not poisonous, but the parasitic Leishmania in its saliva can result in chronic and non-healing sores. This mostly occurs on exposed skin and can lead to itchy skin irritation, and disfiguring and painful ulcers. The burden of the disease is overwhelming and the psychological effect can be disturbing. In some societies, women infected with this disease are stigmatized and deemed unsuitable for marriage and motherhood (Reithinger *et al.*, 2005). The impact of environmental influences on Leishmaniasis cannot be ruled out and human activities play a significant role in the dispersion of the vectors, thereby changing the geographical distribution of the disease.

The rest of the paper is structured as follows. Section 10.2 describes the proposed method. Section 10.3.1, illustrates the method using sample data from a Leishmaniasis study. Results and discussion of the data analysis are given in Section 10.4.

10.2 Model and Methods

Let $y_{st} \equiv y(s,t)$ denote the count of disease at spatial location s and at time t, $s = 1,...,S, t = 1,...,T_s$. Suppose associated with (s,t) are covariates x_{st} that measure the spatial location and time as well as other covariate information. In order to model spatio-temporal correlation and overdispersion, assume there is a nonnegative weakly stationary latent process e_{st} such that conditional on the e's, the y's are independent and are assumed to follow a log-linear model given by

$$\mathrm{E}(y_{st}|e_{st}) = \exp(x_{st}^T\beta)e_{st}, \quad \text{and} \quad \mathrm{var}(y_{st}|e_{st}) = \mathrm{E}(y_{st}|e_{st}),$$

where the $\boldsymbol{\beta}$'s are unknown parameters that capture the association between incidence and the covariates. It is assumed that $\mathrm{E}(e_{st}) = 1$ so $\exp(x_{st}\beta)$ represents the marginal mean of y_{st}. The latent process e_{st} is assumed to have a variance of σ^2 and the covariance between e_{st} and $e_{s't'}$ is given by

$$\mathrm{cov}(e_{st}, e_{s't'}) = \sigma^2\rho(z_{st}, z_{s't'}, \alpha)$$

where $z_{st}, z_{s't'}$ are covariates from $(s,t), (s',t')$ that jointly induce spatio-temporal correlation and α are unknown parameters. This formulation is similar to the model considered by Zeger (1988). Under these assumptions, it can be shown easily that

$$\begin{aligned} \mathrm{E}(y_{st}) &= exp(x_{st}^T\beta) \equiv \mu_{st}(\beta), & (10.1)\\ \mathrm{var}(y_{st}) &= \mu_{st}(\beta) + \mu_{st}(\beta)^2\sigma^2, & (10.2)\\ \mathrm{cov}(y_{st}, y_{s't'}) &= \rho(z_{st}, z_{s't'}, \alpha)\{1 + (\sigma^2\mu_{st}(\beta))^{-1}\}\{1 + (\sigma^2\mu_{s't'}(\beta))^{-1}\}. & (10.3) \end{aligned}$$

If $\rho(z_{st}, z_{s't'}, \alpha) = 0$, then there is an overdispersion Poisson model; furthermore, if $\sigma^2 = 0$ then there is a standard Poisson model at each spatial location and time. Again, the crucial step in modelling spatio-temporal data is to specify the correlation.

For convenience, $y_{s\cdot} \equiv (y_{s1}, ..., y_{st_s})^\tau$ is defined as the vector of counts taken at times $t_1, ..., t_s$ at spatial location s and $y \equiv (y_{1\cdot}, ..., y_{m\cdot})^\tau$, and similar definitions are used for $x_{s\cdot}, x$ and $\mu_{s\cdot}, \mu$.

Following Zeger (1988), a GEE type model can be used to estimate the parameters $\boldsymbol{\beta}$, by solving the following estimating equation:

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \mathbf{D}^\tau\mathbf{V}^{-1}\{y - \boldsymbol{\mu}\} = 0, \qquad (10.4)$$

where $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}^{\tau}$, and $\mathbf{V}$ is the variance covariance matrix of y. With an abuse of notation, $\mathbf{V} = \mathbf{V}_S \otimes \mathbf{V}_{T_S}$, where $\mathbf{V}_S$ represents the ”covariance” matrix for spatial locations and $\mathbf{V}_{T_S}$ represents the ”covariance” matrix between the times at location S. $\mathbf{V}$ can be interpreted as a Kronecker product of the matrices $\mathbf{V}_S$ and $\mathbf{V}_{T_S}$. The dimension of $\mathbf{V}$ is $ST \times ST$, where $T = \sum_{s=1}^{S} T_s$. The matrix $\mathbf{V}$ can be expressed as $\mathbf{A} + \sigma^2\mathbf{A}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}$, where $\mathbf{A} = \text{diag}(\boldsymbol{\mu}_{1\cdot}, ..., \boldsymbol{\mu}_{m\cdot})$ and $\mathbf{R}(\boldsymbol{\alpha})$ is a $ST \times ST$ correlation matrix. Element $s(t_s), s'(t'_s)$ stands for the element in the t_s column in the s block and $t'_{s'}$ column in the s' block and is equal to $\rho(\mathbf{z}_{st}, \mathbf{z}_{s't'}, \boldsymbol{\alpha})$, and σ^2 is the scale parameter used to model overdispersion. The covariance matrix is by construction symmetric and positive definite over (s, t).

Elaborating further on the possible choices of the correlation, one possible candidate would be the independent correlation matrix, another would lead to a model with temporal but no spatial correlation, or a model with spatial but no temporal correlation and finally, a model with both spatial and temporal correlation. Assume all data are independent, then $R(\alpha) = I_{ST}$ where I_K stands for a $K \times K$ identity matrix. To model spatio-temporal correlation, $R(\alpha)$ can be decomposed via a Kronecker product as $R(\alpha) = \Gamma(\zeta) \otimes \Phi_s(\gamma)$ where $\Gamma(\gamma)$ models spatial correlation and $\Phi_s(\zeta)$ models temporal correlation within location s. Since, in the study, there is one count at each time in each location, it is acceptable to assume $\Phi_s(\zeta) = \Phi(\zeta)$ for all locations.

10.3 Illustration: Leishmaniasis Data

10.3.1 Data Sources and Exploration

The study consists of cases of Leishmaniasis reported to the Afghanistan Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) between 2003 and 2009. A total of 148,945 new cases of Leishmaniasis from 20 provinces

Table 10.1: Summary of the Leishmaniasis data

Year	Gender		% Age group			**Total**
	% Male	% Female	0-4	5-14	≥ 15	
2003	51.60	48.40	23.5	38.5	38.0	10944
2004	53.27	46.75	22.7	38.2	39.1	9203
2005	52.98	47.02	22.3	40.2	37.5	12951
2006	50.19	49.81	18.3	56.2	25.5	19689
2007	48.07	51.93	18.2	43.4	38.4	30273
2008	48.41	51.59	21.7	37.6	40.7	24813
2009	50.03	49.97	23.5	37.2	39.3	41,072

were recorded in Afghanistan between 2003 and 2009 (of these, 17,425 occurred in Kabul in 2009). Summary of the data can be found in Table 10.1.

The analysis is begun by exploring the observed cases of Leishmaniasis; a higher disease incidence was observed around the Kabul area (North Eastern) in 2009. Similar patterns were observed in 2003-2008 (maps are not shown here but are available on request). Kabul city accounted for more than 40% of the total new cases in 2009, although this may be attributed to availability of health care facilities, which is crucial for data gathering in public health studies. The rate of infections is about the same for both male and female (around 50% each) while those in age group 5-14 years are more likely to be infected than other ages (Table 10.1).

A striking feature of the data is the high number of cases with zero incidence for many locations (Figure 10.1). Many of the provinces have counts of zero incidence for months, then a sudden jump to a few hundred or thousand cases, then back to zero incidence. Between 2003 and 2006, most of the provinces reported no cases of Leishmaniasis; this claim cannot be verified because this period coincides with the US-led war in Afghanistan and disease reporting may only be possible in a relatively safe environment. The q-q plots (not shown) for the Poisson regression of the counts in each province using GEE with an independent assumption showed similar patterns and lots of

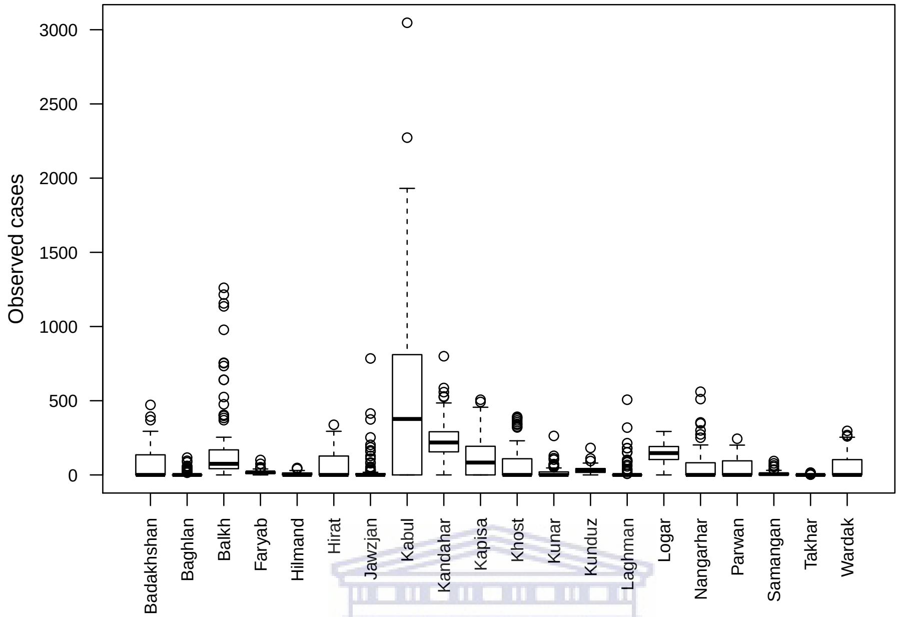


Figure 10.1: Distribution of total cases of Leishmaniasis at provincial level in Afghanistan (2003-2009)

clustering in the lower left corner due to the zero counts in each province. Figure 10.2a also presents the distribution of the observed number of Leishmaniasis cases during this period.

It is very difficult to distinguish between ”true” and ”imputed” zeros, because of the reporting mechanism of disease in Afghanistan (due to security, technical and logistical issues). These problems prompted the consideration of the option of discarding the zeros and model the non-zero data using a Poisson modelling conditional on all cases greater 0. The assumption is made that ”imputed” zeros are a random event. The monthly profile of non-zero Leishmaniasis cases indicate two peaks in the disease occurrence in Afghanistan between 2003 and 2009 – January to March and September to December – which coincide with the cold period. July is the hottest month and March is the wettest month with a monthly average of 1,770 Leishmaniasis cases (horizontal line in Figure 10.2b). The time series plot for the number of Leishmaniasis cases is divided into half; below and above the mean. An upward trend and regularly repeating patterns of highs and lows related to the months of the year was observed in the data sets (see Figure 10.2b); this suggests seasonality in the data. The plot shows a stable pattern and stationarity.

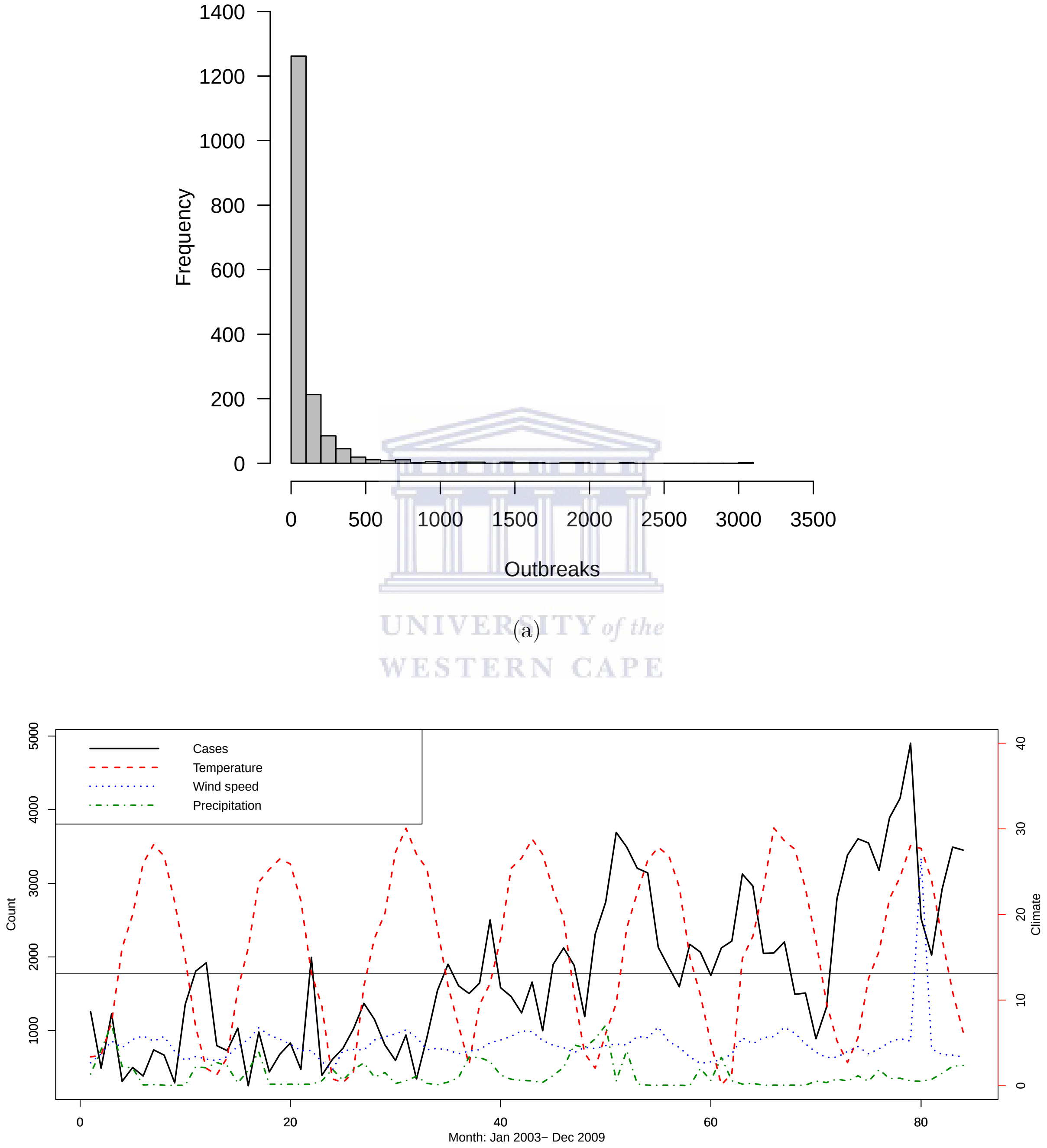
An Augmented Dickey-Fuller test (ADF) (Said and Dickey, 1984; Dickey and Fuller, 1979) rejects the hypothesis of unit-root (statistic=-4.3801, p-value $= 0.01$). The population sizes of Afghanistan between 2003 and 2009 were obtained from the Central Statistics Organization (CSO) of Afghanistan. The incidence of Leishmaniasis occurred in less than 1% of the population except in Kabul province in 2009 with about 2% of the population, infected with the disease.

10.3.2 Correlation Models

Recall that for fixed s, $v_{s,tt'}$, $t, t' = t_1, ..., t_T$ are the elements of the variance covariance matrix of disease counts between times. In the data set, monthly disease counts for each province were captured over 7 years, from 2003-2009, with up to 84 observations per province. However, year is an artificial variable that is not of interest. On the contrary, there might be two different types of temporal correlations: (1) Between months that are nearby and (2) Between the same month in different years (seasonality). To capture these two types of correlations, the following is considered.

We decomposed the $T = 84$ times into months and years. Let $m = 1, ..., 12$ and $M = 1, ..., 7$ be indicators for months of the year and year ($M = 1 \equiv 2003$, etc.). Then $t = t_1, ..., t_T = \{t_{mM}\}$. Define two 84×84 matrices $\mathbf{R}_1(\boldsymbol{\zeta}_1)$ and $\mathbf{R}_2(\boldsymbol{\zeta}_2)$ where $\boldsymbol{\zeta}_1 = \{\zeta_{mM,m'M'}, m, m' = 1, ..., 12, M, M' = 1, ..., 7\}$, $\boldsymbol{\zeta}_2 = \{\zeta_{mM,m'M'}, m, m' = 1, ..., 12, M, M' = 1, ..., 7\}$, such that $\Phi_1(\zeta) = \zeta_{mM,m'M'} = \zeta_1^{|m+12M-('m+12M')|}, \Phi_2(\zeta) = \zeta_{mM,m'M'} = \zeta_2^{|m-m'|}, 0 < \zeta_1, \zeta_2 < 1$.

The spatial correlation $\mathbf{R}_s(\alpha)$, assuming the correlations remain the same across time, is



(a)

(b)

Figure 10.2: (a) Histogram of the number of times Leishmaniasis cases occurred. (b) Times series plot of monthly cases of Leishmaniasis disease in Afghanistan from 2003 to 2009 together with average monthly temperature, precipitation and wind speed

given by $\Gamma(\gamma) = \gamma_{s,s'} = \gamma^{\|s-s'\|}, 0 < \gamma < 1$. Let $\boldsymbol{\alpha} = (\boldsymbol{\zeta}, \boldsymbol{\gamma})^{\tau}$, then we use $\mathbf{R}_1(\boldsymbol{\alpha}) = \mathbf{R}_1 \circ \mathbf{R}_s$ and $\mathbf{R}_2(\boldsymbol{\alpha}) = \mathbf{R}_2 \circ \mathbf{R}_s$ where $\circ$ is the Hadamard product.

10.3.3 Parameter Estimation

Suppose we remove all $y_{st} = 0$, then conditioning on $y_{st} > 0$; (10.1)-(10.3) become

$$\mathrm{E}(y_{st}|e_{st}) = c\mu_{st}(\boldsymbol{\beta})e_{st} \quad \text{and} \quad \mathrm{var}(y_{st}|e_{st}) = [c\mu_{st}(\boldsymbol{\beta}) + c(1-c)\mu_{st}(\boldsymbol{\beta})^2]e_{st}$$

where $c = 1/[1 - \exp(-\mu_{st})]$, leading to

$$\mathrm{E}(y_{st}) = c\mu_{st}(\boldsymbol{\beta}) \equiv \phi_{st}(\boldsymbol{\beta}), \tag{10.5}$$

$$\mathrm{var}(y_{st}) = c\mu_{st}(\boldsymbol{\beta}) + c(1-c)\mu_{st}(\boldsymbol{\beta})^2 + c^2\mu_{st}(\boldsymbol{\beta})^2\sigma^2. \tag{10.6}$$

Let $d = \{d(s,t) = d_{st}\}_{S\times T}$ be a matrix of indicators such that $d_{st} = 1$ if $y_{st} > 0$ and $d_{st} = 0$ otherwise. Note that $y_{st} = 0$ could mean the count was zero or the count was not taken. To resolve the missing counts, it was assumed that the counts were missing completely at random. The problem of counts not missing completely at random can be handled by adding an extra model for the propensity of $d_{st} = 1$. However, it was desirable to illustrate the idea of a spatial GEE and so, it was decided to minimize any distraction to this main idea. For example, for a particular set of spatial working correlations $\tilde{R}^j_{ss'}$, the spatial GEE conditioned only on those observations with $y_{st} > 0$ can be written as

$$\tilde{\mathbf{U}}^j(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \sum_{s=s_1}^{s_S}\sum_{s'=s_1}^{s_S}\sum_{t=t_1}^{t_T}\sum_{t'=t_1}^{t_T} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'}\{y_{st'} - \phi_{st'}\} = 0, \tag{10.7}$$

where $v_{s,tt'}$ is the t,t-th element of $\mathbf{V}_s$, the covariance matrix of $\mathbf{y}_{s\cdot}$. The matrix $\mathbf{V}_s$ can be expressed as $\mathbf{A}_s^{1/2}\mathbf{R}_s(\boldsymbol{\alpha})\mathbf{A}_s^{1/2}$, where $\mathbf{A}_s = \mathrm{diag}[c\mu_{s1}(\boldsymbol{\beta}) + c(1-c)\mu_{s1}(\boldsymbol{\beta})^2 + c^2\mu_{s1}(\boldsymbol{\beta})^2\sigma^2, ..., c\mu_{sT}(\boldsymbol{\beta}) + c(1-c)\mu_{sT}(\boldsymbol{\beta})^2 + c^2\mu_{sT}(\boldsymbol{\beta})^2\sigma^2]$ and $\mathbf{R}_s(\boldsymbol{\alpha})$ is a matrix with its (t,t')-th element representing the correlation between times t and t' at location s.

The primary interest lies in the parameters $\boldsymbol{\beta}$ but the nuisance parameters $\boldsymbol{\alpha}$ must also be dealt with. The parameters were estimated via a Newton-Raphson iteration method and then fitted the data using a standard GEE (10.4) using $\mathbf{R}(\alpha)$ described in 10.3.2. To solve for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ jointly, the method of Prentice (1998) was employed.

Let $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\alpha}}_k$ be the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ at the k-th iteration. A GEE was first fitted with an independence working correlation structure and then the estimating equation for $\boldsymbol{\alpha}$ was solved, iterating until convergence. This step will give the values $v_{s,tt'}$. Denoting $\sum_{s,s',t,t'} \equiv \sum_{s=s_1}^{s_S} \sum_{s'=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T}$, the initial estimate $\hat{\boldsymbol{\beta}}_0$ is found, using (10.7) by assuming an identity matrix for $\mathbf{R}_s(\boldsymbol{\beta})$, equivariance, *i.e.*, $v^{-1}_{s,tt'} = 1$. Then at iteration k,

$$\hat{\boldsymbol{\beta}}_{k+1} = \hat{\boldsymbol{\beta}}_k - \left[\sum_{s,s',t,t'} \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'}(\hat{\boldsymbol{\beta}}_k) \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}} \right]^{-1} \left[\sum_{s,s',t,t'} \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'} \{ y_{st'} - \phi_{st'}(\hat{\boldsymbol{\beta}}_k) \} \right]. \tag{10.8}$$

The next step will be to iterate between (10.7) and (10.8) until convergence.

The standard errors for the $\boldsymbol{\beta}'s$ were obtained using large-sample properties (Liang and Zeger, 1986; Leung *et al.*, 2009; Paul *et al.*, 2013). Under mild regularity conditions, $K^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{GEE} - \boldsymbol{\beta})$ is asymptotically multivariate Gaussian with zero mean and covariance matrix given by:

$$\lim_{K \to \infty} K \left[\sum_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} \right]^{-1} \left[\sum_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'} cov(y_{st'}) \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} \right] \left[\sum_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} v^{-1}_{s,tt'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} \right]^{-1} \tag{10.9}$$

The models were assessed via the Akaike's Information Criterion (AIC) for GEE (Pan, 2001; Leung *et al.*, 2009) given by:

$$\begin{aligned} Q(\boldsymbol{\beta}) &\equiv -\sum_{s,t'} \{ y_{st'} - \phi_{st'}(\hat{\boldsymbol{\beta}}) \}^T v^{-1}_{s,tt'} \{ y_{st'} - \phi_{st'}(\hat{\boldsymbol{\beta}}) \}, && (10.10) \\ AIC &= -2Q(\boldsymbol{\beta}) + 2 \text{ trace}(\hat{\Sigma}_I^{-1} \hat{\Sigma}_R) && (10.11) \end{aligned}$$

where $\hat{\Sigma}_I^{-1}$ is the inverse of the variance of the model coefficients under an independence working correlation and $\hat{\Sigma}_R$ is the variance of the model coefficients under working correlation $\mathbf{R}$ (Leung *et al.*, 2009).

10.4 Results and Discussion

The general introduction of the correlation functions and estimation procedures have been presented in Sections 10.3.2 and 10.3.3 respectively. The proposed method was applied in the study of the influence of environmental factors on Leishmaniasis in Afghanistan, while allowing for different dependencies in the data. The model formulation and detailed methodology were discussed in Section 10.2. Population size was added as offset; the covariates were climatic and environmental variables that included average monthly temperature (Celsius), average monthly precipitation (Inches), average monthly wind speed (Knots) and altitude (Metres). Three different working correlations were explored for **R**, namely: Spatial only, temporal only and spatio-temporal. The parameter estimates, as well as their standard errors from the GEE fitted to the data using the mentioned correlations, are presented in Table 10.2.

Observe that the parameter estimates of the four models are very similar in some cases ($GEE_{Temporal}$ and $GEE_{Spatio-Temporal}$) and very different in others ($GEE_{Spatial}$ and others), likewise their standard errors. The above observation demonstrates that, due to the choice of correlation structures and interest, inferences about functions of the model parameters may be different and erroneous. The value of 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$ was used to ascertain the appropriateness and suitability of the correlation structure. It has been noted that the time correlation function that assumes temporal correlations between months that are nearby, $\Phi_1(\zeta)$, is more appropriate for this data with 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R) =$ 19.38 than the choice that assumed temporal correlations between the same months in different years, $\Phi_2(\zeta)$, with 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R) = 87.054$. Hin and Wang (2009), in Leung *et al.* (2009) have shown that 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$ offers more ability to accurately capture the true correlation structure.

In Table 10.2, perhaps the most distinctive results are from the model with spatial correlation; the model parameter estimates are remarkably different from others. The result may not be surprising as it has been assumed that the correlation remains the same across time. This also suggests that spatial correlation only may not be sufficient for the data, because it involves the specification of spatial correlation across time. The results have shown that the specified spatio-temporal function is more suitable and appropriate for this data (smaller 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$). Moreover, the model with the spatio-temporal correlation function significantly improves the model fit when compared to other specifications, as judged by the smaller AIC. Although the parameter estimates from both temporal and spatio-temporal models are similar, significant differences can be observed in their precision estimation.

The results from this study are similar to that of Adegboye and Kotze (2012); Rajesh and Sanjay (2013); Thompson *et al.* (2002); Valderrama-Ardila *et al.* (2010); Karagiannis-Voules *et al.* (2013). The best model (model with the spatio-temporal correlation function) confirms the significant influence of environmental factors on the incidence of Leishmaniasis. The model indicates that high temperatures are associated with a lower incidence of Leishmaniasis; this is similar to the findings of Rajesh and Sanjay (2013). The survivability of the sand fly (Leishmaniasis vector) has been reported to reduce during high temperatures (Rajesh and Sanjay, 2013). A negative association between precipitation and incidence of Leishmaniasis has been found; this is not surprising as extreme rainfall may have a negative effect on the vector such as flooding (Thompson *et al.*, 2002). The negative effect of temperature and precipitation is also in line with what was observed in the exploratory analysis. Two peaks were observed in the disease occurrence between 2003 and 2009 – January to March and September to December – which coincide with the cold period, while July is the hottest month and March is the wettest month. The results also indicate that low altitudes are associated with an increase in the cases of Leishmaniasis, whereas an increase in the wind speed has a positive effect on the disease.

Table 10.2: Parameter estimates together with the standard errors from GEE with different correlation structures of Leishmaniasis incidence in Afghanistan

Risk factors	$GEE_{Spatial}$	$GEE_{Temporal}(\Phi_1)$	$GEE_{Temporal}(\Phi_2)$	$GEE_{Spatio-temporal-(\Phi_1)}$
Intercept	-0.52289 (0.07342)	-9.09818 (0.02206)	-9.09746 (0.08526)	-8.95422 (0.000024)
Altitude	-0.00012 (0.00023)	0.00026 (0.00001)	0.00026 (0.00001)	-0.00008 (0.000007)
Temperature	-0.42460 (0.00022)	-0.00113 (0.00017)	-0.00118 (0.00035)	-0.01915 (0.000141)
Precipitation	1.58830 (0.00785)	-0.03895 (0.00210)	-0.03920 (0.00066)	-0.02733 (0.000509)
Wind	0.53639 (0.00528)	0.02078 (0.00112)	0.02089 (0.01566)	0.05572 (0.000908)
2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$	69.33	19.38	87.054	16.89
AIC	103.112	46.511	179.39	969.251

This study introduced the method for modelling spatio-temporal correlation based on the platform of GEE. The technique used the correct specification of correlation structures to improve the efficiency of the GEE method. The Leishmaniasis data presented several problems with modelling issues, ranging from correlation/covaraince specification to issues with "imputed" or "non true" zeros. The high percentage of zero disease counts may be the result of no disease incidence or lapse of data collection. Moreover, the dependency in the data may be a result of spatial variation, temporal or both. To resolve this issue, a renowned complex method was used to address the many issues that the data presented in a very novel way. A model truncated at zero was fitted while allowing for dependency in the data via a working correlation matrix using the technique of GEE. The effects of the working correlations were explored and were assessed using AIC and 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$. The study has not just shown the advantages and appropriateness of the choice of correlation/covariance structure, it has also presented a model that enjoys the flexibility of modelling many zeros and true association that leads to more trustworthy inferences, as well as providing a novel approach to parameter estimation.

Chapter 11

General Conclusions

Throughout this thesis, the modelling of correlated data using different illustrations and approaches has been presented. The general framework of the thesis was to explore the existing techniques of analyzing correlated data and then propose a set of new approaches. The idea was to start with simple and known techniques, then build up to new ideas of handling different types of correlation structures, in particular; multilevel, spatial, temporal and spatio-temporal. The techniques were also extended to data sets with a high degree of zeros, where zero counts are not reliable and cannot be validated in public health studies in developing countries.

A few limitations have been observed in this thesis: In Chapter 6 some limitations were: Firstly, the use of survey year as the time factor in cross-sectional data may be seen as a limiting factor; the years of the surveys are not evenly spaced (i.e. 1990, 1999, 2003 and 2008). Secondly, it is difficult to say that the same households or primary sample units were sampled each year, as is the case in longitudinal studies. It would be interesting to be able to identify those households that were sampled over the years. Although the data may have these limitations, the analysis provide extensive information that is crucial for assessing the risk factors for full immunization over the years.

In Chapter 7, the collection of information through VA is based on the respondents' ability to provide an adequate account of the death process. The enquiries about the death may be upsetting, especially in a very sensitive society, thereby raising the issue of the reliability of the data. In developing countries, where death certificates are lacking for the majority of the population, VA is very crucial for investigations into cause of death. Also, although there is a strong correlation between the two disease classifications (COMPLICATIONS and SYMPTOMS), this may not be true for other possible combinations. Besides, there may be other disease classifications that are acting as a confounding factor.

In Chapter 8, the data on the HPAI H5N1 virus was extracted from the OIE database and cases of under reporting cannot be ruled out. Also, Nigeria is an extensive country with very poor data acquisition and the lack of e-data sources may affect timely reporting of disease occurrence. In the study the extent (distance) of the contribution of the wind in the transmission of HPAI could not be assessed.

Although in Chapters 9 and 10, the study focused only on spatial and spatio-temporal data types, the techniques proposed can also be extended to multilevel/hiararchical analysis as well. So far, only a number of correlation structures have been explored in Chapter 9. Other correlation/covariance functions are also plausible. The major limitation in the proposed method for spatio-temporal data in Chapter 10, where the interest was on decoupling the two sources of correlations (space and time), lied in the challenging issue of dealing with many zero counts. In the study, it was assumed that counts were missing completely at random by considering only the cases where $y_{st} > 0$; if the data is not missing at random or the missing responses are conditional on the observed responses, then correct modelling of the missingness probability is required for the method to be valid. This limitation does not really invalidate the findings, as the idea was to illustrate the concept of a spatial or spatio-temporal GEE and to minimize any distraction to this main idea. It may also be of interest to model situations where the counts are not missing at random by adding an extra model for the propensity of the indicator $d_{st} = 1$.

The thesis has two major contributions: The general contributions to the field of public health (including animal health) by providing insightful analyses on several existing techniques for modelling correlated and multilevel data fitted to data from public health studies, though the choice of one or the other depends on the objectives of the study and computational issues. Furthermore, the thesis contains significant contributions to methodological issues in correlated data (especially spatial and spatio-temporal). The methodological contributions include novel techniques to the fitting of spatial, temporal and spatio-temporal correlations, and to cases where data sets contain many zeros.

In the case of spatial data, it was shown that a single choice of the correlation structure may not reflect the true nature of the dependency in the data and hence, a method that optimally combines different choices of correlation structures was proposed. This method was illustrated using Malaria data from Afghanistan. The study introduced the method for modelling spatio-temporal correlation based on the platform of GEE. The technique used the correct specification of a correlation structure to improve efficiency of the GEE method. The Leishmaniasis data presented several problems with modelling issues, ranging from correlation/covaraince specification to issues with ”imputed” or ”non true” zeros. Spatial and temporal effects were modelled separately and were then combined optimally. This approach circumvents the need of inverting the full covariance matrix and simplifies the modelling of complex relationships such as anisotropy, which is known to be extremely difficult or impossible to model in analyzing spatio-temporal data. The high percentage of zero disease counts may be the result of no disease incidence or lapse of data collection. To resolve this issue, a renown complex method was used to address the many issues that the data presented in a very novel way. A model truncated at zero was fitted while allowing for dependency in the data via working correlations using the technique of GEE.

Further research could look at situations where a single spatio-temporal correlation structure would not capture the true correlation in a data. The generic form for $R(\alpha)$ could be used to form different choices, different degrees and types of spatio-temporal correlations which can then be combined optimally. One generic form for $R(\alpha)$ could be

$$R(\alpha) = \nu I_{ST} + (1-\nu)\Gamma(\gamma) \otimes \Phi_s(\zeta), \tag{11.1}$$

where $\nu \in [0,1]$. When $\nu = 1$, there is no spatial or temporal correlations, i.e., all data are independent. When $\Gamma(\gamma) = I_S$, there is no spatial correlation, and when $\Phi(\zeta) = I_{T_s}$, there is no temporal correlation in location s. In general, by varying ν and using different choices, different degrees and types of spatio-temporal correlations can be modelled; hence, the formation of the form $R^1, ...R^J$ by using different ν, $\Gamma(\gamma)$ and $\Phi(\zeta)$. In addition, it may also be of interested to explore situation where the counts are not missing at random, by adding an extra model for the propensity of the indicator $d_{st} = 1$.

The modelling of correlated data was illustrated using several high level techniques to expose the many complications and difficulties the data presented. The techniques proposed for the different scenarios extended known approaches to highlight their suitability.

Appendix A

Poverty and Education

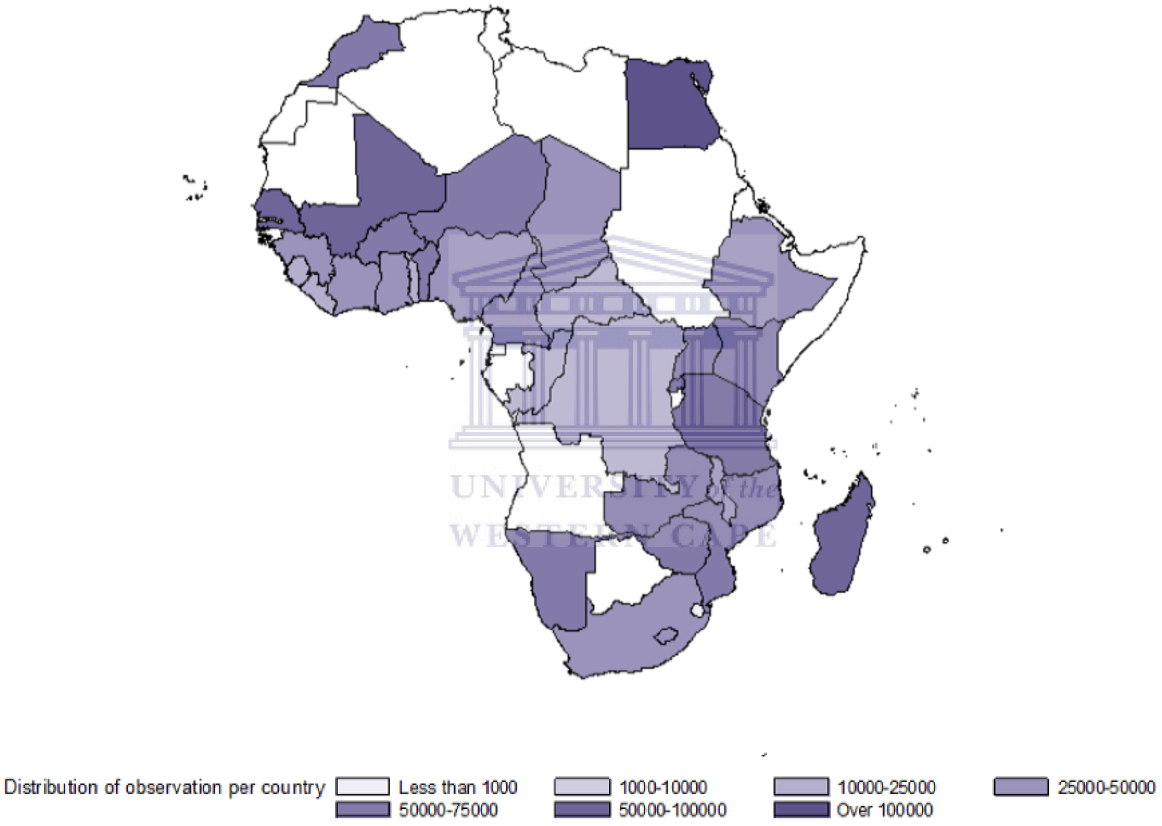


Figure A.1: Map showing the total number of observations used in this study

Table A.1: General variable definitions and description

Variable	Description
Relationship to head	Son/Daughter/ Adopted/Foster/Stepchild Or Other Relation
Single parent	Yes, Both Parent Are Dead Or No, At Least One Is Alive
Mother's education level	No Education, Primary Or Secondary
Father's education level	No Education, Primary Or Secondary
Father's age	Years
Mother's age	Years
Wealth index	Poorest, Poor, Middle, Rich, Or Richest
Total number of family members	Count
Type of residence	Urban Or Rural
Gender	Male Or Female
Educational attainment	No Education, Primary, Or Secondary or Higher
Age	Years
Marital status	Married: Yes Or No
Source of water	Clean Water (Piped Water, Tube Well Or Borehole, Bottled Water ,Cart With Small Tank) Or Unclean Water (Dug Well, Water From Spring, Rainwater, Surface Water)
Type of toilet	Good Sanitation (Flush Or Pour Flush Toilet, Ventilated Latrine) Or Bad Sanitation (None Ventilated Latrine, Bucket, No Facility/Bush, Field)
Sex of head	Male Or Female
Age of head	Years
Literacy gap	Gender Difference In The Proportion Of Male And Female With At Least Primary Education
Education Gini	Calculated

	Compulsory education		Duration of primary and secondary education																
			Age																
Country	Age limits	Duration In years	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Benin	6-11	6			p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Burkina Faso	6-14	9			p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Congo Democratic Republic	6-11				p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Cameroon :General	6-11	6			p	p	p	p	p	p	s1	s1	s1	s1	s1	s2	s2		
:Technical		6			p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Central Africa Republic	6-14	9			p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Chad	6-11				p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Comoros	6-15	10			p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Congo Brazzaville	6-11				p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Cote d'Ivoire																			
Egypt	6-14	9			p	p	p	p	p	p	p	p	p	s	s	s			
Ethiopia	7-16	10				p	p	p	p	p	p	p	p	s1	s1	s2	s2		
Gabon	6-11	6			p	p	p	p	p	p	s1	s1	s2	s2					
Ghana	6-14	9			p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Guinea	7-12	6				p	p	p	p	p	p	s	s	s	s				
Kenya	6-14	8			p	p	p	p	p	p	p	p	s	s	s	s			
Lesotho	6-15	7			p	p	p	p	p	p	p	s1	s1	s1	s2	s2			
Liberia	6-11	6			p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Madagascar	6-13	5			p	p	p	p	p	s1	s1	s1	s1	s2	s2				
Malawi	6-13	8			p	p	p	p	p	p	p	p	s1	s1	s2	s2			
Mali		9				p	p	p	p	p	p	p	p	p	s	s	s		
Mauritania	6-11	6			p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Morocco	4-12	9	p	p	p	p	p	p	p	p	p	s	s	s					
Mozambique	6-12	7			p	p	p	p	p	p	p	s1	s1	s1	s2	s2			
Namibia	6-12	7			p	p	p	p	p	p	p	s1	s1	s1	s2	s2			
Niger	7-12				p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Nigeria	6-14	9			p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Rwanda	6-11	6			p	p	p	p	p	p	s1	s1	s1	s2	s2	s2			
Senegal	7-12	4				p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2	
Sierra Leone																			
South Africa	7-12	9				p	p	p	p	p	p	s1	s1	s1	s2	s2	s2		
Swaziland	6-12	7			p	p	p	p	p	p	p	s1	s1	s1	s2	s2			
Tanzania	7-13	7				p	p	p	p	p	p	p	s	s	s	s	🟥	🟥	
Togo					p	p	p	p	p	p	s1	s1	s1	s1	s2	s2	s2		
Uganda	6-12	7			p	p	p	p	p	p	p	s1	s1	s1	s1	s2	s2		
Zambia	7-13					p	p	p	p	p	p	p	s1	s1	s2	s2	s2		
Zimbabwe	6-12				p	p	p	p	p	p	p	s	s	s	s	🟥	🟥		

Source: World Data on Education, Seventh edition 2010/11

Primary school 🟩 Secondary School: 1st stage 🟫 Secondary School : 2nd stage ■
Secondary School: Straight ⬛ Advanced level: Optional 🟥

Figure A.2: National educational system of the 36 African countries under study

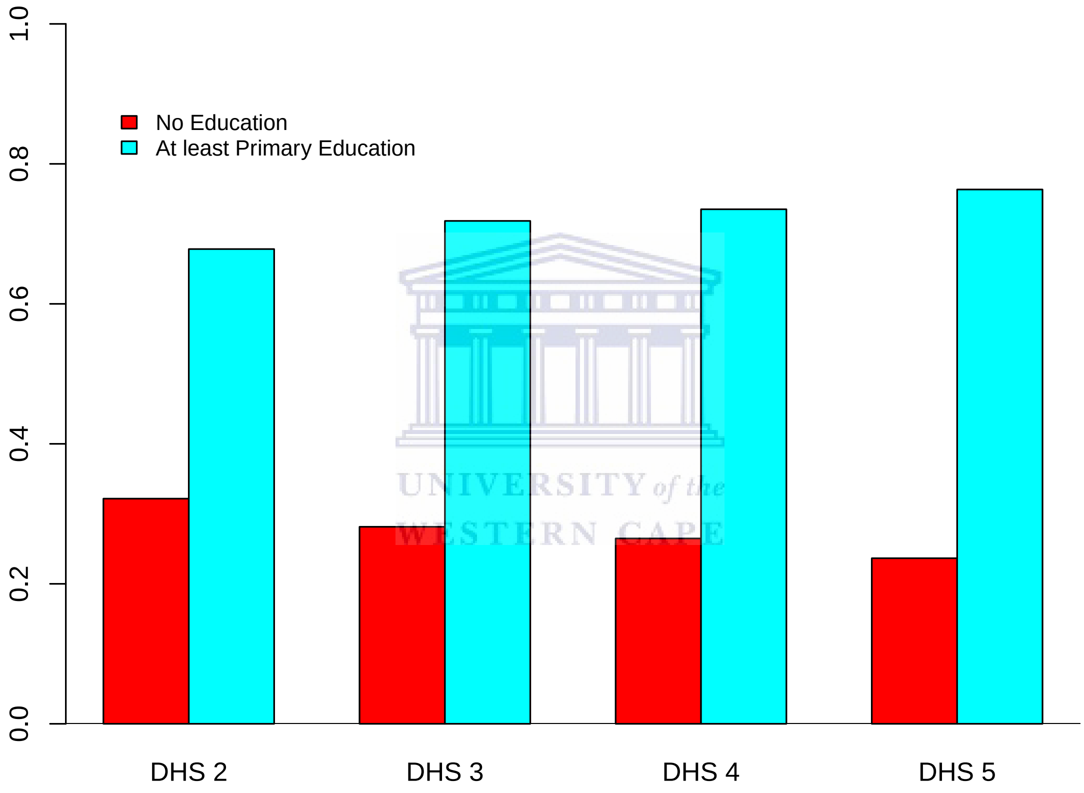


Figure A.3: Distribution of literacy level: Proportion of total population with no education

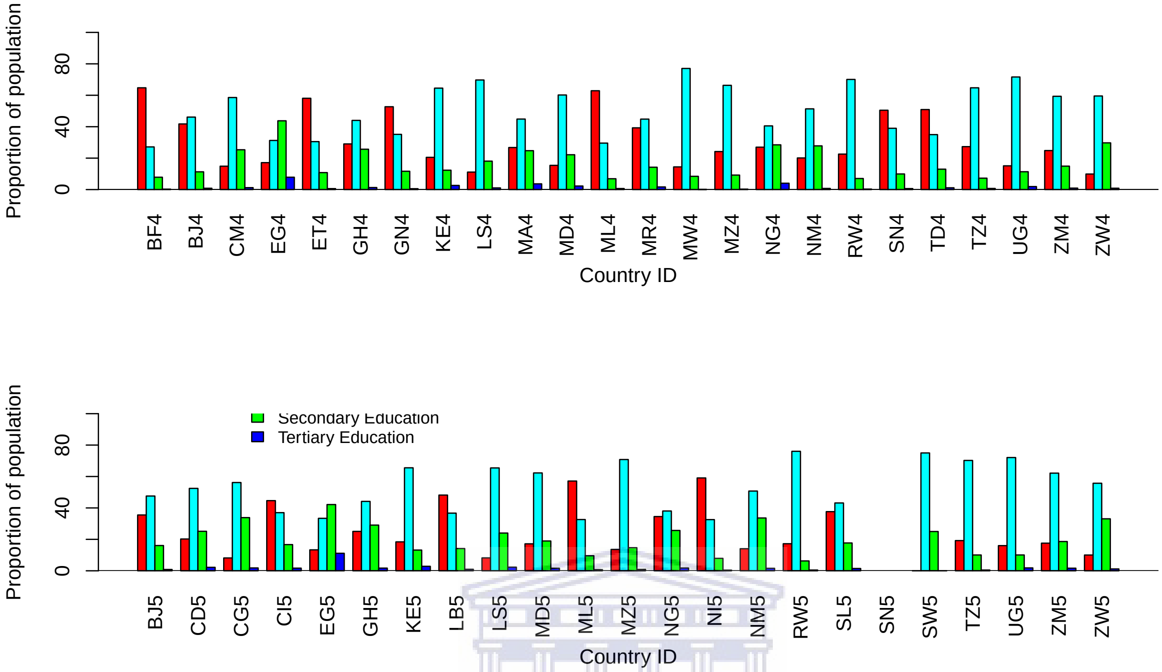


Figure A.4: Educational attainment indicated by highest level of education completed by country. Top is for DHS II (1988-1993) and bottom is for DHS III (1993-1998)

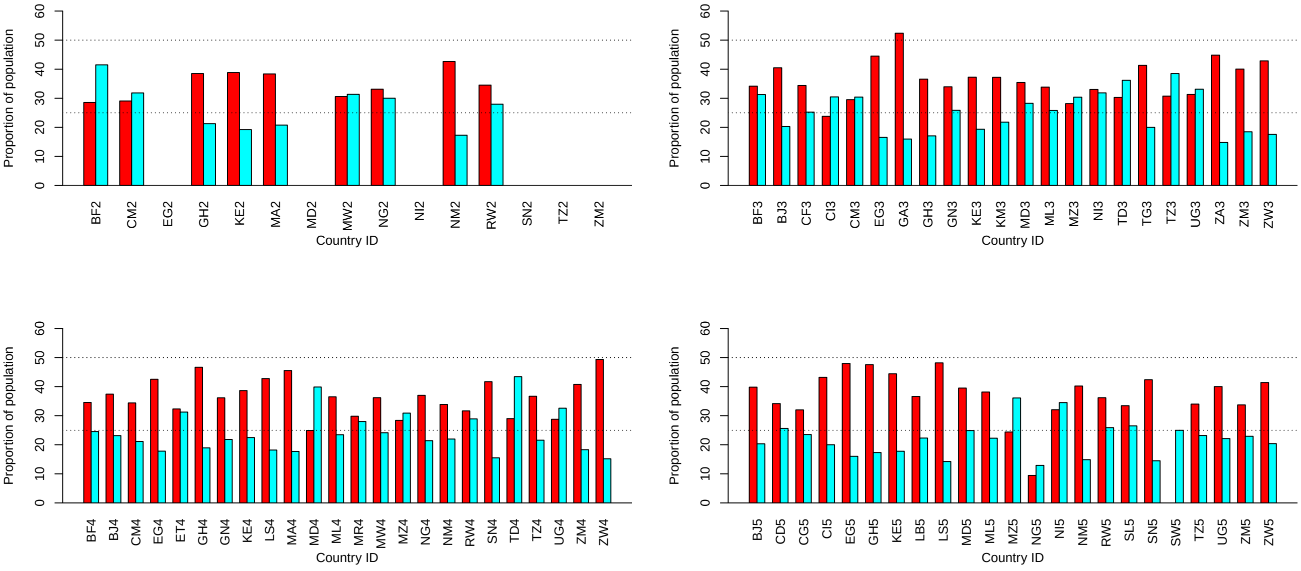


Figure A.5: Wealth index comparisons by proportion (percentage) of the poorest 40% and the richest 20% of the population

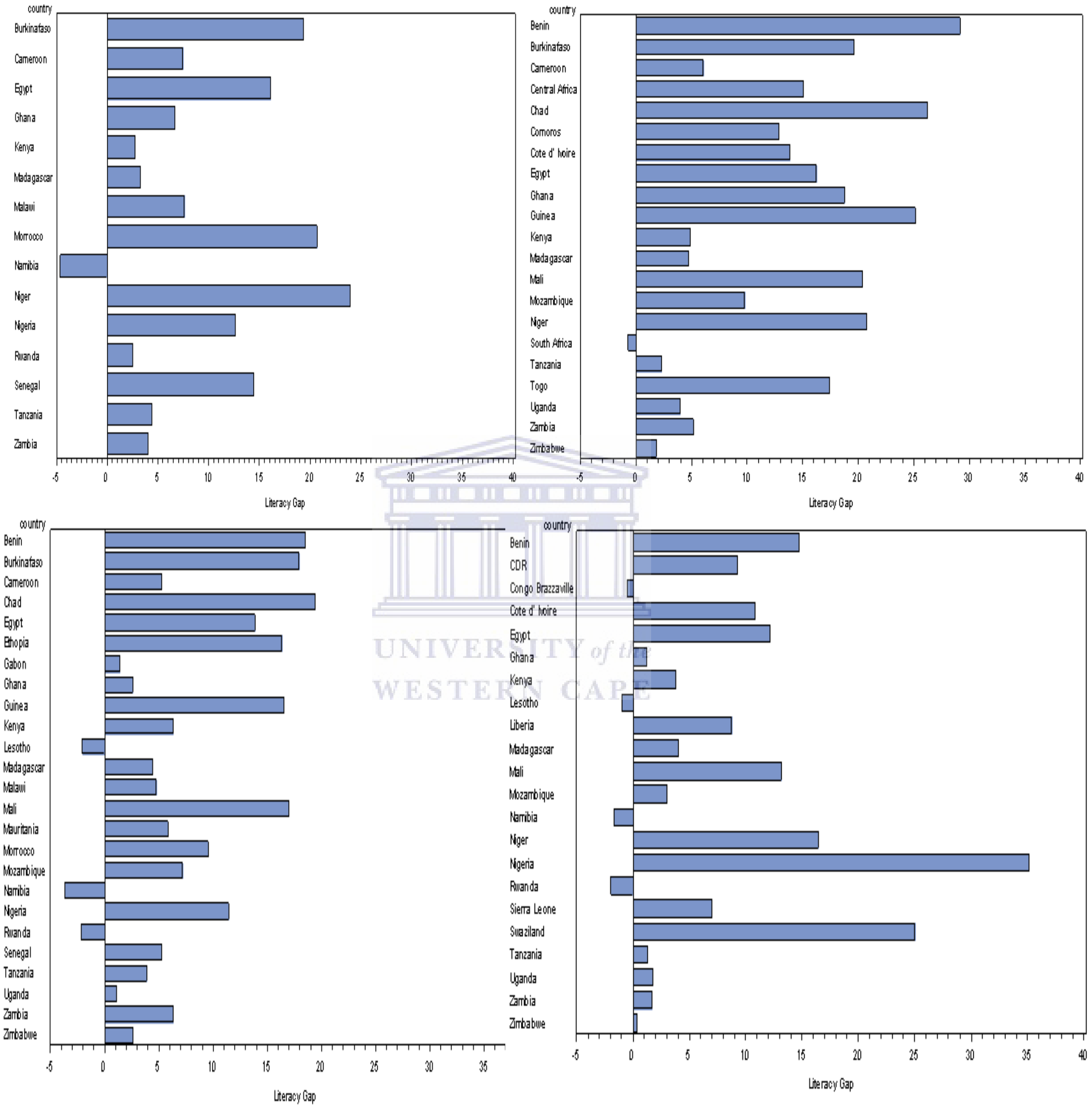


Figure A.6: Gender difference in the proportion of males and females with at least primary education (literacy gap) for countries under study: Top left DHS II (1988-1993), Top right DHS III (1993-1998): Bottom left DHS IV (1998-2003): Bottom right DHS V (2003-2008)

Table A.2: Mean wealth quintile, educational level and number of household by country

Phase	Country	Code	Mean wealth index quintile	Mean wealth index score	Mean educational level	Mean household member	Educational Gini
DHS 2	Burkina Faso	BF2	3.12175	-0.24	0.36582	10.8656	0.73613
	Cameroon	CM2	3.11723	-0.05	0.92111	9.8869	0.39706
	Egypt	EG2	.	.	1.28592	8.1683	0.35319
	Ghana	GH2	3.04678	0.05	0.86135	6.5385	0.25837
	Kenya	KE2	3.01462	-0.06	0.95485	7.5437	0.24306
	Morocco	MA2	3.05350	0.01	0.77534	8.5372	0.54833
	Madagascar	MD2	.	.	0.91253	7.7707	0.35670
	Malawi	MW2	3.09100	-0.05	0.70661	6.8269	0.35776
	Nigeria	NG2	3.09058	-0.17	0.82725	9.2489	0.46639
	Niger	NI2	.	.	0.23965	10.6245	0.82477
	Namibia	NM2	3.01788	-0.06	1.06442	10.4044	0.26745
	Rwanda	RW2	3.11171	-0.06	0.76311	7.0453	0.34351
	Senegal	SN2	.	.	0.43244	13.9561	0.70897
	Tanzania	TZ2	.	.	0.65459	8.7539	0.40539
	Zambia	ZM2	.	.	0.98610	8.7267	0.29361
DHS 3	Burkina Faso	BF3	3.07139	-0.15	0.34334	10.6869	0.74842
	Benin	BJ3	3.10729	0.15	0.56366	9.8640	0.58445
	Central Africa Republic	CF3	3.23216	0.44	0.77470	9.0966	0.42961
	Cote d'Ivoire	CI1	3.54405	0.10	0.69386	12.1208	0.54671
	Cameroon	CM2	3.11010	-0.12	1.03190	9.3842	0.33290
	Egypt	EG3	2.94182	-0.07	1.33779	7.7790	0.33698
	Gabon	GA3	3.07308	0.42	1.28348	9.8022	0.25957
	Ghana	GH3	3.19512	-0.22	0.80488	8.3171	0.24242
	Guinea	GN3	3.15584	0.15	0.48468	10.4004	0.67144
	Kenya	KE3	3.10328	-0.02	1.05447	6.6231	0.19188
	Comoros	KM3	3.08261	0.07	0.72647	8.4148	0.51453
	Madagascar	MD3	3.10537	-0.06	0.89115	7.2983	0.36447
	Mali	ML3	3.20444	0.30	0.37022	9.1818	0.73496
	Mozambique	MZ3	3.199637	0.14	0.71592	7.3056	0.35799
	Niger	NI3	3.13731	-0.03	0.33189	9.5221	0.75904
	Chad	TD3	3.10155	-0.12	0.43864	8.9982	0.65899
	Togo	TG3	3.08636	0.10	0.86816	9.0500	0.37708
	Tanzania	TZ3	3.30029	0.03	0.66468	7.6669	0.40055
	Uganda	UG3	3.05425	-0.15	0.84914	7.5472	0.29210
	South Africa	ZA3	2.93652	-0.01	1.35499	6.4670	0.25848
	Zambia	ZM3	3.13254	0.35	0.96304	8.1218	0.33506
	Zimbabwe	ZW3	2.96414	-0.08	1.19188	7.4533	0.24136

Table A.3: Mean wealth quintile, educational level and number of household by country...contd.

Phase	Country	Code	Mean wealth index quintile	Mean wealth index score	Mean educational level	Mean household member	Educational Gini
DHS 4	Burkina Faso	BF4	3.12462	4910.75	0.41731	10.8159	0.71411
	Benin	BJ4	3.14041	0.13	0.71527	8.7254	0.50227
	Cameroon	CM4	3.12530	7387.10	1.11232	8.6323	0.30780
	Egypt	EG4	2.87794	-0.24	1.42297	7.6476	0.32735
	Ethiopia	ET4	3.02144	-0.27	0.44119	6.7701	0.67873
	Ghana	GH4	3.04563	7149.21	1.06012	6.6546	0.38869
	Guinea	GN4	3.14266	18923.19	0.60180	9.2299	0.60250
	Kenya	KE4	2.95880	-28078.46	0.97666	6.6432	0.29285
	Lesotho	LS4	3.02846	4559.60	1.10788	6.6936	0.23988
	Morocco	MA4	3.00771	3689.54	1.09614	7.4760	0.39959
	Madagascar	MD4	3.05007	-48162.18	0.95676	6.8012	0.31351
	Mali	ML4	3.13058	0.25	0.45506	8.7220	0.68183
	Mauritania	MR4	3.06843	-0.13	0.69678	8.5831	0.53067
	Malawi	MW4	3.04588	0.02	0.92549	6.6184	0.22209
	Mozambique	MZ4	3.20677	10990.04	0.79277	7.3017	0.35151
	Nigeria	NG4	3.08143	1272.51	1.05977	8.2141	0.42435
	Namibia	NM4	2.98757	-0.24	1.10176	7.9422	0.33140
	Rwanda	RW4	3.20230	-0.03	0.82328	6.2545	0.28013
	Senegal	SN4	3.09217	22937.63	0.63625	13.8189	0.57382
	Chad	TD4	3.06476	-20729.57	0.52312	8.4016	0.62446
	Tanzania	TZ4	3.04472	4563.05	0.77570	7.9898	0.33311
	Uganda	UG4	3.14796	-0.12	0.95949	7.6230	0.25296
	Zambia	ZM4	3.11952	0.35	0.94307	7.6539	0.35233
	Zimbabwe	ZW4	2.90483	-0.07	1.24039	6.6598	0.25472
DHS 5	Benin	BJ5	3.07719	10333.80	0.84031	8.0481	0.45246
	CDR	CD5	3.17597	18278.25	1.09356	7.9839	0.34560
	Congo Brazzaville	CG5	3.08829	-438.97	1.28265	8.3163	0.24898
	Cote d'Ivoire	CI5	3.14987	43407.63	0.86219	9.6788	0.50250
	Egypt	EG5	2.86335	-13085.83	1.53136	6.8533	0.30393
	Ghana	GH5	2.99273	2948.10	1.12583	6.2431	0.37010
	Kenya	KE5	2.90630	-27497.42	1.03587	6.4604	0.28328
	Liberia	LB5	3.17953	29069.61	0.67109	7.3763	0.58440
	Lesotho	LS5	2.96773	7519.04	1.24324	6.6768	0.23981
	Madagascar	MD5	3.07144	-3981.36	1.05930	6.7985	0.28398
	Mali	ML5	3.05933	10688.08	0.52081	8.6670	0.64554
	Mozambique	MZ5	3.20113	-2.88	0.96205	6.2247	0.24107
	Nigeria	NG5	3.78448	60061.80	0.94690	11.1638	0.46729
	Niger	NI5	3.05282	-13959.89	0.38767	9.3778	0.70362
	Namibia	NM5	2.98703	-4488.19	1.26252	7.5346	0.29257
	Rwanda	RW5	3.07261	5534.86	0.89843	6.1867	0.22121
	Sierra Leone	SL5	3.11553	-1008.91	0.78013	8.0841	0.50537

Table A.4: General variable definitions and description...contd.

Phase	Country	Code	Mean wealth index quintile	Mean wealth index score	Mean educational level	Mean household member	Educational Gini
DHS5	Senegal	SN5	3.07951	24785.92	.	14.3565	
	Swaziland	SZ5	3.75000	13217.75	1.25000	6.5000	
	Tanzania	TZ5	3.06421	5429.17	0.87784	7.7666	0.25713
	Uganda	UG5	3.10423	7536.75	0.99053	7.5098	0.24440
	Zambia	ZM5	3.16159	17618.25	1.03861	7.0413	0.31351
	Zimbabwe	ZW5	3.02172	-2181.09	1.26180	6.5082	0.25980

Appendix B

D matrix

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \phi_{st}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[\frac{\mu_{st}(\boldsymbol{\beta})}{1 - e^{-\mu_{st}(\boldsymbol{\beta})}} \right] \\
&= \frac{(1 - e^{-\mu_{st}(\boldsymbol{\beta})}) \frac{\partial}{\partial \boldsymbol{\beta}} \mu_{st}(\boldsymbol{\beta}) - \mu_{st}(\boldsymbol{\beta}) e^{-\mu_{st}(\boldsymbol{\beta})} (-1) \frac{\partial}{\partial \boldsymbol{\beta}} \mu_{st}(\boldsymbol{\beta})}{(1 - e^{-\mu_{st}(\boldsymbol{\beta})})^2} \\
&= \frac{(1 - e^{-\mu_{st}(\boldsymbol{\beta})}) + \mu_{st}(\boldsymbol{\beta}) e^{-\mu_{st}(\boldsymbol{\beta})}}{(1 - e^{-\mu_{st}(\boldsymbol{\beta})})^2} \frac{\partial}{\partial \boldsymbol{\beta}} \mu_{st}(\boldsymbol{\beta}) \\
&= \frac{(1 - e^{-\mu_{st}(\boldsymbol{\beta})}) + \mu_{st}(\boldsymbol{\beta}) e^{-\mu_{st}(\boldsymbol{\beta})}}{(1 - e^{-\mu_{st}(\boldsymbol{\beta})})^2} \exp(\mathbf{x}_{st}^{\tau} \boldsymbol{\beta}) \mathbf{x}_{st}^{\tau} \\
&= \frac{(1 - e^{-\mu_{st}(\boldsymbol{\beta})}) + \mu_{st}(\boldsymbol{\beta}) e^{-\mu_{st}(\boldsymbol{\beta})}}{(1 - e^{-\mu_{st}(\boldsymbol{\beta})})^2} \mu_{st}(\boldsymbol{\beta}) \mathbf{x}_{st}.
\end{aligned}$$

Bibliography

Abdulraheem, I. S., Onajole, A. T., Jimoh, A. A. G. and Oladipo, A. R. (2011) Reasons for incomplete vaccination and factors for missed opportunities among rural Nigerian children. *Journal of Public Health and Epidemiology*, **3**, 194–203.

AbdurRab, M., Freeman, T., Rahim, S., Durrani, N., Simon-Taha, A. and Rowland, M. (2003) High altitude epidemic malaria in Bamian Province, central Afghanistan. *East Mediterranean Health Journal*, **9**, 232–239.

Adegboye, O. (2010a) Under-five mortality in Nigeria: Spatial exploration and spatial scan statistics for cluster detection . *International Journal of Statistics and Systems*, **5**, 203–214.

Adegboye, O. and Kotze, D. (2012) Disease mapping of leishmaniasis outbreak in Afghanistan: Spatial hierarchical bayesian analysis. *Asian Pacific Journal of Tropical Disease*, **2**, 253–259.

Adegboye, O. A. (2010b) Spatio-temporal analysis of transmission of avian flu outbreak in Nigeria. In *Proceedings of the 25th International Biometric Conference*. Florinapolis, Brazil: International Biometric Society.

Adeiga, A., Omilabu, S., Audu, R., Sanni, F., Lakehinde, G., Balogun, O. and Olagbaju, O. (2005) Infant immunization coverage in difficult-to-reach area of Lagos metropolis. *African Journal of Clinical and Experimental Microbiology*, **6**, 227–231.

Alabi, R. (2008) Income distribution and accessibility to primary and secondary schools in Nigeria. Tech. rep., Monograph/Discussion Series 114. Institute for World Economics and International Management, University of Bremen.

Albert, P. (1999) Longitudinal data analysis (repeated measures) in clinical trials. *Statistcs in Medicine*, **18**, 1707–1732.

Allepuz, A., Lopez-Quilez, A., Forte, A., Fernandez, G. and Casal, J. (2007) Spatial analysis of bovine spongiform encephalopathy in Galicia, Spain (2000-2005). *Preventive Veterinary Medicine*, **79**, 174–185.

Aluede, R. (2008) Regional demands and contemporary educational disparities in Nigeria. *Journal of Social Sciences*, **14**, 183–189.

Amowitz, L., Reis, C. and Jacopino, V. (2002) Maternal mortality in herat province, Afghanistan, in 2002: An indicator of women's human rights . *JAMA*, **288**, 1284–1291.

Anah, M. U., Etuk, I. S. and Udo, J. J. (2006) Opportunistic immunization with in-patient programme: eliminating a missed opportunity in Calabar, Nigeria. *Annals of African Medicine* , **5**, 188–191.

Antai, D. (2009) Inequitable childhood immunization uptake in Nigeria: A multilevel analysis of individual and contextual determinants. *BMC Infectious Diseases*, **9**.

APHI/MoPH/CSO, Afghanistan and ICF Macro, Calverton, Maryland, USA and IIHMR, India and WHO-EMRO, Egypt (2011) Afghanistan mortality survey 2010. Tech. rep., Afghan Public Health Institute, Ministry of Public Health (APHI-MoPH) [Afghanistan] and Central Statistics Organization (CSO) [Afghanistan] and ICF Macro and Indian Institute of Health Management Research (IIHMR) [India] and World Health Organization Regional Office for the Eastern Mediterranean (WHO-EMRO) [Egypt].

Arnese, R. (2007) Applying ecological niche factor analysis for predicting modelling in the kaulonia field survey. In *Proceedings of the 35th International conference on computer applications and quantitative methods in archaeology*, 428p. Berlin, Germany.

Arnold, B. and Strauss, D. (1991) Bivariate distributions with conditionals in prescribed exponential families. *Journal of Royal Statistical Society B*, **53**, 365–375.

Atlantic Ecology Division (2013) . URL `http://www.epa.gov/aed/html/research/scallop/hsi.html`.

Bai, Y., Song, P. X.-K. and Raghunathan, T. E. (2012) Joint composite estimating functions in spatiotemporal models. *Journal of Royal Statistics Society, B*, **74**, 799–824.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. and Finley, AO amd Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *Journal of Royal Statistics Society, B*, **70**, 825–848.

Bapat, U., Alcock, G., More, N., Das, S., Joshi, W. and Osrin, D. (2012) Stillbirths and newborn deaths in slum settlements in Mumbai, India: a prospective verbal autopsy study. *BMC Pregnancy and Childbirth*, **12**.

Barro, R. and Lee, J. (2010) A new data set of educational attainment in the world, 19502010. Tech. rep., NBER Working Paper No. 15902.

Bartlett, L., Mawji, S., Whitehead, S., Crouse, C., Dalil, S., Ionete, D., Salama, P. and the Afghan Maternal MortalityStudy Team (2005) Where giving birth is a forecast of death: Maternal mortality in four districts of Afghanistan, 1999-2002. *Lancet*, **365**, 864–870.

Basille, M., Calenge, C., Marboutin, E., Andersen, R. and Gaillard, J. (2008) Assessing habitat selection using multivariate statistics: Some refinements of the ecological-niche factor analysis. *Ecological Modelling*, **211**, 233–240.

Bello, M., Lukshi, B. M. and Sanusi, M. (2008) Outbreaks of highly pathogenic avian influenza (H5N1) in Bauchi State, Nigeria. *International Journal of Poultry Science*, **7**, 450–456.

Bernardinelli, L., C. D. and Songini, M. (1995) Bayesian analysis of space-time variation disease risk. *Statistics in Medicine*, **14**, 2433–2443.

Bhutta, Z. (2002) Children of war: the real casualties of the Afghan conflict. *BMJ*, **324**, 349–352.

Bickel, P., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (2006) *Disease and Mortality in Sub-Saharan Africa*. World bank.

Bledsoe, C., John, C., Jennifer, J. K. and H., J. (1999) *Critical Perspectives on Schooling and Fertility in the Developing World*. Washington, DC: National Academy Press.

Bosch-Capblanch, X., Banerjee, K. and Burton, A. (2012) Unvaccinated children in years of increasing coverage: how many and who are they? Evidence from 96 low- and middle-income countries. . *Tropical Medicine and International Health* , **17**, 697710.

Brown, V. and Rohani, P. (2012) The consequences of climate change at an avian influenza 'hotspot. *Biology Letters*.

Burnham, K. and Anderson, D. (2002) *Model selection and multimodel inference : A practical information-theoretic approach*. New York: Springer.

Carey, V., Zeger, S. and Diggle, P. (1993) Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.

Centers for Disease Control and Prevention (2012) URL `http://wwwnc.cdc.gov`.

Chatfield, C. and Collins, A. (1980) *Introduction to Multivariate Data Analysis.* London: Chapman and Hall.

Chiswick, B. (71) Earnings inequality and economic developmen. *Quarterly Journal of Economics*, **85**, 21–39.

Clayton, D. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In *Geographical and environment epidemiology: Methods for small area studies.* (eds. C. J. E. D. Elliott, P. and R. Stern).

Collier, P. (2007) Economic causes of civil conflict and their implications for policy. In *Leashing the Dogs of War: Conflict management in a divided world* (eds. A. Crocker, F. Hampson and P. Aal).

Cressie, N. (1993) *Statistics for Spatial Data.* New York: Wiley.

Cressie, N. and Huang, H. C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of American Statistical Association*, **94**, 13301340.

Cressie, N. and Johannesson, G. (2008) Fixed rankriging for very large spatial data sets. *Journal of Royal Statistics Society, B*, **70**, 209–226.

Deng, L. (2003) Education in southern sudan: War, status and challenges of achieving education for all goals. Tech. rep., EFA Global Monitoring Report 2003/4. URL `http://www.unesco.org/en/efareport/reports/20034-gender`.

Dickey, D. A. and Fuller, W. (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Diddy, A. (2009) Inequitable childhood immunization uptake in Nigeria: a multilevel analysis of individual and contextual determinants. *BMC Infectious Diseases*, **9**.

Diggle, P., Liang, K. Y. and Zeger, S. (1994) *Analysis of Longitudinal Data (second edition).* Oxford:OUP.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*. Oxford:OUP, second edn.

Durr, P., Tait, N. and Lawson, A. (2005) Bayesian hierarchical modeling to enhance the epidemiological value of abattoir surveys for bovine fasciolosis. *Preventive Veterinary Medicine*, **71**, 157–172.

Duru, J. A. (2006) Avian influenza outbreak in Nigeria: The farmers experience . *Animal Production Research Advances*, **2**, 129–133.

Education Policy and Data Center (2012) Namibia core USAID education profile. URL `http://www.unesdoc.unesco.org/images/0012/001297/129777e.pdf`.

EFA Global Monitoring Report (2002) Education for All: Is the world of track? . Tech. rep., UNESCO. URL `http://www.unesdoc.unesco.org/images/0012/001297/129777e.pdf`.

Ekong, P. S., Ducheyne, E., Carpenter, T. E., Owolodun, O. A., Oladokun, A. T., Lombin, L. H. and Berkvens, D. (2012) Spatio-temporal epidemiology of highly pathogenic avian influenza(H5N1) outbreaks in Nigeria, 2006-2008. *Preventive Veterinary Medicine*, **103**, 170–177.

Fan, J., Huang, T. and Li, R. (2007) Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632–641.

Farnsworth, M. and Ward, M. (2009) Identifying spatio-temporal patterns of transboundary disease spread: examples using influenza H5N1 outbreaks. *Veterinary Research*, **40**.

Fasina, F., Bisschop, S., Joannis, T., Lombin, L. and Aboolnik, C. (2009) Molecular characterization and epidemiology of the highly pathogenic avian influenza H5N1 in Nigeria. *Epidemiology and Infection*, **137**, 456–463.

Fasina, F. O. (2008) *Molecular and Spatial-temporal Epidemiology of highly Pathogenic Notifiable Avian influenza H5N1 in Nigeria.* Master's thesis, University of Pretoria, South Africa.

Fasina, F. O., Rivas, A. L., Bisschopd, S. P., Stegemane, A. J. and Hernandezf, J. A. (2011) Identification of risk factors associated with highly pathogenic avian influenza H5N1 virus infection in poultry farms, in Nigeria during the epidemic of 2006-2007. . *Preventive Veterinary Medicine*, **98**, 204–208.

Faulde, M., Hoffmann, R., Fazilat, K. and Hoerauf, A. (2007) Malaria Reemergence in Northern Afghanistan. *Emerging Infectious Diseases*, **13**, 1402–1404.

Fauveau, V. (2005) The study of causes of death in developing countries. In *Demography: analysis and synthesis* (eds. G. Wunsch, G. Caselli and J. Vallin).

Fischer, D., Thomas, S. and Beierkuhnlein, C. (2011) Modelling climatic suitability and dispersal for disease vectors: The example of a phlebotomine sandfly in Europe. *Procedia environemntal sciences*, **7**, 164–169.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008) *Longitudinal Data Analyisi.* FL: Chapmann and Hall.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) *Applied Longitudinal Analysis.* John Wiley and Sons: New York.

Gardner, R. (1998) Education. Tech. rep., DHS Comparative Studies No. 29, Macro International Inc.

Garenne, M. and Fauveau, V. (2006) Potential and limits of verbal autopsies. *Bulletin of the World Health Organization*, **84**.

Gessner, B. D. (1994) Mortality rates, causes of death, and health status among displaced and resident populations of Kabul, Afghanistan. *Journal of the American Medical Association*, **272**, 382–385.

Gibson, J. (2002) Why does the engel method work? food demand, economies of size and household survey methods. *Oxford Bulletin of Economics and Statistics*, **64**, 341–360.

Gilbert, M., Xiao, X., Domenech, J., Lubroth, J., Martin, V. and Slingenbergh, J. (2006) Anatidae migration in the western palearctic and spread of highly pathogenic avian influenza H5N1 virus. *Infectious Disease*, **12**, 1650–1656.

Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data. *Journal of American Statistical Association*, **97**, 590–600.

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hanson, B., Swayne, D., Senne, D., Lobpries, D., Hurst, J. and Stallknecht, D. (2005) Avian influenza viruses and paramyxoviruses in wintering and resident ducks in Texas. *Journal of Wildlife Diseases*, **41**, 624–628.

Hay, S., Guerra, C., Tatem, A., Atkinson, P. and Snow, R. (2005) Urbanization, malaria transmission and disease burden in Africa. *Nature Reviews Microbiology*, **3**, 81–90.

Herfst, S., Eefje, J. A. S., Linster, M., Chutinimitkul, S., Emmie, d. W., Vincent, J. M., Erin, M. S., Theo, M. B., David, F. B., Derek, J. S., Guus, F. R., Albert, D. M. E. O. and Ron, A. M. F. (2012) Airborne transmission of influenza A/H5N1 Virus Between ferrets. *Science*, **336**, 1534–1541.

Hin, L. and Wang, Y. G. (2009) Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine,*, **28**, 642658.

Hirzel, A., Hausser, J., Chessel, D. and Perrin, N. (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data. *Ecology*, **83**, 20272036.

Hjsgaard, S., Halekoh, U. and Yan, J. (2006) The R package geepack for generalized estimating equations. *Journal of Statistical Software*, **15**, 1–11.

Jannie, H. and S., J. (2009) Effects of household and district-level factors on primary school enrolment in 20 developing countries. *World Development*, **37**, 179–193.

Kabir, M. and Iliyasu, Z. and Abubakar, I. S. and Nwosuh, J. I. (2004) Immunization coverage among children below two years of age in Fanshakara, Kano, Nigeria. *Nigerian Journal of Basic and Clinical Sciences* , **1**, 10–13.

Kandala, N., Ji, C., Stallard, N., Stranges, S. and Cappuccio, P. F. (2008) Morbidity from diarrhoea, cough and fever among young children in Nigeria. *Annals of Tropical Medicine and Parasitology*, **102**, 427–455.

Karagiannis-Voules, D., Scholte, R., Guimara, L., Utzinger, J. and P, V. (2013) Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLOS Neglected Tropical Diseases*, **7**.

Kayode, G., Adekanmbi, V. T. and Uthman, O. A. (2012) Risk factors and a predictive model for under-five mortality in Nigeria: evidence from Nigeria demographic and health survey. *BMC Pregnancy and Childbirth,*, **12**.

Kim, Y., Tappis, H., Zainullah, P., Ansari, N., Evans, C., Bartlett, L., Zaka, N. and Zeck, W. (2012) Quality of caesarean delivery services and documentation in first-line referral facilities in Afghanistan: a chart review. *BMC Pregnancy and Childbirth*, **12**.

King, G. and Lu, Y. (2008) Verbal autopsy methods with multiple causes of death. *Statistical Science*, **23**, 78–91.

Korenromp, E. L., F., A., Williams, B., Nahlen, B. and Snow, R. (2004) Monitoring trends in under-5 mortality rates through national birth history surveys. *International Journal of Epidemiology*, **33**, 1293–1301.

Krauss, S., Walker, D., Pryor, S., Niles, L., Chenghong, L., Hinshaw, V. and Webster, R. (2004) Influenza A viruses of migrating wild aquatic birds in North America. *Vector-Borne Zoonotic Disease*, **4**, 177–189.

Kulldorff, M. (1997) A spatial scan statistic. *Commun. Stat. Theory Methods.*, **26**, 1481–1496.

Kulldorff, M., L., H. and Konty, K. (2009) A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, **8**.

Kunle-Olowu, A., Kunle-Olowu, E. O. and Ugwu, M. E. (2011) Immunization coverage of antenatal and immunization clinics attendees in the Niger Delta university teaching hospital. *Journal of Public Health and Epidemiology*, **3**, 90–93.

Lawson, A. (2008) *Hierachical modelling in spatial epidemilogy.* Florida: Chapman and Hall/CRC Press.

Lawson, A. and Zhou, H. (2007) Spatial statistical modeling of disease outbreaks with particular reference to the UK foot and mouth disease (FMD) epidemic of 2001. *Preventive Veterinary Medicine*, **71**, 141–156.

Lee, A., Mullany, L., Tielsch, J., Katz, J., Khatry, S., LeClerq, S., Adhikari, R., Shrestha, S. and Darmstadt, G. (2008) Verbal autopsy methods to ascertain birth asphyxia deaths in a community-based setting in Southern Nepal. *Pediatrics*, **121**.

Lesaffre, E. and Molenberghs, G. (1991) Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine*, **10**, 1391–1403.

Leung, D., Wang, Y.-G. and Zhu, M. (2009) Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics*, **10**, 436–445.

Levine, S. and Roberts, B. (2008) Dynamics of income inequality and poverty in post-independence Namibia. Helsinki, Finland: UNU-WIDER Conference:Frontiers of Poverty Analysis.

Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lorenzo-Luaces Alvarez, P., Guerra-Yi, M., Faes, C., Galn, A. and Molenberghs, G. (2009) Spatial analysis of breast and cervical cancer incidence in small geographical areas in Cuba, 1999-2003. *European Journal of Cancer Prevention*, **18**, 395–403.

Lowe, R., Bailey, T., Stephenson, D., Graham, R., Coelho, C., Carvalho, M. and Barcellos, C. (2010) Spatio-temporal modeling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers and Geosciences*.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Maekawa, M., Douangmala, S., Sakisaka, K., Takahashi, K., Phathammavong, O., Xeuatvongsa, A. and Kuroiwa, C. (2007) Factors affecting routine immunization coverage among children aged 1259 months in LaoPDR after regional polio eradication in Western Pacific Region . *BioScience Trends*, **1**, 43–51.

Mariella, L. and Tarantino, M. (2010) Spatial temporal conditional auto-regressive model: A new autoregressive matrix. *Australian Journal of Statistics*, **39**.

Mayhew, M., Hansen, P., Peters, D., Edward, A., Singh, L., Dwivedi, V., Mashkoor, A. and Burnham, G. (2008) Determinants of skilled birth attendant utilization in Afghanistan: A cross-sectional study. *American Journal of Public Health*, **98**, 1850–1856.

Mbanefo, E., Umeh, J., Oguoma, Y. and Eneanya, C. (2009) Antenatal malaria parasitaemia and haemoglobin profile of pregnant mothers in Awka, Anambra State, SouthEast Nigeria. *American-Eurasian Journal of Scientific Research*, **4**, 235–239.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models.* New York:Chapman and Hall, second edn.

McShane, L. M., Albert, P. S. and Palmatier, M. A. (1997) A latent process regression model for spatially correlated count data. *Biometrics*, **53**, 698–706.

Michael, F., Joachim, S., Gerhard, H., Mohammed, A. and Achim, H. (2008) Zoonotic cutaneous Leishmaniasis outbreak in Mazar-e Sharif, northern Afghanistan: An epidemiological evaluation . *International Journal of Medicial Microbiology*, **298**, 543–550.

Mikkelsen, T., Alexandersen, P., Champion, H., Astrup, P., Donaldson, A., Dunkerley, F., Gloster, J., Srensen, J. and Thykier-Nielsen, S. (2003) Investigation of airborne foot-and-mouth disease virus transmission during low-wind conditions in the early phase of the UK 2001 epidemic. *Atmospheric Chemistry Physics*, **3**, 2101–2110.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modeling of correlated ordinal data using a multivariate Placket distribution. *Journal of the American the Statistical Association*, **89**, 633–644.

Molenberghs, G. and Ryan, L. M. (1999) An exponential family model for clustered multivariate binary data. *Environmetrics*, **10**, 279–300.

Molenberghs, G. and Verbeke, G. (2006) *Models for Discrete Longitudinal Data.* NewYork: Springer-Verlag.

Moran, P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Musa, O., Aderibigbe, S., Salaudeen, G., Oluwole, F. and Samuel, S. (2010) Community awareness of bird flu and the practice of backyard poultry in a North-Central State of Nigeria. . *Journal Preventive Medicine Hygiene*, **51**, 146–151.

National Population Commission (NPC) [Nigeria] and ICF Macro (2009) *Nigeria Demographic and Health Survey 2008*. Abuja, Nigeria: National Population Commission and ICF Macro.

Neelon, B., Anthopolos, R. and Miranda, M. (2012) A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. *Statistical methods in medical research.*

Odoi, A., Martin, S., Michel, P., Holt, J., Middleton, D. and Wilson, J. (2003) Geographical and temporal distribution of human giardiasis in Ontario, Canada. . *International Journal of Health Geographics*, **3**.

Ogunjimi, L. O., Ibe, R. T. and Ikorok, M. M. (2012) Curbing maternal and child mortality: The Nigerian experience. *International Journal of Nursing and Midwifery*, **4**, 33–39.

Ogunjuyigbe, P. O. (2004) Under-five mortality in Nigeria: Perception and attitudes of the Yorubas towards the existence of Abiku. *Demographic research*, **11**, 43–56.

Ojo, O. (2008) *The economic effect of avian influenza in poultry in South-West Nigeria.* Master's thesis, University of Ibadan.

Onyiriuka, A. N. (2005) Vaccination default among children attending a static immunization clinic in Benin city, Nigeria. *Journal of Medicine and Biomedical Research.* , **4**, 71–77.

Oyana, T., Dai, D. and Scott, K. (2004) Spatiotemporal distributions of reported cases of the avian influenza H5N1 (bird flu) in Southern China in early 2004. *Avian Diseases*, **50**, 508–515.

Pan, W. (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–125.

Paul, S., Zhang, X. and Xu, J. (2013) Estimation of regression parameters for binary longitudinal data using GEE: Review, extension and an application to environmental data. *Journal of Environmental Statistics*, **4**, 1–12.

Peng, R. and Bell, M. (2010) Spatial misalignment in time series studies of air pollution and health data. *Biostatistics*, **11**, 720–740.

Peterson, A. (2006) Ecologic niche modelling and spatial patterns of disease transmission. *Emerging infectious diseases*, **12**, 1822–1826.

Philip, W., Osman, S., Chalapati, R., Victoria, A., Colin, M., Yang, G., Yusuf, H., Prabhat, J. and Alan, D. (2005) Sample registration of vital events with verbal autopsy: A renewed commitment to measuring and monitoring vital statistics. Tech. rep., Bulletin of the World Health Organization .

Porcu, E., Mateu, J. and Bevilacqua, M. (2007) Covariance functions that are stationary or nonstationary in space and stationary in time. *Statistica Neerlandica*, **61**, 358382.

Prasad, A. (2006) Disease profile of children in Kabul: the unmet need for health care. *Journal of Epidemiology and Community Health*, **60**, 20–23.

Preisser, J. S., Arcury, T. A. and Quandt, S. A. (2003) Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data. *American Journal of Epidemiology*, **158**, 495–501.

Prentice, R. (1998) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Psacharopoulos, G. and Arriagada, A. (1986) The educational attainment of the labour force: An international comparison. *International Labour Review*, **125**, 561–574.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating functions. *Annals of Statistics*, **22**, 300–25.

Quigley, M. A. (2005) Commentary: Verbal autopsiesfrom small-scale studies to mortality surveillance systems. *International Journal of Epidemiology*, **34**, 1087–1088.

R Development Core Team (n.d.) A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria 2006. URL `http://www.R-project.org`.

Rajesh, K. and Sanjay, K. (2013) Change in global climate and prevalence of visceral leishmaniasis. *International Journal of Scientific and Research Publications*, **3**.

Rammohan, A., Awofeso, N. and Fernandez, R. C. (2012) Paternal education status significantly influences infants measles vaccination uptake, independent of maternal education status. *BMC Public Health* , **12**.

Ravallion, M. (1994) *Poverty comparisons: fundamentals of pure and applied economics.* Switzerland: Harwood Academic Publishers.

Reithinger, R., Aadil, K., Kolaczinski, J., Mohsen, M. and Hami, S. (2005) Social impact of Leishmaniasis, Afghanistan. *Emerging Infectious Diseases*, **11**, 634–6.

Reithinger, R. and Coleman, P. G. (2007) Treating cutaneous Leishmaniasis patients in Kabul, Afghanistan: cost-effectiveness of an operational program in a complex emergency setting. *BMC Infectious Diseases*, **7**.

Reithinger, R., Mohsen, M., Aadil, K., Sidiqi, M., Erasmus, P. and Coleman, P. G. (2003) Anthroponotic cutaneous Leishmaniasis, Kabul, Afghanistan. *Emerging Infectious Diseases*, **9**, 727–729.

Reithinger, R., Mohsen, M. and Leslie, T. (2010) Risk factors for anthroponotic cutaneous

Leishamaniasis at the household level in Kabul, Afghanistan. *PLOS Neglected Tropical Diseases*, **4**.

Renne, E. P. (2010) *The politics of polio in Northern Nigeria.* Indiana University Press: Indiana, USA.

Ribeiro, P. and Diggle, P. (2001) geoR: a package for geostatistical analysis. *R News*, 14–18.

Rutstein, S. O. and Kiersten, J. (1994) The DHS Wealth Index. Tech. rep., DHS Comparative Reports No. 6, Macro International Inc.

Safi, N., Hameed, H., Sediqi, W. and Himmat, E. (2009) NMLCP Annual Report, 2008. *Afghanistan Anuual Malaria Journal*, **1**, 8–14.

Said, S. E. and Dickey, D. A. (1984) Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, **71**, 599–607.

SAS Institute Inc., Cary, N. (2008) Sas/stat 9.2 users guide.

Senn, N., Maraga, S., Sie, A., Rogerson, S., Reeder, J., Siba, P. and Mueller, I. (2010) Population hemoglobin mean and anemia prevalence in Papua New Guinea: New metrics for defining malaria endemicity. *Plos one*, **5**.

Sibai, A. M.and Fletcher, A., Hills, M. and Campbell, O. (2001) Non-communicable disease mortality rates using the verbal autopsy in a cohort of middle aged and older populations in Beirut during wartime, 1983-93. *Journal of Epidemiology and Community Health*, **55**, 271–276.

Snow, R., Craig, M., Deichmann, U. and Marsh, K. (1999) Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bulletin of the World Health Organization*, **77**, 624–640.

Song, C. and Kulldorff, M. (2005) Tango's maximized excess events test with different weights. *International Journal of Health Geographics*, **4**.

Song, C., Kulldorff, M., Gregorio, D., Samociuk, H. and DeChello, L. (2006) Cancer map patterns: Are they random or not? *American Journal of Preventive Medicine*, **30**, 37–49.

Sorungbe, A. (1989) Expanded programme on immunization in Nigeria. *Reviews of Infectious Diseases*, **11**, 509–511.

Spekreijse, D., Bouma, A., Koch, G. and Stegeman, J. (2011) Airborne transmission of a highly pathogenic avian influenza virus strain H5N1 between groups of chickens quantified in an experimental setting. *Veterinary Microbiology*, **152**, 88–95.

Ssematimba, A., Hagenaars, T. and de Jong, M. (2012) Modelling the wind-borne spread of highly pathogenic avian influenza virus between farms. *PloSONE*, **7**.

Stallard, E. (2002) Underlying and multiple cause mortality at advanced ages: United states 1980-1998. *North America Actuarial Journal*, **6**.

Stein, M. L. (2005) Space-time covariance functions. *Journal of American Statistical Association*, **100**, 310321.

Stevenson, M., Morris, R., Lawson, A., Wilesmith, J., Ryan, J. and Jackson, R. (2005) Area-level risk for BSE in British cattle before and after the July 1988 meat and bone meal feed ban. *Preventive Veterinary Medicine*, **69**, 129–144.

Tango, T. (1995) A class of tests for detecting general and focused clustering of rare diseases. *Statistics in Medicine*, **14**, 23232334.

Tatem, A., Guerra, C., Kabaria, C., Noor, A. and Hay, S. (2008) Human population, urban settlement patterns and their impact on plasmodium falciparum malaria endemicity. *Malaria Journal*, **7**, 218.

The Ombudsman (2010) Namibia (NHRI) Submission to the universal periodic review mechanism. URL `http://lib.ohchr.org`.

Thomas, V., Wang, Y. and Fan, X. . (2001) Measuring education inequality: Gini coefficients of education. Tech. rep., World Bank Policy Research Working Paper No. 2525. URL `http://ssrn.com/abstract=258182`.

Thompson, R., De Oliveira Lima, J., Maguire, J., Braud, D. and Scholl, D. (2002) Climatic and demographic determinants of american visceral leishmaniasis in northeastern brazil using remote sensing technology for environmental categorization of rain and region influences on leishmaniasis. *American Journal of Tropical Medicine and Hygiene*, **67**, 648–655.

Toby, L., Sarah, S., Mohammed, S., Ismail, M., Najibullah, M. A., Kathy, F., Annick, L., Rosalynn, O. and Richard, R. (2006) Visceral Leishmaniasis in Afghanistan. *Canadian Medical Association Journal*, **173**, 245–246.

Townsend, P. (1979) *Poverty in the United Kingdom: A survey of household resources and standards of living*. Harmondsworth; Penguin Books.

UNICEF (2010) At a glance: Nigeria statistics. Tech. rep.

UNICEF (2010) Childinfo, monitoring the situation of children and women. Assessed on 16th November 2010. URL `http://www.childinfo.org/mortality_overview.html`.

United Nations (1948) The universal declaration of human rights. URL `http://www.un.org`.

United Nations (n.d.) United Nations millennium development goals. Tech. rep., United Nations.

Valderrama-Ardila, C., Alexander, N., Ferro, C., Cadena, H., Marn, D., Holford, T., Munstermann, L. and Ocampo, C. (2010) Environmental risk factors for the incidence of american cutaneous leishmaniasis in a sub-andean zone of colombia (chaparral, tolima). *American Journal of Tropical Medicine and Hygiene*, **82**, 243–250.

Van der Berg, S. and Moses, E. (2011) The remarkable improvement in functional literacy and numeracy in Namibia: A comparison between SACMEQ II and III. URL `http://www.sacmeq.org/education-namibia.htm`.

Vandegrift, K., Sokolow, S., Daszak, P. and Kilpatrick, A. (2010) Ecology of avian influenza viruses in a changing world. *Annals of the New York Academy of Sciences*, **1195**, 113–128.

Waller, L. and Gotway, C. (2004) *Applied Spatial Statistics for Public Health Data.* New York: John Wiley and Sons.

Wammanda, R., Gambo, M. and Abdulkadir, I. (2011) Age at BCG administration during routine immunization. *Journal of Community Medicine and Primary Health Care*, **16**, 33–35.

WHO (1978) Lay reporting of health information. Geneva: WHO.

WHO (1995) Verbal autopsies for maternal deaths. Geneva: WHO.

WHO (2008) *Global Networks for Surveillance of Rotavirus Gastroenteritis, 2001–2008.* Weekly Epidemiological Record, WHO, Switzerland.

WHO (2011) International statistical classification of diseases and related health problems 10th revision. Geneva: WHO.

WHO (2012) Wolrd health statistics. Tech. rep., WHO. URL `http://www.who.int/countries/afg/en/`.

WHO (2012a) Fact sheet on polio. Tech. rep., WHO. URL `http://www.who.int/mediacentre/factsheets/fs114/en/index.html`.

WHO (2012b) Fact sheet on measles. Tech. rep., WHO. URL `http://www.who.int/mediacentre/factsheets/fs286/en/index.html`.

Willaim, D. (1975) The analysis of binary responses from toxicology experiments involving reproduction and teratogenicity. *Biometrics*, **38**.

Wiysonge, C., Uthman, O., Ndumbe, P. and Hussey, G. (2012) Individual and contextual factors associated with low childhood immunisation coverage in sub-Saharan Africa: a multilevel analysis. *PLoSONE*, **7**.

Woldernicael, G. (2001) Diarrhoea morbidity among young children in Eritrea: Environmental and socio-economic determinants. *Journal of Health, Population, and Nutrition,*, **19**, 83–90.

World Organization for Animal Health (2012) World Animal Health Information Database (WAHID) Interface. URL `http://www.oie.int/animal-health-in-the-world/update-on-avian-influenza/2006/`.

Yan, J. (2002) geepack: Yet another package for generalized estimating equations. *R-News*, **2**, 12–14.

Yan, J. and Fine, J. P. (2004) Estimating equations for association structures. *Statistics in Medicine*, **23**, 859–880.

Yasui, Y. and Lele, S. (1997) A regression method for spatial disease rates: An estimating function approach. *Journal of the American Statistical Association*, **92**, 21–32.

Yohanne, A. G., Streatfield, K. and Bost, L. (1992) Child mortality patterns in Ethiopia. *Journal of Biosocial Sciences*, **24**, 145–155.

Zeger, S., Liang, K. and Albert, P. (1988) Models for longitudinal data: A generalized estimating equation approach. *Biometerics*, **44**, 1049–1060.

Zeger, S. L. (1988) A regression model for time series of counts. *Biometrika*, **75**, 621–629.

Zimmerman, D. (1989) Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *Journal of Statistical Computation and Simulation*, **32**, 1–15.