

https://dspace.org/

Modelling spatiotemporal patterns of childhood HIV/TB related mortality and malnutrition: applications to Agincourt data in rural South Africa

Musenge, Eustasius

2014-02-18
http://hdl.handle.net10539/13847

Downloaded from a DSpace repository. https://dspace.org/

MODELLING SPATIOTEMPORAL PATTERNS OF CHILDHOOD HIV/TB RELATED MORTALITY AND MALNUTRITION: APPLICATIONS TO AGINCOURT DATA IN RURAL SOUTH AFRICA.



By

EUSTASIUS MUSENGE

A Thesis Submitted to the Faculty of Health Sciences,

University of the Witwatersrand in fulfilment of

the requirements for the Degree of

DOCTOR OF PHILOSOPHY by publication in the

Major Area-Subject:

BIOSTATISTICS AND EPIDEMIOLOGY

March, 2013

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award. This thesis is the result of my own independent work, except where otherwise stated. Other sources are acknowledged by explicit references.



Signed **(candidate)** **Date: 19 March 2013**

University of the Witwatersrand, Faculty of Health Sciences,

School of Public Health

7 York road, Johannesburg

March, 2013

Dedication

This PhD thesis is dedicated to my wife Immaculata and my two sons Daniel and Victor as I say what my own dad (who died 12 June 2012) said to me "***You must put your name on the map!***" and the words of the Lord Jesus Christ in John 15 verse 5 ***"…without Me (Jesus) you can do nothing***"

Abstract

Background: South Africa accounts for more than a seventh of the global population living with HIV/AIDS and TB, and ranks highest in HIV/TB co-infection worldwide. Consequent high child mortality is exacerbated by child malnutrition, which is an important indicator of health status and is associated with morbidity as well as mortality. Rural areas usually present with the greatest burden of morbidity and mortality, yet the extent of geographical disparities in child mortality, malnutrition and HIV/TB has hardly been explored. This is a reservoir of information useful for effective public health interventions. In this thesis we investigated the factors associated with childhood HIV/TB mortality and malnutrition, how they interrelate and their spatial distribution in the rural Agincourt sub-district located in north-east South Africa close to the border with Mozambique.

Rationale: Africa at large lacks data that are routinely and reliably collected then validated, to guide policy and intervention programmes. Causes of deaths and even death counts are often misclassified and underestimated respectively, especially for children. To bridge this gap, a health and socio-demographic surveillance systems located in the rural Agincourt sub-district hosts which annually collects and collates data on vital events including fertility, mortality and migration. These data have been collected since 1992 to-date and now cover 80,000 people living in more than 16,000 households situated in 27 villages; all households are fully geo-coded. These hierarchical data allow us to address several epidemiological questions on how person, place (spatial) and time (temporality) have impacted on mortality and malnutrition patterns in children living in the rural Agincourt sub-district.

Objectives: The aims of this thesis were both methodological and applied:

Methodological

(1) To investigate the presence of spatial autocorrelation in the Agincourt sub-district and model this using geographical and geo-statistical procedures

(2) To model large spatial random effects accurately and efficiently

(3) To model hierarchical data with zero inflated outcomes

Applied

(1) To investigate childhood HIV/TB mortality determinants and their geographical distribution using retrospective and cross-sectional data

(2) To determine factors associated with malnutrition outcomes adjusting for their multivariate spatial random effects and selection bias for children under five years

(3) To model how the associated factors were interrelated as either underlying or proximate factors of child mortality or malnutrition using pathway analysis.

Methods: We conducted a secondary data analysis based on retrospective and cross-sectional data collected from 1992 to 2010 from the Agincourt sub-district in rural northeast South Africa. During the period of our study 71,057 children aged 0 to 9 years from 15,703 households were observed. All the data in the thesis were for children aged 1 to under 5 except for the chapter 6 (last paper) who were aged from 0 to 9 years of age. Child HIV/TB death and malnutrition were the outcome measures; mortality was derived from physician-based verbal autopsy. We investigated presence of spatial autocorrelation using Moran's and Geary's coefficients, semi-variograms and estimated the spatial parameters using Bayesian-based univariate and multivariate procedures. Regression modelling that adjusted for spatial random effects was done using linear regression and zero inflated variants for logistic, Poisson and Negative Binomial regression models. Structural equation models were used in

modelling the complex relationships between multiple exposures and child HIV/TB mortality and malnutrition portrayed by conceptual frameworks. Risk maps were drawn based on spatial residuals (posteriors) with prediction (kriging) procedures used to estimate for households where no data were observed. Statistical inference on parameter estimation was done using both the frequentist; maximum likelihood estimation and Bayesian; Markov Chain Monte Carlo (MCMC) directly and sometimes aided with Metropolis Hastings or Integrated Nested Laplace Approximations (INLA).

Results: The levels of child under-nutrition in this area were: 6.6% wasted, 17.3% stunted and 9.9% underweight. Moran's (I) and Geary's (c) coefficients indicated that there was global and local clustering respectively. Estimated severity of spatial variation using the partial-sill-to-sill ratio yielded 12.1%, 4.7% and 16.5%, for weight-for-age, height-for-age and weight-for-height Z-scores measures respectively. Maternal death had the greatest negative impact on child HIV/TB mortality. Other determinants included being a male child and belonging to a household that had experienced multiple deaths. A protective effect was found in households with better socio-economic status and where older children were present. Pathway analyses of these factors showed that HIV had a significant mediator effect and the greatest worsening effect on malnutrition after controlling for low birth-weight selection bias. Several spatial hot spots of mortality and malnutrition were observed, with these regions consistently emerging as areas of greater risk, which reinforces geographical differentials in these public health indicators.

Conclusion: Modelling that adjusts for spatial random effects, is a potentially useful technique to disclose hidden patterns. These geographical differences are often ignored in epidemiological regression modelling resulting in reporting of biased estimates. Proximate

and underlying determinants, notably socioeconomic status and maternal deaths, impacted directly and indirectly on child mortality and malnutrition. These factors are highly relevant locally and should be used to formulate interventions to reduce child mortality. Spatial prediction maps can guide policy on where to best target interventions. Child interventions can be more effective if there is a dual focus: treatment and care for those already HIV/TB infected, coupled with prevention in those geographical areas of greatest risk. Public health population-level interventions aimed at reducing child malnutrition are pivotal in lowering morbidity and mortality in remote areas.

Keywords: **HIV/TB, Child mortality, Child malnutrition, Conceptual framework, Spatial analysis, MCMC, Path analysis, South Africa**

Preface

This thesis is submitted to the University of the Witwatersrand, Faculty of Health Sciences for fulfilment of the requirements for the degree of philosophiae doctorate. Doctoral work has been performed at the School of Public Health, with Professor Kathleen Kahn as main supervisor and with co-supervisor Dr. Penelope Vounatsou.

Financial assistance was received from a doctoral fellowship awarded by the World Health Organization, Tropical Disease Research (WHO/TDR) from 2008 to 2010, and the Swiss South Africa Joint Research (Project Number: JRP IZLSZ3_122926) from 2009 to 2010. The MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt) is funded by the Wellcome Trust UK (Grant numbers 058893/Z/99A, 069683/Z/02/Z and 085477/Z/08/Z), the University of the Witwatersrand and Medical Research Council, South Africa, and the National Institute on Ageing (NIA) of the NIH, Andrew W Mellon and William and Flora Hewlett Foundations, U.S.A. Lastly I am thankful for a technical assistance grant from the South African Medical Research Council (MRC): Research Capacity Development; Sub-Directorate 2011.

Acknowledgements

I would like to give a special thank you to my lovely and caring wife Immaculata who started this journey with me, travelling with me to Basel, Switzerland and repeatedly remaining alone tolerating the messy paperwork and “always on the computer” home lifestyle.

Many thanks also go to Prof. Kathleen Kahn and Dr. Penelope Vounatsou for securing the funding for this work, sacrificial giving of time at the expense of quality time with their lovely families and for the entrusted inclusion of your “big” names with mine on publications.

Special thanks to all my other co-authors on the papers. I gained from your wealth of wisdom into becoming a better researcher: Prof Stephen Tollman, Dr. Mark Collinson, Dr. Elizabeth Kimani-Murage, Prof. Shane Norris and Prof. Tobias Freeman Chirwa.

My former and current heads of school Prof. Sharon Fonn and Prof. Laetitia Rispel you helped me negotiate sharp curves and got me going on breakdowns, a big thank you! I want to thank all the people who directly and indirectly helped me to this place, space does not allow me to cite you by name.

Lastly thanks be to God who always makes us to triumph and daily loads us with benefits for: the gift of life, safe travelling, sustenance in trying times and perpetual motivation and inspiration!.

My contributions in the thesis papers

My contributions as the first author in the thesis papers in chapters 2 to 6 were as follows:

- Chapter 2: Conceptualisation of research question, structuring of the paper, running the statistical analysis, writing of all drafts of the paper, addressing peer reviewers' and co-authors' comments.
- Chapter 3: Conceptualisation of research question, structuring of the paper, running the statistical analysis, writing of all drafts of the paper, addressing peer reviewers' and co-authors' comments.
- Chapter 4: Conceptualisation of research question, structuring of the paper, running the statistical analysis and writing of all drafts of the paper, attending to and addressing reviewer and co-author comments, and checking proofs for the final accepted manuscript.
- Chapter 5: Adopted and modified the concept from a joint Swiss-South African grant proposal, structuring of the paper, running the statistical analysis and writing of all drafts of the paper, attending to and addressing reviewer and co-authors comments and checking proofs for the final accepted and later published manuscript.
- Chapter 6: Conceptualisation of research question, structuring of the paper, running the statistical analysis and writing of all drafts of the paper, attending to and addressing reviewer and co-author comments, and checking proofs for the final accepted manuscript.

  I was also responsible for selection of the journals for submission on papers and was the corresponding author in all the articles.

Contents

Declaration ... **ii**

Dedication ... **iii**

Abstract ... **iv**

Preface ... **viii**

Acknowledgements ... **ix**

My contributions in the thesis papers ... **x**

Contents ... **xi**

List of Figures ... **xv**

List of Tables ... **xvii**

Abbreviations ... **xviii**

Chapter 1 ... **19**

1.0 Introduction *19*
1.1 Background *19*
1.1.1 Under five global, regional and local mortality patterns ... 19
1.1.2 Malnutrition and related mortality patterns ... 20
1.1.3 HIV/TB and related mortality patterns ... 21
1.1.4 Malnutrition, HIV and TB combined burden ... 21
1.2 Problem statement and rationale *22*
1.3 Review of literature *25*
1.3.1 Child HIV/TB mortality and malnutrition studies ... 25
1.3.2 Child mortality studies: sub-Saharan Africa Health and Socio-demographic Surveillance Systems ... 26
1.3.3 Spatial Epidemiology applications in Public Health research ... 27
1.3.4 Childhood HIV/TB and malnutrition mortality determinants ... 27
1.4 Thesis framework *29*
1.5 Aim and specific objectives *33*
1.6 Data, measurements and methods *34*
1.6.1 Agincourt health and socio-demographic surveillance system ... 34
1.6.2 Agincourt's HDSS: data collection and validity ... 36
1.6.2.1 Agincourt's hierarchical data structure ... 36
1.6.2.2 Agincourt resident status, vital events and modules ... 37
1.6.2.3 Verbal autopsy ... 40
1.6.2.4 Anthropometric survey and HIV testing ... 40
1.6.2.5 Agincourt's geographical information system ... 41
1.6.3 Ethical considerations ... 42
1.6.4 Spatiotemporal modelling of mortality and malnutrition ... 42

1.6.4.1 Hierarchical Bayesian spatial models 42
1.6.4.2 Zero Inflated spatiotemporal models and Bayesian inference 45
1.6.4.3 Modelling large geo-statistical random effects 46
1.6.4.4 Spatiotemporal model building and selection 47

Chapter 2 50

2.0 Estimation of multivariable and multivariate spatial variability for child adiposity measures in rural South African households 50
Chapter 2 abstract 51
2.1 Introduction 53
2.2 Materials and methods 56
2.2.1 Study area and ethical clearance 56
2.2.2 Measurement of three adiposity measures 57
2.2.3 Spatial autocorrelation measures 58
2.2.4 Semi-variogram spatial parameter estimations 60
2.2.5 Single outcome spatial random effects Bayesian models 63
2.2.6 Multivariate spatial random effects Bayesian model 65
2.2.7 Software for geo-statistical analyses 66
2.3. Results 66
2.3.1 Simulated data results 67
2.3.2 Spatial autocorrelation analysis results 72
2.3.3 Descriptive statistics for Agincourt child adiposity outcomes and malnutrition explanatory variables 72
2.3.4 Spatial multiple and multivariate results for three adiposity measures. 73
2.4 Discussion 79
2.5 Conclusion 80
2.6 Authors' contributions 81
2.7 Acknowledgements 82
2.8 Competing interests 82

Chapter 3 83

3.0 Spatially adjusted determinants of childhood malnutrition in rural north-east South African households: a multivariate and structural equation modelling approach 83
Chapter three abstract 84
3.1 Introduction 85
3.2 Materials and methods 86
3.2.1 Study site and anthropometric survey description 86
3.2.2 Outcomes and Covariates 88
3.2.3 Multivariate geo-statistical Bayesian model 88
3.2.4 Structural equation modelling 90
3.2.5 Bayesian inference and statistical software used 90
3.3 Results 91
3.3.1 Descriptive characteristics of the survey sample 91
3.3.2 Multivariate spatially adjusted determinants of childhood malnutrition and pathway analyses results 93
3.3.3 Spatial distribution of malnutrition outcomes 96
3.4 Discussion 97
3.5 Acknowledgements 99

Chapter 4 ..... **100**

4.0 The contribution of spatial analysis to understanding and responding to HIV/TB mortality in children: a structural equation modelling approach. *100*
Chapter 4 abstract *101*
4.1 Introduction *103*
4.2 Methods *105*
4.2.1 Agincourt sub-district data and study design 105
4.2.2 Dependent and independent variables 107
4.2.3 Conceptual framework: determinants of child mortality 107
4.2.4 Statistical methodology 109
4.2.5 Ethical clearance and informed consent 110
4.3 Results *111*
4.4 Discussion *117*
4.4.1 Strengths and limitations of the study 119
4.5 Conclusion *121*
4.6 Acknowledgements *121*
4.7 Appendix A: Structural Equation Model for paths analyses *122*

Chapter 5 ..... **124**

5.0 Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spatial and Spatiotemporal Epidemiology 2011; 2, 205-217. *124*
Chapter 5 abstract *125*
5.1 Introduction *126*
5.2 Methods *128*
5.2.1 Agincourt Health and Socio-demographic surveillance system (HDSS) Data 128
5.2.2 Outcome and explanatory variables 130
5.2.3 Spatiotemporal zero inflated Poisson and Negative Binomial models 131
5.2.4 Goodness of fit and best model selection 133
5.2.5 Ethical clearance and informed consent 134
5.3 Results *135*
5.3.1 Exploratory analysis and descriptive statistics 135
5.3.2 Spatiotemporal confounding adjusted determinants of child HIV/TB mortality 137
5.4 Discussion and Conclusions *146*
5.5 Acknowledgements *150*
5.6 Authors' contributions *150*
5.7 Competing interests *150*
5.8 Appendix: Zero Inflated Poisson and Negative Binomial Hierarchical Bayesian models *150*

Chapter 6 ..... **154**

6.0 Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural Northeast South Africa *154*
Chapter 6 abstract *155*
6.1 Introduction *156*
6. 2 Methods *158*
6.2.1 Spatiotemporal model structure 158

6.2.2 Zero Inflated Poisson and Binomial spatiotemporal models 159
6.2.3 Solving the "big m" using the SPDEs to estimate the spatial random effects .. 160
6.2.4 Bayesian inference using INLA 162
6.2.5 Model goodness of fit and convergence diagnostics 164
6.3 Application *165*
6.3.1 Agincourt health and socio-demographic surveillance data (HDSS), rural South Africa, study design and ethics 165
6.3.2 Dependent and independent variables 166
6.3.3 Descriptive statistics 166
6.3.4 Zero Inflated Poisson and Binomial spatiotemporal modelling results 167
6.3.5 Model assumptions, goodness of fit and convergence diagnostics 172
6.4 Conclusions *176*
6.5 Acknowledgements *177*
6.6 Authors' contributions *177*
6.7 Competing interests *177*
6.8 Appendix: Zero inflated Poisson R-code using INLA *178*

Chapter 7 181

7.0 Discussion and conclusions *181*
7.1 Determinants of childhood morbidity: adiposity and malnutrition *182*
7.2 Direct risk factors of child HIV/TB mortality in observational cohorts *185*
7.3 Spatiotemporal modelling of mortality and morbidity in resource scare settings *188*
7.4 Strengths and limitations of the study *190*
7.4.1 Variables recall and selection biases 190
7.4.2 Comparability of findings 191
7.4.3 Confounding and effect modification 191
7.4.4 Analysis and computing 192
7.5 Conclusion *193*
7.5.1 Agincourt's spatiotemporal child malnutrition and mortality hotspots 193
7.5.2 Implications for interventions and policy 193
7.5.3 Scope for future research 194

8.0 References 196

9.0 Appendices: Published and in-press papers and ethics certificate 214

List of Figures

Figure 1.1: Global under five mortality in the year 2010. 20
Figure 1.2: Percentage of global population undernourished. 21
Figure 1.3: Percentage decline in yearly mortality for children under five between 1990 and 2010. Source: The Lancet Volume 375, Issue 9730, Pages 1988-2008 (May 2010) (7). 24
Figure 1.4: Conceptual framework for modelling spatiotemporal patterns of childhood HIV/TB related mortality and malnutrition: adapted from Meade (1977) (67). 28
Figure 1.5: Framework illustrating thesis structure and methods used: adapted from Bjørnson's PhD thesis (68). 32
Figure 1.6: Agincourt Health and socio-Demographic Surveillance system: study site location. 35
Figure 1.7: Agincourt sub-district hierarchical data structure. 37
Figure 2.1: Agincourt sub-district and surveillance area showing households and primary health care facilities. 57
Figure 2.2: Distribution of all pair-wise distances (left) and a semi-variogram diagram (right). 61
Figure 2.3: Box plots of 95% credible intervals and point estimates for simulated data. 68
Figure 2.4: Graphical semi-variogram plots for classical and robust semivariance estimates for simulated data. 69
Figure 2.5: Diagnostic (autocorrelation and density) plots for simulated data parameters. 71
Figure 2.6: Observed (left) and spatial multivariate posterior (right) maps and contour for body mass indices for children in Agincourt 2007. 75
Figure 2.7: Observed (left) and spatial multivariate posterior (right) maps and contour for waist circumferences for children in Agincourt 2007 77
Figure 2.8: Observed (left) and spatial multivariate posterior (right) maps and contour for body fat masses for children in Agincourt 2007. 79
Figure 3.1: Flow diagram of anthropometric survey participants, Agincourt sub-district 2007. 87
Figure 3.2: Spatial distribution of underweight, stunting and HIV positive children in Agincourt rural South Africa. 93
Figure 3.3: Pathway analyses structure of malnutrition outcomes, HIV and maternal factors 95
Figure 3.4: Posterior smoothed maps of malnutrition outcomes in Agincourt, rural South Africa 97
Figure 4.1: Agincourt sub-district in 2004 showing villages (dark grey shading) schools (flags) and clinics and health facilities (circled white and black crosses). 106
Figure 4.2: Conceptual framework for modelling childhood HIV/TB related mortality: adapted from Meade (1977) (67). 108
Figure 4.3: Child HIV/TB mortality: Posterior adjusted Odds Ratio map, Agincourt sub-district, 2004. 113

Figure 4.4: Conceptual framework pathways results for Agincourt sub-district child mortality determinants in 2004 115
Figure 5.1: Agincourt HDSS location, north-east South Africa. Source Kahn et al. (2000) 129
Figure 5.2: Kaplan-Meier (KM) survival function over the years 2000 to 2005 (left) and year specific Kaplan-Meier (KM) curves (right). 136
Figure 5.3: Knots fitted using the Mini-Max space filling criteria from Agincourt households 2000 to 2005. 137
Figure 5.4: Non-spatial ZINB with AR(1) baseline hazard (left) and multivariable spatial ZINB with AR(1) baseline hazard (right). 138
Figure 5.5: Posterior adjusted Hazard rates (AHRs) smoothed map (left) and contour AHRs map (right) for child HIV/TB mortality 2000 to 2005. 146
Figure 6.1: Agincourt sub-district: original household locations (left), triangulation of all households (centre) and triangulation of households within 500m apart (right). 161
Figure 6.2: Hierarchical structure of a zero inflated spatiotemporal model fit using INLA 163
Figure 6.3: Year specific child deaths due to HIV/TB from 1992 to 2010 167
Figure 6.4: Posterior point estimates for ZIP (top) and ZIB (bottom) models 171
Figure 6.5: Posterior mean and 50% (middle), 0.025% (lower) and 0.975% (upper) .. 173
Figure 6.6: Posterior density plots of log-means for fixed effects of the ZIP spatiotemporal model. 174
Figure 6.7: Posterior density plots of log-means for fixed effects of the ZIB spatiotemporal model. 175
Figure 7.1: Vicious cycle depicting inter-linkage of HIV, malnutrition, maternal and geographical locations: adapted from Piwoz and Prebel (2000). 184

List of Tables

Table 1.1: Thesis themes and inter-linkages with papers 30
Table 1.2: Thesis papers titles, status, comparative features and summary of key results. 31
Table 1.3: Agincourt HDSS household or individual modules: objects, events and episodes from 1992 to 2012. Source (66) 39
Table 2.1: Graphical, theoretical and Bayesian derived spatial parameter estimates for simulated data 67
Table 2.2: Graphical, theoretical and Bayesian derived spatial parameter estimates for simulated data 72
Table 2.3: Body mass indices descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results 74
Table 2.4: Waist circumferences descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results 76
Table 2.5: Body fat mass descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results 78
Table 3.1: Bayesian spatial multivariate regression results for malnutrition weight-for-age and height-for-age Z-score outcomes. 92
Table 3.2: Path analyses structural estimation results for variables in figure 3.3 94
Table 3.3: Direct and indirect effects on malnutrition outcomes for children under five from Agincourt sub-district, rural south Africa. 96
Table 4.1: Child HIV/TB mortality: descriptive statistics and results of two multiple logit regression analyses, Agincourt sub-district 2004 112
Table 4.2: Direct, indirect and total effects of child HIV/TB mortality in 2004, Agincourt sub-district 116
Table 5.1: Model goodness of fit non-spatial/spatial zero inflated Poisson and Negative Binomial models 139
Table 5.2: Non spatial and Spatial Multivariable Zero Inflated Poisson model 140
Table 5.3:Non spatial and Spatial Multivariable Zero Inflated Negative Binomial model 143
Table 6.1: Univariate and multiple regression results models using zero inflated Poisson adjusting for spatiotemporal random effects. 168
Table 6.2 : Multiple regression results of four models using zero inflated Binomial adjusting for spatiotemporal random effects 169

Abbreviations

AHR	Adjusted Hazard ratio
AIDS	Acquired Immune Deficiency Syndrome
AOR	Adjusted Odds ratio
ARR	Adjusted Relative Risk
ART	Anti-retroviral therapy
BCI	Bayesian Credible Interval
BMI	Body Mass Index
CI	Confidence Interval
DIC	Deviance Information Criteria
GF	Gaussian Field
GIS	Geographic information System
GMRF	Gaussian Markov Random Field
HAZ	Height-for-Age-Z-score
HDSS	Health and Socio-Demographic Surveillance System
HIV	Human Immunodeficiency Virus
HR	Hazard ratio
ICD10	International Classification of Diseases version 10
INDEPTH	International Network for Demographic Evaluation of Populations and Their Health
INLA	Integrated Nested Laplace Approximations
Km	Kilometres
m	metres
MCMC	Markov Chain Monte Carlo
MDGs	Millennium Development Goals
MH	Metropolis Hastings
MLE	Maximum Likelihood Estimation
MTCT	Mother to child transmission
OR	Odds ratio
PMTCT	Prevention of mother to child transmission
PY	Person Years
RR	Relative risk
SEM	Structural equation modelling
SES	Social Economic Status
SPDE	Stochastic Partial Differential Equations
SQL	Structured Query Language
SSA	sub-Saharan Africa
TB	Tuberculosis
UN	United Nations
VA	Verbal Autopsy
WAZ	Weight-for-Age-Z-score
WHO	World Health Organisation
WHZ	Weight-for-Height-Z-score

Chapter 1

1.0 Introduction

This chapter introduces the thesis by providing a background, rationale, literature review, conceptual framework, thesis framework, specific objectives and methods used in the study. Chapters 2 to 6 are the published and submitted papers. In Chapter 7 I discuss the work according to three themes, relate my methods and findings to other studies, discuss study limitations and provide concluding remarks including the scope for further studies.

1.1 Background

1.1.1 Under five global, regional and local mortality patterns

Child mortality, defined as the death of children under five years of age is an essential indicator of the wellbeing and future of the human race. Figure 1.1 shows the global distribution of under mortality divided into 10 categories. The United Nations (UN) estimated the global population in 2012 to be about 7.0 billion of which 12.5% (0.88 billion) would live in sub-Saharan Africa (SSA) (1). Yet she experiences more than 50% of yearly deaths under five which amounted to 12.4 million deaths in 1990 dropping to 7.4 million in 2010. At the onset of the 21st century, about 10 million children were dying every year in the world, with malnutrition a major underlying cause, and with pneumonia, diarrhoea, neonatal disorders, malaria and AIDS the main causes (2). African children experience the greatest burden of mortality with at least one child less than 5 years dying every 3.6 seconds due to starvation (3). Studies on the causes of these deaths indicated that about 67% of these could be avoided if the main drivers were either dealt with or controlled (4). There are several factors which lead to child mortality, amongst which malnutrition and hygiene-related morbidity accounts for about 50% of the deaths, as well as acquired immunodeficiency syndrome (AIDS) and it's

opportunistic infections like tuberculosis (TB) (2, 5-7). South Africa had an under five mortality of 51 deaths per 1000 in 2010 compared to 58 deaths per 1000 in 1990 (7).

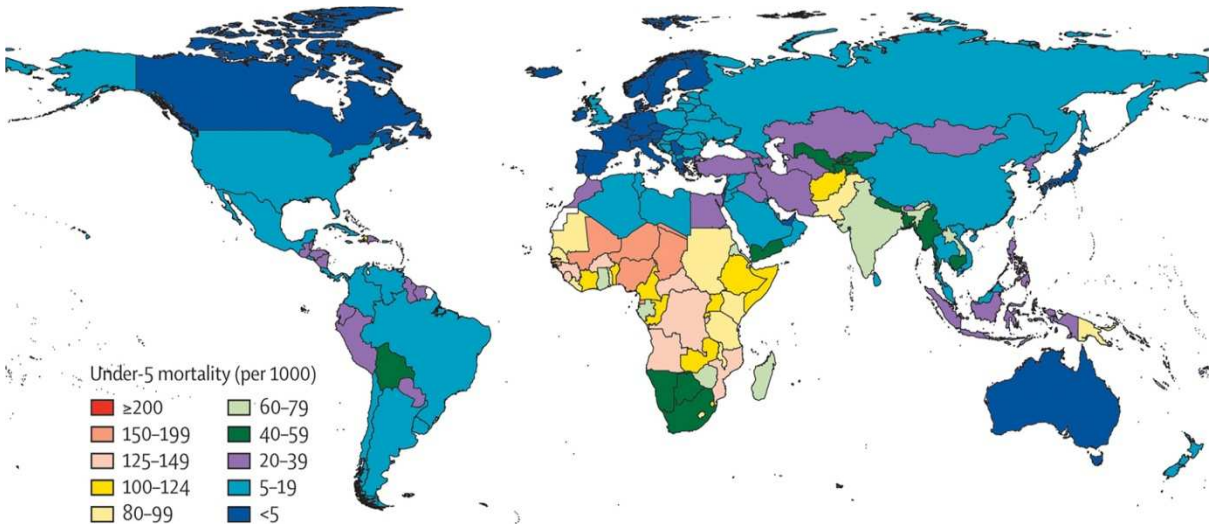


Figure 1.1: Global under five mortality in the year 2010.

Source: The Lancet Volume 375, Issue 9730, Pages 1988-2008 (May 2010) (7).

1.1.2 Malnutrition and related mortality patterns

Childhood malnutrition is a useful proxy for child wellbeing as it associated with mortality, morbidity and child development (8). Malnutrition accounted for more than half of the 10 million child deaths that occurred worldwide yearly between 1990 and 2000 with over 70,000 in South Africa alone (2). Global under-nutrition patterns show similarities with mortality patterns in the sub-Saharan African region as can be seen when comparing Figure 1.1 and 1.2. South Africa's under five malnutrition rates in 2003 were 24.5%, 17.0% and 8.9% for stunting, underweight and wasting respectively (9).

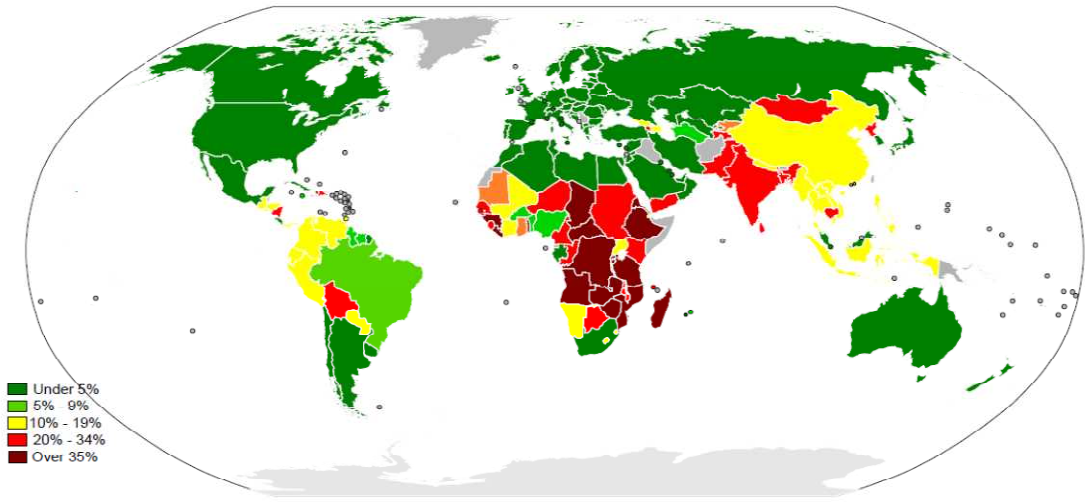


Figure 1.2: Percentage of global population undernourished.

Source: United Nations World Food Program's 2004-06 report under-nutrition percentages

1.1.3 HIV/TB and related mortality patterns

About 88% of children infected with the human immunodeficiency virus (HIV) and 92% of AIDS-related deaths in 2010 occurred in Africa (4). South Africa has the highest absolute number of HIV infected persons in the world (5.5 million) with national and antenatal attendees prevalences of about 18% and 26% respectively in 2010 (4). TB is a growing problem in developing countries, including South Africa. Africa has the highest incidence of TB at 345 per 100,000 person years, with TB being a major cause of death among people living with HIV/AIDS in sub-Saharan Africa (4). South Africa is at present faced with the challenge of implementing more effective TB diagnosis and treatment strategies to curb the incidence of TB which is above 200 per 100,000; according to WHO standards, this is a serious epidemic (4).

1.1.4 Malnutrition, HIV and TB combined burden

HIV and malnutrition are major problems among children in the sub-Saharan African region. The impact of HIV/AIDS on child malnutrition and vice versa is of major concern to the

public health of developing nations. Malnutrition affects the human immune system leading to depletion of disease fighting cells (CD4, T-cells) and hence increasing the viral load for those infected with HIV. Malnutrition also leads to a reduction in essential macronutrients (proteins, fats and carbohydrates) and micronutrients (vitamins A, B, C and zinc) (10-15). HIV on the other hand affects nutrition by loss of appetite, drug side effects such as vomiting and diarrhoea, and nutrient mal-absorption (14, 16-19). This situation is compounded by HIV and TB co-infection and evidence shows that it is difficult to simultaneously manage the treatment of both diseases in malnourished children (20).

1.2 Problem statement and rationale

TB and HIV/AIDS are among the top public health research priorities especially in sub-Saharan Africa. The sub-Saharan African region which constitutes 10% of the world's population is home to more than 75% (25 million) of people living with HIV/AIDS in 2010 (21). South Africa has 6.4% (45 million) of the sub-Saharan African population yet accounts for 22.4% (5.6 million of the 25 million) of people living with HIV/TB in 2010 in sub-Saharan Africa, thus ranking amongst the countries with highest infection rates (21-23). South African statistics on the HIV/TB and malnutrition burden in children under 5 years are scanty, and when available they are often combined with older children, e.g. the prevalence of HIV was about 3% for 2-14 year olds (7). At the beginning of the 21st century over 70,000 children died yearly in South Africa with malnutrition accounting for more than 50% of these deaths (2). HIV accounted for over 40% of child deaths in South Africa in the year 2000, with many of them dying without knowing their status (24).

In public health, identification and quantification of patterns of disease occurrence provide the first steps towards increased understanding towards control of the disease (25).

Public health policy-makers rely on data for allocation of resources and interventions. The African continent at large lacks the facilities as well as the expertise to collect validated, regular and reliable data, as well as the ability to analyse such data. Therefore, lack of reliable evidence to guide the selection of interventions for specific areas and populations is a major drawback for effective control of the diseases and epidemics. More so, ground breaking and successful modelling can only be done based on good data. For this purpose a health and socio-demographic surveillance system (HDSS) was set up in 1992 in the rural Agincourt sub-district of Mpumalanga province, South Africa (26). These demographic surveillance system data, updated once every year, embody the strengths of multilevel (hierarchical) data structures. In addition all households are geo-coded which allows the assessment of geographical differentials.

The need for thorough modelling procedures revealing the relation between person, place (spatial) and time, and communicable as well as non-communicable diseases, is a top public health priority. However, there are analytic challenges, namely the multi-collinearity due to the multilevel structure, spatial auto-correlation and large size of these data. Geo-mapped data showing the geographical regions, associated risk factors as well as the burden of disease, can be more convincing tools to present to policy makers than text results models alone.

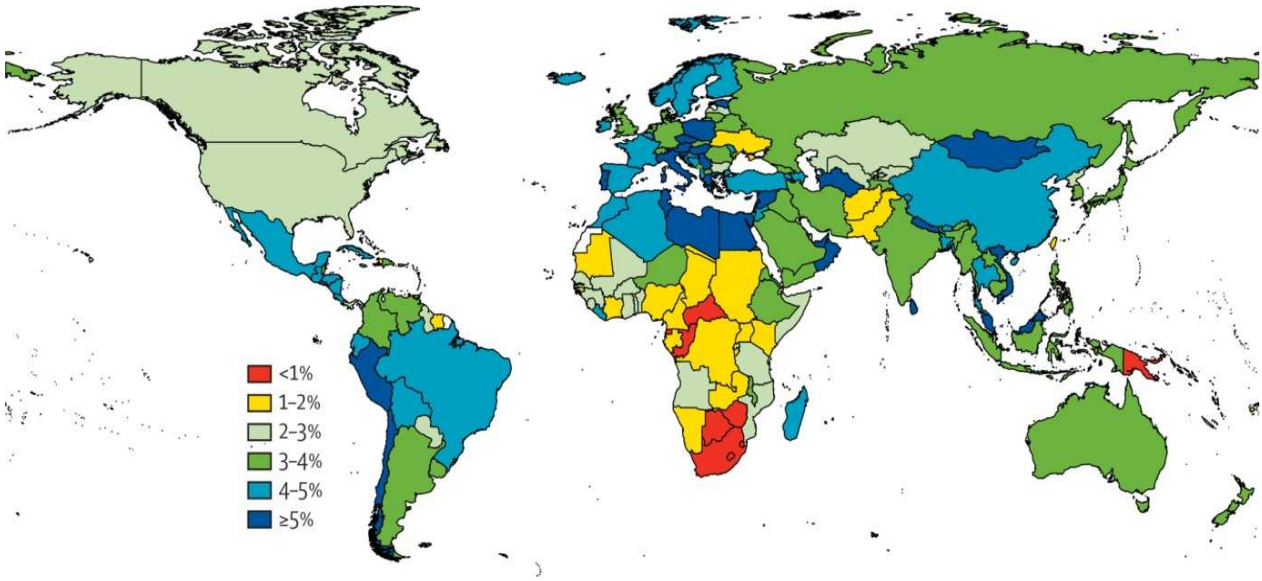


Figure 1.3: Percentage decline in yearly mortality for children under five between 1990 and 2010. Source: The Lancet Volume 375, Issue 9730, Pages 1988-2008 (May 2010) (7).

The public health burden of child malnutrition and HIV/TB mortality in South Africa is critical since mortality has decreased yearly by less than 1% over the period 1990 to 2010 for the country as a whole (figure 1.3) (7). The underlying cause of death is usually malnutrition, as there is a strong two-way interaction between disease and nutritional status (27). Often malnourished children suffer from poorly managed infectious diseases such as HIV/TB, which creates a vicious cycle between communicable and non-communicable diseases (28, 29).

The impact of HIV/AIDS on child malnutrition and vice versa is of concern in the public health of developing nations. Knowledge of inter-linkages between TB and HIV on the child's nutritional status may help focus interventions in different geographic regions of the globe. There is evidence to indicate geographical heterogeneity in HIV/TB mortality and malnutrition both between and within South African provinces, which has resulted in stigmatisation of some high prevalence regions. However, many studies do not cater for this

important independent factor in the analysis (30). I will use spatiotemporal Bayesian modelling techniques in analysing large spatial univariate and multivariate hierarchical data from the Agincourt sub-district. This work can contribute to policy and programme development in rural parts of developing countries globally before and after the end of the 2015 millennium development goals targets.

1.3 Review of literature

1.3.1 Child HIV/TB mortality and malnutrition studies

Scientific evidence gathered from South Africa can be valuable for public health policy and interventions in the country especially if drawn from reliably collected, validated and well analysed data. To date global research on, HIV/TB and malnutrition research has focused largely on drug uptake and adherence, drug- development clinical through clinical trials, societal integration, personal nutrition, physical health and care giving (31-34). Whilst numerous studies have been done in South Africa on childhood HIV/TB mortality and morbidity risk factors (35-42), and on malnutrition (43-47), these two are seldom studied together (48). Moreover an increasing number of studies now incorporate the spatial component descriptively (49, 50) without adjusting for risk factors in regression. Studies that have included the spatial correlation in the regression analysis have focused mainly on village level or cluster analysis (49, 51). In South Africa few studies have been conducted spatiotemporal differentials on childhood HIV/TB and malnutrition cause-specific mortality differentials. Our studies are amongst the first on household location spatial analysis in South Africa, dealing with malnutrition and mortality.

Many epidemiological studies have addressed the person aspect, but little has been done in relation to time and place (spatial or geographical proximity) and even fewer link all three. Therefore, a knowledge gap exists on how to spatially link mortality and disease impacts on rural childhood populations.

1.3.2 Child mortality studies: sub-Saharan Africa Health and Socio-demographic Surveillance Systems

Many articles have been published on HIV/TB mortality in sub-Saharan Africa, but only a few of these report on children. Subjects discussed were mainly on the impact of epidemics, assessing risk factors, measuring efficacy of interventions and identifying high risk clusters. A study done in Niakhar, rural Senegal showed a decreasing trend in child mortality over a 37 year period (52). Spatial analyses on 39 rural villages in Burkina Faso under the Health and Demographic Surveillance System (HDSS) over a nine year period, showed that village and distance from the health facility were among the risk factors for child mortality (53). A seasonal mortality trend was also found in another study in the same HDSS (54). A Tanzanian study that used spatial modelling techniques found that there is a need to adopt both a group-based and a place-based approach in the identification of high-risk HIV groups for interventions. (55). In South Africa's rural KwaZulu-Natal a study found that the prevalence of HIV was associated with ethnicity, urban status, and unemployment based on over 11,000 persons in 700 enumeration areas (56). In the Agincourt sub-district of Mpumalanga province the relative risk (RR) of HIV and TB has increased ten-fold over an eleven year period from 3.2 to 32.3 for the period 1992-1994 to 2001-2003 (57). Methods applied in the analyses were mostly non-Bayesian based, which are limited in catering for large datasets and spatiotemporal random effects (58).

1.3.3 Spatial Epidemiology applications in Public Health research

The discipline of epidemiology deals with the interrelatedness between the person, place and time in the distribution of diseases in the population (59). In the health field, interest in spatial epidemiology began with the recognition of maps as useful tools for understanding potential causes of disease and areas of high risk. Throughout the ages there are examples of diseases that have been eradicated after their spatial patterns were discovered. One classical example of spatial epidemiology is Dr. John Snow's study of London's cholera epidemic in 1854 (60). Snow's belief that cholera was transmitted through drinking water was refuted by some. Brody et al. (61) gives a detailed history of how Snow and others demonstrated the role of maps in the investigation of the 1854 outbreak. Other examples of spatial epidemiology include the study of rickets by Palm in 1890 (62), and Blum in 1948 (63) who showed that sunlight was a causal factor for skin cancer. More recent studies have used spatial Bayesian techniques in malaria mortality, all-cause and cause specific mortality (64-66)

1.3.4 Childhood HIV/TB and malnutrition mortality determinants

Child mortality determinants have been classified into different groupings over the last three decades. In the 1980s researchers explored several interdisciplinary causes of child mortality: medical, socio-economic and demographic (32). In the 1990s these were: fertility behaviour; nutritional status, breastfeeding, and infant feeding; the use of health services by mothers and for children; environmental health conditions; and socioeconomic status (34). And finally in the 21st century these changed to: HIV exposure; poor maternal health; inadequate infant care; increased exposure to infections; parental and caregiver mortality; immune system abnormalities; poor nutrition; reduced breastfeeding; and antiretroviral exposure (33). Rutstein et al. (2000) mentions six main risk factors for child mortality: socio-economic factors; nutritional factors; use of health facilities; environmental health conditions; fertility

behaviour; and infant feeding (34). The explanatory variables for our papers will be selected from these six child specific risk factor categories.

As a way of inter-relating these factors we can formulate a conceptual framework. I adopted a disease ecology model proposed by Meade (1977) (67), looking at individual, population and environmental factors. The multilevel (hierarchical) data structure is considered as the individual factors belong to the population level factors; these are all in turn encompassed or impacted by the environmental factors, which are all inter-related. I combined this with the classical epidemiological causal effects models adjusting for spatial and temporal random effects which were among the exposure variables. Our mediating variable was HIV and the outcomes were HIV/TB mortality or related morbidity such as malnutrition. Figure 1.4 shows how I inter-linked these factors in conceptualising this thesis.

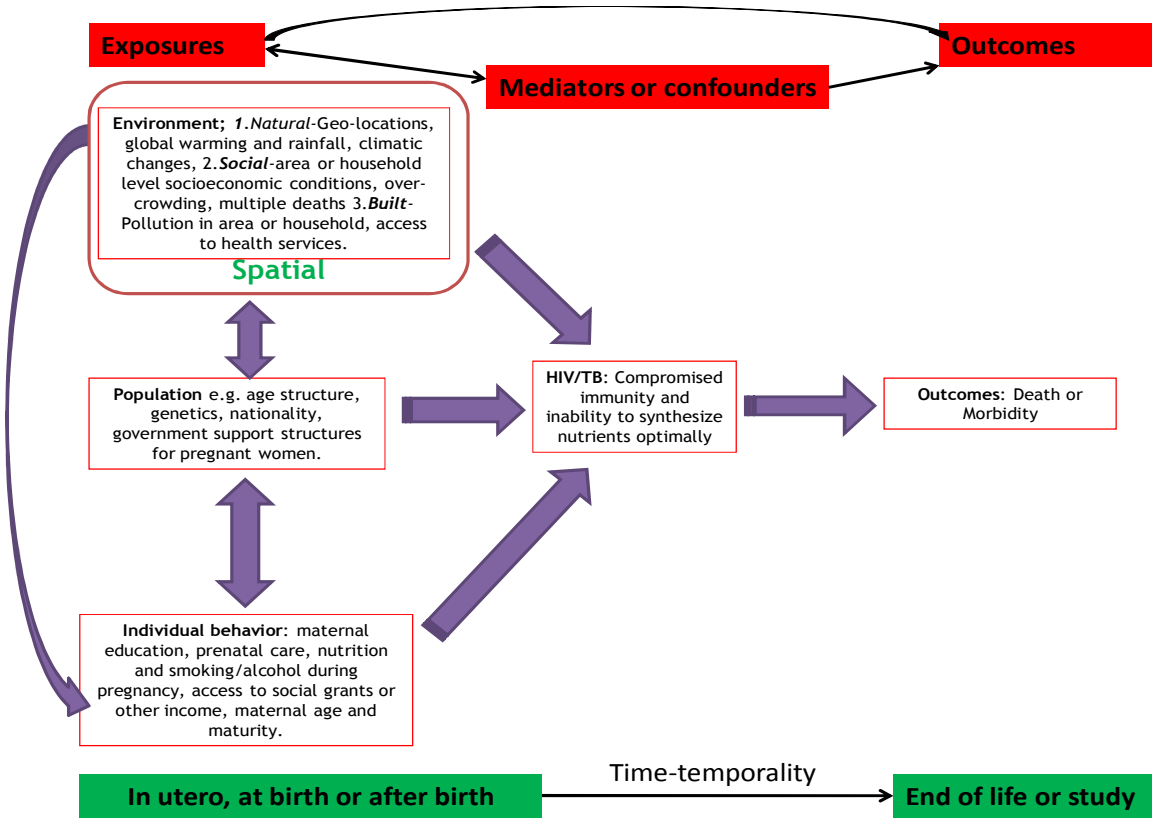


Figure 1.4: Conceptual framework for modelling spatiotemporal patterns of childhood HIV/TB related mortality and malnutrition: adapted from Meade (1977) (67).

1.4 Thesis framework

Based on the conceptual framework, three themes that concern public health policy and interventions have been extracted:

- Determinants of childhood morbidity: adiposity and malnutrition.
- Direct risk factors of child HIV/TB mortality in observational cohorts
- Spatiotemporal modelling of mortality and morbidity in resource scare settings.

Tables 1.1, 1.2 and Figure 1.5 detail how the themes link up with the papers and the fields of Epidemiology and Biostatistics.

Table 1.1: Thesis themes and inter-linkages with papers

Themes	Paper 1: Estimating spatial variability in adiposity	Paper 2: Impact of HIV on malnutrition	Paper 3: Child HIV/TB spatial mortality patterns	Paper 4: Child HIV/TB spatiotemporal mortality patterns	Paper 5: Modelling large spatiotemporal child mortality data
1. Determinants of childhood morbidity: adiposity and malnutrition.	✓	✓ ✓			✓
2. Direct risk factors of child HIV/TB mortality in observational cohorts	✓	✓	✓ ✓	✓ ✓	✓
3. Spatiotemporal modelling of mortality and morbidity in resource scare settings.	✓ ✓	✓	✓	✓ ✓	✓ ✓

Table 1.2: Thesis papers titles, status, comparative features and summary of key results.

Paper	*Title*, type and status	Key comparative features of papers: Design, sample-size, number of years, household locations, outcome and explanatory variables	Analyses methods used and duration of computing	Key result findings
1	-*Estimation of multivariable and multivariate spatial variability for childhood adiposity measures in rural South African households.* - Methodology, **re-submitted to the Spatial Statistics Journal.**	Cross-sectional, 648 children from 648 locations, three obesity indicators: body mass index, waist circumference and percentage body fat..	Semi-variogram analysis, Moran's and Geary's spatial-auto-correlations, Univariate and Multivariate spatial random effects models. Inference done using maximum likelihood estimation (MLE) and Bayesian based Markov chain Monte Carlo (MCMC).One week computing duration for multivariate spatial models.	The multivariable and multivariate analysis showed spatial variation. The multivariable models showed greater variability than the multivariate models, with smaller credible interval estimation bands. Obesity is positively correlated with malnutrition and inversely with poverty.
2	-*Spatially adjusted determinants of childhood malnutrition in rural north-east South African households: a multivariate and structural equation modelling approach.* -Application, **re-submitted to the American Journal of Clinical Nutrition.**	Cross-sectional, 648 children from 648 locations, two malnutrition z-score outcomes with HIV as one of the covariates and mediating variable.	Spatial Multivariate (Bivariate Normal) and structural equation modelling. Inference done using Bayesian MCMC for spatial models (took 3 days computing duration) and MLE for structural equation models.	Spatial multivariate adjusted analyses revealed a negative mean effect due to HIV and protective effects with older aged children and mothers, and maternal education. Pathway analysis of these variables showed that HIV had a significant worsening mediator effect: -0.55 (95% CI: -1.09,-0.02) and -0.61 (95 % CI: -0.98,-0.24) for weight-for-age and height-for-age z-scores respectively.
3	-*The contribution of spatial analysis to understanding HIV/TB mortality in children: a structural equation modelling approach.* Application, **in press in the journal Global Health Action**	Cross-sectional, 6,692 children aged 1-5 years from 5,084 households in 2004, outcome was HIV/TB mortality	A spatial random effects model was done using logistic regression model, catering for individual as well as household random effects. 4 hours computing duration. The determinants were further modelled using structural equation models base on conceptual framework.	A conceptual framework based model of determinants of child HIV/TB mortality. A protective effect on child HIV/TB mortality was found with households that; were male headed, were least poor and had many inhabitants.
4	-*Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa.* -Review and Application, **published in Spatial and Spatiotemporal Epidemiology 2011; 2, 205-217.**	Observational retrospective cohort, 16,844 children aged 1-5, 6 years cumulative follow-up years, 8,863 locations, and outcomes were child HIV/TB mortality with covariates.	Zero inflated Poisson and Negative Binomial models with discrete temporal effects and spatial random effects estimated via lowrank kriging. Inference was done using Bayesian MCMC which took 5 days computing duration.	Mother's death increased the risk of mortality almost threefold [adjusted hazard ratio=2.93, 95 % CI (1.29;6.93)] controlling for individual, household and space-time confounders.
5	-*Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa.* - Methodology, **in-press in the Journal of Applied Earth Observation and Geoinformation**	Observation retrospective cohort, 71,057 children aged 0 -10, 19 years cumulative follow-up years, 15,703 locations, outcomes were child HIV/TB mortality with covariates	Zero inflated Binomial and Poisson models with first order autoregressive temporal effects and spatial random effects. Inference done using integrated nested Laplace approximations (INLA) and 12 hours computing duration.	Greater birth-weight and age had protective effects on HIV/TB mortality with adjusted relative risk=0.73 (0.53; 0.99) and 0.18 (0.14; 0.22) respectively. We employed the triangulation technique to convert a Gaussian field to a Gaussian markov random field, in resolving the numerous household locations. This paper links with paper 3 and handles even larger datasets.

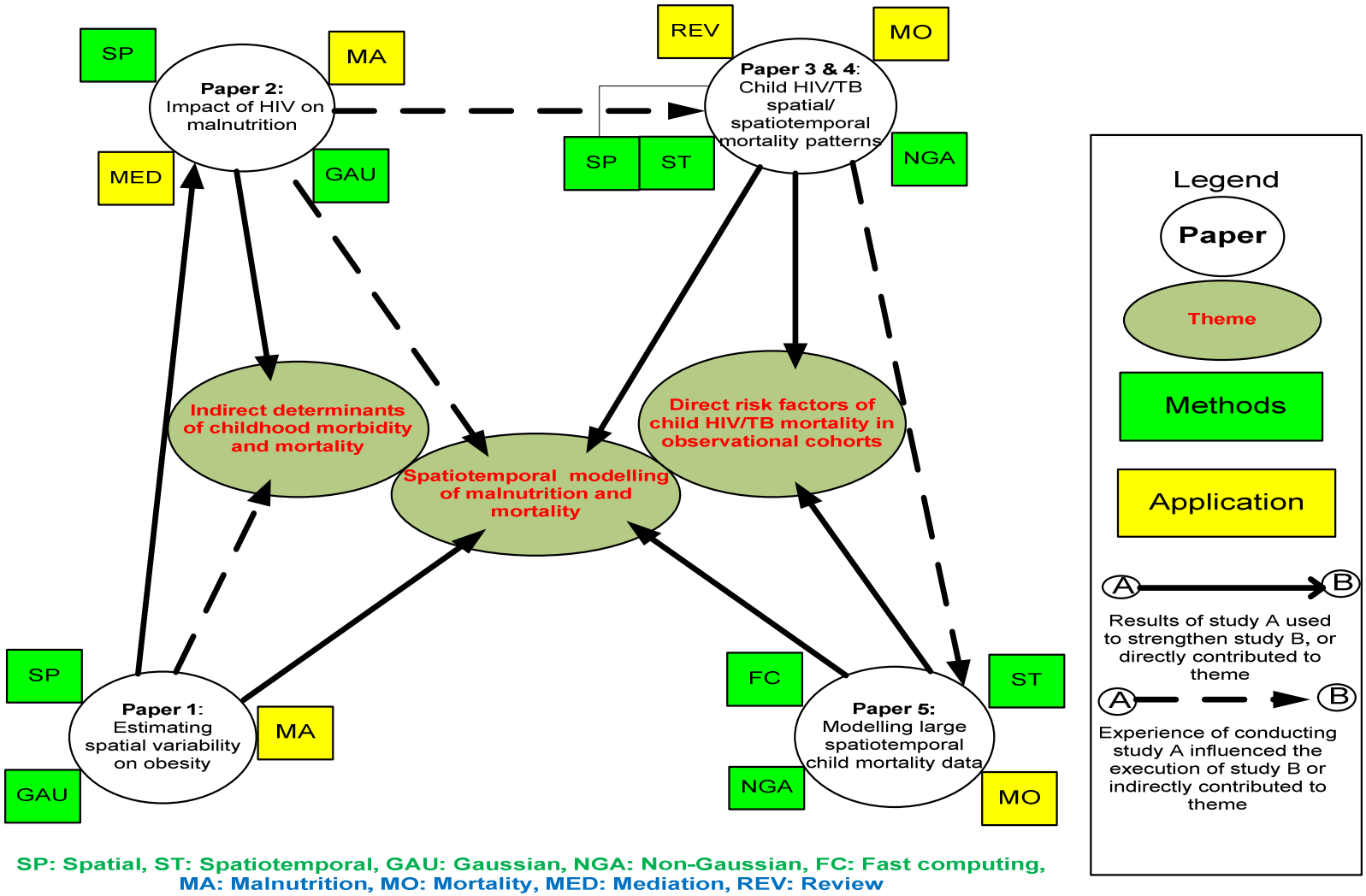


Figure 1.5: Framework illustrating thesis structure and methods used: adapted from Bjørnson's PhD thesis (68).

1.5 Aim and specific objectives

The overall aim was to determine factors associated with childhood malnutrition and HIV/TB mortality using spatiotemporal modelling, based on health and socio-demographic surveillance system data from the rural Agincourt sub-district The modelling procedures adjust for the household-level spatial random effects using Bayesian techniques.

	Specific objectives	**Paper**
1.	To measure the presence and extent of univariate and multivariate spatial variability of adiposity measures for children living in the Agincourt sub-district, Mpumalanga Province in 2007.	➡ **Paper 1**: Estimation of univariate and multivariate spatial variability for childhood malnutrition outcomes from rural South African households.
2.	To determine the spatially adjusted effects of child HIV morbidity on malnutrition in children from the Agincourt sub-district, Mpumalanga Province in 2007.	➡ **Paper 3**: The contribution of spatial analysis to understanding HIV/TB mortality in children: a structural equation modelling approach.
3.	To determine the spatial-temporal trends of HIV/TB related mortality for children in Agincourt sub-district, Mpumalanga Province in 2004, and from 2000 to 2005.	➡ **Paper 2 & 4**: Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa
4.	To develop a modelling procedure that handles, large, zero inflated data adjusting for spatiotemporal random effects.	➡ **Paper 5**: Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa.

1.6 Data, measurements and methods

1.6.1 Agincourt health and socio-demographic surveillance system

The data used in this thesis are from the Agincourt health and socio-demographic surveillance system (HDSS) established in 1992 and located in the rural Agincourt sub-district in Bushbuckridge Municipality, Mpumalanga Province, northeast South Africa (Figure 1.6) (26). The HDSS was set up in that location mainly due to its remote location, availability of several clinics and presence of Mozambican in-migrants who arrived in the early to mid 1980s during the civil war (26). At onset in 1992 more than 57,000 people from nearly 9,000 households located in 20 villages were registered and followed up annually (57). By 2012 the sub-district was comprised of some 80,000 plus people living in more than 16,000 households in 27 contiguous villages. Information at individual level (infants, children, adults and the elderly) is collected once every year together with other annual census updates, that compile demographic, mortality, fertility, migration and health data (57).

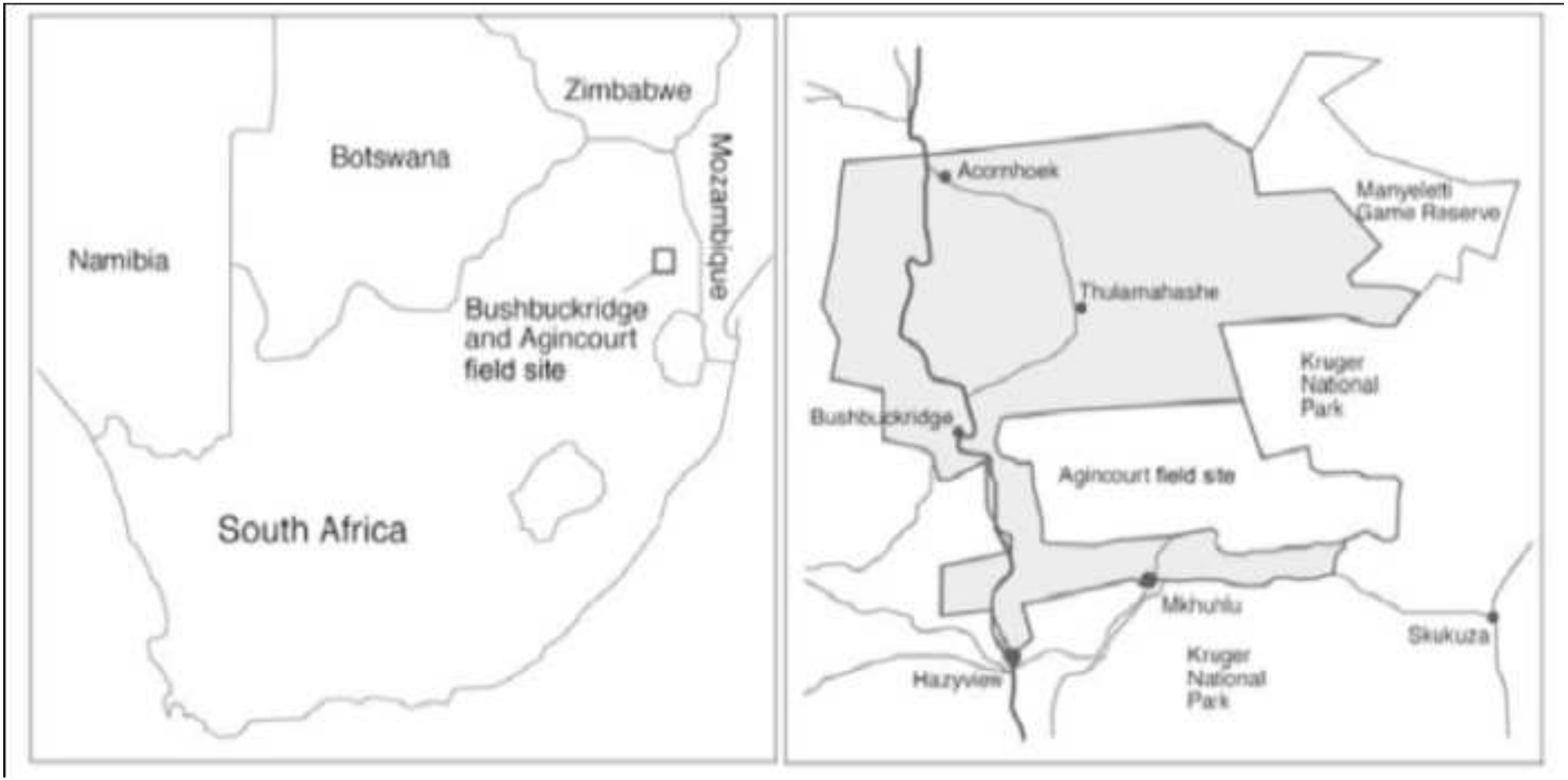


Figure 1.6: Agincourt Health and socio-Demographic Surveillance system: study site location.

1.6.2 Agincourt's HDSS: data collection and validity

1.6.2.1 Agincourt's hierarchical data structure

The Agincourt HDSS data can be viewed as a hierarchical (multilevel) data structure. The data collected from the site are for different objects, events and modules. The object tables are: individual, household and location (households and villages) levels. The events include: births, pregnancies, maternity history, deaths, in- and out-migrations. Modules include: child care grants, child morbidity, education, food security, health care utilisation and household assets (57). In Figure 1.7, I give a diagrammatic overview of the multilevel structure and how this links with the three components of epidemiological research (person, place and time).

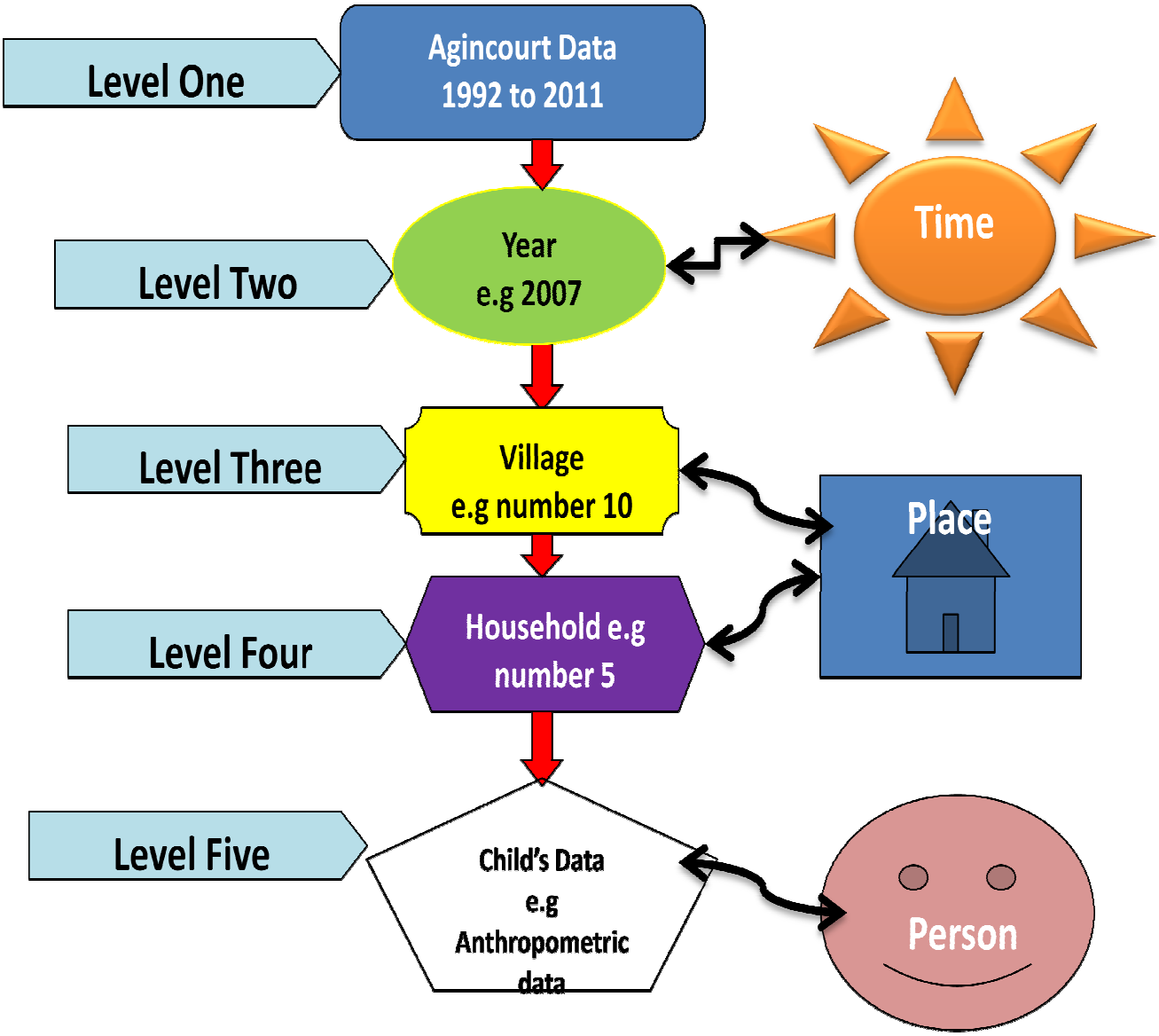


Figure 1.7: Agincourt sub-district hierarchical data structure.

1.6.2.2 Agincourt resident status, vital events and modules

The annual updates are done on resident status and vital events; these include the household roster, pregnancy outcome, death, migration, maternity history and union status (57). Household data collected include the geographical location of each household and village, demographic data on household head, vital events and other modules including household assets and food security. Pregnancy outcome data include

antenatal clinic visits; contraceptive use; duration, outcome and mode of delivery. Death and migration data include dates of events and locations. Other yearly updates focus on union status and maternity history.

The HDSS also conducts other modules less frequently; bi-annually for some and at most five yearly. Table 1.3 lists these modules which include education, labour participation, household assets and adult health. In accordance with pertinent scientific questions, other less regular surveys are carried out as once-off events. Examples include screening questions on chronic cough in 1998 and one-sided weakness in 2000. Relevant to this thesis, a nested cross-sectional anthropometric survey was conducted in 2007 (57).

Table 1.3: Agincourt HDSS household or individual modules: objects, events and episodes from 1992 to 2012. Source (66)

Census year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Household or individual level; target group
Modules																						
Education	■					■					■				■			■			■	All individuals
Labour participation									■				■				■				■	⩾15 years
Household assets										■		■		■		■		■		■		All households
Temporary migrations											■					■					■	All households
Child care grants											■			■			■			■		All households
Health care utilization												■			■				■			All (2003); <5 years (2006); ⩾50 years (2010)
Food security													■			■			■			All households
Adult health															■				■		■	⩾50 years
Unions														■	■	■	■	■	■	■	■	All adults
Father support																■	■					All households
All social grants																				■		All households
Vital documents														■		■						All individuals
National ID numbers																■	■	■	■	■	■	All individuals
Cellphone numbers																				■	■	All individuals
Other names																				■	■	All individuals
Status observations																						
Chronic cough								■														⩾10 years
One-sided weakness										■												⩾15 years
Pregnancy status															■	■	■	■	■	■	■	Females 12–49 years
Epilepsy (seizures)															■	■	■	■	■	■	■	All individuals

1.6.2.3 Verbal autopsy

After every death that occurs in the HDSS, an interview is held with the closest caregiver of the deceased, at least one after death after a month in keeping with cultural mourning customs (69). The interview verbal autopsy is conducted by a team comprised of a supervisor and four specially trained fieldworkers (57). A locally translated and validated instrument adapted from that used in Niakhar, Senegal was used (57). An probing on all symptoms and open ended verbatim detail are collected from the caregiver. Completed questionnaires are given to medical experts to ascertain the probable death cause. The cause of death is independently determined by two medical experts (or three in the case of discrepancies) and classified according to the World Health Organization's International Classification of Diseases (ICD10) (70). The medical experts classify the probable cause as underlying, immediate and contributory. The main outcome variable included in this study was death due to HIV and/or tuberculosis (TB) determined by the WHOs ICD10 verbal autopsy codes A16-A19[1] for TB and B20-B24[2] for HIV.

1.6.2.4 Anthropometric survey and HIV testing

This thesis uses data on 648 deaths of children aged 1-5 years, each coming from a different household. The measurement procedures are described below. The anthropometric survey was conducted between April and July 2007 and involved 4658

[1] A16=Respiratory tuberculosis; not confirmed bacteriologically or histologically ; A17=Tuberculosis of nervous system; A18=Tuberculosis of other organs; A19=Miliary tuberculosis

[2] B20=Human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21=Human immunodeficiency virus (HIV) disease resulting in malignant neoplasms; B22=Human immunodeficiency virus (HIV) disease resulting in other specified diseases ; B23=Human immunodeficiency virus (HIV) disease resulting in other conditions; B24=Unspecified human immunodeficiency virus (HIV) disease

children and adolescents aged 1 to 20 years, randomly selected from the Agincourt sub-district. The sampling targetted at least 100 children per sex per year of age. Those included were permanent residents in the study area at the time of sampling and had lived in the study area at least 80 % of their life time, since birth or since 1992 when the Agincourt HDSS started.

Height was measured using a stadiometer (Holtain, UK) calibrated in millimeters for all children 2-5 years. For children aged less than 24 months, length was measured using an inelastic tape measure in a recumbent position on a flat surface. Weight in kilograms (to one decimal point) was determined using a mechanical bathroom scale (Hanson, UK). Written informed consent was obtained from the parent/caregiver for anthropometric measurements. Anthropometric measurements (WAZ, HAZ and WHZ) were computed from the heights and weights in accordance to the global age specific medians and standard deviations.

HIV testing was carried out using two concurrent rapid tests: Uni-Gold^{TM} (Trinity Biotech, Bray, Ireland) and Determine^{TM} (Abbott, Wiesbaden, Germany), with pre- and post-test counselling. Additional separate written informed consent was obtained from the parent/caregiver for HIV testing.

More details of the sampling, data collection procedures and HIV testing are provided elsewhere (57, 71, 72).

1.6.2.5 Agincourt's geographical information system

Agincourt sub-district is located south of the equator near the Mozambican border (24° 49' 0' South, 31° 15' 0' East). It covers an area of about 420 to 432 square kilometres

with more than 35 kilometres as the maximum straight line distance between two households. In the HDSS every household is geo-coded which enables spatial analysis at household as well as at village level. The household geo-codes and village centroids' longitude and latitude data were measured using the South Oriented Transverse Mercator projection (Gauss-Krüger or Gauss conformal projection) coordinate system (73).

1.6.3 Ethical considerations

The Agincourt HDSS has ethical clearance from the University of the Witwatersrand (reference number M960720). Community consent was obtained from civic and traditional leaders when the site was first established. Individual consent is continuously renewed at every new visit during the annual censuses and at collection of verbal autopsy information (26). This PhD work was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (M081145). The child nutrition data were obtained from an anthropometric survey which had ethical clearance (M070244).

1.6.4 Spatiotemporal modelling of mortality and malnutrition

1.6.4.1 Hierarchical Bayesian spatial models

Epidemiological modelling often involves linking exposures to the outcome variables to explain and/or predict with the goal of producing evidence to inform control measures and interventions (25). Analysis of hierarchical spatial data depends on three main aspects: firstly the type of spatial data, secondly modelling of the spatial covariance

structure, and lastly the distribution of the outcome and manner of adding random effects (Gaussian and non-Gaussian structures).

Three types of spatial data exist: areal (lattice) such as village level, point patterns (locations not fixed but random) and point-level (geo-statistical, locations are fixed over a continuous space) such as household location data (74). Areal models use neighbourhood structures based on arrangements of the blocks in the map. Two very popular models that incorporate such neighbourhood information are the simultaneously and conditionally autoregressive models (abbreviated SAR and CAR respectively) originally developed by Whittle (75) and Besag (76), respectively. An essential property for geo-statistical data is isotropy when the spatial correlation only depends on the distance between the two locations and not directional as with anisotropy (77).

Observed HIV/TB mortality data in houses of close proximity share similar risk exposures as postulated by the first law of geography that near objects are more related than distant objects (78). Spatial exploratory procedures often reveal the presence of such inter-relations (Moran's and Geary's indexes) (79, 80), quantify it (variogram fits for geo-statistical data and Markov Random fields for areal data), while modelling goes further to enable drawing of smooth maps of mortality and morbidity (81, 82). Spatial pair-wise correlations are often modelled based on the Euclidean distance between two geo-locations. The Matérn correlation function is one of several parametric covariance structures with parameters that measure rate of spatial range (the distance beyond which spatial correlation may be ignored) and the spatial variability (83).Other functions are the

exponential (smoothness parameter of Matérn is 0.5), Gaussian (smoothness parameter of Matérn tends to infinity) and the spherical which are discussed in chapter two (84).

Spatiotemporal modelling introduces the time and spatial components either as covariates (measuring trends) or as single or joint random effects such as first order autoregressive or Matérn correlation for the temporal and spatial respectively (85). Geo-statistical models were introduced by Diggle et al. and they blend well with the two commonly used modelling procedures - the general linear models for Gaussian outcomes and the bigger family of Generalized linear mixed models (GLMMs) for both Gaussian and non-Gaussian (86, 87).

Inclusion of the spatial correlation structure in the model is straightforward for Gaussian models as the structure follows a multivariate normal distribution. As for non-Gaussian structures these are fit as random effects, similarly for temporal effects. Hence spatial random effects are fit into GLMMs by assigning a parameter at each location forming a spatial column vector with dimensions equal to the number of geo-locations. Another very important aspect of geo-statistical modelling is spatial prediction commonly known as ‘kriging’ (85), named after a South African mining engineer Krige (88). In HIV/TB or malnutrition research, spatiotemporal modelling aims to find factors that influence the geospatial distributions adjusting for temporal and spatial correlation and spatial prediction (kriging) at unobserved locations.

1.6.4.2 Zero Inflated spatiotemporal models and Bayesian inference

Epidemiological studies with observational count outcome data, where the proportion of zero counts is greater than expected on the basis of the mean of the non-zero counts, are on the rise. Handling such outcomes dealing with the inflation of zeros, requires the use of zero inflated models by introducing a latent probability process that inflates the number of zero points with a weighting to the conditional distribution. The use of zero inflated models in epidemiological studies has been on the increase over the years (23). Modelling zero inflated models using GLMMs requires a linear canonical predictor function (either logit or log) linked to either a count (discrete), dispersed count or binary for Poisson, Negative Binomial or Logistic respectively as the three most commonly used models (89, 90).

Two commonly used statistical inference procedures, frequentist and Bayesian, can be applied to hierarchical data. However, the frequentist based mainly on maximum likelihood estimation (MLE), views the spatial effects as independent and fixed whereas the Bayesian views these as random. For multilevel data the MLE underestimate the standard errors and overestimate the statistical significance (91). Bayesian based inference estimate highly parameterized models with the aid of Markov chain Monte Carlo (MCMC) with variants such as the Gibbs sampling, Metropolis Hastings, slice sampling and reversible jump (20, 92-95). Another Bayesian estimation technique for Gaussian Markov random fields (GMRFs) is the Integrated Nested Laplace Approximation (INLA) (96, 97). Bayesian techniques also have the advantage of simultaneously estimating the outcome-exposure relations together with the random effects, a thing that ML procedures cannot do.

1.6.4.3 Modelling large geo-statistical random effects

Geo-statistical random effect models have a parameter at each location, which gives problems in model estimations when the locations are many due to computer's instability, slower speed (due to order $O(N^3)$ arithmetic operations for m locations), crashing and sometimes it is just not feasible. This problem is also known as the "big m" or "big N" (85, 98). In MCMC for large spatial data structures this problem occurs since the procedure involves repeated inversions (iterations) of a large NxN matrix as many times as the iterations.

In an attempt to reduce the computational burden in estimating the spatial random effects, several approaches have been used: fast Fourier transforms, spectral, lattice, dimension reduction, course fine coupling, predictive process and lowrank kriging and triangulation combined with stochastic partial differential equations (SPDE) (85, 98-101). For geo-statistical data the most commonly used are the predictive process and lowrank kriging approaches, which estimate the spatial parameters using a smaller number of locations and "representative" locations say M which are projected on the full spatial random vector using a linear combination ($M<<N$). Two procedures used in selecting a "representative" sample are the mini-max space filling algorithm and triangulation combined with SPDEs. This procedure, which is also known as sub-sampling, was used in the modelling done by Pardo-Igúzquiza and Dowd (99).

1.6.4.4 Spatiotemporal model building and selection

In this thesis several approaches to fit the models were used. We initially described the covariates stating the measures of central tendency (mean or medians) and measures of variability (standard, deviation, standard errors or ranges). Univariate or bivariate analysis of the non-spatial models followed to determine the risk factors or factors associated with the outcomes. Those that emerged significant were included in the multiple variable models, as well as the potential demographic confounders that showed significance from the 90% and the 95% Bayesian credible intervals. Several multiple variable models were built: firstly adjusting only for other covariates, secondly adjusting for covariates and spatial random effects, thirdly adjusting for temporality and other covariates (only in chapter six) and finally adjusting for spatiotemporal random effects and covariates (chapters 5 and 6). As in chapters 2 and 3, spatial multivariate (two or more outcomes) analysis was performed. Structural equation models utilising robust standard errors were also fit for data for chapters 3 and 4 using only the variables that had come out significant from the spatial multiple variable models. Selection of priors for the Bayesian-based models was mainly centred on maximising the data by use of non-informative priors or weakly informative priors; mostly these were default priors in the software used which occasionally required some adjustments of parameters to fit better to the data.

Deciding on the final model from all those fit involved different approaches. Initially, model convergence was assessed by inspecting the diagnostic graphical outputs (trace-plots and density plots), tests of convergence and closeness of measures of central tendency for Gaussian data. Secondly, the deviance information criteria (DIC) was used

in determining goodness of fit of the proposed models, this can be defined as “classical estimate of fit, plus twice the effective number of parameters”. This is a technique for Bayesian model selection for MCMC derived posteriors – details on this procedure are provided in chapter 5 section 5.2.4. Selection of better model based on the DIC basically favours models with the smaller values. An interpretation of this is quite straight forward: the smaller the effective number of parameters the more parsimonious the model and the smaller the DIC the better the model fit; however the most parsimonious is not always the best model. The introduction of spatial random effects in the thesis increased the number of parameters in the models. The interpretation and selection of “best” models took this into account as the standard errors for the non-spatial and spatial models were different.

Key Introductory points

1. Childhood HIV and malnutrition are key public health problems globally, in developing nations and in South Africa.

2. Determinants of child HIV/TB mortality and malnutrition are interrelated in complex pathways that are often ignored in modelling.

3. Spatial temporal modelling in our analysis can help to adjust for within-groups correlations in hierarchical data structures.

4. Agincourt HDSS data provided high quality and well validated data to produce spatial prediction (kriging) maps for child malnutrition and mortality.

Chapter 2

2.0 Estimation of multivariable and multivariate spatial variability for child adiposity measures in rural South African households

Eustasius Musenge[1,2], Elizabeth Kimani-Murage[1,4], Shane A. Norris[3] and Kathleen Kahn[1,5,6]

1. MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

2. Biostatistics and Epidemiology Division, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

3. MRC/WITS Developmental Pathways for Health Research Unit, Faculty of Health Sciences, University of the Witwatersrand Medical School

4. African Population and Health Research Centre, Nairobi, Kenya

5. Centre for Global Health Research, Umeå University, Umeå, Sweden

6. INDEPTH Network, Accra, Ghana

This paper has been revised and resubmitted to the Spatial Statistics journal.

Chapter 2 abstract

Background: Child adiposity is an important indicator of public health status and associated with morbidity and mortality. South Africa is one of the nations with a high child metabolic syndrome due to among other causes adiposity. In this paper we investigated the spatial distribution of three child adiposity measures in a remote area north-east of South Africa using univariate and multivariate spatial random effects models with covariates.

Methodology: We undertook a secondary analysis of data collected from a cross-sectional study conducted in 2007 on children aged 1-5 years. Three adiposity outcome measure: body mass index, waist circumference and percentage body fat. We investigated presence of spatial autocorrelation using Moran's and Geary's coefficients. Spatial semi-variograms parameter estimation was done diagrammatically (classical and robust) and theoretical approaches. We ran univariate, multiple variable and multivariate spatial regression models with covariate.

Results: Moran's (*I*) and Geary's (*c*) coefficients indicated that there was positive global and local autocorrelation (clustering) respectively. We measured severity of spatial variation using the partial-sill-to-sill ratio which showed a five times smaller non-spatial variability (nugget) in the multiple variable models compared to the multivariate spatial models. There was a negative correlation between adiposity measures and: being males, being wasted and being underweight. However better socio-economic status increased the risk of having resembling higher body mass, wider waist circumferences and higher percentage body fat.

Conclusions: Multivariate modelling adjusting for spatial random effects significantly lowers the unexplained variability and is useful in modelling natural and biological processes. Our analyses reveal that there are spatial differences in child adiposity levels even in this small rural area of South Africa. These geographical differences are often ignored in epidemiological regression modelling resulting in reporting of biased estimates. Accurate estimates of important child growth measures are crucial as we near the end of the millennium development goals in 2015, to guide policy and interventions.

2.1 Introduction

Child malnutrition and adiposity (overweight and obesity) are major global public health problems listed amongst the 1990 Millennium Development Goals (MDGs): target three on goal one is 'To eradicate extreme poverty and hunger'(102). Poverty on the other hand results in nutritionally deprived dietary habits which are usually carried over from childhood to adult life (103-105). Globally developed countries have traditionally the greatest burden of adiposity whereas developing countries have malnutrition-related problems in children. Contrary to the trends observed in the past decades, we are experiencing an epidemiological transition which stipulates changes from communicable to non-communicable causes of morbidity and mortality in children and adults in low and middle-income countries. Eating too little or eating too much comes with a cost as childhood malnutrition and adiposity are associated with development of metabolic syndrome in early adulthood (106-108). Metabolic syndrome is a cluster of key cardiovascular risk factors associated with an increased risk of subsequent diseases of lifestyle such as coronary artery disease, stroke, type 2 diabetes, dyslipidemia and hypertension (109).

South Africa's childhood adiposity and malnutrition rates are 2.4 - 4.8%, 11.0 - 17.0%, 24.5%, 17.0% and 8.9% for obesity, overweight, stunting, underweight and wasting respectively (9, 110). In Mpumalanga province where our study was located rates are 19.5%, 15.5% and 6.7% for stunting, underweight and wasting respectively (9). These figures all show that there are marked malnutrition differences geographically, racially and across age groups. Understanding of child health epidemiology and identifying areas with the greatest risk, can be effective weaponry to saving children's lives. As

there are differences in childhood adiposity between developed and develping countries, urban and rural settlements and also within geographical areas, knowledge of the spatial distribution of child obesity and overweight can help focus interventions. Few studies have catered for the spatial variations in nutritional outcomes analysis (111-113).

Different measures of obesity vary in their accuracy as they either indirectly or directly measure body fat. We used three measures of underweight, obesity and overweigh: the body mass index (BMI), waist circumference and the percentage body fat. The commonly used BMI is a very useful indirect way of assessing obesity but does not distinguish between fat and non-fat weight or indicate where the fat is located (114). Another indirect measure is waist circumference, that measures the amount of fat deposited centrally in the body; this has been shown to be particularly detrimental to cardio-metabolic health. The percentage body fat, a ratio of total body fat mass and total body weight; is an even better and more direct measure of obesity than the former two. Many studies utilize these measures separately (103-108) but some studies have shown that a more wholesome perspective is derived when they are used in combination (114, 115). For example a child who exercises regularly and increases muscle mass may be misclassified as obese using the BMI but properly classified according to waist circumference and percentage body fat (114).

Statistical techniques take into account spatial correlations by introducing in the model either the spatial variables (longitude and latitude) or as spatially structured random effects. For the latter, the random effects assume the spatial data are correlated if this is overlooked as with the former, statistical methods often underestimate the standard

errors hence significance is overestimated, leading to incorrect inference and predictions thereafter (91, 116). It is therefore essential to handle spatial correlation well in analyses to adjust for and disclose the geographical patterns and reduce reporting of erroneous results.

Our main was aim to investigate the presence and then estimate the spatial variability on three adiposity outcomes; separately (univariate and multivariable) and combined (multivariate) while controlling for demographic and potential malnutrition related confounders, for children aged 1-5 years in Agincourt sub-district , rural north–east South Africa in 2007. We utilized classical geographical, geo-statistical, frequentist inference and Bayesian inferences to assess how well the different approaches concur.

The paper is structured as follows: in the methods section we describe the study area and how the outcomes were measured. Then we give a briefing on the theoretical procedures for measuring spatial autocorrelations, elements of theoretical spatial correlation models, frequentist inference and Bayesian inference via univariate and multivariate spatial modelling. We present and discuss our results from using simulated data and real data, and in conclusion we discuss the practical use of our results and scope for further studies.

2.2 Materials and methods

2.2.1 Study area and ethical clearance

The child data used were from the Agincourt sub-district located (24° 49' 0’ South, 31° 15' 0’ East) south of the equator or north-east of South Africa near the Mozambican border as shown in figure 2.1. The study was nested within the Agincourt health and socio-demographic surveillance system (HDSS) established in 1992.The HDSS follows some 80,000 plus people living in more than 12,500 households in 24 contiguous villages covering an area of about 420 square kilometres. This study was conducted between April and July 2007. Included in the study were children who had lived in the area at least 80% of the time since birth. A random sample of children was drawn from each age-sex-village stratum in proportion to the population size of the village. A total of 648 children between 1-5 years of age living in 648 households were used in this paper. Further details of the sampling, sample selection and the Agincourt site are provided elsewhere (71, 117). This work was granted ethical clearances by the University of the Witwatersrand’s Committee for Research on Human Subjects (No. 960720) and (M081145).

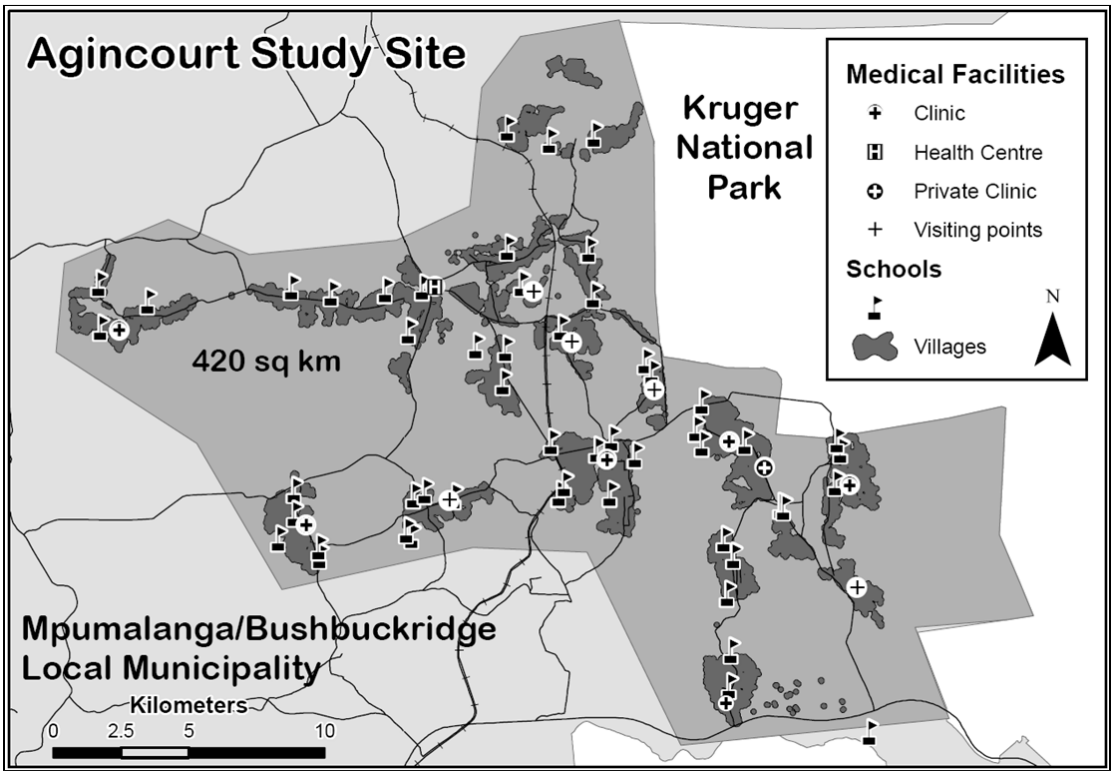


Figure 2.1: Agincourt sub-district and surveillance area showing households and primary health care facilities.

2.2.2 Measurement of three adiposity measures

A team of 12 fieldworkers was carefully trained by experts in anthropometric measurements. To minimise fieldworker variation, each fieldworker specialised in a specific measurement and collected the data on all study participants. Coefficient of variation for anthropometric measurements was determined before and towards the end of the study period; on average, the coefficient of variation towards the end of the study ranged between <1% and 3% for the different measurements (71).

Anthropometric measurements were carried out on all children. Height was measured using a stadiometer (Holtain, UK) calibrated in millimetres. For all children aged less than 24 months, the determination of length was done using an inelastic tape measure in

a recumbent position on a flat surface. Weight in kilograms (to one decimal point) was determined using a mechanical bathroom scale (Hanson, UK). These were then used in computing the body mass index (BMI) using $body\ mass\ index = \left(\frac{weight}{height*height}\right)$. Waist circumference was measured in millimetres using an inelastic tape measure at the natural waist (midway between the tenth rib and the iliac crest) with the participant in a standing position. All measurements were carried out according to standard procedures (71, 118).

The right hand side triceps, biceps, subscapular and suprailiac skinfold thicknesses were measured in triplets with a Harpenden (John Bull British indicators Ltd, London) skin fold calliper to the nearest 0.1 mm and were averaged (119). The four skinfold measurements averages were used to compute the total skinfold by summing the averages. Body density was estimated using the log of the sum of the four skinfold measurements as explanatory variables in gender-specific simple linear regression equations (119-121). Once the body density was estimated, this was then used in computing the percentage body fat using Lohman's equation. This is a child-specific formula similar to the Siri equation, which assumes a constant protein-to-mineral ratio on adult's cadaver body composition (122, 123). The Lohman's formula for prepubescent children, that caters for their lower bone-mineral content and higher water content, used was $Percent\ body\ fat = \left(\frac{5.30}{density} - 4.89\right) \times 100$ (123, 124).

2.2.3 Spatial autocorrelation measures

The Moran's (*I)* and Geary's (*c*) coefficients were used to investigate the presence of global and local clustering or randomness or dispersion. Inference on these two statistics

comes from approximate tests based on the asymptotic distribution of *I* and *c*, which both tend to a normal distribution as n (total sample) increases.

$$I = \frac{n}{W}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad (2.1)$$

where z_i is the deviation of a nutritional score from a location i ($z_i = Y_i - \bar{Y}$), $w_{i,j}$ is the spatial weight between locations i and j, and $W = \sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{i,j}$ is the aggregate of all spatial weights. The null hypothesis (no spatial autocorrelation) is such that the population is normally distributed with mean $E(I) = -1/(n-1)$ and variance $V(I)$ (79, 80). Interpretation of Moran's ($I \in [-1,1]$) coefficient depends on one of three possible outcomes: $I > E(I)$ indicates positive autocorrelation and suggests that neighbouring values have similar deviations (clustering), I close to zero indicating no spatial autocorrelation (random) and negative autocorrelation $I < E(I)$ which implies dispersed neighbouring values (125). The Moran's scatter plot is also used in verifying the results obtained from the computation, since the slope from the line of best fit using method of least squares is also the Moran's coefficient *I* from standardized values (126, 127). There has been some advancement of the Moran's coefficient resulting in a weighting adjusted modified Moran's coefficient, however this has not been used in our study (128).

In computations of the Moran's coefficient shown in (1), the cross-product ($z_i z_j$) is based on the deviations from the mean for the two geo-locations. When this is based on the actual values from the locations divided by two we obtain the Geary's $c \in [0,2]$ statistic for investigating local autocorrelation:

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(Y_i - Y_j)}{2W\sum_{i=1}^{n} z_i^2} \quad (2.2)$$

The null hypothesis of no local spatial autocorrelation is such that it follows the standard normal distribution with sample mean $E(c) = 1$ and variance $V(c)$ (129). Interpretation of Geary's coefficient, $c < E(c)$ for positive local autocorrelation (clustering) with c=0 implying perfect positive, c=1 no local spatial autocorrelation (random) and $c > E(c)$ for the negative local autocorrelation (dispersion) with c=2 for perfect negative (125). Statistical significance for both coefficients can be ascertained using p-values smaller than 0.05 or whether the absolute standard normal (Z) values are greater than 1.96.

2.2.4 Semi-variogram spatial parameter estimations

Spatial correlation between adiposity measures was estimated graphically using parametric semi-variograms. Our standardized nutritional Z-scores had zero means and assumed to be strictly stationary such that covariance between $Y(s_i)$ and $Y(s_i + h)$ was given by the semi-variogram $\gamma(|h|) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Y(s_i) - Y(s_i + h)]^2$ with $N(h)$ the number of sampled points of distance (lag) length h. Figure 2.2 (left) shows the distribution of all the possible distance pairs in our data. Parametric semi-variograms increase monotonically reaching a peak (sill) at a range ($\rho = 1/\phi$) with spatial variance known as the partial-sill (σ_s^2) and non random starting variance (at h>0) known as nugget (τ_s^2) see figure 2.2 (right)..

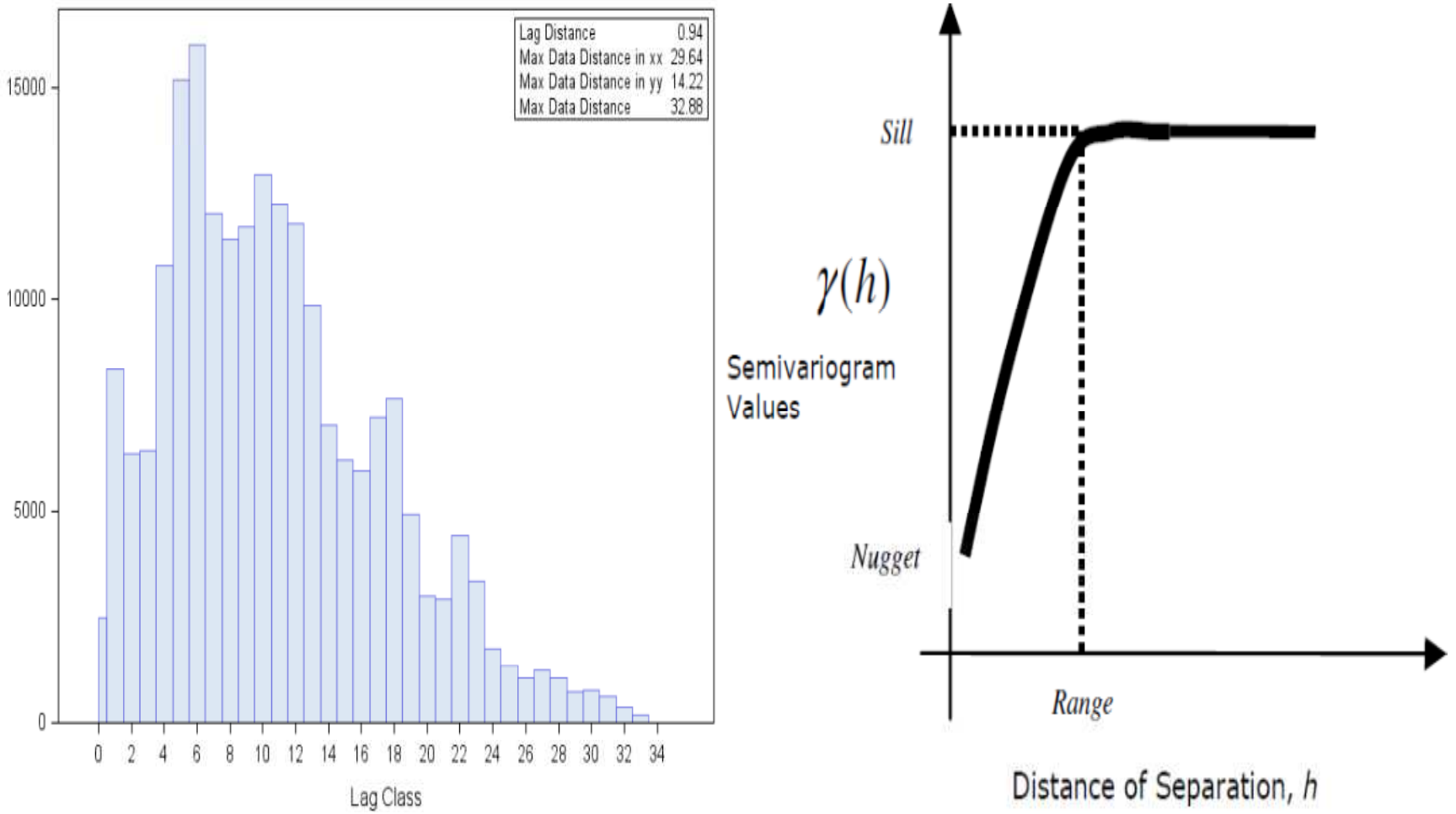


Figure 2.2: Distribution of all pair-wise distances (left) and a semi-variogram diagram (right).

We fit two theoretical semi-variograms, the exponential (equation 2.3) and the Gaussian (equation 2.4). The distance beyond which the spatial dependence would be negligible (95% of the sill) is known as the effective (or practical) range which is $r_{\in} = \frac{3}{\rho} = 3\phi$ and $r_{\in} = \frac{\sqrt{3}}{\rho} = \sqrt{3}\phi$ for the exponential and Gaussian respectively. The effective range is the distance beyond which the correlations are equal to or lower than 0.05

$$\gamma(h) = \begin{cases} 0 \; if \; |h| = 0 \\ \tau_s^2 + \sigma_s^2 \left(1 - exp\left(-\frac{|h|}{\phi}\right)\right) \; if \; |h| > 0 \end{cases} \tag{2.3}$$

where τ_s^2 is the nugget, σ_s^2 is the partial-sill (spatial variance), $|h|$ is the Euclidean distance between two locations and ρ is the range ($\rho = \frac{1}{\phi}$). Similarly for the Gaussian semi-variogram as shown in equation 2.(4):

$$\gamma(h) = \begin{cases} 0 \; if \; |h| = 0 \\ \tau_s^2 + \sigma_s^2 \left(1 - exp\left(-\frac{|h|^2}{\phi^2}\right)\right) \; if \; |h| > 0 \end{cases} \tag{2.4}$$

Semi-variograms were fit to the simulated data (shown in section 2.3) and also to our data using classical and robust approaches. Secondly we fit four theoretical semi-variograms namely; exponential using ordinary least squares (Exp-OLS), exponential using weighted least squares (Exp-WLS), Gaussian using ordinary least squares (Sph-OLS or Gau-OLS) and Gaussian using weighted least squares (Sph-WLS or Gau-WLS).

For reporting on spatial variability, some authors conjecture that the variance ratios $\tau_s^2/(\tau_s^2 + \sigma_s^2)$ = nugget/sill or $\sigma_s^2/(\tau_s^2 + \sigma_s^2)$ = partial-sill/sill are more important for spatial noise estimation than the range which is sometimes fixed in the estimations (81,

82). As such, the partial-sill-to-sill ratios $\sigma_s^2/(\tau_s^2 + \sigma_s^2)$ of <25%, 25-75% and >75% indicate weak, moderate and strong spatial dependences respectively (130, 131).

2.2.5 Single outcome spatial random effects Bayesian models

Adiposity measures; BMI, waist circumference and percentage body fat, provide useful indications of accumulated body fat in the human body and are often measured seperated. These adiposity data for each measure can be denoted as $Y(s_i)$, (i = 1,2, … , N) at a household location (longitude and latitude) $s_i \in \Re^2$, where s_i varies continuously over an area ***D***, a fixed subset of $\Re^2 \subset \boldsymbol{D}$. The spatial and non-spatial random effects model for the ith (i = 1,2, … , N) child at location s_i can be written as:

$$Y(s_i) = X^T(s_i)\boldsymbol{\beta} + \boldsymbol{\varphi}(s_i) \qquad (2.5)$$

where $X^T(s_i)$ is the vector of explanatory variables, $\boldsymbol{\beta}$ the vector of regression coefficients and random effects $\varphi(\boldsymbol{s}) = (\varphi(s_1), \varphi(s_2) \ldots\ldots \varphi(s_N))^T \sim MVN(0, \sum)$ with $\boldsymbol{\Sigma} = \sigma_s^2\rho(\phi) + \tau_s^2 I$ made up of the of spatial (with spatial variance σ_s^2 and decay $\rho(\phi)$ parameters) and non-spatial (τ_s^2 nugget parameter) random effects.

We fit two single outcomes spatial random effects models, these were; lowrank *kriging* (using a sub-sample of the locations) and global *kriging* (using all the locations) via Bayesian inference's Markov chain Monte Carlo (MCMC) parameter estimations. The lowrank kriging model computes latent spatial random effects which follow a multivariate Gaussian random field (GRF) with zero stationary mean which are projected onto an n × 1 sub-sample

$\varphi^*(\boldsymbol{s}) = (\varphi^*(s_1), \varphi^*(s_2) \ldots\ldots \varphi^*(s_n))^T \sim MVN(0, \sum^*)$. To reduce the computational burden associated with many locations, also known as the 'Big N', we estimated $\varphi(\boldsymbol{s}) = Z_{spat}\varphi^*(\boldsymbol{s})$, where Z_{spat} the $N \times n$ matrix is a product of the covariance matrix between the sub-sample and the original locations $c^T(N \times N)$ and it's variance-covariance matrix $\sum^*(n \times n)$ (85). The variance-covariance matrix $\sum^*$ follows an isotropic parametric correlation structure with geographical between-households variance σ_s^2 (partial-sill). For the exponential correlation the multivariate distribution has variance $\sum^* = \sigma_s^{*2} exp\left[-\hat{\rho} d_{ij}\right] + \tau_s^{*2} I$, where d_{ij} is the Euclidean distance between households *i* and *j* from the sub-sample, and $\hat{\rho} = max_{i,j} \left\| s_i - s_j \right\| / 0.001$ is empirically estimated from all the locations and then fixed; rate of spatial correlation decay (82). The 'representative' sub-sample was selected using the mini-max space filling criteria (82, 132). Lowrank kriging predicted for the remainder of 23% (148/648) of the locations after fitting parameters using the 77% (500/648). Whereas the global (universal) kriging model uses all the geo-locations to estimate the latent spatial random effects.

For a full Bayesian model we will require the priors for the parameters and functions. For the fixed effects parameters $\boldsymbol{\beta}$ we assume diffuse priors or weakly informative Normal priors. For spatial random effects priors used in the lowrank *kriging* we had $p(\phi^* | \sigma_{spat}^{*2}) \propto |\sum^*|^{-\frac{1}{2}} exp\left(\frac{-1}{2\sigma_{spat}^{*2}} (\phi^*)' \sum^{*-1} \phi^*\right)$ and the independent non spatial random (nugget) effects had Normal priors and all hyper-priors were Inverse Gamma (IG) (133).

2.2.6 Multivariate spatial random effects Bayesian model

We also jointly modelled the three child adiposity measures assuming they follow a Multivariate (Trivariate) Normal distribution:

$$\underset{\sim}{Y}(\boldsymbol{s}) = X^T \underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\omega}(\boldsymbol{s}) \tag{2.6}$$

where $\underset{\sim}{Y}(\boldsymbol{s}) = [Y_k(\boldsymbol{s})]_{k=1}^{q=3}$ is a vector with BMI, waist circumference and percentage body fat at all households, $\underset{\sim}{\boldsymbol{\beta}} = [\boldsymbol{\beta}_k]_{k=1}^{q=3}$ and $\underset{\sim}{\omega}(\boldsymbol{s}) \sim MVN[\mathbf{0}, K(\underset{\sim}{\theta}; s, s') + I_n \otimes \Psi]$ with vector of spatial decay parameters $\underset{\sim}{\theta} = [\phi_k]_{k=1}^{q=3}$,the qn x qn variance-covariance matrix $K(\theta; s, s')$ which should be symmetric and positive definite.

$$\Psi = \begin{pmatrix} \tau_{1,1}^2 & 0 & 0 \\ 0 & \tau_{2,2}^2 & 0 \\ 0 & 0 & \tau_{3,3}^2 \end{pmatrix} = diag[\tau_{k,k}^2]_{k=1}^{q=3}.$$

To complete Bayesian specification we adopt the inverse Wishart prior $K^{-1} \sim Wish(\Omega^{(K)}, 3)$ for the variance-covariance matrix, non informative Normal priors $\boldsymbol{\beta}_k \sim N(\mu_k^{(\beta)}, \sigma_k^{2(\beta)})$ for the regression coefficients, Uniform priors for the spatial decays $\phi_k \sim U\left(\frac{1}{\max(d_{ij})}, \frac{1}{\min(d_{ij})}\right)$ and an inverse Wishart prior for the matrix of non spatial random effects $\Psi^{-1} \sim Wish(\Omega^{(\Psi)}, 3)$.

To fit this multivariate spatial random effects model we used all the 648 household geo-locations (global *kriging* technique). Estimation of the spatial parameters via Bayesian inference's MCMC employing the Metropolis Hastings (MH) step whenever necessary for conditionals without a closed form.

2.2.7 Software for geo-statistical *analyses*

Preliminary analysis and semi-variogram analysis were done in STATA 12.0 and SAS version 9.2 respectively (134, 135). Bayesian global and lowrank kriging models were fit using R version 2.12.2 packages (spBayes, maptools and coda) and BayesX respectively (136, 137). Semi-variogram plots were done in SAS; simulated data, Moran's scatter plots, diagnostic plots, observed adiposity measures contours and maps were done in R; and box plots of simulated data estimates were generated using STATA. Bayesian model parameters were fit via Gibbs sampling and, where appropriate, the Metropolis Hastings (MH) within Gibbs by Markov chain Monte Carlo (MCMC) (85). Inference was based on a total of 105,000 three chain iterations, the first 5000 were discarded as burn-in and to reduce auto-correlation every 25th value was taken to form a posterior sample of 4000. Model convergence was assessed by inspecting the diagnostic graphical outputs (trace-plots and density plots), tests of convergence and closeness of measures of central tendency for Normally distributed data (138, 139).

2.3. Results

Initially we used univariate simulated data to compare how well all the approaches we discussed were able to capture the true values and to make comparisons between them. We then give results of the spatial autocorrelation investigations using Moran's and Geary's indices for the adiposity measures. Then we ran analysis on the adiposity measures based on the techniques that utilise all the geo-locations in multiple variable as well as multivariate Trivariate models.

2.3.1 Simulated data results

Univariate models were fit to simulated data with a constant of 1.5, spatial variance (partial-sill) 5.00, effective range of 1.80 and non-spatial variance (nugget) 2.00 generated using the geo-locations in our real data. As we assumed our data to follow either the exponential or spherical parametric distributions, model fit was done using the aforementioned graphical and theoretical procedures. Not all of our models were able to capture the true values in the 95% confidence and credible intervals as shown in table 2.1 and figure 2.3.

Table 2.1: Graphical, theoretical and Bayesian derived spatial parameter estimates for simulated data

			Simulated data
			***True value**; Estimate (95% Confidence/Credible interval)*
Partial Sill-spatial variance (σ_s^2)	Exponential	OLS	**5.00**;5.02(2.22;7.92)
		WLS	**5.00**;5.02(4.08;5.98)#
	Gaussian	OLS	**5.00**;3.88(2.02;5.74)
		WLS	**5.00**;3.86(3.14;4.58)
	Bayesian Kriging	**Lowrank**	**5.00;5.03(2.48;8.27)**
		Global	**5.00;4.80(3.02;8.10)**
Range $(\mathbf{3} \times \boldsymbol{\phi})$	Exponential	OLS	**1.80**;1.63(0.05;3.20)
		WLS	**1.80**;1.99(1.16;2.82)#
Range $(\sqrt{\mathbf{3}} \times \boldsymbol{\phi})$	Gaussian	OLS	**1.80**;2.18(0.74;3.62)
		WLS	**1.80**;2.35(1.59;3.12)
Range $(\mathbf{3} \times \boldsymbol{\phi})$	Bayesian Kriging	**Lowrank**	**3.42 (fixed)**
		Global	**1.80;1.68(0.79;2.84)**
Nugget-Non spatial variance (τ_s^2)	Exponential	OLS	**2.00**;1.68(-1.20;4.55)
		WLS	**2.00**;1.94(0.95;2.92)#
	Gaussian	OLS	**2.00**;2.85(1.01;4.68)
		WLS	**2.00**;2.98(2.29;3.68)
	Bayesian Kriging	**Lowrank**	**2.00;2.04(1.73;2.40)**
		Global	**2.00;2.32(1.93;2.76)**

Key:* - significance at 5%, ** - significance 1%, # - graphical 'better' fitting semi-variogram estimates.

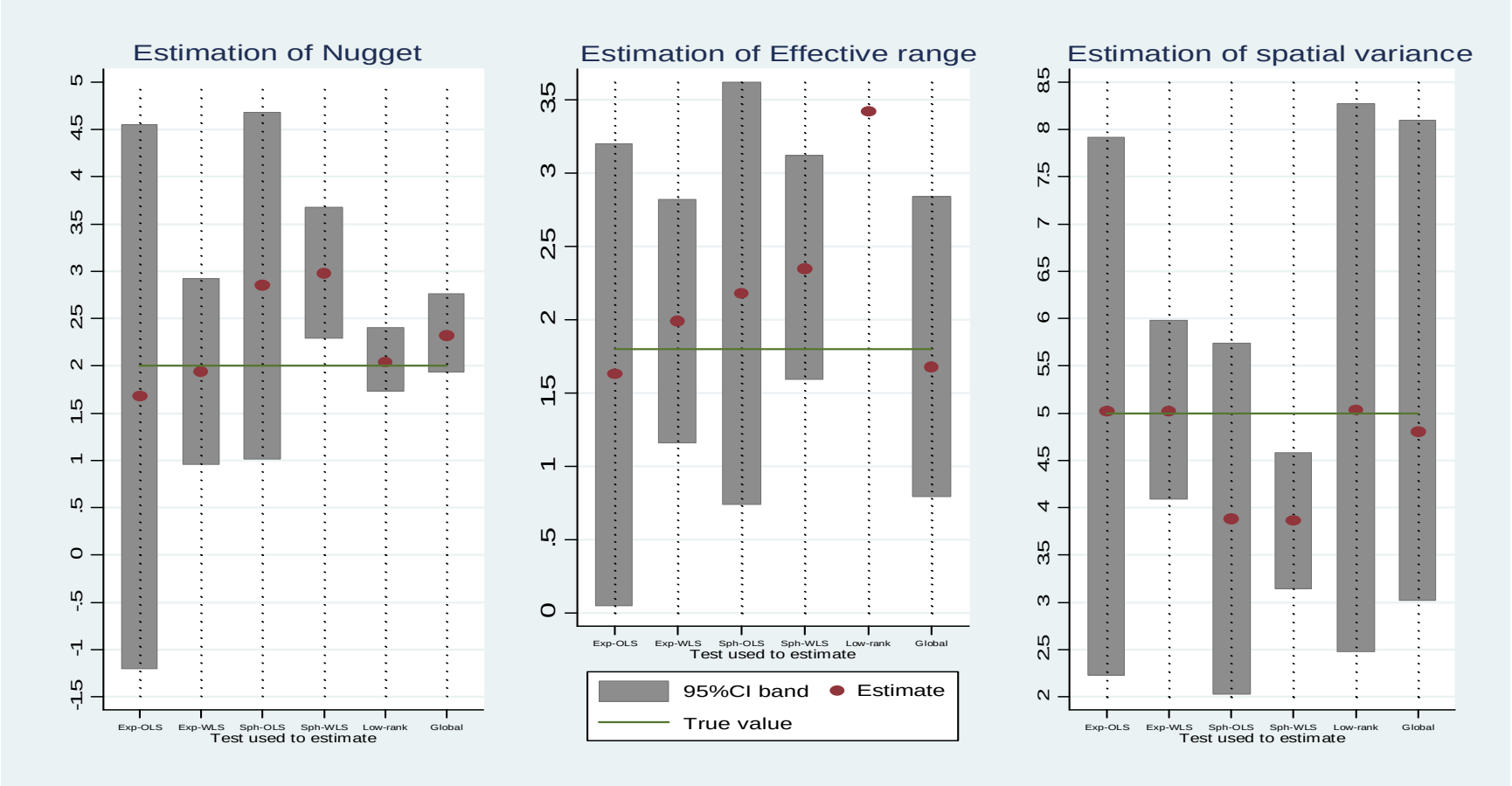


Figure 2.3: Box plots of 95% credible intervals and point estimates for simulated data.

Generally, the exponential models captured all the parameter estimates; the Gaussian models overestimated the nugget and effective range but underestimated the spatial variance. These Bayesian models estimated the nugget well but gave wide credible intervals for the spatial variance. In estimation of the nugget two models' point estimates were very close to the true value and the Bayesian lowrank model gave the narrower credible interval. When estimating the effective range, the model that fixed the range model was the worst and the Bayesian model using all locations (global) had the closest point estimate, later we discuss this approach further. Three models were very close to the partial-sill point estimate and the best was from the exponential weighted least squares (Exp-WLS). Figure 2.3 shows that the Gaussian weighted least square (Sph-WLS or Gau-WLS) performed the worst as it completely missed the true value in two of the three estimates. A similar pattern can be seen in figure 2.4 as it leaves out more of the graphical and robust estimates compared to the others.

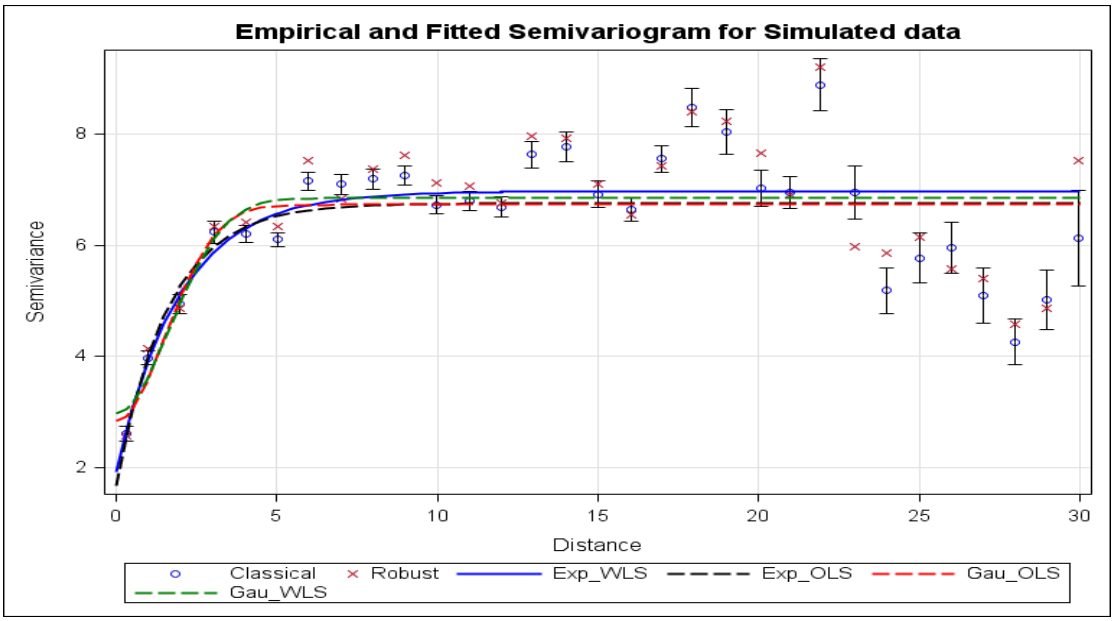


Figure 2.4: Graphical semi-variogram plots for classical and robust semivariance estimates for simulated data.

These results are based on a model fit using the global kriging technique involving all 648 locations. Model diagnostics plots for the full (no burning or thinning) MCMC posterior sequence are shown in figure 2.5, for the Normally distributed constant a perfectly symmetrical pattern was observed. These plots indicated that the model fit did very well in capturing the true values. Upon employing the Heidelberger and Welch's convergence diagnostics test of model convergence, all the posterior sequences passed the stationary test indicating that the samples used were sufficient for convergence and a longer sequence was not necessary (140). Lastly, a close inspection of the summary estimates indicates that the model credible intervals were able to capture the true parameter values and the point estimates were well within the true values.

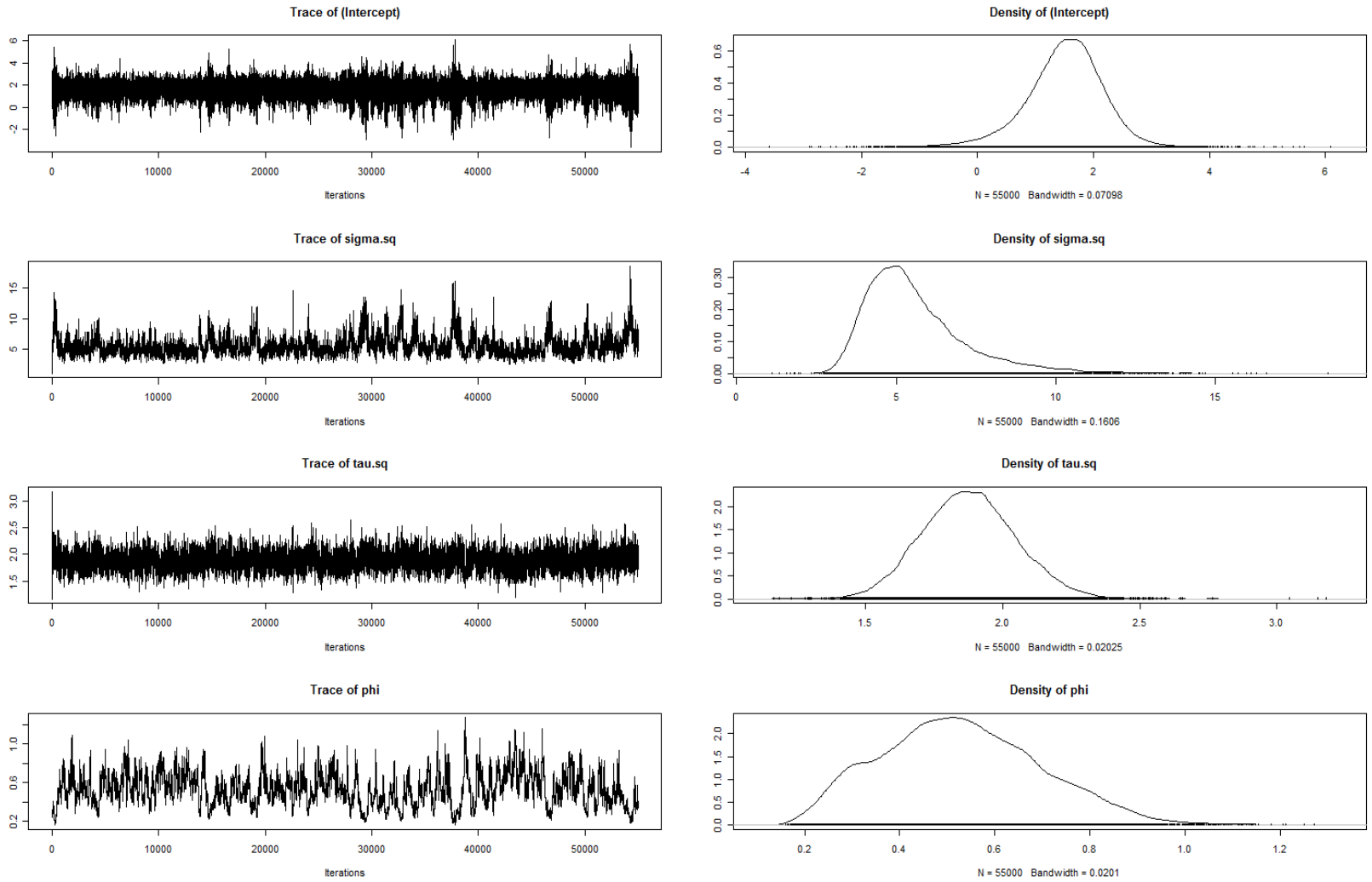


Figure 2.5: Diagnostic (autocorrelation and density) plots for simulated data parameters.

2.3.2 Spatial autocorrelation analysis results

Moran's coefficient for spatial correlation of body mass indices showed a significant positive autocorrelation ($\mathrm{I} = 0.023 > E(\mathrm{I}) = -0.002$) (see table 2.2) hence the spatial distribution of high and/or low values in the BMIs is significantly spatially clustered. Similarly for Geary's sample spatial correlation ($\mathrm{c} = 0.965 < E(\mathrm{c}) = 1.00$) which suggests a trend of neighbouring household values had similar deviations. Analysis of waist circumferences showed statistically significant positive Moran's and Geary's coefficients, hence the data were more prone to be spatially clustered rather than random and significant local spatial clustering, respectively. The body fat masses show more randomness than neither global nor local clustering as indicated by the Moran's and Geary's indices.

2.3.3 Descriptive statistics for Agincourt child adiposity outcomes and malnutrition explanatory variables

We analysed data for 648 children located in the Agincourt sub-district, rural north-east South Africa, whose average age was 3.15 years with slightly more males (50.6%) as shown in table 2.3. Amongst the explanatory variables were child malnutrition Z-scores less than -2 which have health significance: 6.79% were wasted ($\mathrm{WHZ} < -2$), 17.13% were stunted ($\mathrm{HAZ} < -2$) and 10.2% were underweight ($\mathrm{WAZ} < -2$). Socio-economic status quartile indices, computed from household assets using principle component analysis, showed that 19.75% were least poor and 20.68% were in the most poor

Table 2.2: Graphical, theoretical and Bayesian derived spatial parameter estimates for simulated data

Variable	Measure	Index (I)	Expected value E(I)	Standard deviation sd(I)	Z; p-value
Body mass index	Moran's	0.023	-0.002	0.007	3.748; p<0.001
	Geary's	0.965	1.000	0.023	-1.562; p=0.059
Waist circumference	Moran's	0.011	-0.002	0.007	1.820;p=0.034
	Geary's	0.969	1.000	0.015	-2.166;p=0.015
Body fat mass	Moran's	-0.002	-0.002	0.007	0.029;p-0.488
	Geary's	1.000	1.000	0.016	0.029;p-0.488

categories. In the following subsections we give the results from the geostatistical exploratory analysis (autocorrelations) followed by the univariate and multivariate geostatistical Bayesian analysis.

2.3.4 Spatial multiple and multivariate results for three adiposity measures.

The three adiposity measures were correlated such that: BMI and waist circumference ($r=0.2381, p<0.001$); BMI and body fat mass ($r=0.4367$, $p<0.001$) and waist circumference and body fat mass ($r=0.2697, p<0.001$). Spatial investigations using multiple variable and multivariate models showed more variability (12% vs 47% partial-sill-to-sill ratio), smaller nuggets and slower decay with the latter multivariate models.

Table 2.3: Body mass indices descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results

Descriptive measures		Single outcome models						Several outcomes models					
Variable and categories	**Summary statistics**	**Univariate regression Coefficient (95%CI)**			**Spatial multiple regression Coefficient (95%BCI)**			**Multivariate multiple regression Coefficient (95%BCI)**			**Spatial multivariate multiple regression Coefficient (95%BCI)**		
Sex - Male	328(50.62%)	ref			ref			ref			ref		
Female	320(49.38%)	-0.47*	-0.75	-0.18	-0.37*	-0.60	-0.15	-0.4*	-0.64	-0.16	-0.38*	-0.42	-0.34
Age in years	3.15±1.11	-0.54*	-0.67	-0.42	-0.49*	-0.59	-0.39	-0.04	-0.15	0.07	-0.48*	-0.50	-0.46
Wasted - No	604(93.21%)	ref			ref			ref			ref		
Yes	44(6.79%)	-3.84*	-4.34	-3.34	-2.41*	-2.97	-1.84	-2.61*	-3.21	-2.01	-2.54*	-2.64	-2.43
Stunted - No	537(82.87%)	ref			ref			ref			ref		
Yes	111(17.13%)	0.72*	0.34	1.11	0.86*	0.53	1.19	1.12*	0.77	1.47	0.85*	0.79	0.92
Underweight -No	582(89.81%)	ref			ref			ref			ref		
Yes	66(10.19%)	-2.50*	-2.94	-2.05	-1.66*	-2.16	-1.17	-1.69*	-2.22	-1.16	-1.69*	-1.78	-1.60
Socio-economic status - *Most poor*	134(20.68%) ref				ref			ref			ref		
Very Poor	129(19.91%)	0.05	-0.41	0.51	0.10	-0.25	0.46	0.03	-0.34	0.41	0.11*	0.05	0.18
Moderately poor	128(19.75%)	0.09	-0.37	0.55	0.25	-0.11	0.61	0.16	-0.21	0.53	0.25*	0.19	0.32
Poor	129(19.91%)	0.15	-0.31	0.61	0.25	-0.11	0.61	0.15	-0.22	0.53	0.26*	0.19	0.33
Least poor	128(19.75%)	0.37	-0.09	0.83	0.41*	0.05	0.77	0.32	-0.05	0.7	0.38*	0.31	0.45
Spatial variance					0.26*	0.14	0.48				0.58*	0.21	1.21
Nugget					1.92*	1.69	2.17				0.66*	0.15	1.43
Spatial decay					1.03*	0.32	2.89				$3.02*x10^{-3}$	$2.99x10^{-3}$	$4.10x10^{-3}$

*:-significant at the 5% level; ref:-reference (coefficient equal to 0); CI:-confidence interval, BCI:-Bayesian credible interval

.

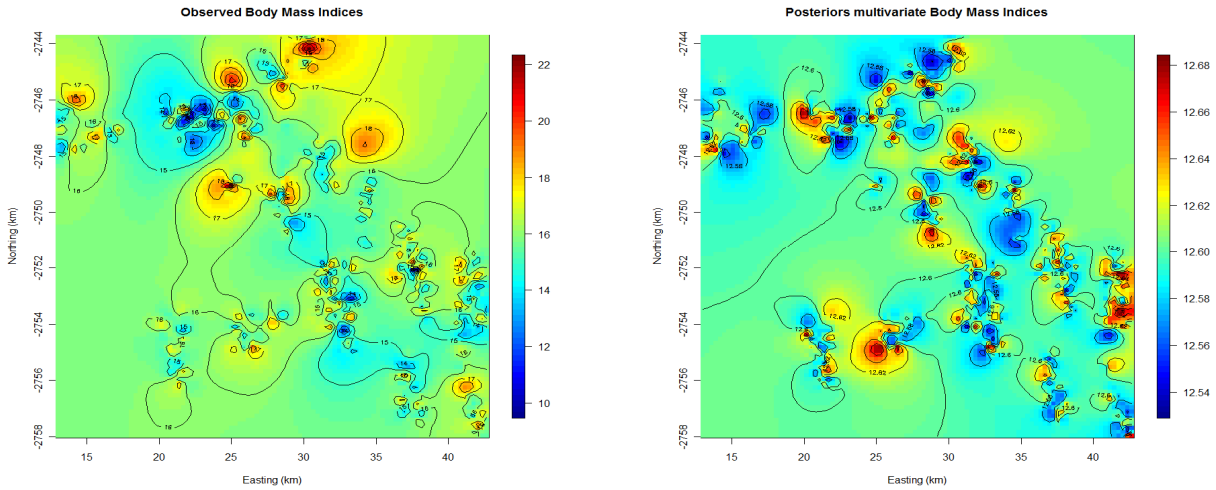


Figure 2.6: Observed (left) and spatial multivariate posterior (right) maps and contour for body mass indices for children in Agincourt 2007.

Several factors were significantly associated with a worsening BMI: being female, older child age, childhood wasting, being underweight and better socio-economic status (table 2.3). Figure 2.6 (left hand side) shows the observed BMIs of two northern areas, with significantly higher and one region north-easterly had significantly low BMI. These patterns are consistent with the low spatial autocorrelations of BMIs highlighted in section 2.3.2. The BMIs posterior maps figure 2.6 (right) adjusting outcomes: waist circumference and percentage body fat and for covariates: child's gender, age, wasting, underweight and better socio-economic status. This indicated several areas predisposed to lower (blue areas – cold spots) BMI related adiposity compared to those with high adiposity (red to yellow areas – hot spots).

Table 2.4: Waist circumferences descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results

Descriptive measures		Single outcome models						Several outcomes models					
Variable and categories	**Summary statistics**	**Univariate regression Coefficient (95%CI)**			**Spatial multiple regression Coefficient (95%BCI)**			**Multivariate multiple regression Coefficient (95%BCI)**			**Spatial multivariate multiple regression Coefficient (95%BCI)**		
Sex - Male	328(50.62%)	ref			ref			ref			ref		
Female	320(49.38%)	-7.97*	-13.10	-2.85	-9.17*	-13.7	-4.63	-8.73*	-13.53	-3.91	-9.17*	-9.28	-9.06
Age in years	3.15±1.11	11.10*	8.93	13.27	10.03*	7.96	12.1	3.26*	1.07	5.44	10.03*	9.98	10.09
Wasted - No	604(93.21%)	ref			ref			ref			ref		
Yes	44(6.79%)	-17.87*	-28.03	-7.71	-3.33	-14.61	8.1	-2.04	-14.15	10.09	-3.28*	-3.55	-3.00
Stunted - No	537(82.87%)	ref											
Yes	111(17.13%)	-22.09*	-28.72	-15.46	-10.42*	-17.12	-3.77	-15.82*	-22.81	-8.8	-10.42*	-10.58	-10.26
Underweight -No	582(89.81%)	ref			ref			ref			ref		
Yes	66(10.19%)	-29.65*	-37.86	-21.44	-22.61*	-32.64	-12.68	-22.78*	-33.4	-12.19	-22.61*	-22.85	-22.38
Socio-economic status - *Most poor*	134(20.68%)	ref			ref			ref			ref		
Very Poor	129(19.91%)	12.31*	4.28	20.34	8.29*	1.17	15.42	9.55*	2.02	17.11	8.29*	8.11	8.46
Moderately poor	128(19.75%)	12.20*	4.14	20.24	8.29*	1.05	15.45	9.61*	2.03	17.17	8.27*	8.10	8.45
Poor	129(19.91%)	12.39*	4.35	20.42	6.47	-0.69	13.63	7.7*	0.09	15.34	6.47*	6.29	6.64
Least poor	128(19.75%)	9.16*	1.11	17.21	5.42	-1.68	12.54	5.84	-1.72	13.41	5.41*	5.24	5.59
Spatial variance					0.64*	0.18	4.03				0.71*	0.28	1.20
Nugget					855.50*	768.46	956.95				775.40*	0.38	908.64
Spatial range					1.75*	0.39	2.96				$3.00*x10^{-3}$	$2.99x10^{-3}$	$3.00x10^{-3}$

*:-significant at the 5% level; ref:-reference (coefficient equal to 0); CI:-confidence interval, BCI:-Bayesian credible interval

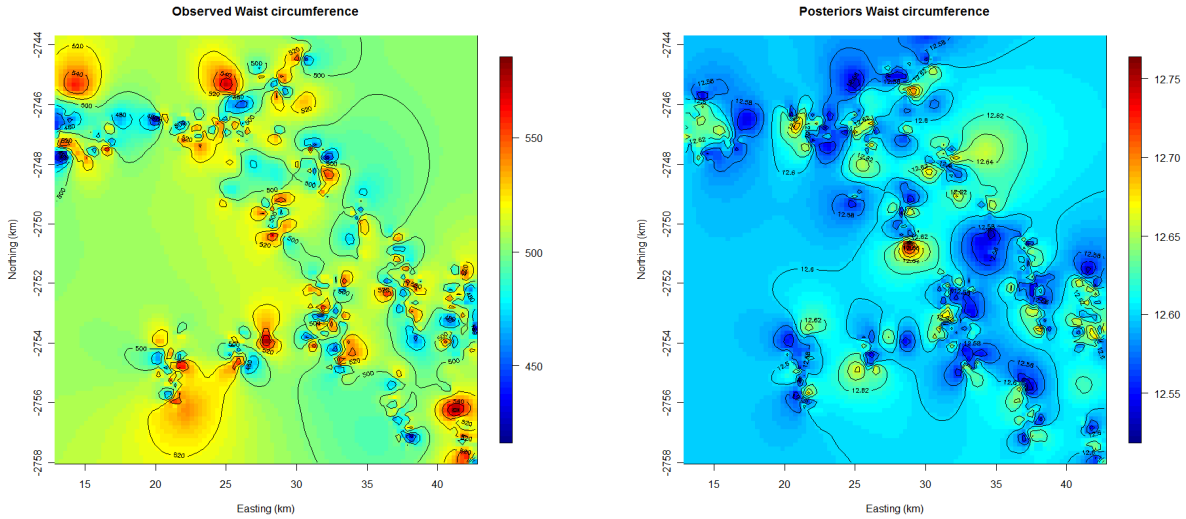


Figure 2.7: Observed (left) and spatial multivariate posterior (right) maps and contour for waist circumferences for children in Agincourt 2007.

Similarly for waist circumference had several factors significantly associated with a worsening adiposity such as being female, elderly child age, childhood wasting, being underweight and better socio-economic status, shown in table 2.4. Figure 2.7 (left hand side) shows the observed waist circumferences show some north-easterly and south-westerly regions with significantly moderate to high adiposity measures. However the prediction posteriors map in Figure 2.7 (right) shows the predicted waist circumferences show several north-easterly and south-westerly regions with significantly lower adiposity measures. It appears that waist circumference adjusting for outcomes BMI and percentage body fat and for covariates: child's gender, age, wasting, underweight and better socio-economic status shows severely lower adiposity in the region.

Table 2.5: Body fat mass descriptive statistics, Univariate, multiple variable and spatial multiple and multivariate results

Descriptive measures		Single outcome models						Several outcomes models					
Variable and categories	**Summary statistics**	**Univariate regression Coefficient (95%CI)**			**Spatial multiple regression Coefficient (95%BCI)**			**Multivariate multiple regression Coefficient (95%BCI)**			**Spatial multivariate multiple regression Coefficient (95%BCI)**		
Sex - Male	328(50.62%)	ref			ref			ref			ref		
Female	320(49.38%)	-6.27*	-7.44	-5.10	-6.12*	-7.22	-5.02	-6.21	-7.36	-5.04	-6.12*	-6.17	-6.07
Age in years	3.15±1.11	-2.17*	-2.72	-1.62	-2.12*	-2.63	-1.62	-0.02	-0.55	0.51	-2.11*	-2.13	-2.09
Wasted - No	604(93.21%)	ref			ref			ref			ref		
Yes	44(6.79%)	-4.63*	-7.12	-2.13	-1.88*	-4.63	0.91	-2.22	-5.12	0.71	-1.86*	-1.97	-1.73
Stunted - No	537(82.87%)	ref			ref			ref			ref		
Yes	111(17.13%)	1.32	-0.36	3.00	0.41	-1.22	2.03	1.59	-0.11	3.27	0.39*	0.32	0.46
Underweight -No	582(89.81%)												
Yes	66(10.19%)	-3.12*	-5.20	-1.03	-3.01*	-5.45	-0.58	-2.99	-5.55	-0.43	-2.99*	-3.10	-2.89
Socio-economic status - *Most poor*	134(20.68%)	ref			ref			ref			ref		
Very Poor	129(19.91%)	0.41	-1.59	2.40	0.30	-1.44	2.04	-0.07	-1.89	1.75	0.32*	0.25	0.40
Moderately poor	128(19.75%)	1.71	-0.29	3.70	2.02*	0.25	3.76	1.55	-0.29	3.38	2.00*	1.92	2.07
Poor	129(19.91%)	0.27	-1.72	2.27	0.55	-1.2	2.29	0.06	-1.78	1.92	0.58*	0.51	0.66
Least poor	128(19.75%)	0.64	-1.35	2.64	0.59	-1.15	2.33	0.27	-1.57	2.1	0.59*	0.52	0.66
Spatial variance					0.72*	0.2	3.33				39.04*	0.25	49.02
Nugget					49.9*	44.37	55.94				1.83*	0.24	4.09
Spatial range					1.66*	0.48	2.95				$3.00*x10^{-3}$	$2.99x10^{-3}$	$3.00x10^{-3}$

*:-significant at the 5% level; ref:-reference with coefficient equal to 0; CI:-confidence interval, BCI:-Bayesian credible interval

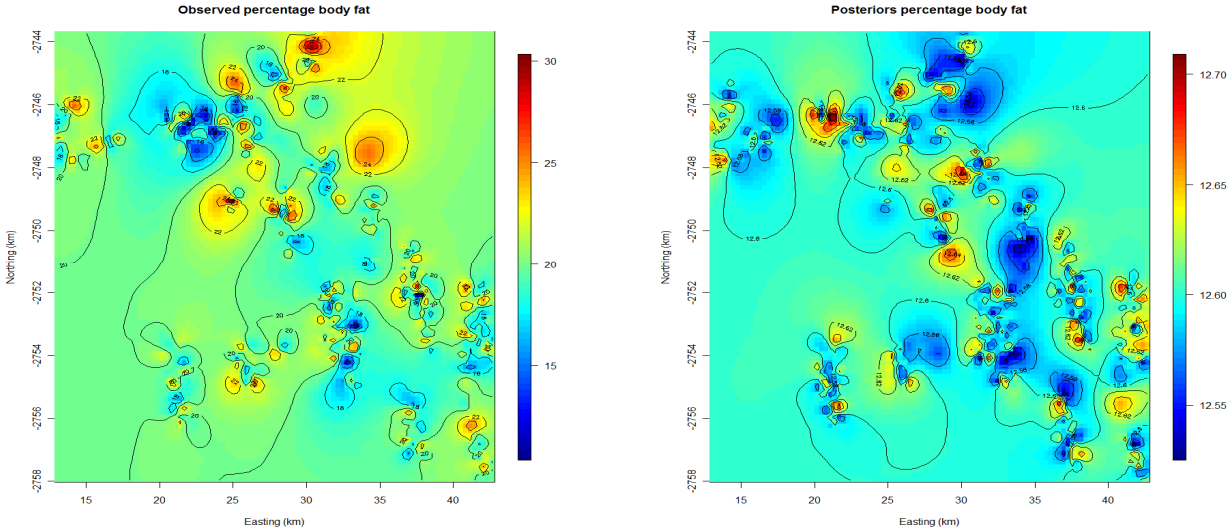


Figure 2.8: Observed (left) and spatial multivariate posterior (right) maps and contour for body fat masses for children in Agincourt 2007.

The observed percentage body fat spatial distribution for hot spots was very similar to that of the BMIs, which also had the better pair-wise correlations. However they generally are less clustered (more random) than the BMIs as shown by the insignificant Moran's and Geary's indices in table 2.2. The percentage body fat adjusted (for BMIs and waist circumference) posterior map shows several cold spots adjusting for covariates: child sex, age, wasting, underweight, malnutrition and socio-economic status.

2.4 Discussion

In this paper we demonstrated that there are differences in spatial variability on three adiposity measures for children aged 1-5 years in a small rural area in South Africa. The multivariate analysis adjusting for the other similar outcome measures as well as covariates indicated that that gender, age, malnutrition indicators (underweight, stunted and wasted) and socio-economic status. Similar findings have also been found in other

studies (141-144), however they did not adjustment for both outcomes and covariates whilst fitting a spatial random effect model. This is an emerging discipline in epidemiological studies.

It is evident in this area that there are severe levels of adiposity mostly for those children residing north-easterly and also on the south-westerly parts. This information can be used to inform how interventions for the area are administered to obtain the most effective public health impact. Such knowledge can also better the manner in which analysis is done and how morbidity/mortality data are reported for this area and beyond. Often regional (country-specific) estimates are reported which mask the true estimates resulting in under or over estimation of diseases and mortality.

We also demonstrated that semi-variogram fitting to data is an interactive process where the modeller has to visually assess how well the theoretical models fit to their data. Often all parameters may not be estimated well enough and sometimes stricter assumptions have to be made to settle for a given theoretical model. The technique we used to estimate then fix the spatial range parameter has been favoured by some authors because often convergence of the range is very difficult to attain (145). This is because its' very sensitive to the fitted theoretical distribution, as our graphical semi-variogram analysis disclosed.

2.5 Conclusion

Geospatial analyses maps can guide to further in-depth studies to determine major causes of adiposity in areas of high HIV/AIDS, related conditions and poverty. With

less than three years to the end of the MDGs (in 2015) there is need for more accurate reporting of morbidity and mortality in member countries as this impacts on allocation of global interventions and accuracy of burden of diseases.

Epidemiological modelling catering for spatial random effects has been getting increasingly more complex due to the nature of the data (large) and interrelatedness of outcomes and explanatory variables. Our study has demonstrated that a univariate as well as multivariate model can be used and the results will be comparable. Lowrank *kriging* is a potentially very useful technique to handle "large" spatial dataset it performed considerably well and yielded similar results to estimations based on all geo-locations (global *kriging*). When several correlated outcomes are in consideration spatial multivariate techniques that can handle such relations are available using software from the public domain. This allows us to model biological or natural processes in accordance with their original state properties. In conclusion we propose that geo-statistical modelling has a great potential in public health research and can be effectively used by researchers to inform policy makers and on high and low risk areas for interventions.

2.6 Authors' contributions

EM drafted the manuscript and did the statistical analysis. EK-M and KK were responsible for the design, primary data collection and data quality of the Agincourt HDSS.

2.7 Acknowledgements

This research was made possible with a technical assistance grant from the South African Medical Research Council (MRC): Research Capacity Development; Sub-Directorate 2011-2; special thanks goes to Dr. Thabi Maitin. We thank Paul Mee and Suleiman Afolabi from the MRC/Wits Agincourt Unit for geo-linking our data.

2.8 Competing interests

The authors declare that they have no competing interests.

Chapter 3

3.0 Spatially adjusted determinants of childhood malnutrition in rural north-east South African households: a multivariate and structural equation modelling approach

Eustasius Musenge[1,2], Elizabeth Kimani-Murage[1,5], Penelope Vounatsou[4], Shane A. Norris[3] and Kathleen Kahn[1,6,7]

1. MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

2. Biostatistics and Epidemiology Division, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

3. MRC/WITS Developmental Pathways for Health Research Unit, Faculty of Health Sciences, University of the Witwatersrand Medical School

4. Swiss Tropical and Public Health Institute, Basel, Switzerland

5. African Population and Health Research Centre, Nairobi, Kenya

6. Centre for Global Health Research, Umeå University, Umeå, Sweden

7. INDEPTH Network, Accra, Ghana

This paper has been re-submitted to the American Journal of Clinical Nutrition.

Chapter three abstract

Background: The extent of geographical disparities on childhood malnutrition and effects of HIV has hardly been explored, yet it is a reservoir of information for effective public health interventions.

Objective: Our primary aim was to determine factors associated with two malnutrition outcomes (weight-for-age and height-for-age) adjusting for their multivariate spatial random effects. Secondarily, pathways linking independent factors, HIV and malnutrition outcomes, were examined. We tested the hypothesis: "HIV retards child growth controlling for other factors and household spatial random effects".

Design: We analyzed data from a cross-sectional anthropometric survey of 648 children aged 1 – 5 years collected in 2007. We employed a Bayesian multivariate spatial model to simultaneously analyze the two indicators of child malnutrition. To assess the mediator impact of HIV on correlated malnutrition outcomes and other explanatory variables, structural equation models were used.

Results: Spatial multivariate adjusted analyses revealed a negative mean effect due to HIV and protective effects with older aged children and mothers, and maternal education. Pathway analyses of these factors showed that HIV had a significant mediator effect and the greatest worsening effect of -0.55 (95% CI: -1.09,-0.02) and -0.61(95% CI: -0.98,-0.24) for weight-for-age and height-for-age z-scores respectively.

Conclusions: Child malnutrition interventions can be more effective if there is a dual focus on those HIV infected and the geographical areas of greatest risk. Public health population level interventions aimed at reducing child malnutrition are pivotal in lowering morbidity and mortality in remote areas.

3.1 Introduction

Childhood malnutrition is a good proxy for child wellbeing and is associated with mortality, morbidity and child development (112). Malnutrition accounts for more than half of the 10 million child deaths yearly worldwide (146). HIV and malnutrition are major problems in children of the sub-Saharan Africa region. The impact of HIV/AIDS on child malnutrition and vice versa is of paramount concern in the public health of developing nations.

Knowledge of inter-linkages between the HIV virus and the child's nutritional status may help focus interventions in different geographic regions of the globe. South Africa has more than two million AIDS orphans (2). Malnutrition indicators in South Africa a 2008, ranged from 8.7% to 23.9 % for children underweight and stunted respectively (147).

The association between HIV and malnutrition is two-way, usually modelled as a vicious cycle. Malnutrition affects the human immune system leading to depletion of disease fighting cells (CD4, T-cells), increased viral load, reduction of essential macronutrients (proteins, fats and carbohydrates) and micronutrients (vitamins A, B, C and zinc) (10, 12-15, 148). HIV on the other hand affects nutrition by loss of appetite, drug side effects such as vomiting and diarrhoea and nutrient mal-absorption (14, 16, 17, 149, 150).

Several studies have demonstrated that anthropometric indices (weight loss, stunting and wasting) were associated with increased risk of opportunistic infections and lower life expectancy regardless of the infection staging (151-157). Analyses of child growth using the anthropometric indicators, height adjusted for age (HAZ), weight adjusted for height (WHZ) and weight adjusted for age (WAZ), are usually carried out independently. Human growth is however multidimensional, the child grows in height, age and weight simultaneously, yet analytic procedures often ignore this correlation.

Geographical variation in child growth has been investigated in the past ignoring the impact of HIV (111, 113, 158). Limited earlier work has also looked at child growth using multivariate Bayesian spatial areal models known as the multivariate conditional autoregressive (MCAR) (159).

To our knowledge no single study has simultaneously examined how malnutrition risk factors interlink (correlations, direct and indirect effects) as well as adjusted for household spatial random effects, hence this study is among the first to do so.

In this paper our primary aim was to determine factors associated with two malnutrition outcomes -WAZ and HAZ - adjusting for their multivariate spatial random effects. Secondarily our aim was to determine the pathways linking the significant factors, HIV/AIDS and malnutrition outcomes.

3.2 Materials and methods

3.2.1 Study site and anthropometric survey description

This study was conducted in rural northeast South Africa, in the Agincourt sub-district, Mpumalanga province, alongside the country's border with Mozambique. Agincourt is a semi-arid setting, situated in the former Gazankulu homeland. The study was nested within the Agincourt health and socio-demographic surveillance system (HDSS), of the University of the Witwatersrand, which was established in 1992 covers the entire Agincourt sub-district. The Agincourt HDSS follows some 80,000 people living in 12,500 households in 27 contiguous villages. Health care services are limited in the area: a network of five primary care clinics refers to a larger public health centre; the nearest district hospital is 25 kilometers away (71).

The study is a secondary data analysis based on data collected between April and July 2007 from 4658 children and adolescents aged between 1 and 20 years, randomly selected from the Agincourt HDSS. The sampling targetted at least 100 children in each sex and year of age. Our analysis was done on 648 children aged 1-5 years; figure 3.1 shows how the sample was obtained.

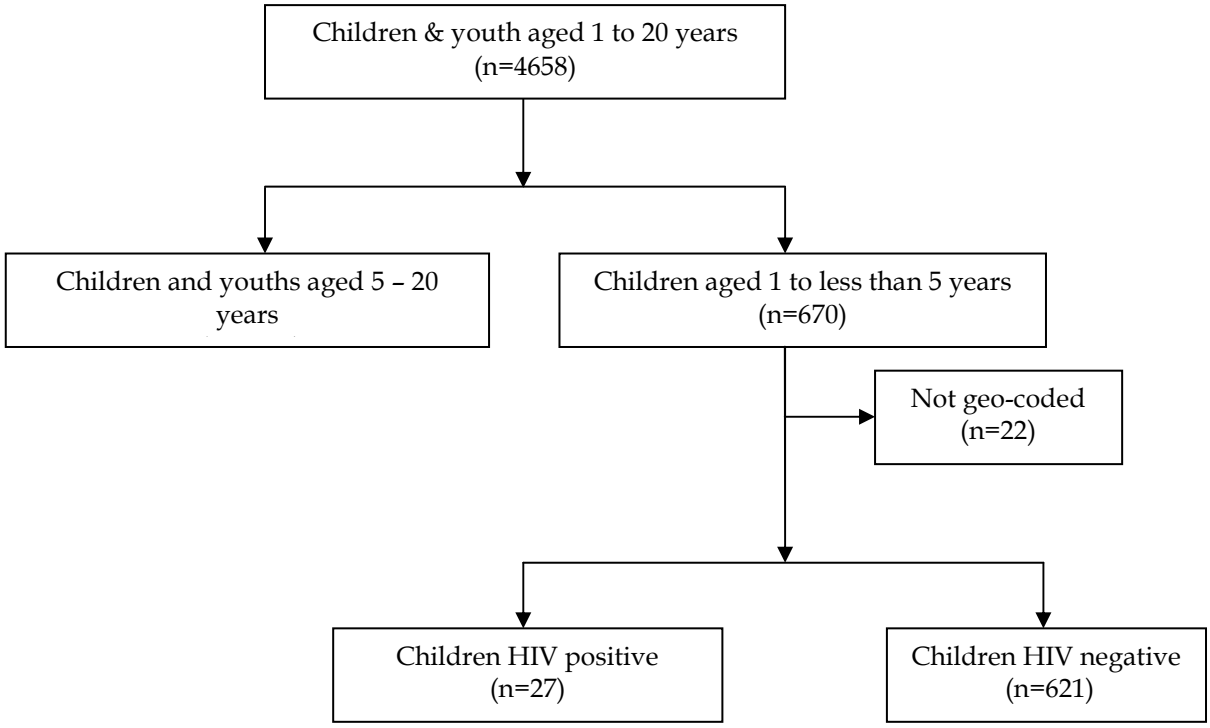


Figure 3.1: Flow diagram of anthropometric survey participants, Agincourt sub-district 2007.

Anthropometric measurements included height and weight: height was measured using a stadiometer (Holtain, UK) calibrated in millimeters. For all children aged less than 24 months, length was measured using an inelastic tape measure in a recumbent position on a flat surface. Weight in kilograms (to one decimal point) was determined using a mechanical bathroom scale (Hanson, UK). HIV testing of children aged 1-5 years was

carried out using two concurrent rapid tests: Uni-Gold™ (Trinity Biotech, Bray, Ireland) and Determine™ (Abbott, Wiesbaden, Germany), with pre- and post-test counselling. Informed consent was obtained from the parent/caregiver for HIV testing of children aged 1-5 years. Ethical clearances were granted by the University of the Witwatersrand Committee for Research on Human Subjects, Medical (M070244 and M081145). Details of the sampling, data collection procedures and HIV testing are provided elsewhere (71, 72, 117). This paper uses data from 648 children aged 1-5 years, each coming from a different household.

3.2.2 Outcomes and Covariates

Our main outcomes were WAZ and HAZ malnutrition indices, generated using the World Health Organization (WHO) 2006 growth standards with the WHO Anthro 2005 program, Beta Version (160). Several explanatory variables were chosen with the notion that they could be confounders, mediators or effect modifiers. These variables were either child specific factors (age in years, gender, HIV status, breastfeeding status, nationality and place of birth), maternal factors (age group, highest education attained, reported HIV status), or household-level factors (access to water, toilet and electricity, socioeconomic status and geo-locations). Geographical longitudes and latitudes were used in approximating the multivariate spatial random effects for the households.

3.2.3 Multivariate geo-statistical Bayesian model

We employed a Bayesian multivariate spatial model to simultaneously analyse two indicators of child malnutrition, WAZ and HAZ; these also provide useful indications of the multi-dimensional growth patterns of the human body and are often modelled singularly. We

modelled the two outcomes of child malnutrition jointly assuming they follow a Multivariate Normal distribution

$$\boldsymbol{Y}_{ik} \sim MVN(\boldsymbol{\mu}_{ik}, \Sigma)$$

with mean $\boldsymbol{\mu}_{ik} = \underset{\sim}{\beta}_j \underset{\sim}{\boldsymbol{X}}_{ik}^T$ where $\boldsymbol{Y}_{ik} = (y_{ik1}, y_{ik2})^T$, y_{ikj} corresponds to nutritional indicator related to WAZ (j=1) and HAZ (j=2) for child $i = 1,..,n$ at household $k = 1,..,m$. The matrix $\Sigma = \{\sigma_{mn}\}$ models the variances and covariances of the nutritional indicators with $\sigma_{mn} = \text{cov}(Y_{im}, Y_{in})$. The parameter $\underset{\sim}{\beta}_j = (\beta_{j1}, \beta_{j2}, ...\beta_{jp})^T$ is the vector of regression coefficients associated with the jth score and $\underset{\sim}{\boldsymbol{X}}_{ik} = (\mathbf{1}, X_{ik}^{(1)}, X_{ik}^{(2)}, ..., X_{ik}^{(p)})$ is the vector of *p* covariates.

To take into account spatial correlation and random effects we introduce household-specific random effects, ϕ_{kj} and ε_{kj} on the mean structure, that is $\boldsymbol{\mu}_{ikj} = \underset{\sim}{\beta}_j \underset{\sim}{\boldsymbol{X}}_{ik}^T + \phi_{kj} + \varepsilon_{kj}$ assuming that they are latent data from an unobserved isotropic spatial Gaussian processes that is

$\underset{\sim}{\phi} \sim MVN[\mathbf{0}, K(\underset{\sim}{\theta}; s, s')]$ with vector of spatial decay parameters $\underset{\sim}{\theta} = [\theta_j]_{j=1}^{q=2}$,the qn x qn variance-covariance matrix $K(\theta; s, s') = \sigma_{mn}[\oplus_{j=1}^{q=2} \rho_j(s, s'; \theta_j)]\sigma_{mn}^T$ which should be symmetric and positive definite.

The random (nugget) effects from normal distributions, $\underset{\sim}{\varepsilon} \sim MVN[\mathbf{0}, \Psi]$

where $$\Psi = \begin{pmatrix} \tau_{1,1}^2 & \tau_{1,2}^2 \\ \tau_{2,1}^2 & \tau_{2,2}^2 \end{pmatrix}$$

To complete Bayesian specification we adopt the inverse Wishart prior $K^{-1} \sim Wish(\Omega^{(K)}, 2)$ for the variance-covariance matrix, non informative Normal priors $\beta_{0j} \sim N(\mu_j^{(\beta)}, \sigma_j^{2(\beta)})$ for

the regression coefficients, Uniform priors for the spatial decays $\theta_j \sim U\left(\frac{1}{\max(d_{ij})}, \frac{1}{\min(d_{ij})}\right)$ and an inverse Wishart prior for the matrix of non spatial random effects $\Psi^{-1} \sim Wish(\Omega^{(\Psi)}, 2)$.

3.2.4 Structural equation modelling

To assess the mediatory impact of HIV on the correlated malnutrition outcomes and other significant explanatory variables, structural equation models were used. Variables which were used as explanatory (exogenous) were mother's age group, mother's education and child's age. Child HIV status was used as both explanatory and outcome (endogenous), whilst the malnutrition indices WAZ and HAZ were endogenous and correlated among them. Three structural equations were fitted with each of the endogenous variables as outcomes together with correlations of some exogenous and endogenous variables. Goodness of global model fit was determined by the chi-square statistic, testing the null hypothesis that the reproduced covariance matrix has the specified structure (model fits the data well) (161).

3.2.5 Bayesian inference and statistical software used

The model fitting was done using Monte Carlo Markov Chain (MCMC) via Gibbs sampling and Metropolis Hastings within Gibbs sampling when the conditionals were numerically intractable (162) (85). Bayesian-based model fitting involved the inversion of a spatial covariance matrix with the same size as the number of geo-locations (households). We did this directly for all our 648 households. A total of 55,000 iterations were done with the first 5000 discarded as burn-in, and to reduce auto-correlation every 25th value was taken to form a posterior sample of 2000. Model fitting was done in R version 2.12.2 (163) using the package spBayes (164) and preliminary analyses and structural equation modelling was done

in STATA 12.0 (134). The multivariate spatial model took 15 days to run on an 8.0 gigabits random access memory (RAM) Intel i7 processor in the Windows environment.

3.3 Results

3.3.1 Descriptive characteristics of the survey sample

Ages were symmetrically distributed and there was balance in the sexes, maternal age groups and socio-economic status (Table 3.1). The majority had breastfed, were HIV negative, of South African origin and born in a health facility. Our investigation of the unsmoothed spatial distribution of those severally underweight (weight-for-age z-score less than -2), stunted (height-for-age z-score less than -2) and HIV positive is summarised in figure 3.2. This figure shows that there is a geographical scatter of malnutrition indicators and HIV of children living in the rural Agincourt sub-district. Almost 10% of the children were underweight; 11% boys and 9% girls. Effect of HIV on being underweight was greater in boys (20%) than girls (11%). More than 17% of the children were stunted, a pooled estimate from 19% boys and 15% girls. HIV had an impact on how tall the children were growing in relation to their age, such that 30% of the boys and 28% of the girls were HIV positive and stunted. The proportion of those underweight was slightly higher than the national level (10% vs 8.7%) whereas those stunted was lower (17% vs 23.9%).

Table 3.1: Bayesian spatial multivariate regression results for malnutrition weight-for-age and height-for-age Z-score outcomes.

	Summary statistics N=648		**Spatial Multivariate results Weight-for-Age Z-scores**			**Spatial Multivariate results Height-for-Age Z-scores**		
Variable	**Category**	**Descriptive**	**2.50%**	**50%**	**97.50%**	**2.50%**	**50%**	**97.50%**
Constant			-1.38	-0.83*	-0.28	-2.51	-1.94*	-1.35
Age (years)[1]		3.14±1.11	-0.09	-0.03	0.03	0.14	0.21*	0.28
Gender[2]	*Male*	326(50.37)						
	Female	322(49.63)	-0.18	-0.05	0.08	-0.04	0.11	0.26
Breastfeeding	*No*	24(3.70)						
	Yes	607(93.67)	-0.41	-0.05	0.33	-0.49	-0.08	0.32
HIV status child	*Negative*	621(95.83)						
	Positive	27(4.17)	-1.23	-0.75*	-0.27	-1.30	-0.78*	-0.28
Child nationality	*South African*	419(64.66)						
	Mozambican	229(35.34)	-0.17	0.06	0.29	-0.20	0.05	0.30
Mother's age group	*15-24*	209(32.25)						
	25-34	257(39.66)	0.07	0.23*	0.40	0.10	0.28*	0.46
	35+	178(27.47)	0.07	0.26*	0.44	0.04	0.24*	0.45
Mother's HIV status	*Negative*	624(96.59)						
	Positive	22(3.41)	-0.13	0.26	0.64	-0.30	0.12	0.53
Mother's nationality	*South African*	418(64.51)						
	Mozambican	219(33.80)	-0.18	0.06	0.30	-0.21	0.06	0.33
Education level mother	*None*	75(11.57)						
	Below metric	356(54.94)	0.01	0.25*	0.49	-0.16	0.09	0.35
	Metric and higher	146(22.53)	0.19	0.47*	0.74	-0.02	0.28	0.57
Type of dwelling	*Non slum*	396(61.11)						
	Slum	252(39.89)	-0.24	-0.08	0.08	-0.22	-0.04	0.14
Birth place	*Health facility*	459(70.68)						
	Home	85(13.12)	-0.13	0.08	0.28	-0.12	0.10	0.33
	Unknown	98(15.12)	-0.31	-0.11	0.08	-0.33	-0.12	0.09
Socio-economic status	*Low*	216(33.33)						
	Middle	206(31.79)	-0.37	-0.08	0.21	-0.34	-0.04	0.26
	High	170(26.23)	-0.24	0.05	0.34	-0.22	0.08	0.39
Multivariate spatial vector estimates			**2.50%**	**50%**	**97.50%**			
Spatial variance WAZ	$\sigma^2_{waz,waz}$		0.04	0.23*	0.43			
Spatial covariance WAZ, HAZ	$\sigma_{waz,haz}$		0.02	0.42*	0.65			
Spatial variance HAZ	$\sigma^2_{haz,haz}$		0.05	0.85*	1.11			
Spatial nugget WAZ	$\tau^2_{waz,waz}$		0.04	0.48*	0.74			
Spatial combined nugget WAZ, HAZ	$\tau_{waz,haz}$		-0.13	-0.07	0.03			
Spatial nugget HAZ	$\tau^2_{haz,haz}$		0.05	0.08*	0.11			
Spatial decay WAZ	$\rho_{waz} \times 10^{-3}$		2.99	3.00*	3.01			
Spatial decay HAZ	$\rho_{haz} \times 10^{-3}$		3.00	3.02*	3.03			

Key: * significant at the 5% level that is 95% credible interval not containing zero, [1]mean ± standard deviation, [2]frequency (percentage), 2.50% lower 95% Bayesian credible band and 97.50% upper 95% Bayesian credible band

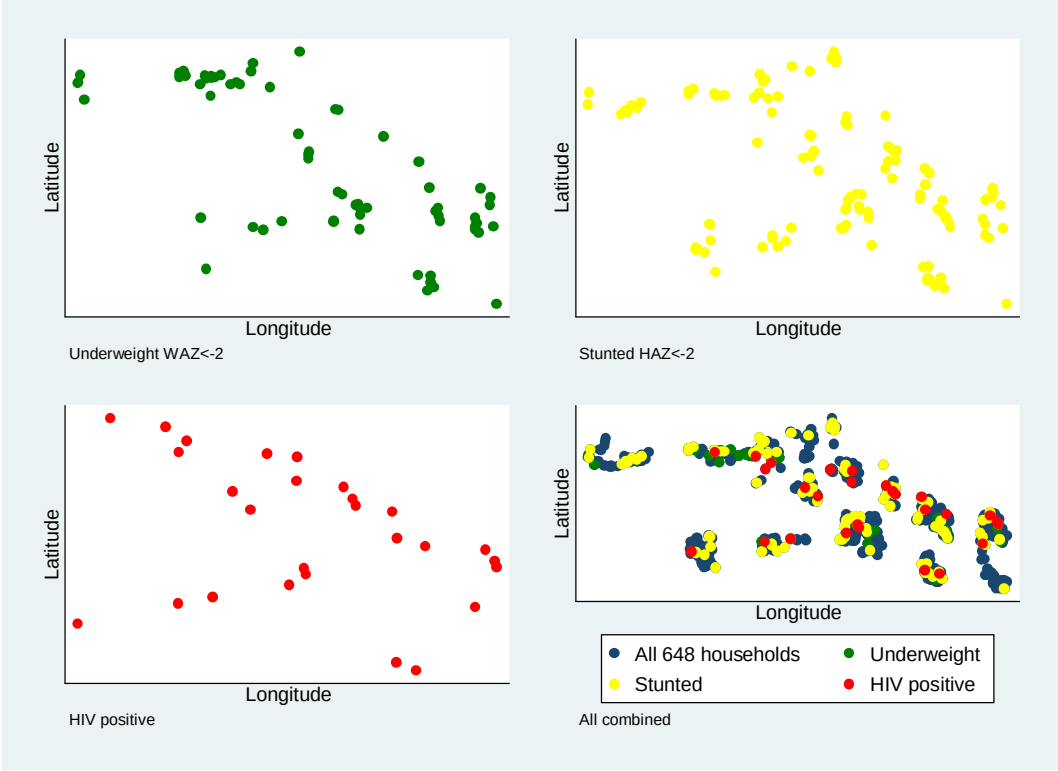


Figure 3.2: Spatial distribution of underweight, stunting and HIV positive children in Agincourt rural South Africa.

3.3.2 Multivariate spatially adjusted determinants of childhood malnutrition and pathway analyses results

We investigated the presence of spatial variability and found significant malnutrition indicators (WAZ and HAZ) and their interaction, with the greatest on the height-for-age z-score (Table 3.1). Spatial decay was very similar for the two malnutrition indicators in this area. Non spatial (nugget) variability was only present for the individual malnutrition levels but the interaction was not significant.

Spatial adjusted multivariate regression results with weight-for-age and height-for-age z-scores as outcomes showed significant mean decreases due to HIV and protective effects with

much older children and women and those whose mothers were educated as shown in Table 3.1. To find out how the significant determinants were inter-linked to each other and how they were related to the outcomes, we performed pathway analyses. The results are shown in **Table 3.2** and **Figure 3.3.**

Table 3.2: Path analyses structural estimation results for variables in figure 3.3

	Summary statistics N=648		**Results**		
Structural Outcome	**Independent variable**	**Category**	**mean**	**95% Confidence interval**	
HIV status child (status)	Constant		0.25**	0.17	0.32
	Education level mother (mumedu)	None (ref)			
		Below and above metric	-0.20**	-0.26	-0.13
	Mother's age group (mumagecat)	*15-24(ref)*			
		25+	-0.075**	-0.11	-0.04
Weight-for-Age Z-score(waz2005)	Constant		*-0.86***	-1.08	-0.63
	HIV status child (status)	*Negative(ref)*			
		Positive	*-0.55**	-1.09	-0.02
	Education level mother (mumedu)	None (ref)			
		Below and above metric	*0.17**	0.005	0.34
	Mother's age group (mumagecat)	*15-24(ref)*			
		25+	0.27**	0.078	0.47
Height-for-Age Z-score(haz2005)	Constant		-1.90**	-2.12	-1.58
	HIV status child	*Negative(ref)*			
		Positive	-0.61**	-0.98	-0.24
	Mother's age group (mumagecat)	*15-24(ref)*			
		25+	0.29**	0.098	0.49
	Child's age in years (ageyear)		0.23**	0.16	0.30
Variances	Variance(HIV Status-Ɛ1)		0.033**	0.024	0.044
	Variance(waz2005-Ɛ2)		1.30**	1.14	1.47
	Variance(haz2005-Ɛ3)		1.30**	1.12	1.49
Covariances	Covariance(waz2005- Ɛ2 ,Error haz2005- Ɛ2)		0.77**	0.65	0.89
	Covariance(mumedu ,mumagecat)		-0.026**	-0.040	-0.013
	Covariance(mumagecat ,ageyear)		0.081**	0.04	0.12

Key: **and* significant at the 1% and 5% level that is p-values 0.01 and 0.05 respectively

Pathway analysis revealed a significant mediatory and direct effect of HIV on malnutrition outcomes. A 60% ($\frac{0.77}{\sqrt{1.3\times1.3}}$) correlation existed between the outcome variables and also -13% and 16% among the explanatory variables mother's education and age, and child's age and mother's age respectively.

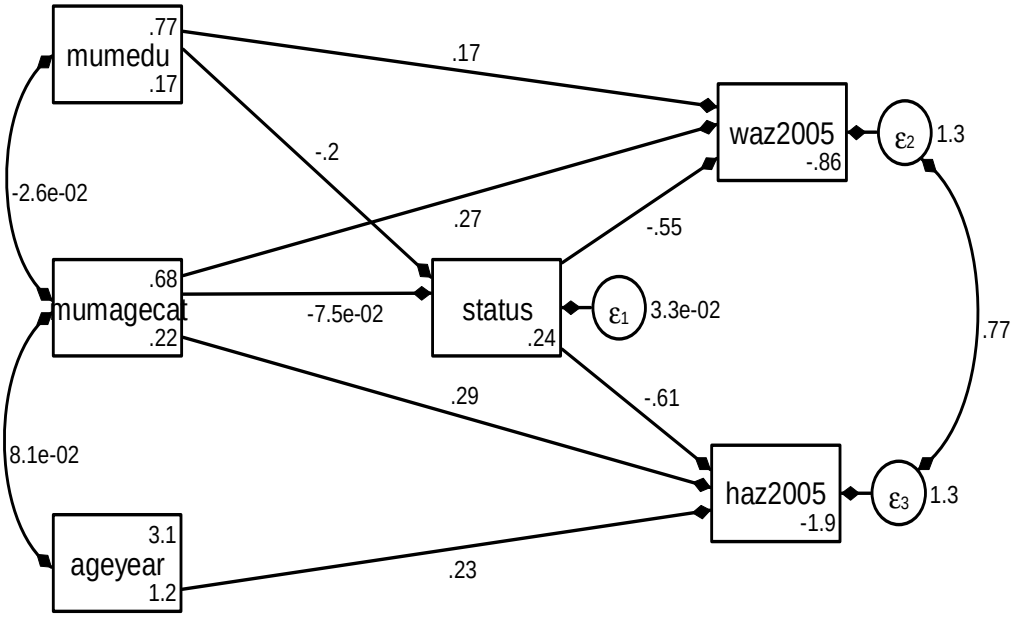


Figure 3.3: Pathway analyses structure of malnutrition outcomes, HIV and maternal factors

We further analysed the data to find out how these significant variables directly or indirectly impacted on the outcomes (Table 3.3). Keeping other variables constant, HIV reduced the mean z-scores by greater than 0.55 and upon adjusting for multivariate spatial random effects the impact was more pronounced to greater than 0.75 (Table 3.1). Maternal education's mean effect on the correlated weight-for-age z-scores was 61% direct and 39% indirect (mediation proportion) whereas mother's age had an 87% direct and 13% indirect impact.

Table 3.3: Direct and indirect effects on malnutrition outcomes for children under five from Agincourt sub-district, rural south Africa.

Outcome	Intermediate	Exogenous	Direct effect (% of Total)	Indirect effect (% of Total)	Total effects
Weight-for-Age Z-score	Childs HIV status (Positive)		*-0.55(100)*	-	-0.55
		Education level mother (Educated)	*0.17(61)*	*0.11(39)*	0.28
		Mother's age group(25 years) plus	*0.28(87)*	*0.04(13)*	0.32
Height-for-Age Z-score	Childs HIV status (Positive)		*-0.61(100)*	-	-0.61
		Education level mother (Educated)	-	*0.12(100)*	0.12
		Mother's age group(25 years) plus	*0.29(84)*	*0.05(16)*	0.34
		Child's age in years	*0.23(100)*	-	0.23

Similarly, on the correlated height-for-age z-scores child's age had a purely direct impact and mother's education purely indirect, whereas mother's age had an 84% direct and 16% indirect effect.

3.3.3 Spatial distribution of malnutrition outcomes

Our results from the multivariate regression and path analyses revealed that HIV had the greatest negative impact on child malnutrition outcomes. From the descriptive map shown in Figure 3.2, those who are HIV positive are scattered all over the area with no apparent visible clustering. However a posterior predictive map Figure 3.4 shows clear patterns of high and low risk areas of underweight and stunting. From these smoothed maps (Figure 3.4), generally stunting is more severe in this area than being underweight; this confirms patterns shown by descriptive statistics and the unsmoothed map in Figure 3.4. The smoothed maps show similar areas of negative impact on childhood malnutrition which are north-westerly area and the diametrically opposite south-easterly. The northern, central and extreme south-

easterly areas indicated more protective effects to underweight, whereas the central area was more protected against stunting.

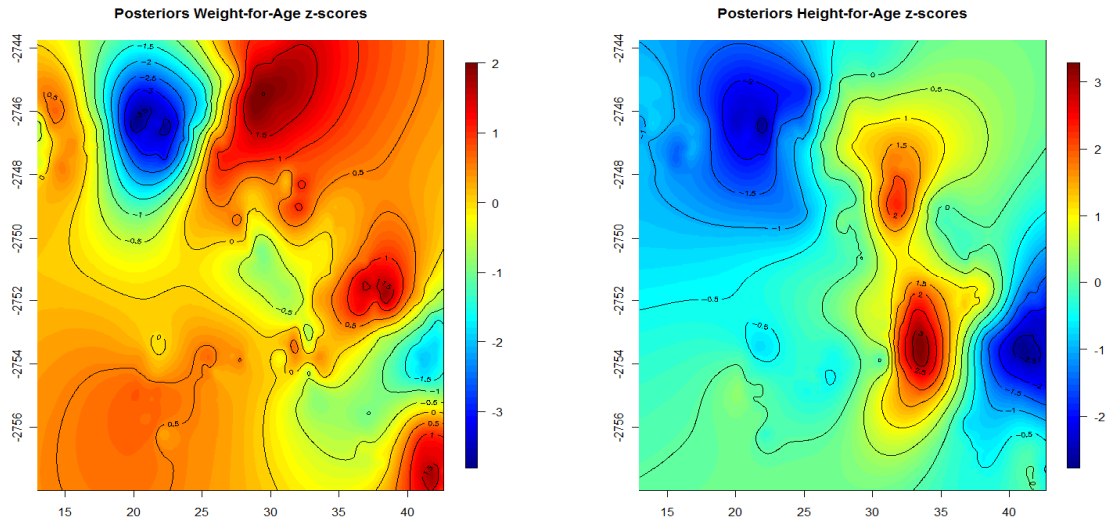


Figure 3.4: Posterior smoothed maps of malnutrition outcomes in Agincourt, rural South Africa

3.4 Discussion

In our study we found that HIV infection was the central direct cause of under-nutrition in rural South African children resulting in retarded growth and weight gain, a finding which is generally well documented in the literature (165-171). This effect was exacerbated when we adjusted for maternal education, maternal age, child's age and spatial random effects. The physiological drivers of this impact area compromised immune system and micro and/or macro nutrient deficient diets (14, 15, 17, 149, 150, 172-174), which are more severe in poor rural settings. Several interventional and observational studies, mainly from urban settings, with higher socio-economic status, demonstrated an improvement in nutritional z-scores upon uptake of antiretrovirals (ARVs) and nutritionally richer diets (175-179). It is noteworthy that

all our study participants were treatment naive as ARVs therapy was only introduced in 2007 in the area.

Our methodological approaches were able to tease out the areas of highest impact and demonstrate how other risk factors indirectly and directly impact on nutritional z-scores, which was not possible by using spatial multivariate regression or structural equation modelling seperately. Several malnutrition and methodological studies have used these modelling techniques (159, 161, 164, 180). However non-have simultaneously disclosed how child malnutrition risk factors interrelate to each other, the mediatory effect of HIV and modeled the seemingly unrelated nature of the outcomes as we have done in our study.

Regardless of the strengths of spatial multivariate and pathway analyses, our study had some limitations. In particular, the cross-sectional study design limited us to discuss associations without extending these to causality. A longitudinal study following individuals over time and then investigating the spatiotemporal patterns and how ARVs have impacted the nutritional status of the cohort is an important future study.

We recommend that for resource-scare areas individual, community level and household based approaches be used to reduce the level of child malnutrition. At community and individual level we advocate that government sources such as child support grants and prevention of mother-to-child transmission (PMTCT) programs be made more accessible especially in areas of high malnutrition. This will directly impact on childhood malnutrition and subsequently on morbidity and mortality. We also observed maternal age and education had both a direct and indirect protective effect on malnutrition, which persisted even after adjusting for spatial random effects. Hence indirect approaches to reducing child malnutrition, such as community health educational campaigns targeting young mothers and

promoting knowledge-sharing between more mature and new mothers, are important. These types of interventions have been effective in resource scarce environments (181-184). A more focused approach, suitable in severely resource-constrained environments, would be to provide health programmes relevant to childhood needs mainly in areas of higher malnutrition. This, coupled with HIV testing of pregnant women, will reduce infections and promote earlier uptake of treatment.

In conclusion, in this study community HIV was at the centre of poor child growth independent of geographical location and maternal factors. Therefore, early interventions to decrease HIV are important to achieve the Millennium Development Goals to reduce the burden of disease, rate of infections and ultimately create an “HIV free generation”.

3.5 Acknowledgements

EM – formulated the research question, writing the literature, undertook data management and analyses and writing of the manuscript; EK-M: formulation of data sampling design, managed data collection contributed to the manuscript; PV gave advise on the statistical modelling and analyses; SAN: acquisition of funds for primary data collection, sampling frame design and revision of manuscript and KK: custodians of the site where data were collected, acquisition of funds for primary data collection, sampling frame design and revision of manuscript. None of the authors declared a conflict of interest.

Chapter 4

4.0 The contribution of spatial analysis to understanding and responding to HIV/TB mortality in children: a structural equation modelling approach.

Eustasius Musenge [1,2§], Penelope Vounatsou [3], Mark Collinson [1,4,5], Stephen Tollman [1,4,5], Kathleen Kahn [1,4,5]

[1] MRC/Wits Rural Public Health & Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[2] Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[3] Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland.

[4] Centre for Global Health Research, Umeå University, Umeå, Sweden.

[5] INDEPTH Network, Accra, Ghana.

This paper is in press the Global Health Action journal and attached in section 9.0

Chapter 4 abstract

Background: South Africa accounts for more than a sixth of the global population of people infected with HIV and TB, ranking her highest in HIV/TB co-infection worldwide. Remote areas often bear the greatest burden of morbidity and mortality, yet there are spatial differences within rural settings.

Objectives: The primary aim was to investigate HIV/TB mortality determinants and their spatial distribution in the rural Agincourt sub-district for children aged 1-5 years in 2004. Our secondary aim was to model how the associated factors were interrelated as either underlying or proximate factors of child mortality using pathway analysis based on a Mosley-Chen conceptual framework.

Design: We conducted a secondary data analysis based on cross-sectional data collected in 2004 from the Agincourt sub-district in rural northeast South Africa. Child HIV/TB death was the outcome measure derived from physician assessed verbal autopsy. Modelling used multiple logit regression models with and without spatial household random effects. Structural equation models were used in modelling the complex relationships between multiple exposures and the outcome (child HIV/TB mortality) relayed on a conceptual framework.

Results: Fifty four of 6,692 children between 1-5 years died of HIV/TB, from a total of 5,084 households. Maternal death had the greatest effect on child HIV/TB mortality (adjusted odds ratio=4.00, 95% confidence interval [1.01 - 15.80]). A

protective effect was found in households with better socio-economic status and when the child was older. Spatial models disclosed that the areas, which experienced the greatest child HIV/TB mortality, were those without any health facility.

Conclusion: Low socioeconomic status and maternal deaths, impacted indirectly and directly on child mortality, respectively. These factors are major concerns locally and should be used in formulating interventions to reduce child mortality. Spatial prediction maps can further guide policy makers to target interventions where they are most needed.

Keywords: **HIV/TB, child mortality, conceptual framework, spatial analysis, path analysis, South Africa**

4.1 Introduction

Tuberculosis (TB) and the HIV epidemic are among the top public health research priorities especially in sub-Saharan Africa. The sub-Saharan African region which constitutes 10% of the world's population is home to more than 67% (22.9 million) of people living with HIV/AIDS (185). TB which is preventable and curable was responsible for 25% of HIV related deaths in 2009 and infected 33% of those living with HIV, as such the World Health Organisation (WHO) targets to reduce TB related deaths by 1 million by year 2015 (186). South Africa's population is about 7.0% (50.6 million (47)) of the sub-Saharan African population, yet accounts for 24.6% (5.63 million of the 22.9 million) of the people living with HIV and an estimated 67% of these have TB , thus ranking as the country with amongst the highest co-infection rates (185).

Children were affected in two ways: a very high proportion of child deaths (60% in the year 2000) were attributed to the HIV epidemic (187) and in addition approximately 1.9 million children were left orphaned by the HIV epidemic in South Africa (185). The nation at the onset of the 21st century had an estimated 70,000 children (about 25% rate of mother-to-child-transmissions (MTCT)) newly infected, whereas 10 years later only 10,000 (4.5% rate of MTCT (188)) which is indicative of an improved and more effective mother-to-child prevention programs(4). In public health, determining patterns of disease occurrence provides the first steps toward increased understanding of determinants and potentially greater control of the disease (25). Understanding the associated factors of child HIV/TB mortality and how these are geographically distributed is key for effective interventions.

Several determinants of child mortality have been reported over the past three decades. In the 1980s, researchers explored causes of child mortality including medical, socio-economic and demographic (32). In the 1990s, fertility behaviour, nutritional status, breastfeeding and infant feeding, the use of health services by mothers and children, ecological variables and socioeconomic status were studied (189). In the 21st century the focus has been on: HIV exposure, poor maternal health, inadequate infant care, increased exposure to infections, deaths of parents or caregivers, immune system abnormalities, poor nutrition, reduced breastfeeding and antiretroviral treatment exposure (33). These determinants are inter-related and can be combined into a conceptual (or theoretical) framework. This combined with spatial aspects forms social ecological models which are a component of the fast developing systems science that examines multiple effects and interrelatedness of social elements in an environment (190). Analysis of such frameworks can now be handled with the aid of structural equation models also known as pathway analysis.

The motivation for spatial analysis in epidemiological modelling is based on the notion that people living in a household and those in close proximity share similar exposures which impact on the observed outcomes. In HIV/TB epidemiological research the focus has largely been on drug development, societal integration, personal nutrition, physical health and care giving (32, 33, 189, 191). Country-wide or region-specific pooled results are often reported, disregarding geographical confounding that often exists. Little has been done to inter-link factors associated with child HIV/TB mortality or model their impact on areas of residence (households) (192). The few studies implementing spatial analysis have looked across different infectious diseases (64, 66, 193), used different study designs (cross sectional and

longitudinal) (50, 194) and on varying levels of analysis (village, household and individual) (64, 194, 195). As this is an emerging field few studies have focused on HIV/TB cause-specific mortality assessing geographical differentials in South Africa over time (195).

The primary aim of this paper was to investigate the cause-specific (HIV/TB) mortality determinants and their spatial (geographical) distribution in the Agincourt sub-district of rural northeast South Africa for children aged 1-5 years in 2004. Our secondary aim was to model how the associated factors were interrelated as either direct (underlying) or indirect (proximal) factors of child HIV/TB mortality using pathway analysis based on the Mosley-Chen conceptual framework that incorporates social, biological and ecological determinants of child mortality (32).

4.2 Methods

4.2.1 Agincourt sub-district data and study design

The Agincourt health and socio-demographic surveillance system was established in 1992 in a remote rural sub-district in the Bushbuckridge Municipality of Mpumalanga Province, northeast South Africa near the Mozambique border (196). There are annual census updates that collect socio-demographic information including in-migration and out-migration, mortality, fertility, health and behavioural data (196).

In this study we extracted data for 2004 describing a cross sectional study of households in the study site. In that year, the Agincourt health and socio demographic surveillance system covered a sub-district population of over 70,000 persons living in approximately 12,000 households scattered throughout 21 neighbouring villages.

There were several health facilities, clinics and private clinics in the area as shown in figure 4.1.

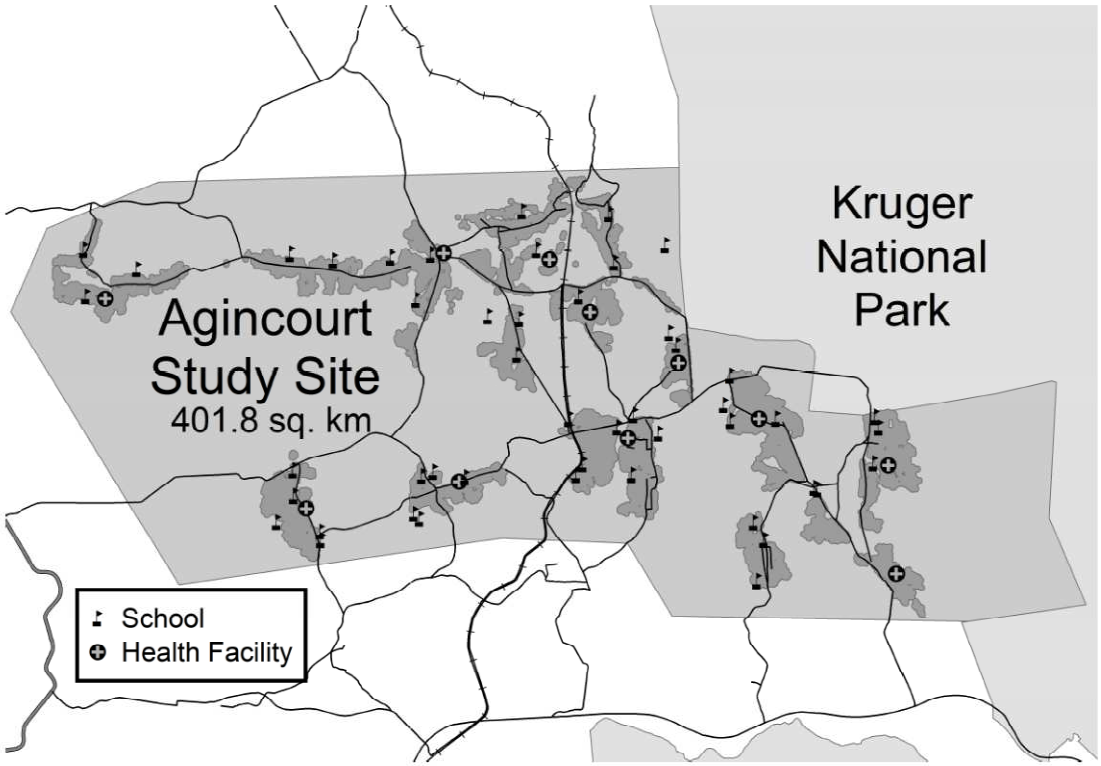


Figure 4.1: Agincourt sub-district in 2004 showing villages (dark grey shading) schools (flags) and clinics and health facilities (circled white and black crosses).

Cause of death data were obtained through verbal autopsies conducted on every recorded death during the year (197). Interviews were conducted by trained lay field workers, within one to eleven months after a death, with the closest caregiver of the deceased in their mother tongue. Cause of death was independently determined by two medical practitioners; a third was sought to resolve discrepancies. Their consensus cause of death was classified according to the World Health Organization's International Classification of Diseases 10 (ICD10) (198). HIV/TB mortality in children was ascertained by the reported signs and symptoms, and in some instances

this was substantiated by the mother's cause of death (198). Over 90% of the HDSS households were geo-coded thus enabling spatial analyses at household level.

4.2.2 Dependent and independent variables

The study sample included all children aged between one and under five years who were registered in the Agincourt HDSS between January and December 2004. The independent variables used were: child's gender, nationality and age, mother's age, parity and death, gender of household head, cumulative deaths in household, number of household dwellers, minimum distance to health facility and household socio-economic status. The latitude and longitude of each household were used to construct the variables for the spatial correlation (random effects), based on the straight line (Euclidean) distances between all pairs of households. The dependent variable was death due to HIV and tuberculosis (TB) determined by the WHOs ICD10 codes A16-A19[3] for TB and B20-B24[4] for HIV. The verbal autopsy cannot reliably distinguish HIV and TB due to low sensitivity and specificity, thus these were combined in our analysis (199).

4.2.3 Conceptual framework: determinants of child mortality

As a way of interrelating child mortality determinants we formulated a hybrid conceptual framework for child mortality determinants and disease ecology as proposed by Mosley-Chen and Meade (32, 67). This approach integrates ecological, social and biological variables in the analysis, with the idea that these determinants

[3] A16=Respiratory tuberculosis; not confirmed bacteriologically or histologically ; A17=Tuberculosis of nervous system; A18=Tuberculosis of other organs; A19=Miliary tuberculosis

[4] B20=Human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21=Human immunodeficiency virus (HIV) disease resulting in malignant neoplasms; B22=Human immunodeficiency virus (HIV) disease resulting in other specified diseases ; B23=Human immunodeficiency virus (HIV) disease resulting in other conditions; B24=Unspecified human immunodeficiency virus (HIV) disease

operate through a common set of either 'proximate' (intermediate or indirect) or 'underlying' (direct) variables that impact on HIV/TB child mortality (32, 200-202). Variables are either endogenous (outcome variables) or exogenous (explanatory) or both and these are modeled using structural equation models. Details of the structural equation model are given in Appendix A.

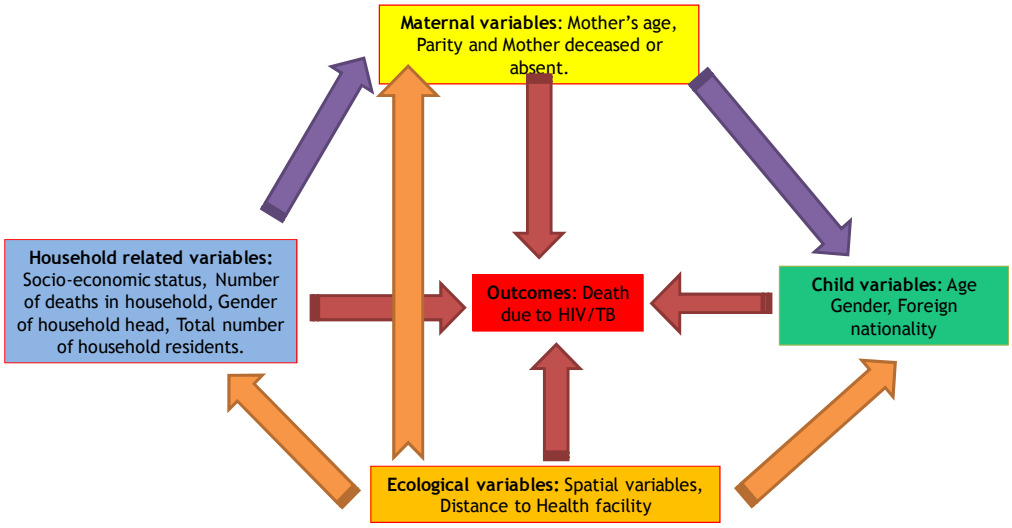


Figure 4.2: Conceptual framework for modelling childhood HIV/TB related mortality: adapted from Meade (1977) (67).

Note: The red arrows reflect a direct effect on mortality and the purple (explanatory variables) and orange (spatial random effects) arrows depict an indirect effect.

We used the observed variables: child specific, mother specific, household related; and unobserved (latent) variables: socio-economic quintiles and spatial (ecological) variables. We tested the pathways shown in figure 4.2, to measure the direct and indirect determinants of child mortality for the Agincourt sub-district in 2004. Two

approaches that complemented each other were used in our modelling. Firstly we modelled the HIV/TB mortality determinants for children aged 1-5 years in 2004 and adjusted for their spatial random effects. In relation to our figure 4.2 we estimated the direct effects pathways (red arrows) from all categories of the explanatory (exogenous) variables controlling for confounders amongst them and their spatial or geographical effects (orange arrows); results for these are presented later in table 4.1 and figure 4.4. Lastly using the significant variables as outcomes (endogenous) we determined the indirect (proximal denoted by purple arrows) within and between categories to establish those factors impacting on child HIV/TB mortality.

4.2.4 Statistical methodology

We modelled the child HIV/TB mortality data to find associated factors and determine geographic disparities. Statistical techniques take into account spatial correlations by introducing in the model spatially structured random effects. These are usually based on independent observations however, with multilevel correlated structures commonly used, maximum likelihood estimation (MLE) methods often underestimate the standard error and thus the statistical significance is overestimated (91). Analysis was done using a geospatial logit regression model which catered for individual as well as household level random effects (85). Inference was done using classical MLE as well as Bayesian inference techniques to fit the non-spatial and spatial models respectively. Bayesian estimation of the many parameters was done using Markov Chain Monte Carlo (MCMC) simulations with the aid of Metropolis Hastings (MH) steps (85). Geospatial prediction maps showing the geographical

coordinates, adjusting for child HIV/TB mortality determinants were used in the analysis.

Data extraction was carried out using Structured Query Language (SQL), data management was done in STATA 10.0 (203), and structural equation modelling was peformed in STATA 12.1 (Stata Corporation, College Station, Texas USA). The Bayesian models were fit using software available on the public domain: BayesX version 2.1 for the spatial modelling and R-cran version 2.12.2 for geospatial mapping using the analysis and mapping packages; maptools, maps, spBayes and Multilevel B-spline Approximation (MBA) (33, 136).

4.2.5 Ethical clearance and informed consent

While this study used secondary data the original study on which it was based was careful to consider ethical issues. Verbal informed consent was obtained when the Agincourt health and socio-demographic update rounds were conducted and also when verbal autopsy data were collected from a close relative of the deceased. To avoid emotional harm to the interviewee a culturally accepted period of one month was observed before the interview was conducted. Risks of stigmatisation were minimised through special training of fieldworkers on the importance of confidentiality. This work was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (No. 960720 and M081145).

4.3 Results

A total of 6,692 children between 1 and less than 5 years, with a mean age of 2 years 8 months of which 54 HIV/TB-related deaths were recorded in the HDSS in 2004. There were 5,084 fully geo-coded households with children living or dead. Just over half the children were female (51%) and more than a third (36%) were born to former Mozambican refugees. The mean age of mothers was 29 years and the mean distance to the nearest health facility was 2.33 kilometres. Table 4.1 shows the results for the descriptive, univariate and multivariable logit regression analyses. Maternal death had the greatest effect on child HIV/TB mortality (logit=1.386 (0.707), equivalent to an adjusted odds ratio (AOR) of 4.00 (95% Bayesian credible interval (BCI) [1.01 - 15.80]) controlling for spatial correlation and keeping all other variables constant.

Multivariable analysis showed three further significant predictors of child mortality in 2004; these were child, maternal and household specific. Adjusting for spatial correlation and keeping all other variables constant there were several significant determinants of child HIV/TB mortality. Child specific differentials showed that boys were almost twice as likely to die compared to girls (logit=0.641(0.291) AOR=1.89, 95% BCI [1.07 – 4.36]). As the children got older there was a protective effect against death; for every year older the child got the likelihood of death decreased by about 50% (logit=-0.630(0.137), AOR=0.53, 95% BCI [0.40 – 0.70]). Male-headed households had a 60% lower risk of child HIV/TB mortality compared to female headed households (logit=-0.895(0.30), AOR=0.41, 95% BCI[0.23 - 0.74]).

Table 4.1: Child HIV/TB mortality: descriptive statistics and results of two multiple logit regression analyses, Agincourt sub-district 2004

Variable	Summary Number (%)	Univariate logit coefficient (standard error)	Non Spatial Multiple logit coefficient (standard error)	Spatial Multiple logit coefficient (standard error)
Age of child in years[1]	2.77 ±1.20	-0.617(0.131)**	-0.601(0.146)**	-0.630(0.137)**
Gender				
Female	3,379 (50.52)	0	0	0
Male	3,310 (49.48)	0.556(0.283)*	0.698(0.303)*	0.641(0.291)*
Child's parents former refugee				
Yes	2,388 (35.71)	0		
No	4,299 (64.29)	-0.137(0.280)		
Mother's age in years[1]	28.70 ±7.59	-0.003(0.018)		
Cumulative household deaths[2]	0.68±0.96;(0-6)	0.684(0.086)**		
Mother deceased by 2004				
No	6,634 (99.13)	0	0	0
Yes	58 (0.87)	1.508(0.732)*	1.504(0.867)*	1.386(0.707)*
Minimum distance to health facility[1]	2.33 ±1.61	0.078(0.083)		
Household size[1]	4.88±2.92	-0.011(0.048)		
Gender of household-head				
Female	2,434(36.37)	0	0	0
Male	4,258 (6 4.63)	-0.865(0.418)**	-0.880(0.302)**	-0.895(0.30)**
Parity[2]	2.12±1.43;(1-10)	0.085(0.090)		
Socio-economic status quintiles				
Most poor	1,823(27.24)	0		
Very Poor	1,073(16.03)	-0.112(0.393)		
Moderately poor	1,120(16.74)	-0.516(0.444)		
Poor	1,316(19.67)	-0.425(0.406)		
Least poor	1,360(20.32)	-0.458(0.406)		
Non-spatial variance			0.10;95%BCI(0.001;0.68)	0.04;95%BCI(0.001;0.40)
Spatial variance estimate				0.04;95%CI(0.001;0.28)
Deviance information criteria			457.26±15.56	555.87±11.65

*Significance :*p*<0.05; ** *p*<0.01.Notes: [1] mean ±standard deviation, [2] mean ± standard deviation (range)). 95% BCI =95 percent Bayesian credible interval*

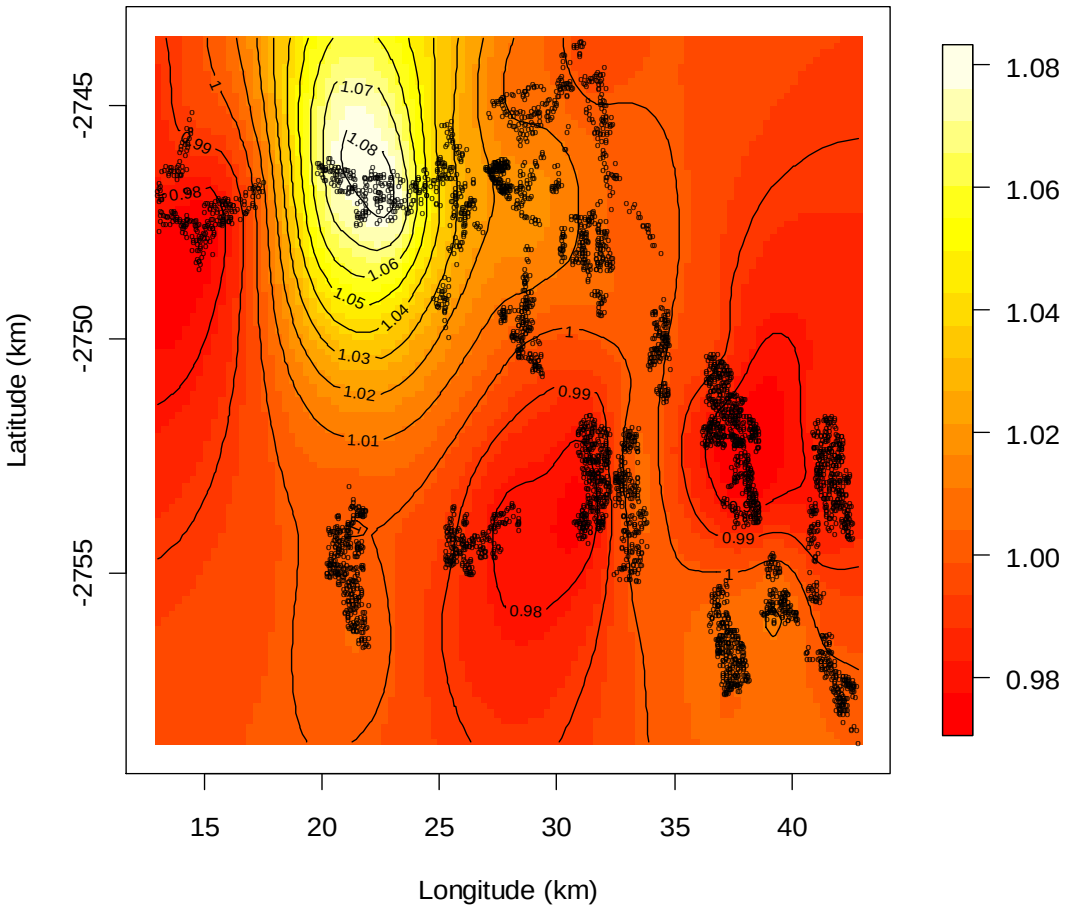


Figure 4.3: Child HIV/TB mortality: Posterior adjusted Odds Ratio map, Agincourt sub-district, 2004.

Spatial disparities can be viewed on maps which are able to show unexpected relationships that may have been overlooked in standard regression analysis. We obtained household specific posterior odds ratios which were adjusted for child, maternal and household related variables in the multivariable models. One main hotspot emerged with high mortality in the north and there were two southerly areas with low mortality. The odds ratio posterior map in figure 4.3 shows the hotspot area between longitude 20 degrees and 25 degrees and latitude -2,750 degrees and -2,745 degrees. This area is characterised by the following: higher rate of maternal deaths,

more male child mortality, lack of a public or private health facility and being furthest from any health facility as shown in figure 4.1.

The results from the conceptual framework analysis of child HIV/TB mortality are shown in figure 4.4 and table 4.2. All the connected arrows show pathways that were statistically significance whether directly or indirectly. Being male (logit=5.2 x 10^{-3}) and older child age (logit=-4.6 x 10^{-3}) both impacted directly on HIV/TB mortality controlling for potential confounders related to the maternal and households variables. Upon treating male household headship as an exogenous variable this directly impacted negatively on child HIV/TB deaths (logit=-7.6 x 10^{-3}) adjusting for household variables. Moreover when male headship was an endogenous variable the following household variables indirectly impacted mortality: better socio-economic status and larger household size had a protective effect (logit=7.7 x 10^{-3}) and increase in household deaths resulted in greater child HIV/TB mortality (logit=-8.3 x 10^{-2}).

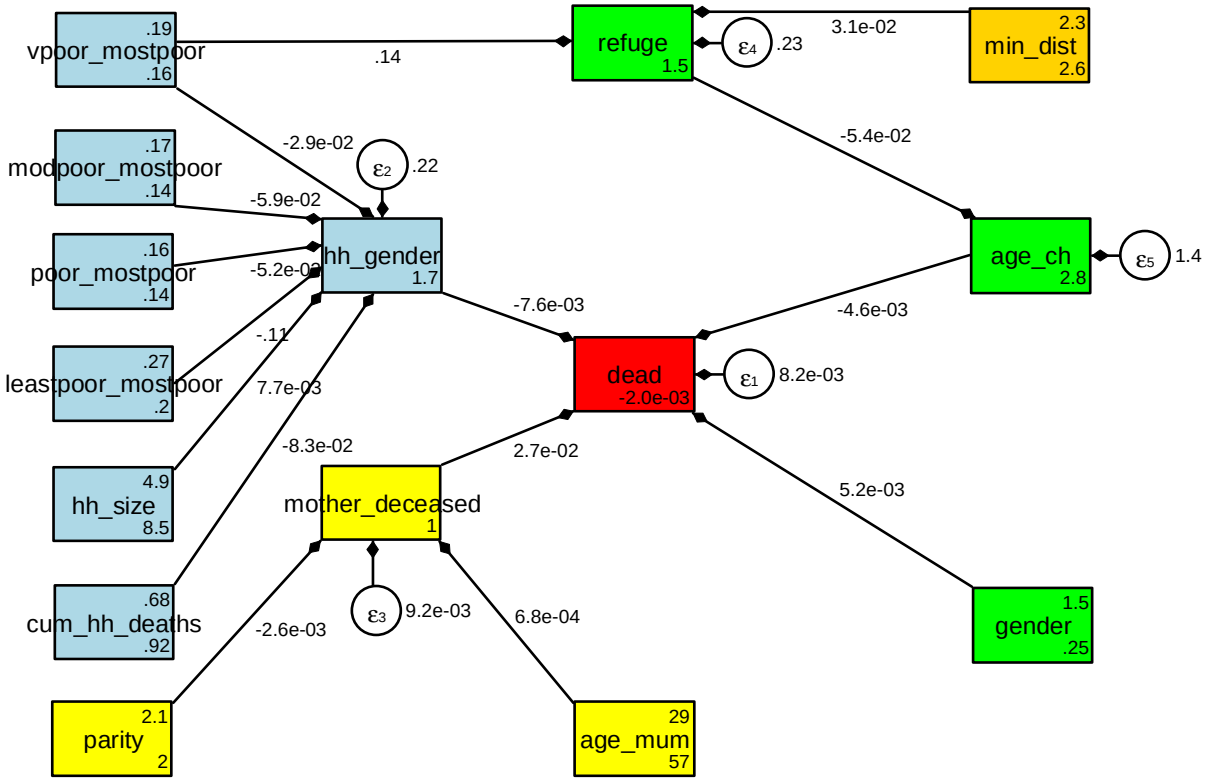


Figure 4.4: Conceptual framework pathways results for Agincourt sub-district child mortality determinants in 2004

*Key for variable names: Very poor versus Most poor (**vpoor_mostpoor**), Moderately poor versus Most poor (**modpoor_mostpoor**), Least poor versus Most poor (**leastpoor_mostpoor**), Poor versus Most poor (**poor_mostpoor**), Age of child in years(**age_ch**), Household size (**hh_size**), Cumulative household deaths (**cum_hh_deaths**), Gender of household head (**hh_gender**), mother age in years (**age_mum**), mother deceased by 2004 (**mother_deceased**), child born to former refugee parents(**refuge**), Gender (**gender**),Minimum distance to health facility (**min_dist**) Parity (**parity**), Child deceased in 2004 (**dead**). Other; the arrows pointing from the exogenous (explanatory) to endogenous (dependent) variables and the error terms (ε) placed on all five endogenous variables.*

Table 4.2: Direct, indirect and total effects of child HIV/TB mortality in 2004, Agincourt sub-district

	Direct effects on child deceased as shown in figure 4.2 conceptual framework results					**Indirect effects on child HIV/TB death**	**Total effects on child HIV/TB death**
	Gender of household-head	**Child a former refugee**	**Child deceased**	**Mother deceased by 2004**	**Age of child in years**		
Very Poor vs Most Poor	-0.029(0.019)	0.140(0.015)**				**2.6 x 10^{-4}** (1.6x 10^{-4})$^{ns!}$	2.6x 10^{-4}(1.6x 10^{-4})^{ns}
Moderately poor vs Most Poor	-0.059(0.020)**					4.5x 10^{-4}(2.1x 10^{-4})*	4.5x 10^{-4}(2.1x 10^{-4})*
Poor vs Most Poor	-0.052(0.020)**					4.0x 10^{-4} (2.0x 10^{-4})*	4.0x 10^{-4} (2.0x 10^{-4})*
Least poor vs Most Poor	-0.111(0.018)**					8.4x 10^{-4}(4.0x10^{-4})**	8.4x 10^{-4}(4.0x10^{-4})**
Household size vs Most Poor	0.008(0.002)**					-5.9x 10^{-5}(2.5x 10^{-5})*	-5.9x 10^{-5}(2.5x 10^{-5})*
Cumulative household deaths	-0.083(0.006)**					6.4x 10^{-4}(2.1x 10^{-4})**	6.4x 10^{-4}(2.1x 10^{-4})**
Minimum distance to health facility		0.031(0.004)**				7.72x 10^{-6} (4.88x 10^{-6})^{ns}	7.72x 10^{-6} (4.88x 10^{-6})^{ns}
Male household-head			-0.0076(0.002)**				-0.0076(0.002)**
Mother deceased by 2004			0.027(0.012)*				0.027(0.012)*
Age of child in years			-0.0046(0.001)**				-0.0046(0.001)**
Gender			0.0052(0.0023)*				0.0052(0.0023)*
Mother's age (years)				0.001(0.000)**		1.9x 10^{-5}(9.99x 10^{-6})$^{+}$	1.9x 10-05(9.99x 10^{-6})$^{+}$
Parity				-0.003(0.001)*		-7.0x 10^{-5}(4.0x 10^{-5})^{ns}	-7.0x 10-05(4.0x 10^{-5})^{ns}
Parents former refugee					-0.054(0.032)+	-2.5x 10^{-4}(1.5x 10^{-4})$^{+}$	-2.5x 10^{-4}(1.5x 10^{-4})$^{+}$
Intercept	1.719(0.017)**	1.537(0.011)**	-0.002(0.014)	0.995(0.005)**	2.814(0.054)**		

+ p<0.10; * p<0.05; ** p<0.01, ns-not significant;! Indirect effects computed as the product along the related pathways i.e -0.029x-0.0076+0.14x-0.054x-0.0046=**2.55 x 10^{-4}≈2.6 x 10^{-4}**

Table 4.2 details the direct pathways for child HIV/TB deaths (arrows directly linked to the red square) and the indirect pathways all possible routes for child HIV/TB deaths as shown in figure 4.4. In our analysis poverty indirectly impacted on HIV/TB child mortality as conferred by two pathways: firstly, male headship of households (protective) and secondly, refugee status (a risk) controlling for child's age (as shown on table 4.2 foot note). The greatest direct impact on child HIV/TB mortality was maternal death (logit=2.7 x 10^{-2}) adjusting for the indirect effects of parity (logit=-2.6 x 10^{-3}) and mother's age (logit=6.8 x 10^{-4}).

4.4 Discussion

In this study we showed that HIV/TB child mortality is influenced by a combination of child, maternal, household and spatial determinants. Over the years 2002 to 2005 HIV/TB alone accounted for 34% of all cause-mortality among the children living in the area (199). While being male was a risk factor for HIV/TB mortality as children got older, irrespective of sex they were less likely to die. A major finding in our study was that maternal death had the greatest direct impact on HIV/TB child deaths. A comparable study in the same area, adjusting for spatial random effects at village level, showed that infants whose mother's had died (presumptively of HIV) were at greater risk of death during the first 12 months of life (64).

The main source of TB transmission among children are adults, as shown by a study on a high risk spatial cluster in South Africa with an incidence three times higher than the WHO levels (50). The most likely cause of HIV infection was mother-to-child

transmission. Prevention of mother-to-child transmission (PMTCT), using a single-dose of Nevirapine (sdNVP) administered to the child and mother, was found to reduce child mortality by about 9% in South Africa (204). From 2004 the guidelines for PMTCT and ART eligibility have changed compared to 2012. There is also a seven fold lower (from 25% to 4.5%) rate of infection from mother to child over the period 2000 to 2010 (188). Kimani-Murage et al. (2009) found that knowledge of child HIV status influences the manner of care-giving as well as acceptance and up-take of community-wide interventions (205). The mother's survival significantly increases the likelihood of the child's survival, a finding that strengthens the current policy focus on maternal health. In South Africa a comprehensive three stage process has been set up to improve maternal, foetus, infant and ultimately child survival during antenatal care, labour and delivery and postnatal care (48). Reduction of TB morbidity and mortality is the focus of the WHO 2015 goal of zero new TB infections and zero TB related death (186). The impact of PMTCT aided by sdNVP, ziduvudine (AZT) and ART has also been demonstrated from evidence gathered from another site in rural South Africa (206).

In Agincourt, female headed households had a higher child mortality compared to those headed by males. Studies in Tanzania, Zambia and Limpopo province South Africa showed similar results (207-209). The pathway analysis (figure 4.4) demonstrates that male headship does not independently lead to lower mortality, but is a proxy for better socio-economic status of households. Main bread winners are usually male with females having fewer opportunities in the labour market and consequently earning less. The absence from the household of a male due to death or permanent out-migration tended to impact negatively on household socio-economic status, leading to 90% greater child

mortality odds (210). In other African countries similar patterns were observed as men temporarily migrate within a country or to neighbouring countries due to either pull or push factors (211-213). HIV/TB deaths have lowered the life expectancy of working men from the area who often "return home to die" from their jobs (197). A recent qualitative study found that a father's unemployment, inability to work and reluctance to provide support were common factors in HIV infected households (205). From our pathway analysis (table 4.2 and figure 4.4) we observed an indirect impact on child HIV/TB mortality: as the number of household deaths increased so the risk of child deaths was higher. This suggests that households with a greater burden of mortality are more vulnerable socio-economically through loss of bread winners and high health care and funeral costs. This compromises a family's ability to feed and care for all children, including those who are HIV positive. We also observed that former Mozambicans are at greater risk of child HIV/TB deaths, indicating the need for health care interventions that target these marginalised and vulnerable households.

4.4.1 Strengths and limitations of the study

The strengths and limitations of this paper are discussed in four categories: classification of exposures and outcomes; local or global relevance of results; confounding and effect modification and analysis deviance.

A major strength in our observed variables is that the data are consistently updated, corrected and checked for inconsistencies. Nevertheless potentially there are two variable-related biases namely selection and recall bias. Selection bias may occur if permanent out-migrants differ from residents or temporary migrants which results in

incomparable groups. Recall bias on the outcome variable mortality is unlikely as death is a major household event, however it may affect the quality of the verbal autopsy (VA) interview, conducted at least one month but even up to 11 months after death, and hence the ability of medical experts to assign a valid cause of death .

Our study focused on one sub-district only, but demonstrated geographical differentials in child HIV/TB mortality. There was one main hotspot with highest mortality and a visible downward trend in mortality moving southwards. These patterns are comparable to those found in other studies done on the area (64, 195). While many rural areas are not geo-coded to enable direct comparison of trends and patterns, our findings likely reflect similar patterns in other rural settlements at least in northeast South Africa.

To deal with confounding variables we used multiple regression and added the spatial random effects in our modelling and also performed pathway analysis. Mediation was tested for in the pathway analysis as we determined factors directly and indirectly related to mortality and malnutrition. However we were unable to model time varying covariates as this was a cross-sectional study design.

The robustness of our analysis procedures was demonstrated in that after catering for the spatial household random effects we obtained narrower confidence bands compared to the non-spatial analogue. Therefore catering for the spatial random effects further improves the model and parameter estimation compared to classical statistical models. Bayesian and pathway analysis combined were able to handle the complex structure of hierarchical data from the Agincourt health and socio-demographic surveillance system. Other advantages of Bayesian modelling are the ability to adjust for spatial

confounding, ability to handle many parameters and inclusion of prior known information on the distribution of data.

4.5 Conclusion

The growing field of spatiotemporal epidemiology brings in person, place and time aspects in the risk factor analysis. Spatial Bayesian models provide better fitting models compared to non-spatial models. This procedure also enables drawing of risk maps which can be used by policy makers to target and develop community-based interventions.

Going forward, further studies investigating trends over time using spatiotemporal Bayesian analysis will strengthen the evidence on associated factors and trends found in this study and demonstrate the value of this analytic approach. Studies comparing deaths due to other causes with those for HIV/TB in children would be important in disclosing if there is any selection bias between the two groups. Also adopting a secondary analysis of a cohort study design would have enabled temporality to be investigated in greater detail and survival analysis. Moreover a combination of qualitative and quantitative analysis methods is needed to fully understand the proximal and distal factors that influence child HIV/TB mortality. Findings that can be used for policy nationally would require sentinel sites located strategically across the country.

4.6 Acknowledgements

This work was made possible by doctoral fellowship support from the World Health Organization, Tropical Disease Research (WHO/TDR) and Swiss-South Africa Joint Research Programme. The MRC/Wits Rural Public Health and Health Transitions

Research Unit (Agincourt) is funded by the Wellcome Trust UK (grants 058893/Z/99/A and 069683/Z/02/Z and 085477/Z/08/Z), the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew W Mellon and William and Flora Hewlett Foundations, U.S.A. Special thanks goes to Benn Sartorius for data extraction, David Rees for his valuable review of an early draft, community leaders, field workers, supervisors and the Agincourt community.

4.7 Appendix A: Structural Equation Model for paths analyses

The structural equation model fit in this paper for given individual level outcome (endogenous) variables y_i (observed; child HIV/TB deaths) and/or η_i (for latent endogenous) which then form the vectors of exogenous variables $\boldsymbol{y}$ and $\boldsymbol{\eta}$. The individual level explanatory variables x_j (observed; maternal variables, child variables and household related variables) and/or ξ_j (latent; spatial and household related variables such as socio-economic status) makes up the vectors of endogenous variables $\boldsymbol{x}$ and $\boldsymbol{\xi}$. These can be expressed in a multivariate equation as follows:

$$\boldsymbol{Y} = \boldsymbol{BY} + \boldsymbol{\Gamma X} + \boldsymbol{\alpha} + \boldsymbol{\zeta}$$

where $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{\eta} \end{pmatrix}$ is the vector of all endogenous variables with observed and latent variables $\boldsymbol{y}$ and $\boldsymbol{\eta}$ respectively, the vector of exogenous variables $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\xi} \end{pmatrix}$ with observed and latent variables $\boldsymbol{x}$ and $\boldsymbol{\xi}$ respectively and the vector of error terms $\boldsymbol{\zeta} = \begin{pmatrix} e.\boldsymbol{y} \\ e.\boldsymbol{\eta} \end{pmatrix}$. $\boldsymbol{B} = [\beta_{ij}]$ is the matrix of coefficients on endogenous variables predicting other endogenous variables which were; gender of household head, mother's death,

former Mozambican and age of child's in years. $\boldsymbol{\Gamma} = [\gamma_{ij}]$ is the matrix of coefficients on exogenous variables which were; parity at child's birth, mother's age, socio-economic status, total deaths in household by 2004, household size and minimum distance to health facility in kilometres. $\boldsymbol{\alpha} = [\alpha_i]$ is the vector of intercepts from the endogenous variables, whereas $\boldsymbol{\zeta}$ is assumed to have mean zero and covariance $cov(\boldsymbol{X}, \boldsymbol{\zeta}) = \boldsymbol{0}$ such that $\boldsymbol{\zeta} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$ where

$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} = Var(\boldsymbol{X}) & \boldsymbol{\Sigma}_{X\zeta} = \text{cov}(\boldsymbol{X}, \boldsymbol{\zeta}) \\ \boldsymbol{\Sigma}'_{X\zeta} = \text{cov}(\boldsymbol{X}, \boldsymbol{\zeta}) & \boldsymbol{\Sigma}_{\zeta\zeta} = Var(\boldsymbol{\zeta}) \end{pmatrix} = \begin{pmatrix} [\phi_{ij}] = \boldsymbol{\Phi} & \boldsymbol{0} \\ \boldsymbol{0} & [\psi_{ij}] = \boldsymbol{\Psi} \end{pmatrix}$. Let $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix}$ be the vector of all endogenous and exogenous variables then $\boldsymbol{Z} \sim MVN(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$ where

$\boldsymbol{\mu}_Z = \begin{pmatrix} E(\boldsymbol{Y}) = \boldsymbol{\mu}_Y \\ E(\boldsymbol{X}) = \boldsymbol{\kappa} \end{pmatrix}$ and $\boldsymbol{\Sigma}_Z = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}'_{YX} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}'_{YX} & \boldsymbol{\Phi} \end{pmatrix}$

with expectation vectors $\boldsymbol{\mu}_Y = E[(\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\Gamma X} + \boldsymbol{\alpha} + \boldsymbol{\zeta})] = (\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\Gamma\kappa} + \boldsymbol{\alpha})$ and $\boldsymbol{\kappa} = [\kappa_j]$, variance matrix with variables $\boldsymbol{\Sigma}_{YY} = Var(\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi})\{(\boldsymbol{I} - \boldsymbol{B})^{-1}\}'$ and $\boldsymbol{\Sigma}_{YX} = \text{cov}(\boldsymbol{Y}, \boldsymbol{X}) = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Gamma\Phi}$ (134, 214-216). The parameter estimation was done using maximum likelihood in

STATA 12.1 (Stata Corporation, College Station, Texas USA).

Chapter 5

5.0 Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spatial and Spatiotemporal Epidemiology 2011; 2, 205-217.

Eustasius Musenge[1,2§] , Penelope Vounatsou[3] and Kathleen Kahn[1,4,5]

[1]MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, University of the Witwatersrand , 7 York Road, Parktown 2195, Johannesburg, South Africa.

[2]Biostatistics and Epidemiology Division, School of Public Health, University of the Witwatersrand , 7 York Road, Parktown 2195, Johannesburg, South Africa.

[3]Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland.

[4]Centre for Global Health Reasearch, Umeå University, Umeå, Sweden

[5]INDEPTH Network, Accra, Ghana.

This paper has been published in the Journal Spatial and Spatio-temporal Epidemiology 2 (2011) 205–217, attached in the appendices

Chapter 5 abstract

South Africa has the greatest burden of HIV/TB, with a prevalence of 17% and 2.1 million AIDS orphans. We used data from Agincourt located in rural northeast South Africa, collected longitudinally over the years 2000 and 2005. A total of 187 deaths were observed from 16,844 children aged 1-5 years coming from 8,863 households. In this paper we employ two zero inflated models adjusting for household spatiotemporal random effects using Bayesian inference. Bayesian zero inflated spatiotemporal models were able to detect hidden patterns within the data. Our main finding was that maternal orphans were almost thrice at greater risk of HIV/TB death compared to those with living mothers (AHR=2.93, 95% CI[1.29;6.93]). Risk factor analyses which adjusts for person, place and time enables policy makers to use estimates and maps for interventions. We conclude saying child survival is dependent on the mother's survival and advocate for policies that promote maternal longevity.

Key words: (Spatiotemporal confounding, HIV/TB, zero inflated, over-dispersed, Agincourt rural South Africa)

5.1 Introduction

The years 2000 to 2005 mark the middle of the United Nation's Millennium Development Goals (MDG) set out by the 192 member states and 23 international organizations (217). Goal 4.1 aims at reducing child (under five) mortality by two thirds over the years 1990 and 2015. The Sub-Saharan African region has the greatest burden of disease and highest mortality of HIV/TB. In 1990 the Sub-Saharan Africa region had 184 child deaths per every 1000 live births, by the year 2006 twenty seven of the countries in the region had made no progress in reducing child deaths (218). Globally 2.1 million children were living with HIV/TB in the year 2008 of which 90% were from the Sub-Saharan region (219). South Africa's HIV/TB orphans in 2008 were about 1.4 million, the highest absolute number globally (220). In 1998 the year Demographic Health Surveys were started in South Africa, she had an under five mortality rate (U5MR) of 59.4 deaths per 1000 live births, aiming a reduction to 20 per 1000 by 2015 (221). The main cause of child deaths is the HIV epidemic, mainly due to mother-to-child-transmissions (MTCT) with over 70,000 new infections yearly (222).

Public Health priorities such as disease control and intervention are made difficult in many developing countries due to lack of good data mainly due to poor resources, poor management and inadequate skills. The analyses of Public Health data often build from the statistical notion of each person having a risk or probability of contracting a disease. The main goals in analyses are to identify and quantify any exposures, behaviours and characteristics that may modify an individuals' or a population's risk. However HIV/TB interventions are hampered by the high rate of infectiousness, which is a factor of geographical proximity.

Many articles have been published on HIV/TB mortality in sub-Saharan Africa. The issues discussed focused mainly on the impact of epidemics, assessing for risk factors, measuring efficacy of interventions and identifying high risk clusters. A study done in rural Senegal showed a decreasing trend in child mortality over a 37 year period (52). Spatial analyses on 39 rural villages in Burkina Faso under the Health and socio-Demographic Surveillance System (HDSS) over a nine year period, showed that village and distance from the health facility where among the risk factors for child mortality (53). A seasonal mortality trend was also disclosed by the same authors in another study of the same HDSS (223). A Tanzanian study used spatial modeling techniques and found that there is a need to adopt both a group-based and a place-based approach in identification of high-risk HIV groups for intervention. (55). In South Africa rural Kwa-Zulu-Natal a study found that prevalence of HIV was associated with ethnicity, urban status, and unemployment based on over 11,000 persons in 700 enumeration areas (224). In our area of study the relative risk (RR) of HIV and TB has increased ten-fold over an eleven year period, 32.3 and 3.2 for the period 2001-2003 compared to 1992-1994, for the Agincourt HDSS (225). Covariates that have mainly been used to model child mortality in the 21st century are: HIV exposure, poor maternal health, inadequate infant care, increased exposure to infections, increased mortality, immune system abnormalities, poor nutrition, reduced breastfeeding and antiretroviral exposure (33). Methods applied in the analyses were mostly non-Bayesian based, which are limited in catering for large datasets, time and spatial random effects (58).

In epidemiological research, two study designs have been used for follow-up studies: time series studies, these estimate acute risk associated with short-term exposure and cohort studies, these estimate chronic effects associated with long term exposure. These approaches have been criticized for potential confounding by time varying covariates and individual risk

factors or household level traits, respectively (226). Failure to account for extra-variation induced by temporal and spatial autocorrelation can lead to an understatement of the uncertainty of the risk of mortality (227). Several approaches can be used to address confounding, at the design level;-randomization, during the study;-use of experts and validated instruments and during analyses;-multivariable regression, standardization and stratification (228). Bayesian and spatial statistical techniques can be used to evaluate differences in rates observed from different geographical locations, separate noise from patterns and identify disease clusters.

To date HIV/TB research has largely focused on drug development, societal integration, personal nutrition, physical health and care giving (189). Most of the research focuses on individuals and treatment cohorts. Little has so far been done on diseases' impact on populations and their areas of residence. In this paper we did an exploratory analyses looking for risk factors for childhood HIV/TB mortality using spatial modelling. The modelling procedure adjusts for the household level spatial random effects for child HIV/TB mortality. Some related earlier work has looked at all-cause mortality adjusting for village (areal) level random effects (51, 64). Our findings can aid in understanding the household differentials adjusted risk factors of child HIV/TB mortality in a rural South African set-up for cohort data with rare occurrence count of the outcomes.

5.2 Methods

5.2.1 Agincourt Health and Socio-demographic surveillance system (HDSS) Data

Agincourt HDSS was set up in 1992 in a rural sub-district in Bushbuckridge Municipality of Mpumalanga Province, northeast South Africa. The Agincourt HDSS site was set up in its

current location mainly due to availability of several clinics and presence of Mozambican in-migrants (229). The Agincourt HDSS has a population of over 70,000 persons, living in approximately 12,000 households all scattered in 21 villages. The area covered is 432 square kilometres, with 34.2 kilometres as the maximum distance between two households. Information at individual level (infants, children and adults) is collected once every year together with other annual census updates, that compile migration, health, demographic, mortality and fertility data (230).

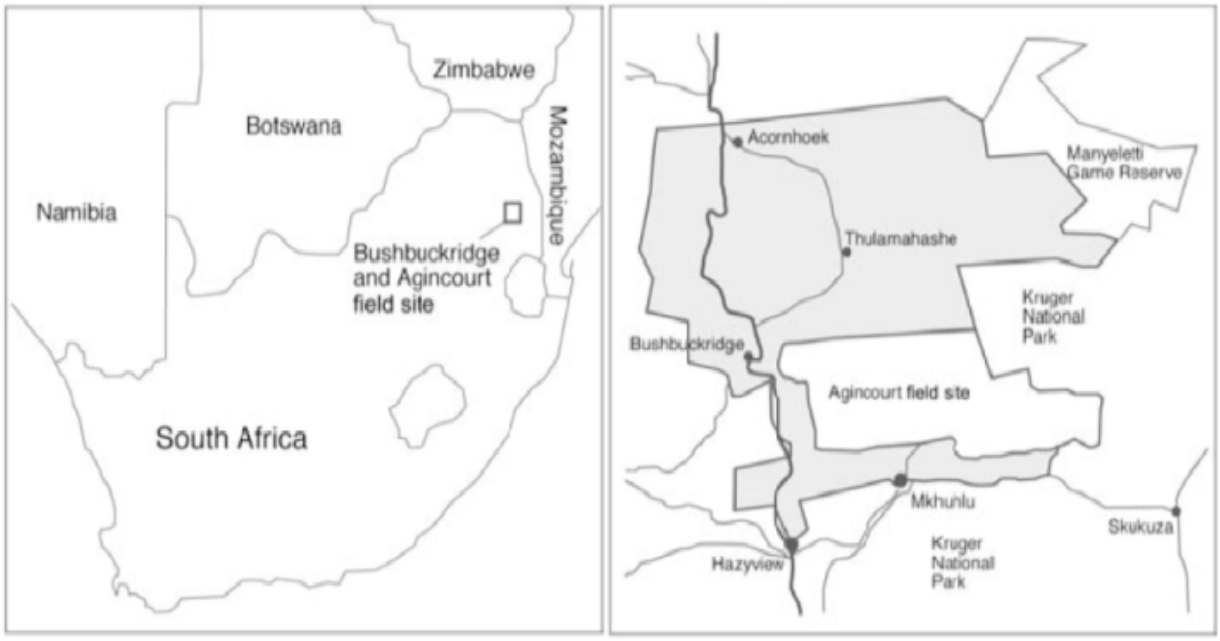


Figure 5.1: Agincourt HDSS location, north-east South Africa. Source Kahn et al. (2000)

Special modules are run periodically such as the socioeconomic status and food security. Cause of death data are obtained through verbal autopsies conducted on every recorded death by trained lay fieldworkers (69). The cause of death is independently determined by two at most three medical experts and classified according to the World Health Organization's International Classification of Diseases (ICD10). The HDSS' every household is geo-coded thus enabling spatial analysis at household as well as at village level. Agincourt HDSS longitudinal data are correlated in time and space (spatiotemporal), since they are collected

yearly on the same population. This makes it difficult to analyze the data due to their collinearity and also their large size.

5.2.2 Outcome and explanatory variables

The child mortality independent variables were selected in accordance with the approach proposed by Rutstein (2000) who mentions six main risk factors for child mortality : socio-economic status, nutritional status, use of health facilities, environmental health conditions, maternal factors and infant feeding (189). The persons included in the study were all the children aged between 1 and under 5 years who lived or had lived in the Agincourt HDSS between January 2000 and December 2005. The explanatory variables used were: child's gender, child's and mother's nationality, mother's age, parents' dead or alive, gender for household head, antenatal clinic visit, mother's parity at pregnancy and birth, cumulative deaths in household, number of household dwellers, number of deaths in household and mother's total live births. The individual child's age was treated as an offset variable. The variable years ranging from 2000 to 2005 was used in determining the temporal random effects and the household's latitude and longitude were used construction of the variable to determine spatial random effects.

The outcome variable was death due to HIV and or tuberculosis (TB) determined by the WHOs ICD10 verbal autopsy codes A16-A19[5] for HIV and B20-B24[6] for TB. The outcome death was treated as a discrete (count) variable since this provides epidemiologically better

[5] A16=Respiratory tuberculosis; not confirmed bacteriologically or histologically ; A17=Tuberculosis of nervous system; A18=Tuberculosis of other organs; A19=Miliary tuberculosis

[6] B20=Human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21=Human immunodeficiency virus (HIV) disease resulting in malignant neoplasms; B22=Human immunodeficiency virus (HIV) disease resulting in other specified diseases ; B23=Human immunodeficiency virus (HIV) disease resulting in other conditions; B24=Unspecified human immunodeficiency virus (HIV) disease

estimates of prevalence and relative risk compared to the odds ratio which results from treating it as binary outcome (90, 231). Treating the outcome as such allowed us to employ the zero inflated Poisson (ZIP) and zero inflated Negative Binomial (ZINB) which yields more epidemiologically plausible relative risks for these cohort data. These models which are members of the exponential family have a log link function which has less identifiability issues in the generalized linear model compared to the logit link function which results from treating the outcome as binary and using logistic regression techniques (23).

5.2.3 Spatiotemporal zero inflated Poisson and Negative Binomial models

Agincourt child HIV deaths data had several problems that made it difficult to use standard statistical procedures: over dispersion, caused by unobserved heterogeneity or temporal correlation, large data, excess of zeros, and household spatial random effects. Conventional statistical methods applied to spatiotemporal data often underestimate the standard error and thus the statistical significance is overestimated (91). Bayesian analyses of sophisticated multilevel models for complex geographically referenced data can be done with the aid of the Markov Chain Monte Carlo (MCMC) parameter estimation procedure (85). Bayesian data models that can deal with the highlighted problems simultaneously and within a unified inferential approach were employed. These are models for over dispersed, zero-inflated data combined with semi-parametrically structured geo-additive predictors, resulting in flexible spatial regression models (232).

Geo-additive spatiotemporal zero inflated Poisson (ZIP) as well as Negative Binomial (ZINB) models (see appendix A) were used in the analysis. The ZIP and ZINB models have

many parameters and are hierarchical (multilevel), thus we resort to full Bayesian inference with the computationally efficient MCMC techniques. Multiple variables models were used to assess the effect of different covariates in the presence and absence of geographical heterogeneity (66). We investigated the presence of over-dispersion (when $\delta \to \infty$ it's absent) and zero inflation (when $\theta = 0$, it's absent) (232). Spatial dependence $f_{spat}(\boldsymbol{s})$ was modelled by assuming that the random effects follow a zero mean stationary multivariate Gaussian random field (GRF) with spatial variance σ^2_{spat} and isotropic Matern correlation functions. For the exponential correlation which is a special case of the Matern function, the multivariate GRF $f_{spat}(\boldsymbol{s}) \sim MVN(\mathbf{0}, \sum)$ has variance-covariance matrix $\sum = \sigma^2_{spat} exp\,[-\hat{\rho} d_{ij}]$, where d_{ij} is the Euclidean distance between households *i* and *j*, σ^2_{spat} is the geographic (spatial) variability known as the sill and $\hat{\rho} = max_{i,j} \|s_i - s_j\| / 0.001$ is the empirically estimated rate of spatial correlation decay (233). Multivariate Gaussian random field (GRF) priors $f_{spat}(\boldsymbol{s}) = p(\beta^{spat} | \sigma^2_{spat}) \propto |\sum|^{-\frac{1}{2}} exp\left(\frac{-1}{2\sigma^2_{spat}} (\beta^{spat})' \sum^{-1} \beta^{spat}\right)$, where β^{spat} is the vector of household specific structured spatial coefficients, whereas the non-structured utilize the independent and identically distributed (i.i.d) Gaussian priors with Inverse Gamma (IG) hyper-priors (133).

Bayesian based model fitting requires the inversion of the variance-covariance matrix ($\sum$) with the same size as the number of geo-locations (households). Due to the large number of observations in our dataset, the estimation of model parameters becomes unstable and unfeasible; this problem is popularly known as the "big N". We addressed this problem by using low rank kriging to approximate the stationary Gaussian random field (234). Firstly the spatial correlation decay ($\hat{\rho}$) is estimated from a "representative" subsample obtained using the mini-max space filling criteria, hence only parameter $\sigma^2_{spat}(1/\tau^2_{spat})$ needs to be

estimated. This reduces the computational burden since the spatial decay is fixed and the large Euclidean distance matrix is also constant, resulting in estimation of the spatial variance and parameter vector of household spatial effects.

Full Bayesian inference was done via MCMC simulation based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data. Gibbs sampling was used for closed conditionals and for the numerically intractable a slightly modified form of the Metropolis-Hastings (MH) algorithm based on iteratively weighted least squares (IWLS) (85). A total of 55,000 iterations on a single chain were run with the first 5000 discarded as burn-in and to reduce auto-correlation every 25th value was taken to form a posterior sample of 2000. MCMC convergence was assessed using diagnostic procedures proposed by Raftery and Lewis (1992), Geweke (1991) and Heidelberger and Welch (1983) (235-237). These convergence diagnostics techniques were done in the R-package, "convergence diagnosis and output analysis" (CODA) (138**).** Additionally a graphical assessment was done using trace-plots, histogram density plots and auto-correlations of parameters and the summary statistics assessing the closeness of the mean and median posterior estimates (139).

5.2.4 Goodness of fit and best model selection

The deviance information criteria (DIC) was used in determining goodness of fit of the proposed models, this can be defined as "classical estimate of fit, plus twice the effective number of parameters" (238). This is a technique for Bayesian model selection for MCMC derived posteriors. The procedure works on the assumption that the posterior distribution is multivariate normal and it's un-standardised deviance is given as: $\mathrm{D}(\Omega) = -2\log\big(\mathrm{p}(\mathrm{y}|\Omega)\big)$, where y are the data, Ω are the unknown parameters of the model, $\mathrm{p}(\mathrm{y}|\Omega)$ is the likelihood.

After addition of the standardising constant $C = 2\log f(y)$ we get the Bayesian or saturated deviance $D(\Omega) = -2\log(p(y|\Omega)) + C$. The expectation of this deviance $\overline{D} = E^{\Omega}[D(\Omega)]$ measures how well the model fits the data, the smaller it is the better. The measure of model complexity or parsimony is the effective number of parameters; $p_D = \overline{D} - D(\overline{\Omega})$,where $\overline{\Omega}$ is the expectation of Ω the smaller it is the more parsimonious. From these measures we get the deviance information criteria; $DIC = p_D + \overline{D}$ and upon substitution of $\overline{D}$ we get $DIC = 2p_D + D(\overline{\Omega})$, which penalises on complexity and lack of fit. Selection of better model based on the DIC basically favours models with the smaller values.

The data extraction was done using Structured Query Language (SQL). Preliminary data analyses, trend analyses, survival graphs and data management were done using STATA 10.0 (203). The Bayesian MCMC analysis was done using a software package called BayesX (239). This package utilises computationally efficient C++ algorithms (240). The computations were done on a quad core processor with 4.0 gigabytes of random access memory (RAM) within a Linux environment. The routines used to estimate the parameters of the models via MCMC which perform Gibbs sampling and Metropolis Hastings within Gibbs sampling (162). Multivariable analyses were performed to determine patterns for HIV/TB mortality adjusting for possible household confounders. Posterior risk maps were produced using mapping packages in R (163).

5.2.5 Ethical clearance and informed consent

The Agincourt HDSS site was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (No. 960720). This work was

also granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (M081145). Verbal informed consent was obtained when the census rounds were conducted and also when verbal autopsy data were collected from a close relative of the deceased.

5.3 Results

5.3.1 Exploratory analysis and descriptive statistics

The data used were for children aged 1 to 5 years residing in Agincourt HDSS from 2000 to 2005 whose geo-location data were available, these totalled 16,844 children from 8,863 households, with a range of 138 to 835 households across the 21 villages. A total 187 deaths were HIV /AIDS (including HIV/Tuberculosis) related from 59,448.15 person years yielding a 1 to 5 years mortality rate (1-4MR) of 3.15 deaths per 1000 person years. There were slightly more females than males (51% vs 49%) and 37% were former Mozambican in origin. The mothers had an average age of 29 years , 65% were South African and about 1% had deceased compared to 2% of the males. The women in the area on average visited the antenatal facility four times during their pregnancy and 34% were heads of their households.

An investigation of the survival experience using the Kaplan-Meier (KM) curves (Figure 5.2, left) showed differences in the survival experiences over the years 2000 to 2005. To statistically test the differences in survival experiences over the years using the generalized Fleming-Harrington test (241) (weighted for later years failures) was significant (p-value=0.048). However there was no significant trend over the years (p-value=0.728). This

procedure of stratifying over the years was used to assess for the possibility of temporal random effects.

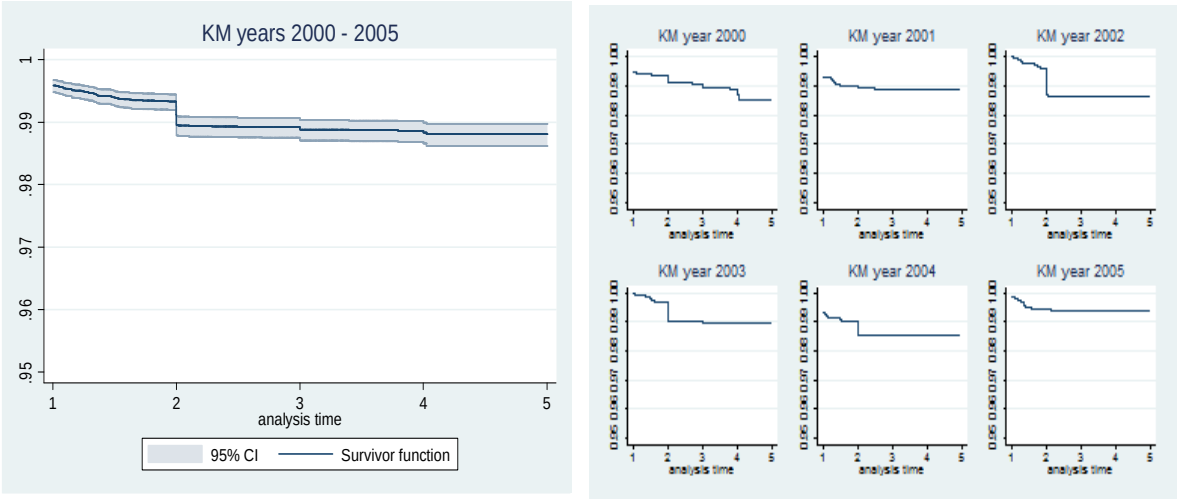


Figure 5.2: Kaplan-Meier (KM) survival function over the years 2000 to 2005 (left) and year specific Kaplan-Meier (KM) curves (right).

The year specific child 1-4MRs per 1000 person years from 2000 to 2005 were 3.57, 2.41, 3.33, 2.54, 4.44 and 1.55 respectively, consistent with the KM curves (5.2, right) for the respective years. There was a statistically significant difference in the 2002 mortality rates compared to the 2000 from the spatial and non-spatial analyses (see Table 5. 2). The knots used in the low rank kriging were fit using the mini-max space filling criteria, figure 5.3 shows distribution of 50 knots. The small faint (red) dots are the locations and the squares are the selected 50 knots, since the $\hat{\rho}$ (rate of spatial correlation decay) is fixed the number of knots only impacts on the computational speed and not on the accuracy of the estimates. The spatial variances were estimated for each spatial model using Bayesian MCMC with inverse gamma prior and hyper-parameters of 0.001. The ZIP model was used to cater for zero inflation and to cater for both the excess zeros and over-dispersion, the ZINB model was used. The zero inflated Negative Binomial spatial temporal model showed an over-dispersion estimate $\delta = 1.79 \pm 1.87[95\%\text{CI}(0.041; 6.935)]$ and the zero inflation estimate was

$\theta = 0.077 \pm 0.062$ [95%CI(0.010; 0.217)]. These values show that there is some over-dispersion and zero inflation, thus ignoring these entirely in the modelling may cause model instability.

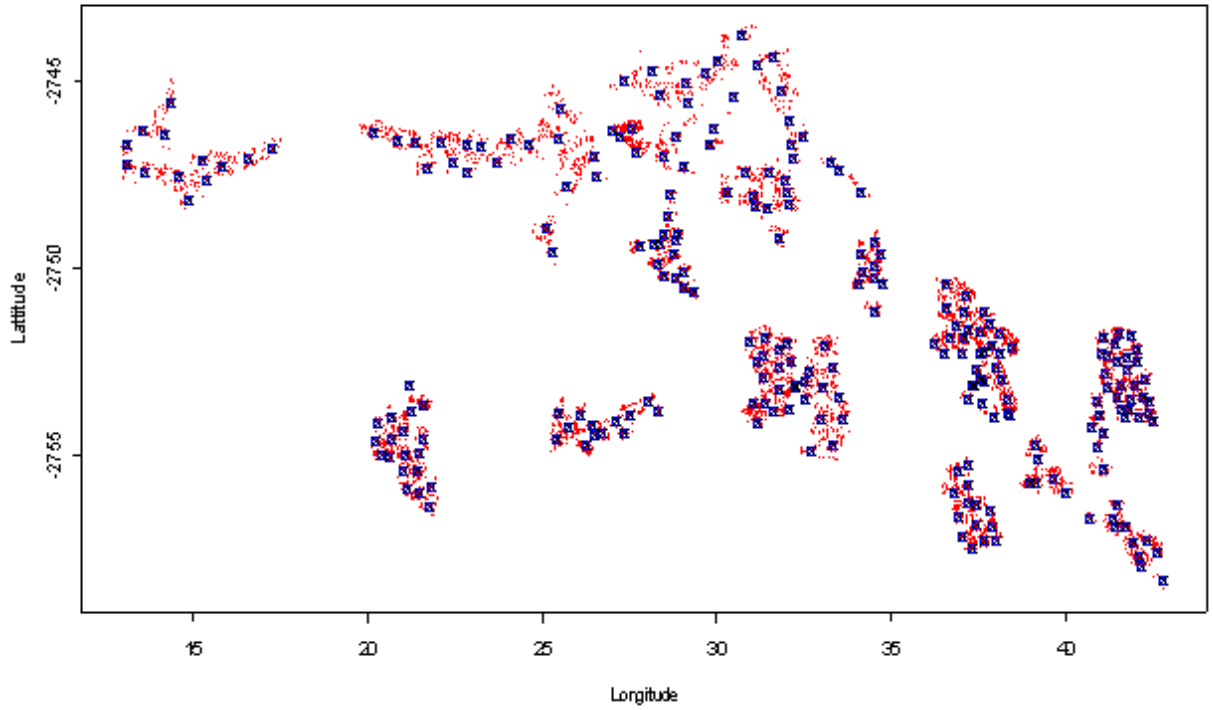


Figure 5.3: Knots fitted using the Mini-Max space filling criteria from Agincourt households 2000 to 2005.

5.3.2 Spatiotemporal confounding adjusted determinants of child HIV/TB mortality

Multivariable analysis was done using non-spatial and spatial models for child specific variables, maternal variables and household variables. Bayesian semi-parametric geo-statistical models with location-specific random effects were fit to estimate the degree of zero inflation, over-dispersion, temporal and spatial correlation in the child HIV/TB mortality data. Spatial models showed more stable results with smaller standard errors in comparison to the non-spatial, the mean-based adjusted hazard ratios [AHR=exp(mean)] estimations were

reported. The ages (Interval start ages) were used in the analysis as the duration of follow-up (offset variables) until death or survival between 1 and 5 years of age over the period 2000 to 2005. The outcome variable was treated as a discrete count of child death due to HIV/TB (0 = Alive and 1 = Death). The outcome was treated as such to get a more epidemiologically intuitive estimates for cohort data and to avoid identifiability problems associated with the logit link functions as highlighted earlier. Multivariable regression models were fit to account for spatial and temporal correlation; co-variables were added to the models to control for potential confounding in determine factors associated with child HIV/TB mortality (58). Poisson and Negative Binomial models were fit to the data as well as their zero inflated variants. The offset variable was fit using the piece wise exponential (p.e.m) with a first order autoregressive model temporal prior (see figure 5.4). Figure 5.4 shows the log-baseline hazard function for selected models, showing a narrower confidence interval for the spatial multivariable model in comparison to the non-spatial.

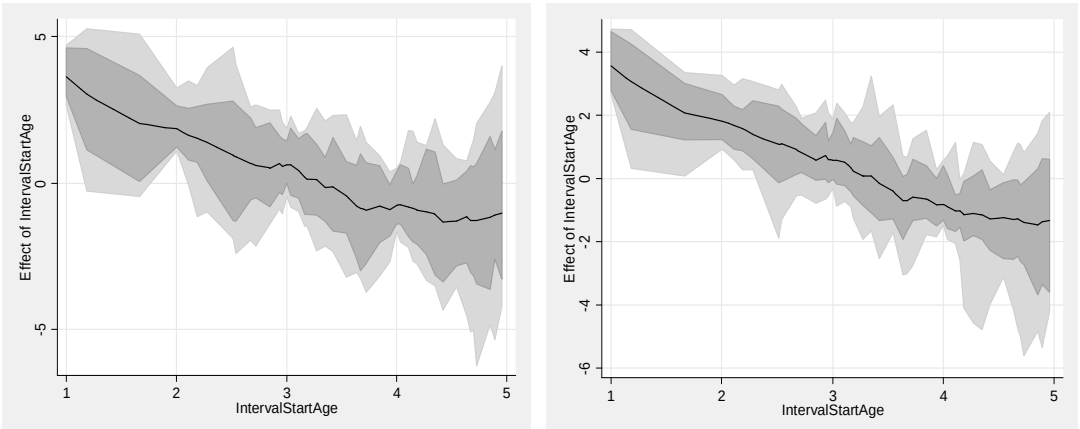


Figure 5.4: Non-spatial ZINB with AR(1) baseline hazard (left) and multivariable spatial ZINB with AR(1) baseline hazard (right).

Four models were fit in the multivariable analyses these were the non-spatial zero inflated Poisson (ZIP) model, spatial zero inflated Poisson model, non-spatial zero inflated Negative

Binomial (ZINB) model and spatial zero inflated Negative Binomial model. The first two were to cater for zero inflation and the last two were catering for zero inflation as well as over dispersion. The deviance information criteria (*DIC*) was used to measure model goodness of fit and the effective number of parameters (p_D) to measure model complexity (238). Generally the spatial models had higher DIC because of the spatial parameters and also the Negative Binomial models had greater DICs compared to the Poisson; this made it difficult to judge the best model based on the DIC alone. The effective number parameters p_D which assesses model simplicity (parsimony) was then used to determine the most parsimonious model amongst these four. The zero inflated Negative Binomial spatial model had the least number of effective parameters and selected as the "best" model (see table 5.1).

Table 5.1: Model goodness of fit non-spatial/spatial zero inflated Poisson and Negative Binomial models

	Non-spatial Zero Inflated Poisson	**nsZIP**	Spatial Zero Inflated Poisson	**sZIP**	Non-spatial Zero Inflated Negative Binomial	**nsZINB**	Spatial Zero Inflated Negative Binomial	**sZINB**
	Un-standardized	**Saturated**	**Un-standardized**	**Saturated**	**Un-standardized**	**Saturated**	**Un-standardized**	**Saturated**
$\bar{D}$	649.16	273.16	664.3	288.33	1674.77	1252.4	1865.41	1228.41
pD	246.51	246.51	247.7	247.7	83.86	104.72	47.95	36.23
DIC	1142.2	766.19	1159.79	783.79	1842.49	1461.86	1961.32	1300.88

The spatial multivariable analyses results showed several determinants of the child HIV/TB mortality in the categories consistent with the non-spatial analyses. Several variables were statistically significant at the 5% level, these were either child, mother or household related. The results from the analyses are shown in tables 5.2 and 5.3 for the ZIP and ZINB respectively. The ZIP models both had five out of the sixteen variables significant in the non-spatial and spatial analysis respectively, whereas the ZINB had eight and nine variables for non-spatial and spatial models respectively. The final spatiotemporal ZINB model was considered the best model which caters well for zero inflation, over-dispersion and spatiotemporal confounding.

Table 5.2: Non spatial and Spatial Multivariable Zero Inflated Poisson model

			Non-Spatial Zero Inflated Poisson					nsZIP	Spatial Zero Inflated Poisson					sZIP
Variable	**Category**	**Descriptive,n(%)**	**mean**	**sdev**	**2.5%.**	**median**	**97.50%**	**AHR**	**mean**	**sdev**	**2.5%.**	**median**	**97.50 %**	**AHR**
Gender	Female (ref)	8506 (50.50)												
	Male	8838 (49.50)	0.415	0.25	-0.12	0.42	0.93	1.51	0.38	0.25	-0.11	0.39	0.88	1.46
Year	2000(ref)	3360 (19.95)												
	2001	2826 (16.78)	0.49	0.30	-0.14	0.50	1.10	1.63	0.50	0.30	-0.10	0.52	1.02	1.65
	2002	2633 (15.63)	0.837*	0.37	0.16	0.86	0.54	2.31	0.81*	0.333	0.17	0.81	1.46	2.25
	2003	2496 (14.82)	0.56	0.43	-0.21	0.53	1.43	1.75	0.56	0.40	-0.32	0.58	1.27	1.75
	2004	3002 (17.82)	0.15	0.34	-0.50	0.15	0.80	1.16	0.17	0.32	-0.52	0.17	0.80	1.19
	2005	2527 (15.00)	-0.22	0.47	-1.14	-0.20	0.61	0.80	-0.21	0.45	-1.07	-0.20	0.65	0.81
Child's Nationality	Mozambican(ref)	6242 (37.06)												
	South African	10,602 (62.94)	-0.82	0.65	-2.11	-0.86	0.34	0.44	-0.70	0.53	-1.65	-0.69	0.36	0.50
Mother's Nationality	South African(ref)	10,883 (64.61)												
	Mozambican	5961 (35.39)	-0.11	0.67	-1.56	-0.11	1.25	0.90	-0.05	0.54	-1.07	-0.05	1.00	0.95
Mother’s Age at Birth	Mean years ±SD	29.00 ±7.60)	0.01	0.02	-0.02	0.01	0.04	1.01	0.01	0.02	-0.03	0.01	0.04	1.01
Mother's Status	Alive(ref)	16,736 (99.36)												
	Deceased	108 (0.64)	2.665*	0.81	0.90	2.74	3.90	14.37	2.456*	0.63	1.15	2.46	3.74	11.66
Father's Status	Alive(ref)	16,531 (98.14)												
	Deceased	313 (1.86)	0.98	0.66	-0.33	1.00	2.33	2.66	1.081*	0.49	0.06	1.09	2.04	2.95
Household head Gender	Female (ref)	3012 (33.98)												
	Male	5851 (66.02)	-0.49	0.26	-1.04	-0.50	0.01	0.61	-0.45	0.25	-0.96	0.44	0.02	0.64
Antenatal Visits	Mean visits ±SD	4.04±2.42	-0.185*	0.05	-0.29	-0.19	-0.09	0.83	-0.184*	0.05	-0.29	-0.18	-0.90	0.83
Pregnancy Parity	Mean ±SD	2.104 ±1.396	0.25	0.22	-0.14	0.24	0.67	1.28	0.25	0.21	-0.16	0.25	0.67	1.28
Socio-economic status	Most poor(ref)	1566 (17.67)												
	Very poor	1701 (19.19)	-0.17	0.32	-0.83	-0.17	0.41	0.84	-0.22	0.30	-0.82	-0.22	0.36	0.80
	Poor	1833 (20.68)	-0.20	0.35	-0.84	-0.20	0.51	0.82	-0.28	0.29	-0.85	-0.29	0.27	0.76

	Moderately poor	1905 (21.49)	-0.26	0.33	-1.03	-0.27	0.36	0.77	-0.30	0.32	-0.95	-0.29	0.29	0.74
	Least poor	1858 (20.97)	-0.856*	0.42	-1.60	-0.84	-0.04	0.42	-0.941*	0.42	-1.83	-0.94	-0.13	0.39
Cumulative household deaths	Mean ±SD	0.252 ±0.553	-0.16	0.25	-0.68	-0.17	0.29	0.85	-0.10	0.24	-0.62	-0.13	0.30	0.90
BirthParity_2	Mean ±SD	1.569 ±0.856	-0.369*	0.20	-0.77	-0.35	-0.03	0.69	-0.37*	0.18	-0.74	-0.37	-0.02	0.69
Total living household	Mean ±SD	1.569 ±0.884	0.01	0.02	-0.04	0.01	0.06	1.01	0.00	0.02	-0.04	0.00	0.04	1.00
Number of deaths in household	Mean ±SD	0.047 ±0.222	0.854	0.50	-0.11	0.86	1.80	2.35	0.89	0.48	-0.17	0.92	1.79	2.44
Total live births	Mean ±SD	2.030 ±1.855	-0.21	0.18	-0.60	-0.21	0.13	0.81	-0.22	0.18	-0.61	-0.21	0.12	0.80

* significant at 5% level of significance

The final zero inflated spatial Negative Binomial model results showed these nine predictors of child mortality: gender, nationality of child, death of mother, gender of head of household, year, antenatal clinic visits, socio economic status, birth order position and number of deaths in household. The boys were 36% more at risk of death compared to the girls adjusting for other variables and space-time confounders ($1.36 = e^{0.311}$,95%CI [1.08,1.88]). There was temporal effect on HIV/TB mortality risk for the year 2002 which was more than twice that of 2000 ($2.02 = e^{0.719}$,95%CI[1.13,4.00]). South African children were almost twice better protected to HIV/TB death compared to the Mozambican children after adjusting for several variables ($0.51 = e^{-0.683}$, 95%CI[0.24,0.96]). Orphans with a deceased mother were almost thrice at risk of HIV/TB death compared to those whose mother's were alive keeping other variables constant ($2.93 = e^{1.075}$,95%CI[1.29;6.93]). Adjusting for other factors male headed households were 42% less likely to experience a child death in comparison to female headed households ($0.58 = e^{-0.548}$,95%CI[0.45;0.75]).

Table 5.3:Non spatial and Spatial Multivariable Zero Inflated Negative Binomial model

			Non-Spatial Zero Inflated Negative Binomial					nsZINB	Non-Spatial Zero Inflated Negative Binomial					sZINB
Variable	**Category**	**Descriptive,n(%)**	**mean**	**sdev**	**2.5%.**	**median**	**97.50%**	**AHR**	**mean**	**sdev**	**2.5%.**	**median**	**97.50%**	**AHR**
Gender	Female (ref)	8506 (50.50)												
	Male	8838 (49.50)	0.31*	0.14	0.07	0.28	0.62	1.32	0.326*	0.17	0.08	0.31	0.63	1.36
Year	2000(ref)	3360 (19.95)												
	2001	2826 (16.78)	0.46	0.27	-0.05	0.48	1.05	1.62	0.55	0.23	-0.06	0.58	0.92	1.78
	2002	2633 (15.63)	0.74*	0.24	0.16	0.79	1.14	2.20	0.762*	0.31	0.12	0.72	1.38	2.05
	2003	2496 (14.82)	0.39	0.32	-0.19	0.40	1.01	1.48	0.45	0.31	-0.14	0.45	1.02	1.56
	2004	3002 (17.82)	0.00	0.25	-0.42	0.00	0.46	1.00	0.03	0.24	-0.40	0.03	0.49	1.03
	2005	2527 (15.00)	-0.54	0.41	-1.27	-0.53	0.26	0.59	-0.43	0.32	-1.08	-0.41	0.12	0.67
Child's Nationality	Mozambican(ref)	6242 (37.06)												
	South African	10,602 (62.94)	-0.41	0.32	-1.02	-0.35	0.09	0.71	-0.643*	0.31	-1.43	-0.68	-0.04	0.51
Mother's Nationality	South African(ref)	10,883 (64.61)												
	Mozambican	5961 (35.39)	0.05	0.32	-0.59	0.06	0.57	1.06	-0.12	0.33	-0.68	-0.19	0.57	0.83
Mother’s Age at Birth	Mean years (SD)	29.00 ±7.60)	0.01	0.01	-0.01	0.01	0.05	1.01	0.01	0.01	-0.01	0.01	0.03	1.01
Mother's Status	Alive(ref)	16,736 (99.36)												
	Deceased	108 (0.64)	1.479*	0.36	0.63	1.52	2.15	4.55	1.09*	0.47	0.25	1.08	1.94	2.93
Father's Status	Alive(ref)	16,531 (98.14)												
	Deceased	313 (1.86)	0.17	0.32	-0.55	0.24	0.70	1.27	0.11	0.44	-0.82	0.16	0.82	1.17
Household head Gender	Female (ref)	3012 (33.98)												
	Male	5851 (66.02)	-0.449*	0.16	-0.70	-0.49	-0.15	0.62	-0.54*	0.14	-0.79	-0.55	-0.29	0.58
Antenatal Visits	Mean visits ±SD	4.04±2.42	-0.156*	0.03	-0.21	-0.16	-0.09	0.85	-0.17*	0.03	-0.22	-0.17	-0.12	0.84
Pregnancy Parity	Mean ±SD	2.104 ±1.396	-0.04	0.11	-0.24	-0.03	0.18	0.97	0.08	0.11	-0.14	0.08	0.26	1.09
Socio-economic status	Most poor(ref)	1566 (17.67)												
	Very poor	1701 (19.19)	-0.32	0.18	-0.65	-0.32	0.08	0.72	-0.36	0.24	-0.81	-0.31	0.09	0.73

	Poor	1833 (20.68)	-0.37	0.20	-0.77	-0.38	0.05	0.68	-0.41	0.21	-0.79	-0.43	0.02	0.65
	Moderately poor	1905 (21.49)	-0.23	0.24	-0.78	-0.22	0.18	0.80	-0.28	0.24	-0.75	-0.26	0.22	0.77
	Least poor	1858 (20.97)	-0.945*	0.33	-1.57	-0.97	-0.28	0.38	-0.99*	0.27	-1.44	-0.97	-0.49	0.38
Cumulative household deaths	Mean ±SD	0.252 ±0.553	-0.10	0.14	-0.38	-0.09	0.16	0.92	-0.14	0.16	-0.48	-0.15	0.14	0.86
BirthParity_2	Mean ±SD	1.569 ±0.856	-0.291*	0.12	-0.54	-0.31	-0.04	0.74	-0.314*	0.11	-0.55	-0.31	-0.12	0.73
Total living household	Mean ±SD	1.569 ±0.884	0.01	0.01	-0.01	0.01	0.04	1.01	0.01	0.01	-0.01	0.01	0.04	1.01
Number of deaths in household	Mean ±SD	0.047 ±0.222	0.703*	0.27	0.14	0.71	1.16	2.03	0.681*	0.29	0.12	0.69	1.20	2.00
Total live births	Mean ±SD	2.030 ±1.855	0.03	0.09	-0.11	0.02	0.23	1.02	-0.06	0.09	-0.26	-0.07	0.08	0.94

* significant at 5% level of significance.

For every increase by one visit to the antenatal clinic the adjusted risk of death reduced by 16% ($0.84 = e^{-0.173}$, 95%CI[0.81;0.89]). The least economically poor households were 62% less likely to experience a child death compared to the most poor households ($0.38 = e^{-0.966}$ adjusted, 95%CI[0.24;0.61]). The birth order position was protective such that as this increased by one level the adjusted risk of child death decreased by 27% ($0.73 = e^{-0.311}$, 95%CI[0.58;0.89]). Lastly as the number of household deaths increased by one the adjusted risk of child HIV/TB death increased by two fold ($2.00 = e^{0.694}$, 95%CI[1.12;3.33]).

Posterior risk estimates were shown on smoothed maps showing the smoothed areas where child deaths occurred. Maps of posterior mean adjusted hazard rates (AHRs) show point estimates and extreme high and low risk areas (242). Risk factors of child HIV/TB mortality adjusting for potential spatiotemporal confounders were: being a boy, a former foreigner, losing your mother, having a female head of house, being socio-economically disadvantaged, being born earlier and coming from households were many deaths occurred. Protective to child HIV\TB mortality were increasing number of clinic visit and coming from households with more people. Figure 5.5 shows the maps of the households and the mean AHRs smoothed posterior maps adjusting for the aforementioned risk factors, spatial and temporal potential confounders.

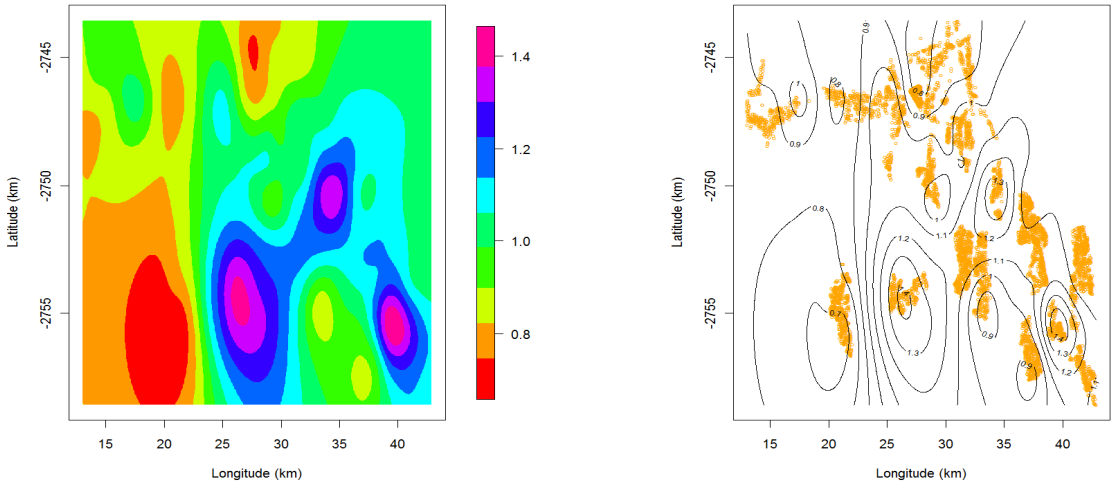


Figure 5.5: Posterior adjusted Hazard rates (AHRs) smoothed map (left) and contour AHRs map (right) for child HIV/TB mortality 2000 to 2005.

5.4 Discussion and Conclusions

In our study we found that maternal orphans experience a three-fold greater risk of death due to HIV/TB than children whose mothers were still living (AHR=2.93(95%CI[1.26;6.93])). This was adjusted for potential spatiotemporal confounding, number of household deaths, birth order position and parity, child and mother's nationality, gender of house head and gender of child, which were all significant risk factors. Increased antenatal clinic visits and cumulative total of adults in the house were protective of the child's death. These risk factors are interrelated and fall into either child specific, mother related, household socio-economic status, temporal or spatial. Risk map analysis showed that there were three hotspots central, south easterly and south-westerly over the years 2000 to 2005, similar hot spots were seen for the year 2004 alone and also in another study considering only the areal (village) level and temporal random effects (51). The trends seen in this paper are consistent with other studies in South Africa, in the same site and elsewhere. A study in Agincourt over the past two decades showed that HIV/TB accounted for about a third of the under-five deaths from 2002 to 2005 (243). A similar non-spatial study on the impact of parental presence on child

mortality in Agincourt, 1997 to 2005 showed that death of a mother was the most predictive of the child's death (OR=5.22,95%CI[2.37;11.54]) (244). The study also showed that former Mozambican mother's children were more than a third at greater risk of mortality than the South African's children (OR=1.34,95%CI[1.09;1.65]). Similarly in Mtubatuba of rural KwaZulu-Natal a South African Province, the mother's death increased the mortality risk four-fold adjusting for child, household and mother's variables (245).

There are different schools of thoughts on how the maternal orphan HIV/TB mortality trends can be explained, mainly the biological and socio-economically. Mothers who die due to the AIDS, would most likely have passed the virus to their children, regardless of the stage at which this occured (in utero, intrapartum and postnatal). Policies that are in place currently favour local South African mothers, thus making it difficult for breast feeding foreign mother's to get access to freely available resources. Socially if the mother (who will most likely be the remaining bread winner) passes on, this leaves AIDS orphans with no guardian or their grandparents who can no longer work. This culminates in the deaths of under fives due to hunger, malnutrition, weakened immune system and lack of adequate support due to the double barrelled impact of being an HIV positive refugee and orphan. Although we concentrated on HIV to have been the main cause of death for these under five children we acknowledge that TB could have played a significant role leading to death, however the modes of transmissions are more complex than simply linking child and mother deaths. Also the ICD10 technique used to determine cause of death does not distinguish these very well for children. Although Agincourt is in rural South Africa, the use of available health facilities seems to be increasing using antenatal clinic visits as a proxy.

The other risk factors were mainly related to poverty as direct proxies indirect consequences of it. We found that poorer socio-economic households had higher child HIV/TB mortality rates. A finding which corresponds with a study done in rural Kwazulu-Natal Province (246). Poverty is a major contributor to child deaths in Africa, every 3.6 seconds one person dies of starvation, usually a child under the age of 5. One of the Millennium Development Goals (MDGs) is to: "Eradicate extreme poverty and hunger" (199, 247). The child grant provided by the South African government is an amount of ZAR240 p/m (US$30) and can be all that some families in the study area have to live on. The elderly also get an old-age pension of ZAR940p/m (US$115) which becomes the main source of livelihood of the families often averaging nine persons per household with a range of 1-43. Case et. al (2000) showed that female old-age pensioners cared for the children and improved the girl child's likelihood to go to school in Agincourt (248). These latter studies are also revealing some gender differentials in access of health services, schooling and paternal treatment which are still prevalent in this rural part of South Africa.

Controlling for potential confounding in epidemiological analyses gives strength to the estimates derived from such modelling procedures. Our modelling approach caters for spatial and temporal confounders and also for large zero inflated spatiotemporal sparse data. Spatiotemporal Bayesian modelling provided good estimates and maps which can aid interventional programmes to be more effective in the fight against HIV/TB. There are some modelling limitations which if addressed in future research will improve on the epidemiological modelling. The temporal component was discretized into whole years; this could have been modelled as continuous and allow for formulation of time varying covariates and spatiotemporal interactions. As a way of dealing with identifiability issues and getting more epidemiologically correct estimates; members of the exponential (log link) family were

used, other less flexible family of distributions could be also used for sensitivity testing. Mostly non-informative priors have been used; experts' opinion could be sought to build and apply more informative priors in our modelling. In dealing with the "big N" problem several approaches can be used as well. The package BayesX whilst it cannot address all the modelling issues but provides a very useful Bayesian spatiotemporal modelling algorithms which are applicable even in a resource scare environment.

Our results are additional facets of evidence useful for policy makers advocating foreigners better access to health access and child support faculties, increase and diversification of HIV mother to child transmission awareness programmes in rural parts of developing countries. This study provides more evidence towards the South African government's revised prevention of mother to child transmission (PMTCT) policy effected on 1 April 2010. The policy allows HIV positive mothers to be on ante-retro-virals (ARV) with a CD4 count of 350 (national cut off is 200) or less, symptomatic HIV regardless of CD4 and also pregnant mothers from 14 gestational weeks onwards. Infants can now be tested without the mother's consent to save the child (249). Child survival is dependent on the survival of the mother, hence the need to strengthen antiretroviral treatment. South African policy advocating for maternal longevity need to be strengthened. In an address to the nation the South African president said: "One of the most important results of the roll-out of antiretroviral therapy among the general population will be the extension of the lives of AIDS sick parents leading to a dramatic decline in the number of orphans" (250). The main challenge as with any other policy is implementation. We suggest that targeting to reduce mother-to-child transmissions by implementing PMTCT interventions and campaigns these need to reach rural villages and households which are usually neglected.

5.5 Acknowledgements

This work was made possible by doctorial fellowship support from World Health Organisation, Tropical Disease Research (WHO/TDR) and Council for Scientific and Industrial Research (CSIR), statistical support from the Swiss Tropical and Public Health Institute, Basel Switzerland. The Agincourt HDSS is funded by the Wellcome Trust UK (Grant number 058893/Z/99/A and 069683/Z/02/Z) and the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew Mellon and Hewlett Foundations, U.S.A. Special thanks goes to Benn Sartorius for data extraction, community leaders, field workers, supervisors and the Agincourt community.

5.6 Authors' contributions

EM drafted the manuscript, did the statistical analysis and wrote this paper. PV and KK gave their expert advice on methodology and epidemiology aspects of this paper respectively.

5.7 Competing interests

The author declares that he have no competing interests.

5.8 Appendix: Zero Inflated Poisson and Negative Binomial Hierarchical Bayesian models

Observational outcome data can be modelled by different known statistical distributions including the flexible exponential family. Mortality outcomes of the count of deaths can be modelled using the basic log-linear Poisson model were the outcomes $y_i|X_i \sim Po(\lambda_i)$ are conditionally independent with mean and variance $\lambda_i = exp(\eta_i)$ with linear predictor $\eta_i = X_i\beta$. The commonly applied logistic regression models which have logit link functions have

been seen to cause identifiability problems, resulting in unstable model especially for spatial models (23). We therefore treated the outcome death as a discrete (count) variable since this provides better estimates of prevalence and relative risk compared to the odds ratio which results from treating it as binary outcome (90, 231).The Poisson model can also incorporate the exposure time by introducing piece-wise exponential offsets o_i resulting in $y_i|X_i \sim Po(\mu_i), \mu_i = exp(o_i + \eta_i)$ and in log-linear form $y_i|\eta_i \sim Po(\mu_i)$ (133). In this paper we extend this model in three dimensions to address different inherent problems in our hierarchical data structure. Firstly the problem of over-dispersion when the mean is less than the variance which can be addressed by use of conditionally independent Negative Binomial (NB) distribution (251). We then extend our model incorporate the dispersion parameter $\delta > 0, y_i|\eta_i, \delta \sim NB(\mu_i, \delta)$. Secondly we have the problem of excess of zeros known as zero inflation, this is

addressed by use of a latent process that inflates the number of zero points with a weighting θ for extra zeros and $1 - \theta$ for the ZIP and ZINB where $0 < \theta < 1$. We then obtain the conditionally independent Zero Inflated Poisson $y_i|\eta_i, \theta \sim ZIP(\mu_i, \theta)$ and Zero Inflated Negative Binomial (ZINB) $y_i|\eta_i, \delta, \theta \sim ZINB(\mu_i, \delta, \theta)$ whose hierarchical model structures and likelihoods are shown below. Thirdly modelling which is flexible to include; time varying covariates, temporal random effects, spatial random effects, non-linear covariates and desired interactions. This was done by extending the linear predictor $\eta_i = X_i\beta$ to a structured additive predictor:

$$\eta_i = \sum_{j=1}^{p} f_j\left(z_{ij}\right) + X_i\beta + f_{spat_i}(s_k) + \varepsilon_{s_k}$$

where $f_j(z_{ij})$ are nonlinear effects of continuous covariate z_{ij}, X_i is the linear (fixed) effects design matrix and vector of parameters β, $f_{spat_i}(s_k)$ is the structured random spatial at household $s_k \in$ s (all households with children in the area) and ε_{s_k} being the unstructured

non-spatial random effect. Due to the complexity of the spatiotemporal data (correlated in time and space) and their large size. Statistical methods applied to spatial data which have spatial autocorrelation often underestimate the standard error and thus the statistical significance is overestimated (91). We used the Bayesian estimation called the Markov Chain Monte Carlo (MCMC), which require Gibbs sampling and Metropolis Hastings within Gibbs (85). Modelling spatial longitudinal data which are highly auto-correlated has been made possible by introduction of the Hierarchical Bayesian (HB) modelling approach. This decomposes a complex problem into a series of simpler conditional levels such that at any given level in the hierarchy inference conditionally relies on hypotheses made at higher levels (85). The HB model constitutes of four levels: firstly the data level which specifies the conditional distribution of data given parameters i.e $y_i|.$; secondly a process level which specifies the conditional distributions of process given their parameters (these are usually logit and log functions of Generalised Linear Models); thirdly a parameter level with conditional parameter distributions given priors and lastly hyper-parameter level which specifies the prior distributions for hyper-parameters. The posterior distribution can then be used for full Bayesian inference of the ZIP and ZINB models given the following full conditional distributions. To estimate the parameters Gibbs sampling is used by sampling from the full conditional distributions for posteriors with known closed forms(162). For those without the closed form we use the Metropolis Hastings (MH) random walk algorithms (252).

Zero Inflated Poisson model

Level 1: Data

The ZIP with a structures semi-parametric predictor offers a flexible way to model the zero inflation parameter (θ).

$y_i|\eta_i \sim Po(\mu_i)$ with probability density function

$$P(y_i|\eta_i,\theta) = \begin{cases} \theta + (1-\theta)\dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \; if \; y_i = 0 \\ (1-\theta)\dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \; if \; y_i \neq 0 \end{cases}$$

With expectation $E(y_i|\eta_i,\theta) = (1-\theta)\mu_i$ and variance

$V(y_i|\eta_i,\theta) = (1-\theta)(\mu_i) + \theta(1-\theta)\mu_i^2$

Level 2: Processes

$\eta_i = \boldsymbol{X}_i\beta + f_{spat_i}(s_k) + \varepsilon_{s_k}$;

$\theta \sim Uniform[0,1]$

$\boldsymbol{X}_i$ is the linear (fixed) effects design matrix and vector of parameters β, $f_{spat_i}(s_k)$ is the structured random spatial effect and ε_{s_k} being the unstructured non-spatial random effect.

Level 3: Parameters

Priors for $P(\beta), P(f_{spat_i}|\tau^2_{spat}), P(\varepsilon_{s_k}|\tau^2_{error}), P(\theta)$. Uniform priors for $P(\beta)$ and $P(\theta)$. Multivariate Gaussian random field (GRF) priors for $P(f_{spat_i}|\tau^2_{spat})$ and $P(\varepsilon_{s_k}|\tau^2_{error})$ are Gaussian independent and identically distributed (i.i.d) random effects

Level 4: Hyper-parameters and likelihood

Hyperpriors for $P(\tau^2_{spat}), P(\tau^2_{error})$ and $g(\tau^2)$ are all Inverse Gamma ~ $IG(a^*, b^*)$; with hyper parameters a*=b*=0.001(133). After some computations the likelihood of the ZIP models is:

$$L(y_i|\eta_i,\theta) = \prod_{i=1}^{n} P(y_i|\eta_i,\theta)$$
$$\propto \exp\left\{\sum_{y_i=0} \log(\theta + (1-\theta)\exp(-\mu_i)) + w_0 \log(1-\theta) + \sum_{y_i\neq 0}(-\mu_i + y_i\log(\mu_i))\right\}$$

Where w_0 is the number of units with strictly positive responses.

Zero Inflated Negative Binomial model

Level 1: Data

The ZINB with a structures semi-parametric predictor offers a flexible way to model the variability of the data catering for over-dispersion (δ), zero inflation(θ).

$y_i|\eta_i,\delta,\theta \sim ZINB(\mu_i,\delta,\theta)$ with probability density function

$$P(y_i|\eta_i,\delta,\theta) = \begin{cases} \theta + (1-\theta)\left(\dfrac{\delta}{\delta+\mu_i}\right)^{\delta} \; if \; y_i = 0 \\ (1-\theta)\dfrac{\Gamma(y_i+\delta)}{\Gamma(y_i+1)\Gamma(\delta)}\left(\dfrac{\delta}{\delta+\mu_i}\right)^{\delta}\left(\dfrac{\mu_i}{\delta+\mu_i}\right)^{y_i} \; if \; y_i \neq 0 \end{cases}$$

With expectation $E(y_i|\eta_i,\theta) = (1-\theta)\mu_i$ and variance

$V(y_i|\eta_i,\delta,\theta) = (1-\theta)\left(\mu_i + \frac{\mu_i^2}{\delta}\right) + \theta(1-\theta)\mu_i^2$.

Level 2: Processes

$\eta_i = \boldsymbol{X}_i\beta + f_{spat_i}(s_k) + \varepsilon_{s_k}$;

$\delta|b \sim Gamma(a,b)$; $\theta \sim Uniform[0,1]$

$\boldsymbol{X}_i$ is the linear (fixed) effects design matrix and vector of parameters β, $f_{spat_i}(s_k)$ is the structured random spatial effect and ε_{s_k} being the unstructured non-spatial random effect. We set a = 1 and assume a Gamma hyper-prior for $b \sim Gamma(\alpha 1, \alpha 2)$.

Level 3: Parameters

Priors for $P(\beta), P(f_{spat_i}|\tau^2_{spat}), P(\varepsilon_{s_k}|\tau^2_{error}), P(\theta), P(\delta|b)$. Uniform priors for $P(\beta)$ and $P(\theta)$, Gamma priors for $P(\delta|b)$. Multivariate Gaussian random field (GRF) priors for $P(f_{spat_i}|\tau^2_{spat})$ and $P(\varepsilon_{s_k}|\tau^2_{error})$ are Gaussian independent and identically distributed (i.i.d) random effects

Level 4: Hyper-parameters and likelihood

Hyperpriors for $P(\tau^2_{spat}), P(\tau^2_{error}), P(b)$ and $g(\tau^2)$ are all Inverse Gamma ~ $IG(a^*, b^*)$; with hyper parameters a*=b*=0.001(133). After some computations the likelihood of the ZINB models is:

$$L(y_i|\eta_i,\delta,\theta) = \prod_{i=1}^{n} P(y_i|\eta_i,\delta,\theta)$$
$$\propto \exp\left\{\sum_{y_i=0} \log\left(\theta + (1-\theta)\left(\frac{\delta}{\delta+\mu_i}\right)^{\delta}\right) + w_0\left(\log(1-\theta) - \log(\Gamma(\delta)) + \delta\log(\delta)\right) + \sum_{y_i\neq 0}\left(\log(\Gamma(y_i+\delta)) + y_i\log(\mu_i) - (y_i + \delta)\log(\delta+\mu_i)\right)\right\}$$

Where w_0 is the number of units with strictly positive responses.

Chapter 6

6.0 Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural Northeast South Africa

Eustasius Musenge [a,b§], Tobias Chirwa [b], Kathleen Kahn [a,d,e], Penelope Vounatsou [c]

[a] MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[b] Biostatistics and Epidemiology Division, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[c] Swiss Tropical and Public Health Institute, Basel, Switzerland.

[d] Centre for Global Health Reasearch, Umeå University, Umeå, Sweden

[e]I NDEPTH Network, Accra, Ghana.

This paper is in press in the International Journal of Applied Earth Observation and Geoinformation and attached in appendix 9.2

Chapter 6 abstract

Longitudinal mortality data with few deaths usually have problems of zero-inflation. This paper presents and applies two Bayesian models which cater for zero-inflation, spatial and temporal random effects. To reduce the computational burden experienced when a large number of geo-locations are treated as a Gaussian Field (GF) we transformed the field to a Gaussian Markov Random Fields (GMRF) by triangulation. We then modelled the spatial random effects using the Stochastic Partial Differential Equations (SPDE). Inference was done using a computationally efficient alternative to Markov chain Monte Carlo (MCMC) called Integrated Nested Laplace Approximation (INLA) suited for GMRF. The models were applied to data from 71,057 children aged 0 to 9 years from rural northeast South Africa living in 15,703 households over the period 1992 to 2010. We found protective effects on HIV/TB mortality due to greater birth weight, older age and more antenatal clinic visits during pregnancy (adjusted RR (95% CI)): 0.73(0.53;0.99), 0.18(0.14;0.22) and 0.96(0.94;0.97) respectively. Therefore childhood HIV/TB mortality could be reduced if mothers are better catered for during pregnancy as this can reduce also mother-to-child transmissions and contribute to improved birth weights. The INLA and SPDE approaches are computationally good alternatives in modelling large multilevel spatiotemporal GMRF data structures.

6.1 Introduction

Public health data on mortality have been growing increasingly rich as more accurate information on “who”, “where” and “when” becomes available. These form hierarchical (multilevel) data structures which are correlated such that person-level (“who”) information can be repeated, geo-statistical (“where”) data often has spatial correlation, and temporal (“when”) data can be auto-correlated. Statistical techniques are usually based on independent observations, but when applied to multilevel data structures they often underestimate the standard errors. As a result of this the statistical significance is overestimated leading to erroneous results and subsequent inferences (91). This defeats the main goal in epidemiological analysis, which is to identify and quantify correctly any exposures, behaviours and characteristics that may modify a population’s or individual’s risk and use these to implement more appropriate interventions (253).

In modelling hierarchical data we can take into account spatial and temporal correlations by introducing in the model spatiotemporal random effects. Several other hurdles have to be overcome when modelling hierarchical mortality data such as: zero inflation when there is a greater proportion of non-occurrence for an outcome, handling large data structures, repeated measures and estimating many parameters rapidly and accurately. Bayesian techniques with the aid of the Markov chain Monte Carlo (MCMC) simulation methods have successfully overcome these hurdles and fit spatiotemporal random effects for reasonably sized geo-locations of Gaussian fields (GF) (20, 92-95). However as the number of geo-locations increases MCMC computations of a dense GF $m \times m$ spatial correlation matrix become infeasible or extremely slow in the order of power three $\left(\mathrm{O}(m^3)\right)$, this problem is popularly known as the “big m” or “big N” (254). Several approaches have been used to resolve the "big m". Banerjee et al. (2003) give brief summaries of these: sub-sampling, spectral, lattice,

dimension reduction and course fine coupling methods (234, 254-258). Generally these techniques attempt to reduce the dimension of the GF by selecting a "representative" sub-sample or fixing some parameters or changing the scale from continuous to discrete with the aim of reducing the computational burden in running MCMC simulations.

We addressed this problem using techniques proposed by Rue et al. (2005) who changed the continuous scale GF to a discrete scale Gaussian Markov Random Field (GMRF), for the Matérn family of covariance structures (96). More recently Lindgren et al. (2011) provide the detail of how the GF and GMRF relate via Stochastic Partial Differential Equations (SPDE) using basis functions (259, 260). Secondly we performed inference and prediction using Integrated Nested Laplace Approximation (INLA) well suited for GMRF as opposed to the commonly used MCMC (261). Hence we greatly reduced the computational burden and could run in hours what usually took days having reduced the computational operations for a spatiotemporal model from power 3 to power 1.5 $\left(\mathrm{O}(m^3) \rightarrow \mathrm{O}(m^{3/2})\right)$.

The aim of this paper is to discuss and apply a Bayesian model that can handle large zero-inflated spatiotemporal observational data on mortality producing reliable estimates speedily. In the next section we explore the Bayesian methods and model fitting inference, prediction and goodness of fit. In Section 6.3 we apply the discussed Bayesian spatiotemporal model to data from the Agincourt health and socio-demographic surveillance system in rural South Africa which has 71,057 children aged 0 to 9 years living in 15,703 households over the period 1992 to 2010. In the concluding section, we discuss the merits of our model and distil the public health implications of our results in interventional studies.

6. 2 Methods

6.2.1 Spatiotemporal model structure

The outcome $y_i(s_j,t)$ was the observed HIV/TB related death of a child $i=1,.....,N$ from a given household $j=1,.....,m$ in a specific year $t=1,.....,T$ which is a realisation of the spatio-temporal process $y(.,.)\in Y(.,.)$. Assuming the outcomes distribution belongs to the exponential family of distributions, we can fit flexible structural additive models belonging to the generalized linear mixed models (GLMM) (86, 262). Our data may be represented by the equation:

$$y_i(s_j,t)=X(s_j,t)\boldsymbol{\beta}+\sum_{j=1}^{n_f} f^{(j)}(u_{ji},t)+\varepsilon_i(s_j,t) \tag{6.1}$$

where $X(s_j,t)$ is the design matrix with fixed p covariates, $\boldsymbol{\beta}=(\beta_0,.....\beta_p)'$ is the regression coefficients vector, $f(.)$ which is one of the $f^{(j)}$ used to relax the linear relationship or introduce random effects or both and $\varepsilon_i(s_j,t)\sim N(0,\sigma_\varepsilon^2)$ are the error terms which are neither temporally nor spatially correlated (260). As our data are spatially and temporally correlated we can introduce random effects $f(.)=f(s_j,t)$ a Gaussian random field with a first order autoregressive temporal effect $\xi(s_j,t-1)$ and coefficient ϕ and zero mean multivariate normal (temporally independent) spatial effects $\omega(s_j,t)\sim MVN\left(0,\Sigma=\sigma_\omega^2 C\left(\left\|s_j-s_k\right\|=h\right);\ j\neq k\right)$ resulting in the equation:

$$f(s_j,t)=\phi\xi(s_j,t-1)+\omega(s_j,t) \tag{6.2}$$

where $|\phi|<1$ in case of stationarity, $\xi(s_j,1)\sim N\left(0,\sigma_\omega^2/(1-\phi^2)=1/\tau_\omega(1-\phi^2)\right)$ and the spatial effect is second order stationary and isotropic. When the spatial correlation follows a

Matérn covariance structure we obtain $C(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^{\nu} K_{\nu}(\kappa h)$ for the Euclidean distance lags *h*. The parameter ν measures the degree of smoothness and also the order of the modified Bessel function (when $\nu > 0$) of second kind K_{ν} and finally $\kappa > 0$ is the scaling parameter with a range $\rho = \frac{\sqrt{8\nu}}{\kappa}$ where the spatial correlation is close to 0.1 for each ν (259, 261)·

6.2.2 Zero Inflated Poisson and Binomial spatiotemporal models

Observational binary outcome data are commonly analysed using the logistic regression model, which has a logit linear predictor in the GLMMs canonical link structure. However this model has problems of instability especially with spatial random effects, which would be exacerbated due to zero inflation (23). In epidemiological cohort studies a relative risk is preferable to an odds ratio as this caters for temporality and also a more intuitive measure of burden of morbidity or mortality (90, 231). In light of this and in an endeavour to have better fitting models, two models that can handle zero inflation were employed. The two conditionally independent models fit were the zero inflated Poisson and zero inflated Binomial with a log and a logit canonical link functions respectively: $\boldsymbol{y}_t|\boldsymbol{\eta}_t, \theta_t \sim ZIP(\boldsymbol{\mu}_t, \theta_t)$ and $\boldsymbol{y}_t|\boldsymbol{\eta}_t, p_t, \theta_t \sim ZIBin(n_t, p_t, \theta_t)$.

The mortality outcome (count/binary) data $\boldsymbol{y}_t(s_j)$ observed at the households in the Agincourt area are zero inflated and assumed to follow either a Poisson (count) or Binomial distribution (binary). We will occasionally drop the s_j for notational convenience in the rest

of the article. We therefore resorted to the zero inflated models to cater for the imbalance due to many zeros. The model that takes care of zero inflation (θ_t) can then be represented as:

$$\pi(\boldsymbol{y}_t|\boldsymbol{\eta}_t,\theta_t)=\begin{cases}\theta_t+(1-\theta_t)\pi(0|\boldsymbol{\eta}_t,\theta_t) & \text{if } \boldsymbol{y}_t=0\\ (1-\theta_t)\pi(1|\boldsymbol{\eta}_t,\theta_t) & \text{if } \boldsymbol{y}_t\neq 0\end{cases} \tag{6.3}$$

With canonical links of the expected means: $g\left(E(y_t(.))\right)=\log(\boldsymbol{\mu}_t)=\boldsymbol{\eta}_t$ and

$g\left(E(y_t(.))\right)=\text{logit}(p_t)=\boldsymbol{\eta}_t$ with means are $\boldsymbol{\mu}_t=\exp(\boldsymbol{\eta}_t)$ and $p_t=\dfrac{\exp(\boldsymbol{\eta}_t)}{1+\exp(\boldsymbol{\eta}_t)}$ for the Poisson and Binomial distributions, respectively. A spatiotemporal canonical link (linear predictor) model can be expressed as

$$\begin{aligned}\boldsymbol{\eta}_t(s_j)&=X_t(s_j)\boldsymbol{\beta}+f_t(s_j)+\boldsymbol{\varepsilon}_t(s_j)\\ f_t(s_j)&=\phi\xi_{t-1}(s_j)+\omega_t(s_j)\end{aligned} \tag{6.4}$$

where $\boldsymbol{\varepsilon}_t(s_j)\sim N(\boldsymbol{0},\sigma_\varepsilon^2 I_m)$ with identity matrix I_m of dimension $m\times m$, $\omega_t\sim N(\boldsymbol{0},\Sigma=\sigma_\omega^2\tilde{\Sigma})$ with a stationary AR(1) process $\boldsymbol{\xi}_1\sim N\left(\boldsymbol{0},\Sigma/(1-\phi^2)\right)$ (260). The Σ is a dense GF m by m dimensional matrix from a Matérn distribution with scale and smoothness parameters κ and ν (which is fixed in all our computations) respectively. As the size of m increases computations become increasingly more difficult due to the “big m” as previously highlighted.

6.2.3 Solving the "big m" using the SPDEs to estimate the spatial random effects

To resolve the computational burden associated with the GF Matérn covariance function we used a technique that changes this to a GMRF proposed by Rue and Held (2005)(96). Briefly

the locations are converted into areal triangulations firstly by making them the initial triangle vertices before adding more vertices for proper triangulation which extends the grid and are very useful for prediction. Figure 6.1 shows how we employed triangulation to our data, the diagram on the right was made assuming households within 500 m were similar we fit 488 vertices and 938 triangles.

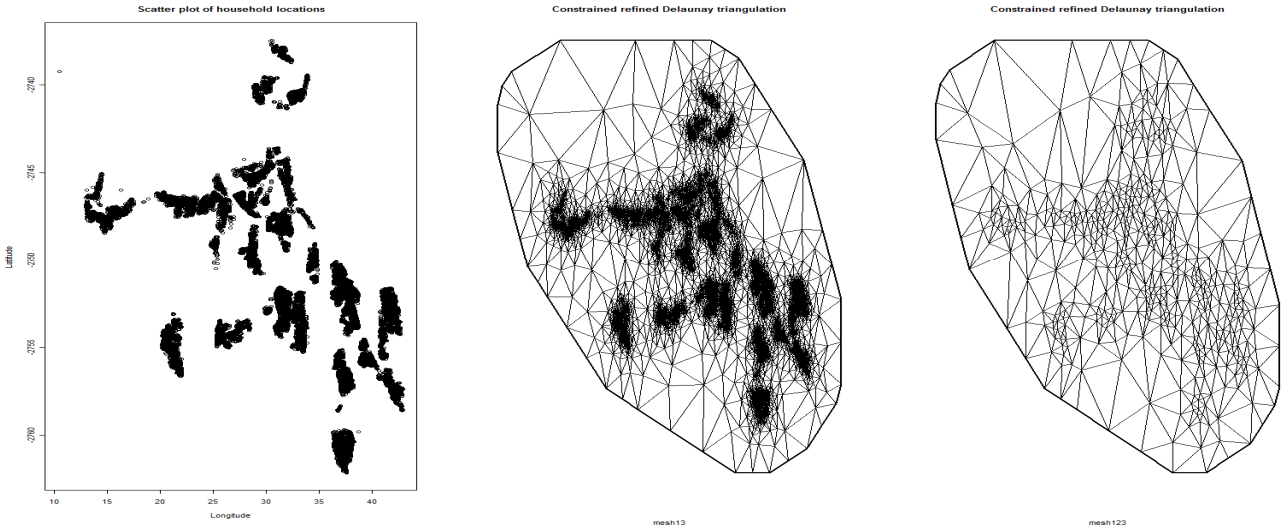


Figure 6.1: Agincourt sub-district: original household locations (left), triangulation of all households (centre) and triangulation of households within 500m apart (right).

The SPDE technique redefines the Matérn field as a finite ”representative” linear combination of basis functions on a triangulation of the locations (260). Hence for our spatiotemporal random effects in GMRF representation we get $\tilde{\omega}_t \sim N(\mathbf{0}, \Sigma = Q_s^{-1})$ and

$\xi_1 \sim N\left(\mathbf{0}, \frac{Q_s^{-1}}{1-\phi^2} = Q_T^{-1}\right)$ where Q_s is a sparse time-invariant precision matrix with dimension m^* vertices from the triangulations. A joint spatiotemporal GMRF

$\tilde{f}_t \sim N\left(\mathbf{0}, Q^{-1} = \left(Q_T \otimes Q_s\right)^{-1}\right)$ whose precision matrix is the Kronecker product of the temporal and spatial precision matrices; is such that $\tilde{f}_t \approx \boldsymbol{B} f_t$ where the basis $\boldsymbol{B}$ is a sparse

matrix with unit elements for matching triangle vertices and zero's elsewhere (259).

Therefore equation (6.4) becomes $\boldsymbol{\eta}_t = X_t\boldsymbol{\beta} + \boldsymbol{B}f_t + \boldsymbol{\varepsilon}_t$, where we let $\boldsymbol{x}_t = \{\boldsymbol{\beta}, \boldsymbol{B}, \theta_t\}$ be the vector of latent Gaussian fields and $\boldsymbol{\varphi}_t = \{\sigma_\omega^2, \phi, \kappa, \sigma_\varepsilon^2\}$ being a vector of unknown parameters. We can thus express our model into a hierarchical Gaussian latent variable fashion as follows, stage 1-observational equation, stage 2- Latent Gaussian field and stage 3- parameter model (263):

$$\begin{aligned} &\text{stage } 1: \boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\varphi}_t \sim N\left(X_t\boldsymbol{\beta} + \boldsymbol{B}f_t; \sigma_\varepsilon^2 I_m\right) = N\left(\boldsymbol{\mu}_{y_t}(\boldsymbol{\varphi}_t) = A\boldsymbol{x}_t, Q_{y_t}^{-1}(\boldsymbol{\varphi}_t)\right) \\ &\text{stage } 2: \boldsymbol{x}_t \mid \boldsymbol{\varphi}_t \sim N\left(\boldsymbol{\mu}_{x_t}(\boldsymbol{\varphi}_t); Q_{x_t}^{-1}(\boldsymbol{\varphi}_t)\right) \\ &\text{stage } 3: \boldsymbol{\varphi}_t \sim \pi(\boldsymbol{\varphi}_t) \end{aligned} \tag{6.5}$$

where the precision matrices $Q_{\cdot}^{-1}(.)$ are either small enough (for easier multiple factorisation) or sparse (264). These models cover a wide range of models and can easily be estimated using INLA as shown in the next subsection.

6.2.4 Bayesian inference using INLA

In accordance with the Bayesian paradigm we aim to find the posterior distribution of the processes and parameters updated by data (94).This could be expressed as:

Probability (process, parameters | data) $\propto$ *Likelihood (data| process, parameters)*$\times$ *Probability(process| parameters)* $\times$ *Probability(parameters)*

Applying this expression to our model, letting $\Theta = \left\{\boldsymbol{\beta}, \phi, \theta_t, \sigma_\omega^2, \kappa, \sigma_\varepsilon^2\right\}$ denote the vector of all parameters and dropping the subscripts to present in vector form, $\boldsymbol{\xi} = \{\xi_t\}$ and data $\boldsymbol{y} = \{y_t\}$ (260, 261), their joint posterior distribution is thus:

$$\pi(\Theta,\boldsymbol{\xi},\boldsymbol{\eta}\,|\,\boldsymbol{y}) \propto \pi(\boldsymbol{y}\,|\,\Theta,\boldsymbol{\xi},\boldsymbol{\eta})\,\pi(\boldsymbol{\eta}\,|\,\boldsymbol{\xi},\Theta)\,\pi(\boldsymbol{\xi},\Theta) \quad \text{.....(6.6)}$$

Figure 6.2 gives a simplified pictographic view, where level 1 are the data and assumed distributions, level 2 is a process, level 3 and 4 are parameters and level 5 gives the default hyper-parameters used in the INLA package.

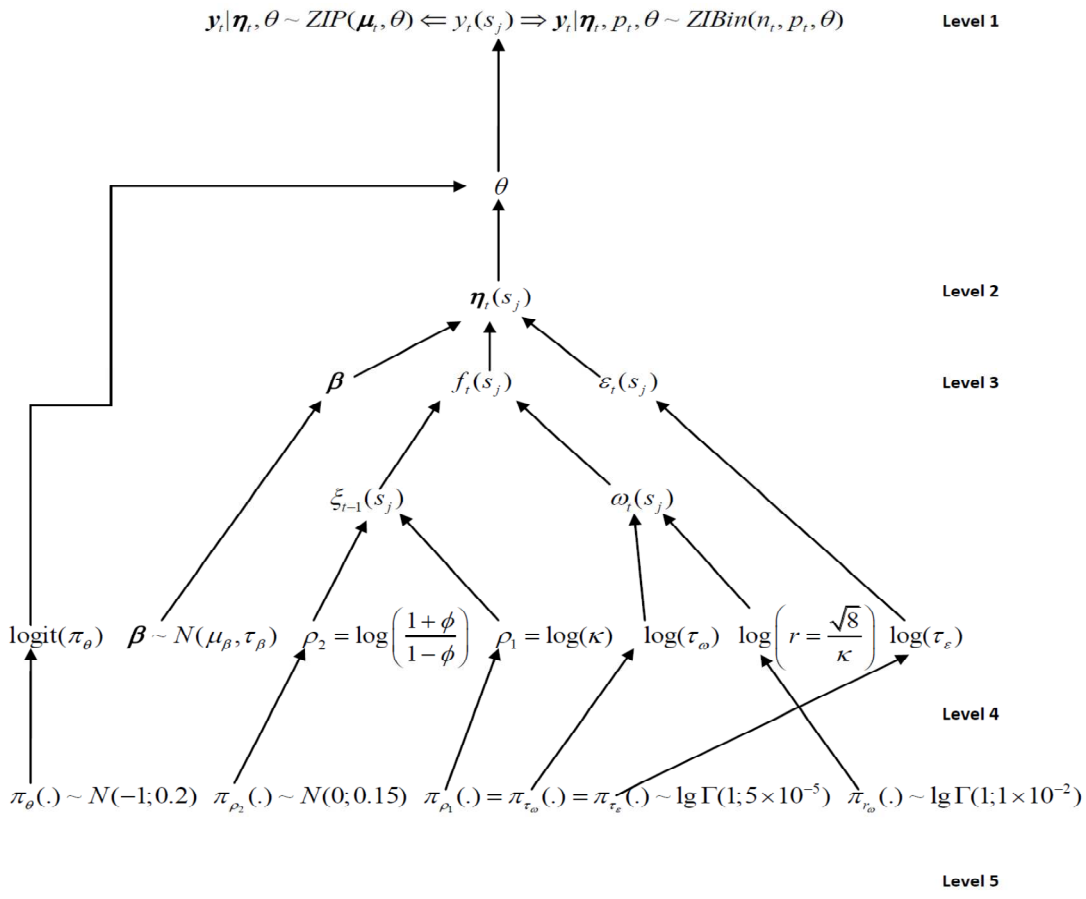


Figure 6.2: Hierarchical structure of a zero inflated spatiotemporal model fit using INLA

The posterior marginals are required, standard (i), nested approximation (ii) and numerical integrations (iii) for latent fields $\boldsymbol{x} = \{\beta, f, \theta\}$ and hyper-parameters $\varphi = \left\{\phi, \sigma_\omega^2, \kappa, \sigma_\varepsilon^2\right\}$ respectively (96, 261):

$$
\begin{aligned}
&i)\ \pi(\boldsymbol{x},\boldsymbol{\varphi}\mid\boldsymbol{y})=\int\pi(x_i\mid\boldsymbol{\varphi},\boldsymbol{y})\pi(\boldsymbol{\varphi}\mid\boldsymbol{y})\,d\boldsymbol{\varphi} && \pi(\varphi_j\mid\boldsymbol{y})=\int\pi(\boldsymbol{\varphi}\mid\boldsymbol{y})\,d\boldsymbol{\varphi}_{-j}\\
&ii)\ \tilde{\pi}(\boldsymbol{x},\boldsymbol{\varphi}\mid\boldsymbol{y})=\int\tilde{\pi}(x_i\mid\boldsymbol{\varphi},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\varphi}\mid\boldsymbol{y})\,d\boldsymbol{\varphi} && \tilde{\pi}(\varphi_j\mid\boldsymbol{y})=\int\tilde{\pi}(\boldsymbol{\varphi}\mid\boldsymbol{y})\,d\boldsymbol{\varphi}_{-j} \qquad (6.7)\\
&iii)\ \tilde{\pi}(\boldsymbol{x},\boldsymbol{\varphi}\mid\boldsymbol{y})=\sum_k\tilde{\pi}(x_i\mid\boldsymbol{\varphi}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\varphi}_k\mid\boldsymbol{y})\Delta_k && \tilde{\pi}(\varphi_j\mid\boldsymbol{y})\propto\int\left.\frac{\pi(\boldsymbol{x},\boldsymbol{\varphi}\mid\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}\mid\boldsymbol{\varphi},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\varphi})}d\boldsymbol{\varphi}_{-j}
\end{aligned}
$$

Using this technique we aim to initially get the terms “nested” inside the integrand in equation 6.7 *ii)* left hand side without integration (264). To do so we firstly estimate the marginal $\tilde{\pi}_G(\boldsymbol{x}\mid\boldsymbol{\varphi},\boldsymbol{y})$ which is a Gaussian approximation of $\boldsymbol{x}$ with mode $\boldsymbol{x}^*(\boldsymbol{\varphi})$, for a given $\boldsymbol{\varphi}$. Secondly we estimate $\tilde{\pi}(x_i\mid\boldsymbol{\varphi},\boldsymbol{y})=N\left(x_i;\mu_i(\boldsymbol{\varphi}),\sigma_i^2(\boldsymbol{\varphi})\right)$ using either Gaussian or Laplace or a simplified Laplace approximations (261). We computed these marginals using the R package Integrated Nested Laplace Approximation (INLA) which uses the simplified Laplace approximations (96, 163, 261) The INLA procedure also enables easier spatial prediction since it computes posterior conditionals for the spatial random effects on all triangulation vertices including the extensions as shown in figure 6.1.

6.2.5 Model goodness of fit and convergence diagnostics

We assessed the accuracy of $\tilde{\pi}(\boldsymbol{\varphi}\mid\boldsymbol{y})$ using the effective number of parameters, which can be approximated as the difference between the dimension of the normalised integral $\left\{\tilde{\pi}(\boldsymbol{\varphi}\mid\boldsymbol{y})\right\}\quad n$ and the trace of the product of the prior precision matrix and the posterior covariance (238):

$$p_D(\boldsymbol{\varphi})\approx n-tr\left\{Q(\boldsymbol{\varphi})Q^*(\boldsymbol{\varphi})^{-1}\right\} \qquad (6.8)$$

The deviance information criteria (DIC) was also used which is defined as the difference between twice the mean of the deviance and the deviance of the mean according to Spiegelhalter et al. (2002) (238) and expressed as:

$$D(x,\boldsymbol{\varphi}) = -2\sum_{i}\log\left\{\tilde{\pi}\left(y_i \mid x_i,\boldsymbol{\varphi}\right)\right\} + constant \tag{6.9}$$

An interpretation of this is quite straight forward: the smaller the effective number of parameters the more parsimonious the model and the smaller the DIC the better the model fit, more-so most parsimonious is not always the best model.

6.3 Application

6.3.1 Agincourt health and socio-demographic surveillance data (HDSS), rural South Africa, study design and ethics

The Agincourt health and socio-demographic surveillance system (HDSS) site was set up in the Agincourt sub-district in 1992 due to its remote location, availability of several clinics and presence of Mozambican in-migrants who arrived as refugees during the early to mid-1980s during the civil war (265). By 2010, the Agincourt HDSS had a population of over 84,000 persons living in approximately 15,000 households scattered throughout 27 neighbouring villages. Cause-of-death data were obtained through verbal autopsies conducted on every recorded death (197). Interviews were conducted by trained field workers within one year after a death, with the closest caregiver of the deceased in their mother tongue. Cause-of-death was independently determined by two medical practitioners with the third serving as arbitrator when consensus could not be reached. The consensus cause-of-death was classified according to the World Health Organization's International Classification of Diseases (ICD10) (198). HIV/TB mortality in children was ascertained by reported signs and symptoms, and in some instances supported by the mother's cause of death (198). Over 90%

of the HDSS households were geo-coded by 1992 and by 2010 all the households were geo-coded, thus enabling spatial analyses at household as well as village level. The study design was a retrospective cohort study covering households observed from the establishment of the site to December 2010. Both the Agincourt HDSS and this study were granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects, M960720 and M081145 respectively. Verbal informed consent was obtained for each census update round and for verbal autopsy data collection. .

6.3.2 Dependent and independent variables

The persons included in the study were all children aged between 0 and 10 years who lived or had lived in the Agincourt HDSS between 1992 and December 2010. The independent variables used were: child's gender, birth-weight category, age category, socio-economic status (availability of water, electricity and toilet) and number of mother's antenatal clinic visits during pregnancy and year of observation. The household's latitude and longitude were used to construct the latent variables for the spatial random effects and year the AR(1) temporal effects. The dependent variable was death due to HIV and/or tuberculosis (TB) determined by the WHOs ICD10 codes A16-A19 and B20-B24. These data were extracted using Structured Query Language (SQL). Preliminary data analyses and data management were done using STATA 10.1 (203). The Bayesian analysis was done using an R software package called INLA (163, 261) .

6.3.3 Descriptive statistics

There were similar proportions of boys and girls (50.3%) with an average age of 6.14 years (standard deviation of 3.38). Almost 9% (8.89%) of the births were of low birth weight, the majority (60.3%) of the children were only observed after their 5th birthday. Only 8.09% had access to all of clean tap water, flush toilet inside house and electricity, which is expected in

this rural area. Health check-ups during pregnancy averaged 6 antenatal care visits, with the majority (73%) not attending at all. A total of 456/71,057 (0.64%) deaths were due to AIDS (including HIV/Tuberculosis) therefore showing presence of zero inflation. Mortality was moderate in the early to mid 1990s before gradually increasing from 1999 and reaching a peak in 2001, remaining high for several years before a gradually declining from 2007 as shown in figure 6.3.

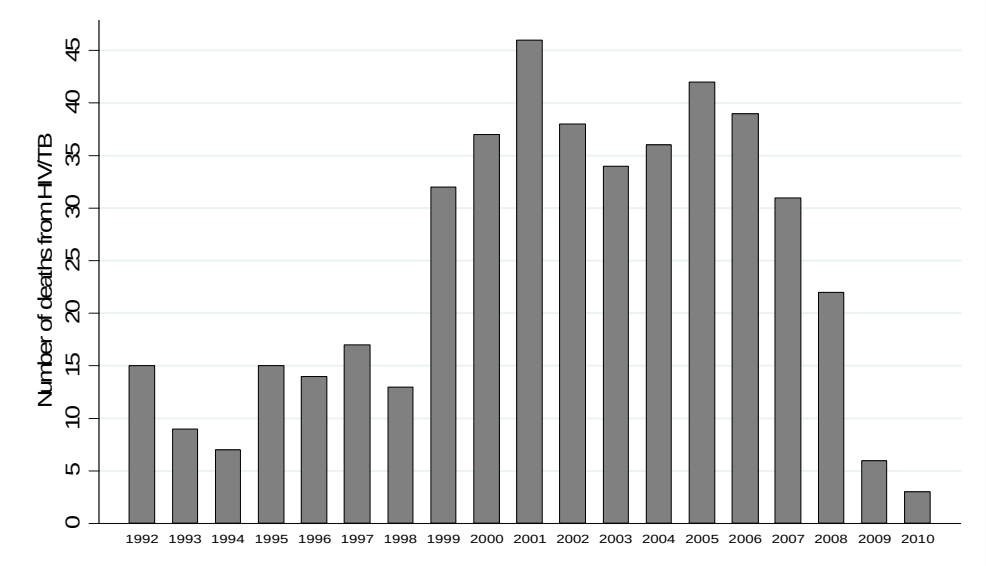


Figure 6.3: Year specific child deaths due to HIV/TB from 1992 to 2010

6.3.4 Zero Inflated Poisson and Binomial spatiotemporal modelling results

We performed univariate and multiple variable analyses using both the ZIP and ZIB models. Multiple variable models were fit systematically starting with one without random effects, then spatial only, temporal only and finally spatiotemporal random effects. The R-codes using INLA are given in appendix 6.8. From our results shown in tables 6.1 and 6.2, our discussion will be centred on the ZIP spatiotemporal model which was the best fitting model since this had the lowest DIC (4532.31) and catered for spatiotemporal random effects, as later indicated in section 6.3.5.

Table 6.1: Univariate and multiple regression results models using zero inflated Poisson adjusting for spatiotemporal random effects.

Variable	Summary, n(%) or mean±sd	Univariate results, RR(95% credible interval)	Non-spatial multiple variable model, adjusted RR(95% credible interval)	Temporal multiple variable model, adjusted RR(95% credible interval)	Spatial multiple variable model, adjusted RR(95% credible interval)	Spatiotemporal multiple variable model, adjusted RR(95% credible interval)
Sex						
Male	35,317(49.70)	1.00				
Female	35,740(50.30)	0.89(0.74;1.07)				
Birth weight category						
Low weight	6,320(8.89)	1.00	1.00	1.00	1.00	1.00
Moderate weight	15,241(21.45)	0.81(0.61;1.07)	0.70*(0.53;0.93)	0.85(0.63;1.16)	0.71*(0.53;0.93)	0.85(0.63;1.16)
High weight	49,496(69.66)	0.39*(0.30;0.51)	0.57*(0.42;0.76)	0.73*(0.53;0.99)	0.58*(0.43;0.77)	0.73*(0.53;0.99)
Age category						
0 – 1 years	8,580(12.07)	1.00	1.00	1.00	1.00	1.00
1 - 5 years	19,619(27.61)	0.21*(0.17;0.26)	0.20*(0.16;0.25)	0.18*(0.23;0.64)	0.20*(0.16;0.25)	0.18*(0.14;0.22)
5 – 9 years	42,858(60.31)	0.05*(0.04;0.07)	0.04*(0.03;0.05)	0.04*(0.05;0.54)	0.04*(0.03;0.05)	0.04*(0.03;0.05)
Slum (electricity, water and toilet)						
None of the three	10,051(14.14)	1.00	1.00		1.00	
At least one	27,329(38.46)	0.78(0.60;1.01)	0.79(0.61;1.02)		0.77(0.60;1.00)	
At least two	21,871(30.78)	0.74*(0.56;0.97)	0.73*(0.56;0.96)		0.71*(0.54;0.93)	
All three	5,748(8.09)	0.57*(0.37;0.86)	0.55*(0.35;0.83)		0.53*(0.34;0.81)	
Antenatal clinic visits	6.00(±1.50)	0.98*(0.97;0.99)	0.98*(0.97;0.99)	0.96*(0.94;0.97)	0.98*(0.96;0.99)	0.96*(0.94;0.97)
Zero inflation parameter			*0.06*(0.002;0.25)*	*0.05*(0.003;0.18)*	*0.07*(0.003;0.28)*	*0.04*(0.002;0.15)*
Precision for year				*0.91*(0.23;2.19)*		*0.86*(0.16;2.32)*
Rho for year				*0.88*(0.72;0.97)*		*0.89*(0.70;0.98)*
Kappa					1.98*(1.09;3.66)	*1.55*(1.41;1.63)*
Tau					1.37*(0.18;2.600	*2.31*(1.19;9.08)*
Moran's Indexes: Observed (Expected ± standard deviation)						*0.32*(2.1±3.33x10^{-3})*
Effective number of parameters			*12.05(0.05)*	*25.61(1.10)*	*23.18(5.04)*	*29.16(6.33)*

Table 6.2 : Multiple regression results of four models using zero inflated Binomial adjusting for spatiotemporal random effects

Variable	Non-spatial multiple variable model, adjusted OR (95% credible interval)	Temporal multiple variable model, adjusted OR (95% credible interval)	Spatial multiple variable mode, adjusted OR (95% credible interval)	Spatiotemporal multiple variable model, adjusted OR (95% credible interval)
Sex				
Male				
Female				
Birth weight category				
Low weight	1.00	1.00	1.00	1.00
Moderate weight	0.70*(0.52;0.92)	0.86(0.62;1.20)	0.70*(0.52;0.93)	0.005*(0.003;0.07)
High weight	0.62*(0.46;0.83)	0.67*(0.48;0.95)	0.63*(0.47;0.85)	0.05*(0.007;0.04)
Age category				
0 – 1 years	1.00	1.00	1.00	1.00
1 - 5 years	0.07*(0.05;0.09)	0.05*(0.04;90.07)	0.07*(0.06;0.09)	0.18*(0.14;0.22)
5 – 9 years	0.006*(0.004;0.008)	0.005*(0.003;0.95)	0.006*(0.004;0.008)	0.04*(0.03;0.05)
Slum (electricity, water and toilet)				
None of the three	1.00		1.00	
At least one	0.77(0.60;1.01)		0.76*(0.58;0.99)	
At least two	0.71*(0.54;0.95)		0.69*(0.52;0.93)	
All three	0.52*(0.34;0.81)		0.51*(0.32;0.78)	
Antenatal clinic visits		0.94*(0.92;0.96)	0.97*(0.95;0.98)	0.94*(0.93;0.96)
Zero inflation parameter	*0.22*(0.042;0.66)*	*0.43*(0.19;0.79)*	*0.15*(0.031;0.45)*	*0.42*(0.18;0.67)*
Precision for year		*0.78*(0.17;1.96)*		*0.80*(0.20;1.93*
Rho for year		*0.87*(0.67;0.97)*		*0.87*(0.69;0.97)*
Kappa			1.98*(1.14;3.55)	*1.54*(1.41;1.62)*
Tau			1.37*(0.27;2.54)	*2.23*(1.22;6.91)*
Moran's Indexes: Observed(Expected ± standard deviation)				*0.30*(2.1±3.33x10^{-3})*
Effective number of parameters	*11.10(0.080)*	*24.58(1.03)*	*22.97(5.09)*	*31.12(5.18)*
DIC	*5179.43*	*4754.60*	*5180.50*	*4803.93*

Two variables consistently showed significant associations with child HIV/TB mortality, firstly the age category of the child, which showed a decrease in mortality with increase in age (Chi-Square p-value<0.001). Those aged 1 to 4 (127/19,619) were 82% {0.18 (0.14; 0.22)} less likely to experience a death due to HIV/TB compared to those under 1(262/8,580), and those 5 to 9 years (67/42,858) were 96%{0.04 (0.03;0.05)} less likely to die from HIV/TB relative to the under 1s keeping all other variables constant. Secondly the category of birth-weight was significantly associated with deaths due to HIV/TB, showing a decreasing trend with greater birth weight (Chi-Square p-value<0.001). Those who had a birth weight of 3.5 kilograms and above (232/49,496) were 27% (0.73 (0.53;0.99)} less likely to die due to HIV/TB relative to those who were born with a birth weight of less than 2.5 kilograms (76/6,329) independent of other risk factors. The only other significant protective factor was the number of antenatal clinic visits made by the mother; for every additional visit the mother attended the risk of losing the child from HIV/TB decreased by 4% (0.96(0.94;0.97)} keeping other variables constant as shown by the spatial only and no-random effects models. Those who had access to all three resources – electricity, water and flush toilet (the more affluent in this area) experienced fewer deaths due to HIV/TB compared to those who had none of these resources. We show the posterior marginal point estimates from the ZIP and ZIB models in figures 6.4 top and bottom respectively.

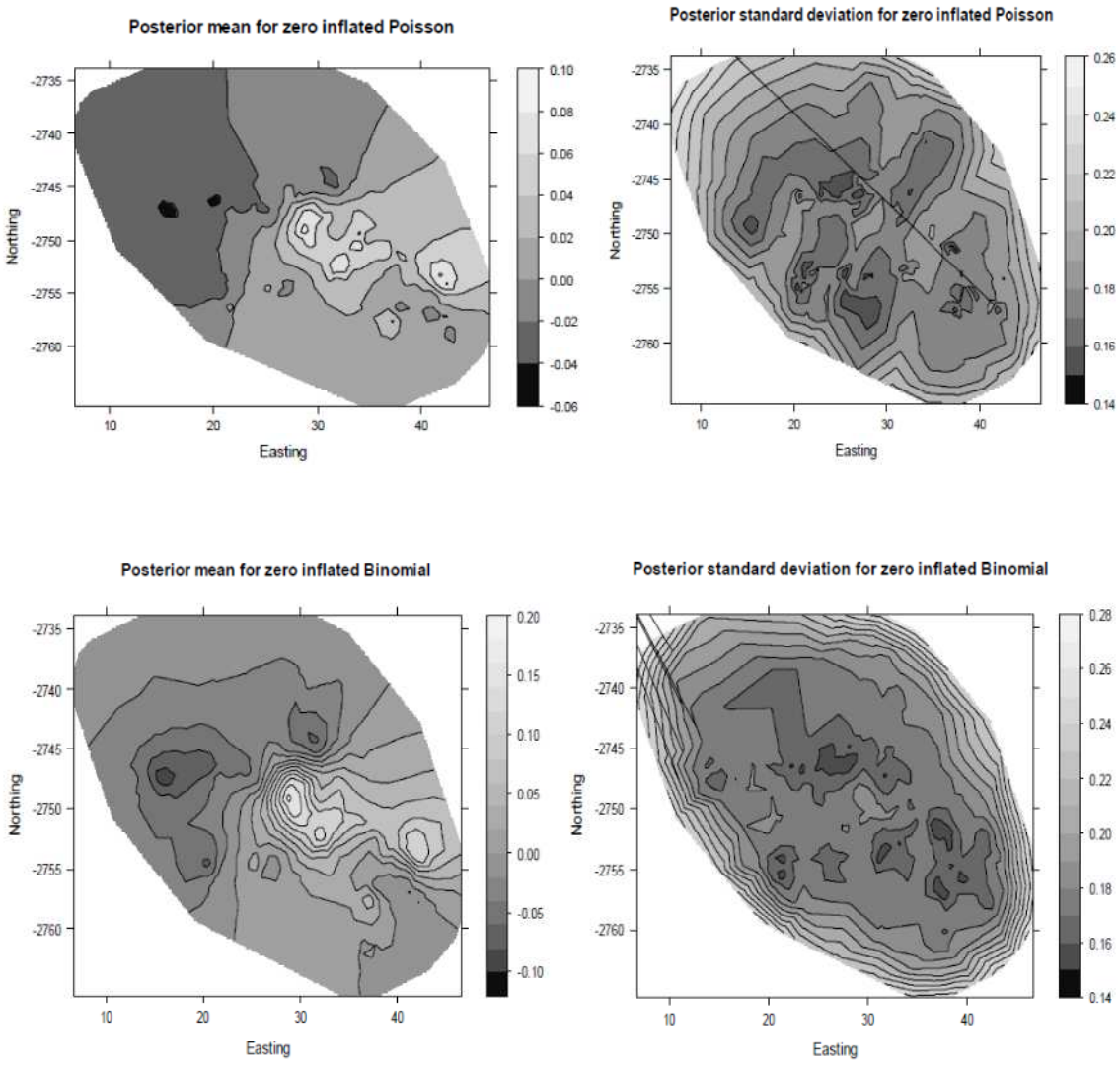


Figure 6.4: Posterior point estimates for ZIP (top) and ZIB (bottom) models

As shown in figure 6.4, the area with the greatest risk of child HIV/TB mortality is the central region with a trend towards the south-west. This region is where households had lower birth weights, greater infant mortality and fewer visits to the health facility compared to households in other regions. The lowest risk region is the top north-east with the ZIP model showing fewer clusters than the ZIB model. The standard error maps indicate the greatest errors on the peripherals which are also the regions where the extensions of the triangulation shown in figure 6.1. Also consistent with our results in tables 6.1 and 6.2, there is a greater margin of error with the ZIB model as compared with the ZIP model.

6.3.5 Model assumptions, goodness of fit and convergence diagnostics

Using the final spatiotemporal models in both ZIB and ZIP we investigated the presence of zero inflation, temporal and spatial correlation. The null hypothesis for zero inflation was that the inflation parameter θ_t is zero (no zero inflation) our results in tables 6.1 and 6.2 indicate significant presence of zero inflation of 0.04 and 0.42 for ZIP and ZIB respectively. Investigation of temporality was done by testing the ϕ coefficient of the first order autocorrelation model with the null hypothesis that the process was not stationary $|\phi| > 1$. Both our models indicated significant presence of stationarity that is 0.86 and 0.87 for ZIP and ZIB respectively. Figure 6.5 below shows the posterior means plus 95% credible bands, time series plots from the modelling which resembles the observed data shown in figure 6.3. Lastly on checking model assumptions we investigated the presence of spatial effects by inspecting the significance of the components (κ and τ) of the Matérn covariance (see tables 6.1 and 6.2). Also this was investigated with the classical Moran's indices on the spatial residuals testing the null hypothesis that there is zero spatial autocorrelation which from both tests yielded significant results, hence rejecting the null.

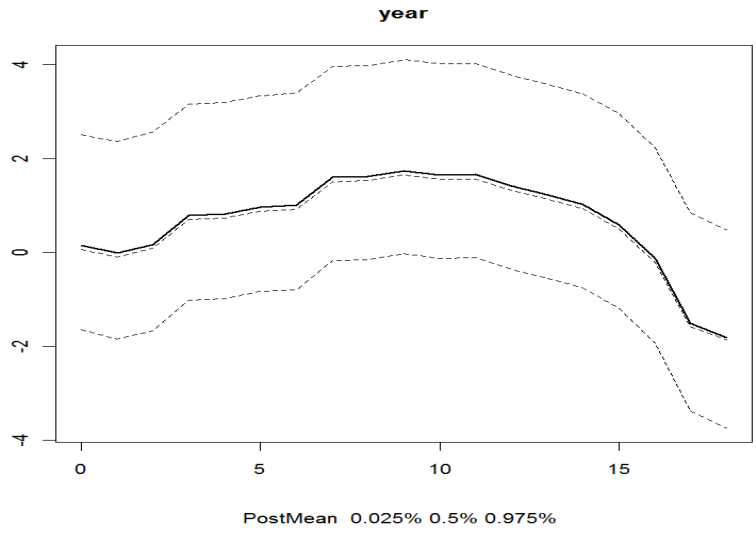


Figure 6.5: Posterior mean and 50% (middle), 0.025% (lower) and 0.975% (upper)

We compared the two models, the ZIB and ZIP, and found that the latter fitted better in all the variants from the no-random effects model to the spatiotemporal model as this yielded smaller DICs and model parsimony. Some of the parameter estimates were similar, except for the zero inflated parameter, age categories and birth weights. The ZIP model also showed more stable estimates and consistent results in the adjusted relative risks (adjRR) compared to the ZIB and narrower credible intervals. As a way of showing the convergence pictographically for some parameter estimates, we show the fixed effects logarithmic form posterior means density plots in figures 6.6 and 6.7.

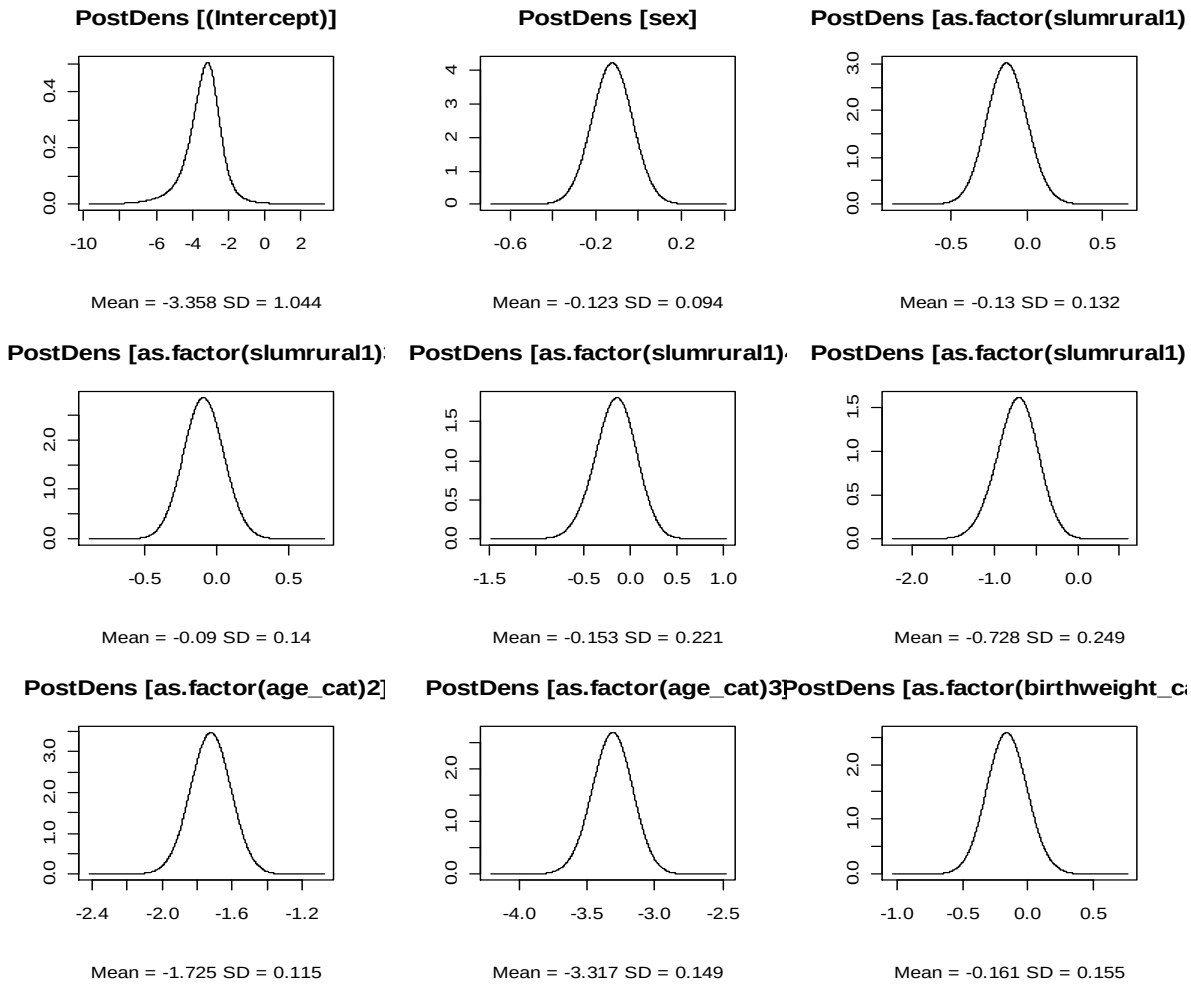


Figure 6.6: Posterior density plots of log-means for fixed effects of the ZIP spatiotemporal model.

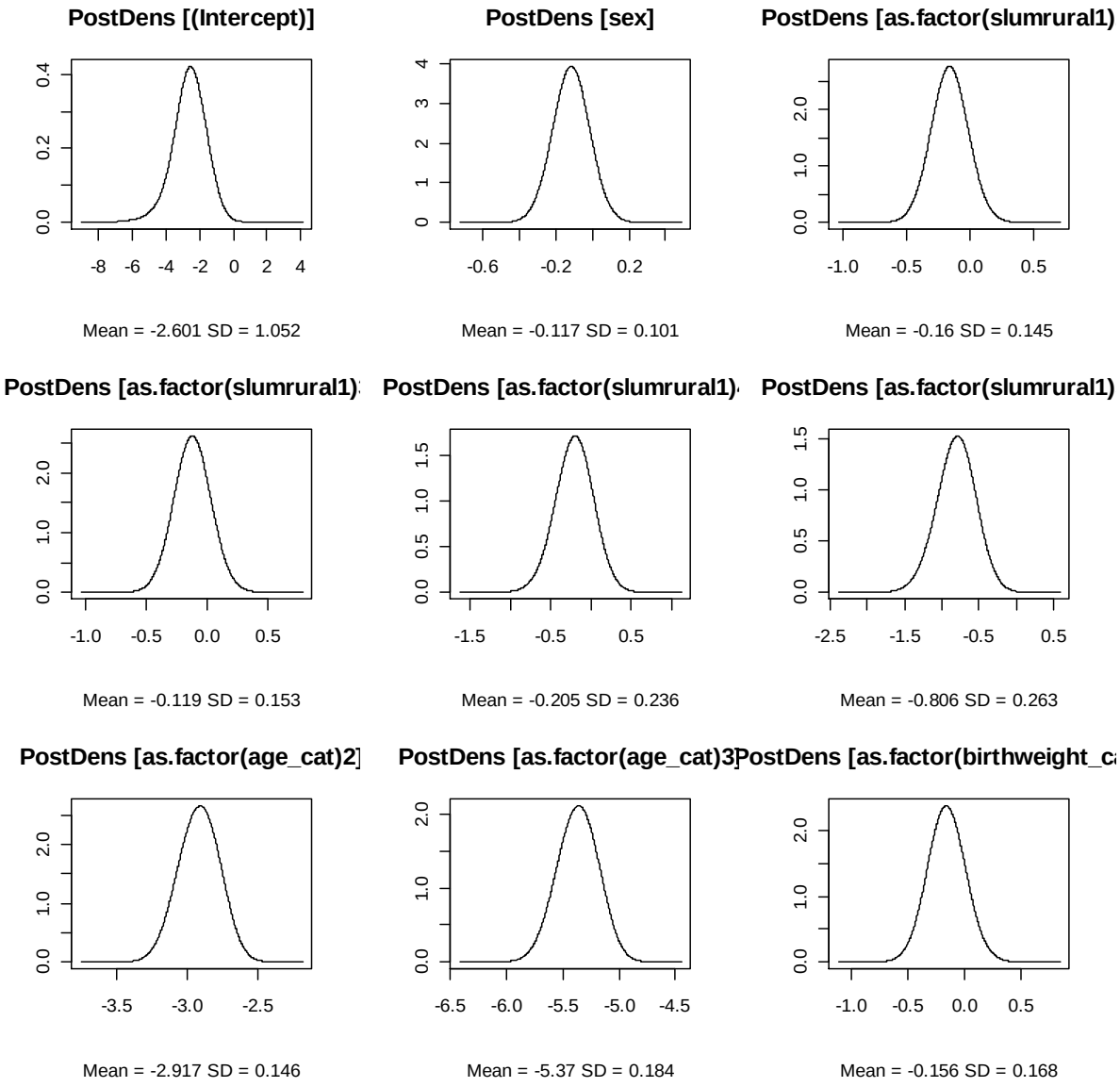


Figure 6.7: Posterior density plots of log-means for fixed effects of the ZIB spatiotemporal model.

The plots in figures 6.6 and 6.7 show that the parameters fitted well as GMRF since they are symmetrical about thus the mean equals the mean and mode.

6.4 Conclusions

The public health implications of our results centre on inappropriate health seeking behaviour in rural areas. Interventions could target two groups of mothers (or other caregivers) – those not yet infected and those already infected. For both groups, antenatal clinic attendance could possibly improve birth weight and reduce the risk of infant mortality. For those mothers' already HIV-infected, there need to be widely accessible and effective prevention of mother-to-child programmes, with provision of anti-retro therapy for HIV-infected mothers and infants.. A significant decline in mortality can already be observed in our cohort post 2007, the year when ARVs were first offered within the rural sub-district. Poverty, however, still remains the greatest challenge in low- and middle-income countries and care must be taken to ensure interventions reach those most in need. Our results from spatiotemporal Bayesian modelling coupled with maps can be very useful in allocating scarce resources.

Modelling that controls for potential confounding is an add-on to epidemiological studies and gives strength to the estimates derived from such modelling procedures. Our modelling approach caters for potential spatial and temporal confounders. This approach was also able to cater for "large" zero inflated spatiotemporal data.

A major limitation in low- and middle-income nations is lack of reliable and user friendly analytic software, which was addressed by using INLA a package available on the public domain. Zero inflated Poisson (ZIP) and zero inflated Negative Binomial (ZINB) models were used for log link function linear predictors and also catered for spatial random effects (266, 267). It has also been noted from other studies that the log link function linear predictor models were more stable than logit link function models when they contained a spatial component (266). This finding was also confirmed in our study as we observed that models

from the exponential family with a log-linear predictor were more stable and able to converge better than the logit-link function model. This motivated us to treat the outcome as a discrete measure as opposed to binary mostly used for logistic regression(23). We demonstrated that the “big m” can be resolved with great ease for public health mortality hierarchical structures with the aid of SPDEs and INLA.

6.5 Acknowledgements

This work was supported by a doctoral fellowship from the World Health Organization’s, Tropical Disease Research (WHO/TDR), 2008 to 2010, and the Swiss-South African Joint Research Program (SSAJP) project number JRP IZLSZ3122926, 2009 to 2010. The Agincourt HDSS is funded by the Wellcome Trust UK (Grant number 058893/Z/99/A, 069683/Z/02/Z, and 085477/Z/08/Z), the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew Mellon and Hewlett Foundations, U.S.A. Special thanks goes to Benn Sartorius for data extraction, community leaders, field workers, team supervisors and the Agincourt communities. Avery special thanks to Finn Lindgren for answering all my questions promptly whilst I developed models using INLA.

6.6 Authors' contributions

EM drafted the manuscript and did the statistical analysis. KK provided critical intellectual content, and TC and PV the statistical content. KK was also responsible for the design and quality of the Agincourt verbal autopsy data. All authors read and approved the final manuscript.

6.7 Competing interests

The authors declare that they have no competing interests.

6.8 Appendix: Zero inflated Poisson R-code using INLA

adapted from http://www.r-inla.org/ (259).

```
#1.#IMPORTING LIBRARIES
library(foreign)
library(INLA)
library(coda)
library(fields)

#2.# Read data:
decdata = read.dta("c:/jagnovredo/childleveldec.dta")
attach(decdata)
names(decdata)

#3.# Build triangular mesh of locations within 0.5 km see figure 6.1
mesh123 = (inla.mesh.create(cbind(decdata$coordlongx/1000,
decdata$coordlaty/1000),refine=TRUE,cutoff=0.5))
plot(mesh123)


#4.# Store the data-->vertex mapping:
## ( mesh$idx$loc[m] is the mesh vertex for data location number m )
decdata$hhid = mesh123$idx$loc

#5.# Create the SPDE/GMRF model, (kappa^2-Delta)(tau x) = W:
spde123 = inla.spde.create(mesh123, model="matern")

#6.# Functions of four models multivariate models fit
#6a.No random effects model with household and individual random effects
norand=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(birth
weight_cat)+antenatalvisits+f(hhid, model="iid")+f(idno, model="iid")

#6b.Temporal random effects model with household and individual random
effects
temponly=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(bir
thweight_cat)+antenatalvisits+f(year, model="ar1")+f(hhid,
model="iid")+f(idno, model="iid")

#6c.Spatial random effects model with individual random effects
spatonly=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(bir
thweight_cat)+antenatalvisits+f(idno,model="iid")+f(hhid,model=spde123)

#6d.Spatiotemporal random effects model with individual random effects
spatiotemp=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(b
irthweight_cat)+antenatalvisits+f(year, model="ar1")+f(idno,
model="iid")+f(hhid,model=spde123)


#7.# Running INLA models on the four stated in step 6
norand = inla(norand, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals
.predictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(norand)

spatonly = inla(spatonly, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals
.predictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(spatonly)

temponly = inla(temponly, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals
.predictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(temponly)
```

```
spatiotemp = inla(spatiotemp, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals
.predictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(spatiotemp)

#Saving output for model to store posterior estimates and other
save(spatiotemp, file="c:/resultsdec/spatiotemp.Rdata")

#-----------------------------------------------------------------#

#8.#Loading data from previous analyses in new session
spatiotemp<-load(file="c:/jagnovredo/spatiotemp.Rdata")
names(spatiotemp)

#9.#Assessing for posteriors Spatial auto-correlation (Moran's I)
library(ape)
D.dists <- as.matrix(dist(cbind(mesh123$loc[,1], mesh123$loc[,2])))
D.dists.inv <- 1/D.dists
diag(D.dists.inv) <- 0
Moran.I(spatiotemp$summary.random$hhid[,"mean"], D.dists.inv)


#10.# Diagnostic plots see figures 6.6 or 6.7
plot.inla(spatiotemp,
 plot.fixed.effects = TRUE,constant=FALSE,
 plot.lincomb = TRUE,
 plot.random.effects = TRUE,
 plot.hyperparameters = TRUE,
 plot.predictor = TRUE,
 plot.q = TRUE,
 plot.cpo = TRUE)


#11.# Extract the SPDE parameters, point and interval estimates
tauconf = exp(spatiotemp$summary.hyperpar[2,])
kappaconf = exp(spatiotemp$summary.hyperpar[3,]/2)

tau = exp(spatiotemp$summary.hyperpar[2,"mean"])
kappa = exp(spatiotemp$summary.hyperpar[3,"mean"]/2)

#12a.# Get the spatial precision matrix (Qs) see equation section 6.2.1 and
2.3
Q = inla.spde.query(spde123, precision=list(tau=tau,
kappa2=kappa^2))$precision


#12b.# Reference point for covariance/correlation comparisons:
ref.s = (which.min((mesh123$loc[,1]-mean(range(mesh123$loc[,1])))^2 +
 (mesh123$loc[,2]-mean(range(mesh123$loc[,2])))^2))

#12c.# Calculate covariances (S) and correlations (SS):
S = solve(Q)
SS = diag(1/sqrt(diag(S))) %*% S %*% diag(1/sqrt(diag(S)))
D = as.matrix(dist(mesh123$loc))

#12d.# Theoretical Matérn correlations and covariances:
dd = (0:1000)/1000
SS.theory = (dd*kappa)*besselK(dd*kappa,1)
SS.theory[1] = 1
S.theory = SS.theory/(4*pi*kappa^2)/tau^2

#-----------------------------------------------------------------------#


#13.# Prediction, plotting posterior prediction maps figure 6.4

## Calculate mapping between triangulation vertices and grid points:
## Resolution for gridded output was dims=c(300,300)
proj = inla.mesh.projector(mesh123, dims=c(300,300))
```

```
## Construct greyscale palette function:
my.grey.palette = function (n,...) { return
(grey.colors(n,0.05,0.95,...))}
## Use it:
my.palette = my.grey.palette

## Construct map data appropriate for easy plotting:
mm = calc.map(map)

## Plot results:

par(mfrow=c(1,2))
## Map resulting posterior mean field to a grid:
plotdata = inla.mesh.project(proj,
spatiotemp$summary.random$hhid[,"mean"])
## Plot PM contours:

postmap = (levelplot(row.values=proj$x, column.values=proj$y, x=plotdata,
 mm=mm, panel=levelplotmap,
 col.regions=my.palette,
 xlim=range(proj$x), ylim=range(proj$y), aspect="iso",
 contour=TRUE, cuts=11, labels=FALSE, pretty=TRUE,
 xlab="Easting",ylab="Northing", main="Posterior mean for zero inflated
Poisson"))
plot(postmap)


## Map resulting posterior standard deviation field to a grid:
plotdata = inla.mesh.project(proj, spatiotemp$summary.random$hhid[,"sd"])
## Plot std.dev. contours:
#dev.new()
postmap2 = (levelplot(row.values=proj$x, column.values=proj$y, x=plotdata,
 mm=mm, panel=levelplotmap,
 col.regions=my.palette,
 xlim=range(proj$x), ylim=range(proj$y), aspect="iso",
 contour=TRUE, cuts=11, labels=FALSE, pretty=TRUE,
 xlab="Easting",ylab="Northing",main="Posterior standard deviation for
zero inflated Poisson"))

plot(postmap2)
```

Chapter 7

7.0 Discussion and conclusions

In this thesis I have categorised our findings into indirect and direct determinants or risk factors of morbidity and mortality adjusting for geographical locations as a proxy for socio-economic and environmental factors. A consistent pattern of results was observed when using different analysis techniques coherent with the different study designs, multi-level structures, amount of data and spatiotemporal correlations. As I conclude this thesis I provide a discussion of the results in relation to the global scientific body of knowledge. These can be grouped according to three thematic areas: determinants of childhood morbidity: adiposity and malnutrition; direct risk factors of child HIV/TB mortality in observational cohorts; and spatiotemporal modelling of mortality and morbidity in resource scare settings (shown in table 1.1).

This thesis uses data over the period 1992 to 2010, a span of 19 years that makes up almost three quarters of the United Nation's Millennium Development Goals (MDG) period of 25 years. Goal 4.1 aims to reduce under five child mortality by two thirds over that period (3). Our results overall demonstrate a significant decline in mortality due to HIV/TB after the introduction of antiretroviral treatment in the rural Agincourt sub-district in 2007.

7.1 Determinants of childhood morbidity: adiposity and malnutrition

Theme 1: Key summary points

1. Spatial differentials exist in childhood adiposity and malnutrition measures.

2. HIV morbidity retards child growth and plays a mediatory role.

3. Child morbidity due to either malnutrition or HIV and vice-versa results in a vicious cycle.

In this thesis, through the manuscripts in Chapters 2 and 3, we identified determinants of childhood adiposity: malnutrition and mortality. Similar determinants have been found in other studies as well (167-169, 172, 173, 268-270), however none of these studies have simultaneously examined how the determinants interlink (directly or indirectly) nor have they adjusted for household spatial random effects. The chapters in this thesis highlight spatial differences within the Agincourt sub-district a relatively small geographical area. Firstly, childhood adiposity: there is evidence of spatial differences in those who are either overweight or obese (Chapter 2). Secondly, child malnutrition measures: geographical variations in the study area are shown in Chapter 3. Finally, child HIV/TB mortality: we demonstrated clear spatial patterns for either a single year (Chapter 4) or over several years (Chapter 5) and also including infants and those above five years of age (Chapter 6). The findings indicate similar geographical patterns in Agincourt; areas with no health facility had significantly greater malnutrition and adiposity, and areas with a higher concentration of former Mozambican refugees and households without biological parents were at greater risk

of mortality. Access to affordable health care is currently a major global debate; yet the issue for rural dwellers in South Africa is not only that of consultation costs (primary health care is free) but includes transport costs as a major barrier, health awareness and a health seeking behaviour and the dual use of traditional and western medicine in treating and preventing childhood diseases. The free primary health care is often overburdened due to political and economic refugees from neighbouring countries.

Other key findings were that those children infected with HIV were more malnourished and were likely more vulnerable to acute illnesses. Figure 7.1 graphically presents what is currently known of the effects of HIV on malnutrition and how our study results contribute to this body of knowledge. The association between HIV and malnutrition is usually modelled as a vicious cycle as supported by several studies (10-19).

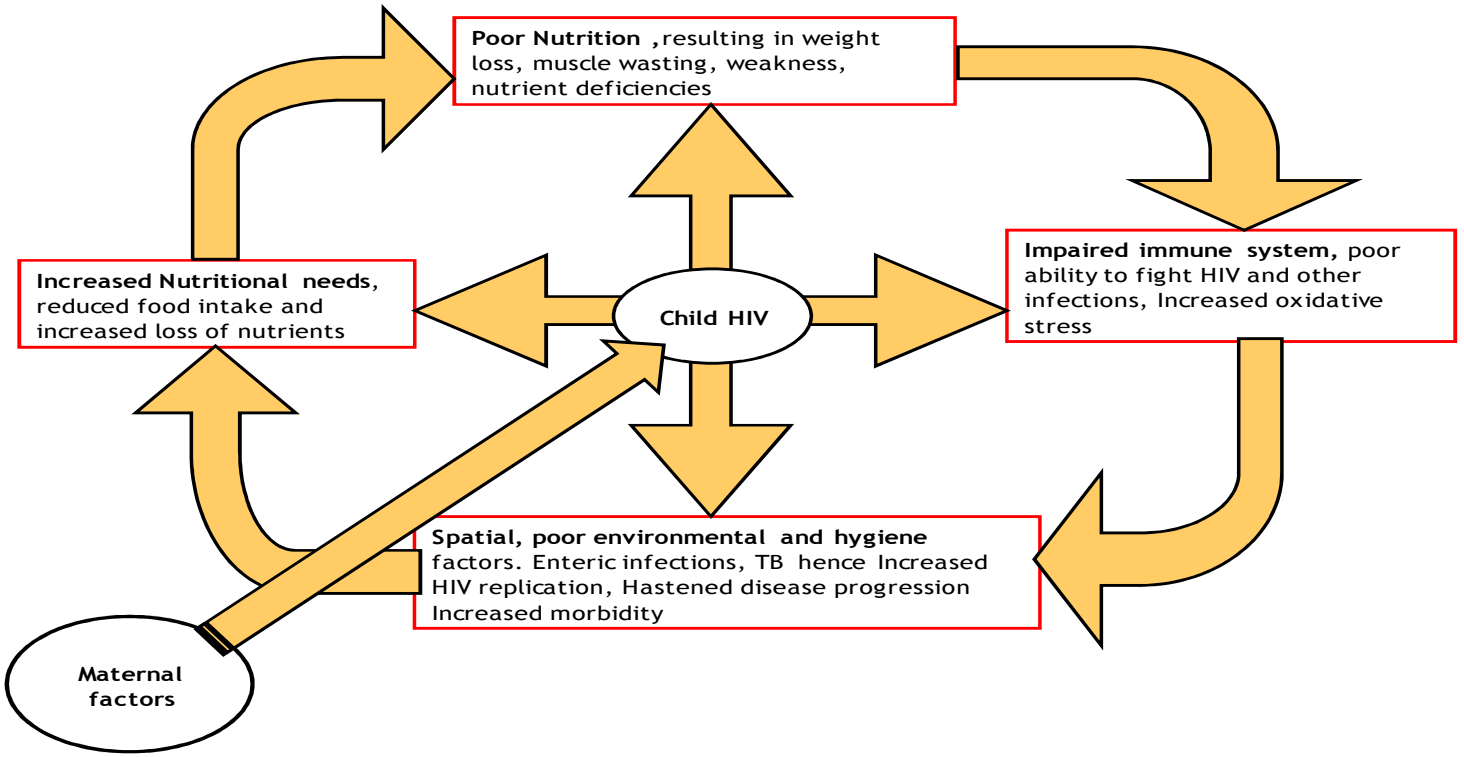


Figure 7.1: Vicious cycle depicting inter-linkage of HIV, malnutrition, maternal and geographical locations: adapted from Piwoz and Prebel (2000).

Moreover, our study provides information on the precise areas where interventions are most needed, which is a powerful tool for policy-makers aiming for effective control. Many studies have demonstrated that lower anthropometric indices are associated with an increased risk of opportunistic infections (151-156, 271). The impact on children starts in utero: among pregnant women that are HIV infected there are higher rates of preterm births, lower-birth weight, slower growth and greater risk of severe malnutrition (272-275). As such, interventions have centred on prevention of mother-to-child transmission, maternal uptake of ART, supplementation of macro and micronutrients, stimulation of breastfeeding, and home-based post-testing counselling and support (276-282). The efficacy of counselling was demonstrated in a follow up qualitative study of those who were infected; this revealed that when women knew the status of their children there was a change to improved health seeking behaviour (205).

7.2 Direct risk factors of child HIV/TB mortality in observational cohorts

Theme 2: Key summary points

1. Childhood mortality due to HIV/TB is exacerbated by poverty.

2. Maternal deaths increase the risk of child death threefold.

3. Spatial differentials in mortality are found between rural households and over time.

In Chapters 4, 5 and 6 we showed that HIV/TB child mortality was influenced by household factors, multiple deaths, male household headship and, most importantly, mother's survival

and presence. Common patterns found in our analyses can be grouped into three main categories of child HIV/TB mortality determinants in the Agincourt subdistrict: gender, maternal and household.

Gender: boys were at least twice as likely to die from HIV/TB compared to girls, and households headed by females had a higher child mortality compared to those headed by males. Some government policies have female affirmative policies for university entry and employment to address gender differentials.

Maternal death had the greatest impact on HIV/TB child deaths. The most likely cause of HIV infection was mother-to-child transmission. Prevention of mother-to-child transmission (PMTCT), using a single-dose of Nevirapine (sdNVP) administered to the child and mother, was found to reduce child mortality by about 9% in South Africa (204). Unborn and newborn babies acquire the virus from their mothers either in utero, at birth (intra-partum) or after birth (postnatal) (98, 283). Many studies report pooled results from these clearly distinguished stages of child HIV infection and hence provide inaccurate estimates of the risk factors (284). The mother's survival significantly increases the likelihood of the child's survival, hence antiretroviral therapy (ART) for the mother is vital for the health of her child(ren) (284). The impact of ART has been demonstrated in a study done in rural South Africa (206). The main source of transmission of TB among children are adults, as shown by a study on a high risk spatial cluster in South Africa with an incidence three times higher than WHO levels (50). Kimani-Murage et al. (2009) found that knowledge of child HIV status does influence the manner of care-giving and acceptance of community-wide interventions (205).

In Chapters 5 and 6 we show that the larger the number of household members the lower the risk of child death. However, as the number of household deaths increased, the risk of child deaths was also higher. These findings imply that households with healthy adults may have a higher socio-economic status, effective childcare, and more and diverse livelihood strategies which are protective of HIV/TB child mortality. On the other hand, households with more deaths are more likely to degenerate socio-economically through loss of a bread winner, high health care and funeral costs. This then compromises their ability to cater for HIV positive children. Former Mozambican refugees remain at greater risk of child HIV/TB deaths, indicating the need for specific health care interventions to marginalised and vulnerable households.

The Agincourt sub-district, covering over 400 square kilometres is a tiny portion of the vast 1.2 million square kilometres of South Africa. Yet there is clear evidence of spatial variability and differentials in child HIV/TB mortality and malnutrition dependent on household location. This approach to analysis can provide policy makers with visual maps relecting burden of disease its determinants that are useful for developing and targeting interventions.

7.3 Spatiotemporal modelling of mortality and morbidity in resource scare settings

Theme 3: Key summary points

1. Modelling of large spatiotemporal sparse data and multi-level (hierarchical) data using zero inflated distributions

2.Use of system science approach in modelling a conceptual framework using structural equations.

3. Adjusting for confounding and biases using multiple variable Bayesian random effects models and multivariate spatial regression.

4. Analysis of complex large spatiotemporal data using freely available and user friendly statistical software.

This thesis has used some novel applications of well established and emerging statistical methodologies. In Chapters 2, 3, 4, 5 and 6 I provided detail (often in the appendices) on: multiple and multivariate spatial regression, structural equation modelling, zero inflated models and faster estimating procedures respectively. The growing field of spatiotemporal epidemiology brings in the person, place and time aspects in the risk factor analysis. I demonstrated that spatial Bayesian models provide better fitting models compared to non-spatial models. Therefore catering for the spatial random effects further improves the model and parameter estimates obtained compared to classical statistical models. Bayesian methods were also capable of handling the complex structure of hierarchical data from the Agincourt sub-district. Prior known information on the distribution of data can also be utilised in Bayesian modelling. Another advantage of such modelling is adjusting for space-time related

confounding in the analyses. I was also able to estimate household specific posterior hazard ratios which were adjusted for socio-economic and confounding variables in the multivariable model.

Zero inflated models such as the Poisson (ZIP) and zero inflated Negative Binomial (ZINB) models were used with a log link function of linear predictors which also catered for spatial random effects (23, 267, 285). It has also been noted from other studies that the log link function converges faster than the logit link when they contained a spatial random effect (23). This finding was also confirmed in our study as I observed that models from the exponential family with a log-linear predictor converged better and faster than the logit-link function model (Chapters 4 to 6) (23).

Full and lowrank kriging (spatial prediction) also enabled drawing of smoothed risk maps which can be used by policy makers to better target community interventions. This can be this can accelerate achieving selected MDGs such as reducing child mortality by two thirds by 2015, since rates can be mapped to specific areas that may be lagging behind public health goals and targets.

Another emerging area in epidemiological research is system science analysis of complex social, psychosocial and medical inter-relations. We used this technique in Chapters 3 and 4 when we modelled the inter-relations between HIV and malnutrition and HIV/TB mortality and it's determinants, respectively.

Several approaches were used to address confounding and biases in this thesis. In Chapter 2 we used multivariate Bayesian procedures to simultaneously model different adiposity

measures, and in Chapter 3 the same procedures were used for malnutrition growth indices adjusting for age, weight and height of the child. In Chapter 4, structural equation models were used to adjust for confounding. Chapters 5 and 6 adjusted for space-time confounding in the multiple variable models.

A major limitation in developing nations is lack of reliable and user friendly analysis software, a hurdle we overcame by using R packages and BayesX freely available in the public domain. I demonstrated that the "big n" (big m or big N) can be resolved with great ease for public health mortality hierarchical structures with the aid of SPDEs and INLA (Chapter 6) and moderately using lowrank kriging (Chapters 4 and 5).

7.4 Strengths and limitations of the study

The findings of this thesis have both strengths and limitations which I will discuss under four main categories:

1) Variable recall and selection biases
2) Comparability of findings
3) Confounding and effect modification
4) Analysis and computing.

7.4.1 Variables recall and selection biases

A major strength in our observed variables is that the data are consistently updated, corrected and checked for inconsistencies (annually for nearly 20 years at the time of undertaking this thesis). Nevertheless two potential variable-related biases remain, namely selection and recall bias. Selection bias may occur if permanent out-migrants differ from residents or temporary migrants. Measurement bias on the outcome variable mortality is unlikely as death is a major

household event, however it may affect the quality of the verbal autopsy (VA) interview, conducted at least one month after death. Furthermore, bias can be introduced through the procedure of medical experts that assign a cause of death.

7.4.2 Comparability of findings

The Agincourt sub-district patterns are comparable to similar rural settings in South Africa, for example those found in rural northern KwaZulu-Natal province (206). While many rural areas are not geo-coded to enable direct comparison of trends and patterns, our findings likely reflect similar patterns in other rural settings. However, our study may differ somewhat by the presence former Mozambican refugees, a particularly vulnerable sub-group.

Global and regional comparability is still a challenge as a typical rural area in South Africa may differ from other countries. Comparison with other HDSS sites that have geo-location data isn a potential future study, that would shed light on similarities and differences across different geographic, cultural and socio-economic settings.

7.4.3 Confounding and effect modification

Several approaches can be used to address confounding: at the design level (e.g. randomisation), during the study (e.g. through the use of subject experts and validated instruments), and during analysis (e.g.multivariable regression, standardisation and stratification). In dealing with confounding we predominately used multiple variable regression models in the all the Chapters 2 to 6. The modelling approach used caters for spatial and temporal confounders and also for large zero inflated spatiotemporal sparse data. This was done directly as covariates, as in Chapter 5, or as specified correlation structures as in Chapter 6. However, I did not model time varying covariates and spatiotemporal

interactions. Mediation on the other hand was tested for in Chapter 3 as we determined factors directly and indirectly related to HIV and malnutrition in Chapter 3.

7.4.4 Analysis and computing

The analysis techniques brought me into contact with several fields of research beyond my existing areas of expertise. Social science – conceptual framework formulation and modelling via structural equation models; computing – modelling large datasets and handling data structures in different platforms, use of clusters, elementary coding and decoding; Bayesian statistics – conditional model formulation and translating models to programs. I used programs that were freely available and got to interact with the creators whilst I did my analyses. Since mostly non-informative priors have been used, experts' opinions were sought to build and apply more informative priors in our modelling which allowed us to fully utilise the data.

A limitation of using the non-user specific algorithms (BayesX and R) was that some changes that were necessary for our work could not be incorporated without in-depth knowledge of programming in C++ and Fortran. Design of some of the studies in the thesis, in particular the cross-sectional study design (in Chapters 2, 3 and 4) limited us to discuss associations without extending these to causality. Statistical misclassification was also a possibility as mentioned, because all the variables (except for child's age in Chapter five) were treated as linear in the modelling yet they could have been non-linearly related to the outcome and modelled as splines or higher order polynomials instead.

7.5 Conclusion

7.5.1 Agincourt's spatiotemporal child malnutrition and mortality hotspots

Throughout the thesis three areas consistently emerged as having the greater burden of HIV/TB mortality and malnutrition, and in chapters 2, 3 and 4 common patterns of malnutrition and adiposity were observed. One of these areas in the north was found to have the fewest health facilities in the sub-district. In chapters 5 and 6 consistent temporal patterns were observed, with households further south many with former Mozambican refugees – having the greatest risk of HIV/TB mortality. There appear to be some very clear spatial patterns relating poverty to poor nutritional and mortality outcomes. Interestingly, similar areas were observed using maps without adjusting for covariates, but became more focussed upon adding covariates. This strengthens the need for modelling as opposed to deductions from purely observational maps, and confirms that the models are not producing a picture very different from what was actually observed. This, more detailed information on the distribution of mortality and the causes driving it should be useful to government policy makers, non-governmental organisations and donors to target short-term or a long-term interventions.

7.5.2 Implications for interventions and policy

During the past decades South Africa has experienced a shift in disease burden yet policies have not been timeously adjusted to meet new challenges (286). For resource scare areas, individual, community-level and household-based approaches should be employed to provide information that can be used to inform programme planners on strategies to reduce the level of malnutrition in children. At the individual child level government services should be made more accessible, especially in areas of high child malnutrition and mortality. Increasing access to mobile clinics, community health workers, child support grants and agricultural

support could help target childhood malnutrition and consequently reduce associated morbidity and mortality more effectively.

Maternal age and education had both a direct and indirect protective effect on malnutrition, which persisted even after adjusting for spatial random effects. Hence an indirect approach could help reduce child malnutrition, for example launching community health education campaigns targeting women and promoting knowledge-sharing between more mature mothers and younger ones. These types of interventions have proven to be effective in resource-scarce environments elsewhere (181-184). A focused approach could include health education relevant to childhood needs, growing food in areas of high malnutrition and improving access to clean water and hygiene. This coupled with promotion of HIV testing for pregnant women will help prevent HIV transmission and promote earlier uptake of treatment. Efforts should be channelled towards strengthening functional family structures and social support for HIV-affected households and to improve access to antiretroviral therapy, food supplements and psycho-social support.

This study provides evidence that supports the South African government's revised policies on prevention of mother-to-child transmission (PMTCT) effected on 1 April 2010. The policy allows HIV positive pregnant women to be on antiretroviral (ARVs) from 14 gestational weeks onwards regardless of their CD4 count. I addition, all people with a CD4 count of 350 and symptomatic HIV or TB regardless of CD4 count are eligible for treatment. Infants can now be tested without their mother's consent in accordance with the new guidelines (249).

7.5.3 Scope for future research

Going forward, further studies investigating trends over time using spatiotemporal Bayesian analysis will strengthen the evidence on risk factors and trends found in this study. A combination of integrated, qualitative and quantitative analysis will be useful to fully

understand the proximal and distal factors that influence child malnutrition and HIV mortality. A longitudinal study following individuals' over time, investigating the spatiotemporal patterns and impact of ARVs on nutritional status and mortality, would be valuable future work.

However, generalizable results would be most useful for policy makers. These could be obtained by setting up sentinel sites located strategically across the country. Improving mother's or caregiver's health and socio-economic opportunities would have a direct impact on child survival. Research results informing government efforts to implement and monitor structures in rural South Africa that provide high quality PMTCT, ante-retroviral therapy, access to child support grants and anti-stigmatisation campaigns will contribute to reducing child malnutrition and HIV/TB mortality.

The end !.

8.0 References

1. Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008 - 2030): a population-based study. The Lancet Oncology. 2012;Accepted.
2. Black RE, Morris SS, Bryce J. Where and why are 10 million children dying every year? The Lancet. 2003;361(9376):2226-34.
3. Sachs J. Sustainable Developments-Millennium Development Goals at 10. Scientific American Magazine. 2010;302(6):30
4. UNICEF. Young child survival and development. 2010; Available from: http://www.unicef.org/childsurvival/index.html: [cited 15 August 2012].
5. Bergquist R. New tools for epidemiology: a space odyssey. Memorias Do Instituto Oswaldo Cruz. 2011;106(7):892-900.
6. Steinbrook R. Global health - The AIDS epidemic in 2004. New England Journal of Medicine. 2004;351(2):115-7.
7. Rajaratnam JK, Marcus JR, Flaxman AD, Wang H, Levin-Rector A, Dwyer L, et al. Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970 - 2010: a systematic analysis of progress towards Millennium Development Goal 4. The Lancet. 2010;375(9730):1988-2008.
8. Khatab K. Childhood Malnutrition in Egypt using Geoadditive Gaussian and Latent Variable Models. American journal of tropical medicine and hygiene. 2010;82(4):653.
9. Zere E, McIntyre D. Inequities in under-five child malnutrition in South Africa. International Journal for Equity in Health. 2003;2(1):7.
10. Gorbach SL, Knox TA, Roubenoff R. Interactions between nutrition and infection with human immunodeficiency virus. Nutrition Reviews. 1993;51(8):226-34.
11. Scrimshaw NS, SanGiovanni JP. Synergism of nutrition, infection, and immunity: an overview. American Journal of Clinical Nutrition. 1997;66(2):464S.
12. Scrimshaw NS. Historical concepts of interactions, synergism and antagonism between nutrition and infection. The Journal of Nutrition. 2003;133(1):316S-21S.
13. Beisel WR. Nutrition and immune function: overview. The Journal of Nutrition. 1996;126(10):2611S-5S.
14. Semba RD, Tang AM. Micronutrients and the pathogenesis of human immunodeficiency virus infection. British Journal of Nutrition. 1999;81(1):181-9.
15. Kupka R, Fawzi P. Zinc nutrition and HIV infection. Nutrition Reviews. 2002;60(3):69-79.
16. Babameto G, Kotler DP. Malnutrition in HIV infection. Gastroenterology Clinics of North America. 1997;26(2):393-415.
17. Macallan DC. Wasting in HIV infection and AIDS. The Journal of Nutrition. 1999;129(1):238S.

18. Beach RS, Mantero-Atienza E, Shor-Posner G, Javier JJ, Szapocznik J, Morgan R, et al. Specific nutrient abnormalities in asymptomatic HIV-1 infection. AIDS. 1992;6(7):701.
19. Bogden JD, Kemp FW, Han S, Li W, Bruening K, Denny T, et al. Status of selected nutrients and progression of human immunodeficiency virus type 1 infection. American Journal of Clinical Nutrition. 2000;72(3):809.
20. Berliner LM, Wikle CK, Cressie N. Long-lead prediction of Pacific SSTs via Bayesian dynamic modelling. Journal of climate. 2000;13(22):3953-68.
21. Anderson BA, Phillips H. Adult mortality (age 15-64): based on death notification data in South Africa, 1997-2004: Statistics South Africa; 2006.
22. WHO. Report 2009: Global Tuberculosis Control, Epidemiology, Strategy, Financing. World Health Organization, Geneva, Switzerland. 2009.
23. Agarwal DK, Gelfand AE, Citron-Pousty S. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics. 2002;9(4):341-55.
24. Bradshaw D, Bourne D, Nannan N. What are the leading causes of death among South African children? Congenital heart disease. 2003;1238:1-2.
25. Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Atlanta, Georgia: Wiley; 2004.
26. Tollman SM. The Agincourt field site – Evolution and current status. South African Medical Journal. 1999;89(8):853-8.
27. Cutler DM, Deaton AS, Lleras-Muney A. The determinants of mortality. National Bureau of Economic Research. 2006.
28. Boutayeb A. The double burden of communicable and non-communicable diseases in developing countries. Transactions of the Royal society of Tropical Medicine and Hygiene. 2006;100(3):191-9.
29. Bryce J, Boschi-Pinto C, Shibuya K, Black RE. WHO estimates of the causes of death in children. The Lancet. 2005;365(9465):1147-52.
30. Feltbower RG, Manda SOM, Gilthorpe MS, Greaves MF, Parslow RC, Kinsey SE, et al. Detecting small-area similarities in the epidemiology of childhood acute lymphoblastic leukemia and diabetes mellitus, type 1: A Bayesian approach. American Journal of Epidemiology. 2005;161(12):1168-80.
31. Hargreaves J, Collinson M, Kahn K, Clark S, Tollman S. Childhood mortality among former Mozambican refugees and their hosts in rural South Africa. International Journal of Epidemiology. 2004;33(6):1271 - 8.
32. Mosley WH, Chen LC. An analytical framework for the study of child survival in developing countries. Population and development review. 1984;10:25-45.
33. Filteau S. The HIV-exposed, uninfected African child. Tropical Medicine & International Health. 2009;14(3):276-87.
34. Rutstein SO. Factors associated with trends in infant and child mortality in developing countries during the 1990s. Bulletin of the World Health Organization. 2000;78(10):1256-70.
35. Mash B, Fairall L, Adejayan O, Ikpefan O, Kumari J, Mathee S, et al. A morbidity survey of South African primary care. PloS one. 2012;7(3):e32358.
36. Zanoni BC, Phungula T, Zanoni HM, France H, Feeney ME. Impact of tuberculosis co-treatment on viral suppression rates among HIV-positive children initiating HAART. AIDS. 2011;25(1):49-55.

37. Wiseman CA, Schaaf HS, Cotton MF, Gie RP, Jennings T, Whitelaw A, et al. Bacteriologically confirmed tuberculosis in HIV-infected infants: disease spectrum and survival. International Journal of Tuberculosis and Lung Disease. 2011;15(6):770-5.
38. Frigati LJ, Kranzer K, Cotton MF, Schaaf HS, Lombard CJ, Zar HJ. The impact of isoniazid preventive therapy and antiretroviral therapy on tuberculosis in children infected with HIV in a high tuberculosis incidence setting. Thorax. 2011;66(6):496-501.
39. du Plessis V, Andronikou S, Struck G, McKerrow N, Stoker A. Baseline chest radiographic features of HIV-infected children eligible for antiretroviral therapy. South African Medical Journal. 2011;101(11):829-34.
40. Soeters M, de Vries AM, Kimpen JLL, Donald PR, Schaaf HS. Clinical features and outcome in children admitted to a TB hospital in the Western Cape - the influence of HIV infection and drug resistance. South African Medical Journal. 2005;95(8):602-6.
41. Madhi SA, Huebner RE, Doedens L, Aduc T, Wesley D, Cooper PA. HIV-1 co-infection in children hospitalised with tuberculosis in South Africa. International Journal of Tuberculosis and Lung Disease. 2000;4(5):448-54.
42. Hesseling AC, Westra AE, Werschkull H, Donald PR, Beyers N, D Hussey G, et al. Outcome of HIV infected children with culture confirmed tuberculosis. Archives of Disease in Childhood. 2005;90(11):1171-4.
43. Pretorius B, Schoenfeldt HC. Vitamin A content of fortified maize meal and porridge as purchased and consumed in South Africa. Food Research International. 2011;47(2):128-33.
44. Oldewage-Theron W, Egal AA. Impact of a soy nutrition education programme on knowledge in a low-income community in Qwa-Qwa, South Africa. International Journal of Consumer Studies. 2012;36(4):480-5.
45. Kruger HS, Steyn NP, Swart EC, Maunder EMW, Nel JH, Moeng L, et al. Overweight among children decreased, but obesity prevalence remained high among women in South Africa, 1999-2005. Public Health Nutrition. 2011;15(4):594-9.
46. Kindra G, Coutsoudis A, Esposito F. Effect of nutritional supplementation of breastfeeding HIV positive mothers on maternal and child health: findings from a randomized controlled clinical trial. BMC Public Health. 2011;11:946.
47. Heidari S, Mofenson LM, Hobbs CV, Cotton MF, Marlink R, Katabira E. Unresolved Antiretroviral Treatment Management Issues in HIV-Infected Children. Journal of Acquired Immune Deficiency Syndromes. 2012;59(2):161-9.
48. De Maayer T, Saloojee H. Clinical outcomes of severe malnutrition in a high tuberculosis and HIV setting. Archives of Disease in Childhood. 2011;96(6):560-4.
49. Kanjala C, Alberts M, Byass P, Burger S. Spatial and temporal clustering of mortality in Digkale HDSS in rural northern South Africa. Global health action. 2010;3(S1).
50. Middelkoop K, Bekker LG, Morrow C, Zwane E, Wood R. Childhood tuberculosis infection and disease: a spatial and temporal transmission analysis

in a South African township. South African Medical Journal. 2009;99(10):738-43.
51. Sartorius B, Kahn K, Collinson MA, Vounatsou P, Tollman SM. Survived infancy but still vulnerable: spatial-temporal trends and risk factors for child mortality in rural South Africa (Agincourt), 1992-2007. Geospatial Health. 2011;5(2):285-95.
52. Delaunay V, Etard JF, Preziosi MP, Marra A, Simondon F. Decline of infant and child mortality rates in rural Senegal over a 37-year period (1963-1999). International Journal of Epidemiology. 2001;30:1286-93.
53. Kynast-Wolf G, Sankoh OA, Gbangoun A, Kouyate B, Becher H. Mortality Patterns, 1993-98, in a rural area of Burkina Faso, West Africa, based on the. Nouna demographic surveillance system. Tropical Medicine and International Health. 2002;7:349-56.
54. Becher H, Muller O, Jahn A, Gbangou A, Kynast-Wolf G, Kouyate B. Risk factors of infant and child mortality in rural Burkina Faso. Bulletin of the World Health Organization. 2004:265-73.
55. Msisha WM, Kapinga SH, Earls FJ, Subramanian SV. Place matters: multilevel investigation of HIV distribution in Tanzania. AIDS. 2008;22(6):741-8.
56. Kleinschmidt I, Pettifor A, Morris N, MacPhail C, Rees H. Geographic distribution of human immunodeficiency virus in South Africa. American journal of tropical medicine and hygiene. 2007;77(6):1163-9.
57. Kahn K, Collinson MA, Gomez-Olive FX, Mokoena O, Twine R, Mee P, et al. Profile: Agincourt Health and Socio-demographic Surveillance System. International Journal of Epidemiology. 2012;41:988-1001.
58. Pacheco AG, Tuboi SH, Faulhaber JC, Harrison LH, Schechter M. Increase in Non-AIDS Related Conditions as Causes of Death among HIV-Infected Individuals in the HAART Era in Brazil. Plos One. 2008;3(1):7.
59. Last JM. Re: "A dictionary of epidemiology, fourth edition, edited by John M. Last, Robert A. Spasoff and Susan G. Harris". American Journal of Epidemiology. 2001;154(4):389-.
60. Snow JR. On the mode of Communication of Cholera. London: John Churchill; 1855.
61. Brody H, Rip MR, Vinten-Johanson P, Paneth N, Rachman S. Map-making and myth-making in Broad Street: The London Cholera Epidemic, 1854. Lancet. 2000;356:64-9.
62. Palm TA. The geographical distribution and aetiology of rickets. Practitioner. 1890;45:270 -9.
63. Blum HF. Sunlight as a causal factor in cancer of the skin of man. Journal of the National Cancer Institute. 1948;9:247-58.
64. Sartorius B, Kahn K, Vounatsou P, Collinson M, Tollman S. Young and vulnerable: Spatial-temporal trends and risk factors for infant mortality in rural South Africa (Agincourt), 1992-2007. BMC Public Health. 2010;10(1):645.
65. Musenge E, Vounatsou P, Kahn K. Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spatial and Spatio-temporal Epidemiology. 2011;2:205-17.
66. Gosoniu L, Vounatsou P, Tami A, Nathan R, Grundmann H, Lengeler C. Spatial effects of mosquito bednets on child mortality. BMC Public Health. 2008;8:356.

67. Meade MS. Medical geography as human ecology: the dimension of population movement. Geographical Review. 1977:379-93.
68. Bjørnson FO. Knowledge management in software process improvement: Norwegian University of Science and Technology; 2007.(Unpublished, PhD thesis)
69. Kahn K, Tollman SM, Garenne M, Gear JS. Validation and application of verbal autopsies in a rural area of South Africa. Tropical Medicine and International Health 2000;5(11):824-31.
70. Organization WH. The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research: World Health Organization; 1993.
71. Kimani-Murage E, Kahn K, Pettifor J, Tollman S, Dunger D, Gomez-Olive X, et al. The prevalence of stunting, overweight and obesity, and metabolic disease risk in rural South African children. BMC Public Health. 2010;10(1):158.
72. Kimani-Murage E, Norris S, Pettifor J, Tollman S, Klipstein-Grobusch K, Gomez-Olive X, et al. Nutritional status and HIV in rural South African children. BMC Pediatrics. 2011;11(1):23.
73. Kennedy M. Understanding map projections. 1994 [cited 2012 11 September]; 110]. Available from: http://php.auburn.edu/academic/classes/fory/7470/lab08/understanding%20map%20projections.pdf.
74. Gemperli A. Development of spatial statistical methods for modeling point referenced spatial data in malaria epidemiology. Basel: University of Basel; 2003. (Unpublished, PhD thesis)
75. Whittle P. On stationary process in the plane. Biometrika. 1954;41:434–49.
76. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society,Series B. 1974;36:192–236.
77. Zimmerman DL. Another look at anisotrophy in geostatistics. Mathematical Geology. 1993;25:453-70.
78. Sui DZ. Tobler's first law of geography: A big idea for a small world? Annals of the Association of American Geographers. 2004;94(2):269-77.
79. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1/2):17-23.
80. Walter S. The analysis of regional patterns in health data. American Journal of Epidemiology. 1992;136(6):730.
81. Nychka DW. Spatial process estimates as smoothers. Smoothing and Regression. 2000:393-424.
82. Kammann E, Wand MP. Geoadditive models. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2003;52(1):1-18.
83. Matérn B. Spatial Variation. 2nd ed. Berlin: Spinger-Verlag.; 1960;reprinted 1986.
84. Handcock MS, Wallis J. An approach to statistical spatial temporal modeling of meteorological fields (with discussion). Journal American Statistical Association. 1994;89:368-90.
85. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. Boca Raton, Fla.: Chapman & Hall/CRC; 2004.

86. Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis. 2006;50(4):967-91.
87. Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. Journal of the Royal Statistical Society Series C-Applied Statistics. 2001;50(2):201-20.
88. Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal Chemical,Metallurgical and mining society of South Africa. 1951 52:119-39.
89. Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology. 2003;3:21.
90. Fekedulegn D, Andrew M, Violanti J, Hartley T, Charles L, Burchfiel C. Comparison of Statistical Approaches to Evaluate Factors Associated With Metabolic Syndrome. Journal of Clinical Hypertension. 2010;12(5):365-73.
91. Cressie NAC. Statistics for spatial data. New York: Wiley; 1993.
92. Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice: Chapman & Hall/CRC; 1996.
93. Casella G, Robert CP. Monte Carlo statistical methods. Springer-Verlag, New York; 1999.
94. Wikle CK. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology. 2003;84(6):1382-94.
95. Wikle CK, Berliner LM, Cressie N. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics. 1998;5(2):117-54.
96. Rue H, Held L. Gaussian Markov random fields: theory and applications: Chapman & Hall; 2005.
97. Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modelling of particulate matter concentration through the SPDE approach. Advanced Statistical Analysis, Accepted. 2011.
98. Andiman W, Boucher M, Burns D, Bryson Y, Farley J, Fowler H, et al. The mode of delivery and the risk of vertical transmission of human immunodeficiency virus type 1 - A meta-analysis of 15 prospective cohort studies. New England Journal of Medicine. 1999;340(13):977-87.
99. Pardo- Igúzquiza E, Dowd PA. AMLE3D: A computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. Computers and Geosciences. 1997;23(7):793-805.
100. Feuntes M. Spectral methods for nonstationary spatial processes. Biometrika. 2002;89:197-210.
101. Monahan J. Numerical methods of statistics. Cambridge: Cambridge University Press; 2001.
102. Sachs J. Sustainable Developments-Millennium Development Goals at 10. Scientific American Magazine. 2010;302(6):30.
103. Gregory JR, Lowe S, Bates CJ, Prentice A, Jackson L, Smithers G, et al. National Diet and Nutrition Survey – young people aged 4 to 18 years, Vol. 1: Report of the diet and nutrition survey. London: The Stationery Office: 2000.
104. Pomerleau J, Lock K, McKee M, Altmann DR. The challenge of measuring global fruit and vegetable intake. J Nutr. 2004;134(5):1175-80.

105. Labadarios D, Maunder E, Steyn N, MacIntyre U, Swart R, Gericke G, et al. National food consumption survey in children aged 1-9 years: South Africa 1999. Forum Nutr. 2003;56:106-9.
106. Gordon-Larsen P, Adair LS, Nelson MC, Popkin BM. Five-year obesity incidence in the transition period between adolescence and adulthood: the National Longitudinal Study of Adolescent Health. Am J Clin Nutr. 2004;80(3):569-75.
107. Guo SS, Roche AF, Chumlea WC, Gardner JD, Siervogel RM. The predictive value of childhood body mass index values for overweight at age 35 y. Am J Clin Nutr. 1994;59(4):810-9.
108. St-Onge MP, Keller KL, Heymsfield SB. Changes in childhood food consumption patterns: a cause for concern in light of increasing body weights. Am J Clin Nutr. 2003;78(6):1068-73.
109. Eckel RH, Grundy SM, Zimmet PZ. The metabolic syndrome. The Lancet. 2005;365(9468):1415-28.
110. Armstrong M, Lambert M, Sharwood K, Lambert E. Obesity and overweight in South African primary school children in the Health of the Nation Study. Journal of Endocrinology, Metabolism and Diabetes of South Africa. 2008;11(2):53-63.
111. Kandala NB, Lang S, Klasen S, Fahrmeir L. Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two african countries. 2001.
112. Khatab K. Childhood Malnutrition in Egypt using Geoadditive Gaussian and Latent Variable Models. The American journal of tropical medicine and hygiene. 2010;82(4):653.
113. Mueller I, Vounatsou P, Allen B, Smith T. Spatial patterns of child growth in Papua New Guinea and their relation to environment, diet, socio-economic status and subsistence activities. Annals of Human Biology. 2001;28(3):263-80.
114. Schmidt M, Dwyer T, Magnussen C, Venn A. Predictive associations between alternative measures of childhood adiposity and adult cardio-metabolic health. International Journal of Obesity. 2010;35(1):38-45.
115. Ledoux TA, Watson K, Barnett A, Nguyen NT, Baranowski JC, Baranowski T. Components of the Diet Associated with Child Adiposity: A Cross-Sectional Study. Journal of the American College of Nutrition. 2011;30(6):536-46.
116. Hoeting JA. The importance of accounting for spatial and temporal correlation in analyses of ecological data. Ecological Applications. 2009;19(3):574-7.
117. Kahn K, Tollman S, Collinson M, Clark S, Twine R, Clark B, et al. Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt Health and Demographic Surveillance System. Scand J Public Health. 2007;35(Suppl 69):8 - 20.
118. Lohman T, Roche A, Martorell R. Anthropometric standardization reference manual.: Champaign, IL: Human Kinetics Books; 1991.
119. Johnston J, Leong M, Checkland E, Zuberbuhler P, Conger P, Quinney H. Body fat assessed from body density and estimated from skinfold thickness in normal children and children with cystic fibrosis. American Journal of Clinical Nutrition. 1988;48(6):1362.

120. Durnin J, Womersley J. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. British Journal of Nutrition. 1974;32(01):77-97.
121. Durnin J, Rahaman M. The assessment of the amount of fat in the human body from measurements of skinfold thickness. British Journal of Nutrition. 1967;21(03):681-9.
122. Siri W, Brozeck J, Henschel A, editors. Body composition from fluid spaces and density: analysis of methods1961: National Academies.
123. Lohman T, Slaughter M, Boileau R, Bunt J, Lussier L. Bone mineral measurements and their relation to body density in children, youth and adults. Human biology. 1984;56(4):667-79.
124. Boileau R, Lohman T, Slaughter M, Ball T, Going S, Hendrix M. Hydration of the fat-free body in children during maturation. Human biology; an international record of research. 1984;56(4):651.
125. Anselin L. Chapter eight: The Moran scatterplot as an ESDA tool to assess local instability in spatial association. Spatial analytical perspectives on GIS. 1996;4:121.
126. Anselin L. Local indicators of spatial association-LISA. Geographical Analysis. 1995;27(2):93-115.
127. Anselin L, Syabri I, Kho Y. GeoDa: An introduction to spatial data analysis. Geographical Analysis. 2006;38(1):5-22.
128. Jackson MC, Huang L, Xie Q, Tiwari RC. Research A modified version of Moran's I. 2010.
129. Geary RC. The contiguity ratio and statistical mapping. The incorporated statistician. 1954;5(3):115-46.
130. Cambardella C, Moorman T, Novak J, Parkin T, Karlen D, Turco R, et al. Field-scale variability of soil properties in central Iowa soils. Soil Science Society American Journal. 1994;58:1501-11.
131. Guo X, Fu B, Ma K, Chen L. Spatial variability of soil nutrients based on geostatistics combined with GIS--a case study in Zunghua City of Hebei Province]. Ying yong sheng tai xue bao= The journal of applied ecology/Zhongguo sheng tai xue xue hui, Zhongguo ke xue yuan Shenyang ying yong sheng tai yan jiu suo zhu ban. 2000;11(4):557.
132. Kneib T, Fahrmeir L. A mixed model approach for geoadditive hazard regression. Scandinavian Journal of Statistics Theory and Applications. 2007;34(1):207-28.
133. Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics & Data Analysis. 2006;50(4):967-91.
134. StataCorp. Stata Statistical Software Release 12. College Station. Texus,USA: StataCorp LP; 2011.
135. Institute S. SAS user's guide: statistics: version 9.2: SAS Inst; 2009.
136. Belitz C, Brezger A, Kneib T, Lang S, Fronk E, Heinzl F, et al. Bayesx-software for bayesian inference in structured additive regression models. Version 2.01.Available from http://www.stat.uni-muenchen.de/~bayesx; 2011.

137. R-Cran. R: A Language and Environment for Statistical Computing. R Development Core Team. 2.12.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2011.
138. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7-11.
139. Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 2008;24(4):581.
140. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. Operations Research. 1983:1109-44.
141. Duarte Bonini Campos JA, Melanda EA, Antunes JdS, Rozato Foschini AL. Dental caries and the nutritional status of preschool children - a spatial analysis. Ciencia & Saude Coletiva. 2011;16(10):4161-8.
142. Fraser LK, Clarke GP, Cade JE, Edwards KL. Fast Food and Obesity A Spatial Analysis in a Large United Kingdom Population of Children Aged 13-15. American Journal of Preventive Medicine. 2012;42(5):E77-E85.
143. Pawloski LR, Curtin KM, Gewa C, Attaway D. Maternal-child overweight/obesity and undernutrition in Kenya: a geographic analysis. Public Health Nutrition. 2012;15(11):2140-7.
144. Choh AC, Demerath EW, Goring HH, Dyer TD, Haack K, Blangero J, et al. Suggestive linkage for QTL influencing visceral adiposity on chromosomes 13 and 17: The southwest ohio family study. Circulation. 2007;115(8):E299-E.
145. Irvine KM, Gitelman AI, Hoeting JA. Spatial designs and properties of spatial correlation: effects on covariance estimation. Journal of agricultural, biological, and environmental statistics. 2007;12(4):450-69.
146. Black RE, Morris SS, Bryce J. Where and why are 10 million children dying every year? Lancet. 2003;361(9376):2226-34.
147. World Health Organization. Department of Nutrition for Health and Development. WHO global database on child growth and malnutrition. Geneva2010.
148. Scrimshaw NS, SanGiovanni JP. Synergism of nutrition, infection, and immunity: an overview. The American journal of clinical nutrition. 1997;66(2):464S.
149. Beach RS, Mantero-Atienza E, Shor-Posner G, Javier JJ, Szapocznik J, Morgan R, et al. Specific nutrient abnormalities in asymptomatic HIV-1 infection. AIDS (London, England). 1992;6(7):701.
150. Bogden JD, Kemp FW, Han S, Li W, Bruening K, Denny T, et al. Status of selected nutrients and progression of human immunodeficiency virus type 1 infection. The American journal of clinical nutrition. 2000;72(3):809.
151. Wheeler DA, Gibert CL, Launer CA, Muurahainen N, Elion RA, Abrams DI, et al. Weight loss as a predictor of survival and disease progression in HIV infection. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology. 1998;18(1):80-5.
152. Malvy D, Thiebaut R, Marimoutou C, Dabis F, Grp Epidemiologie Clin Sida A. Weight loss and body mass index as predictors of HIV disease progression to AIDS in adults. Aquitaine cohort, France, 1985-1997. Journal of the American College of Nutrition. 2001;20(6):609-15.

153. Melchior JC, Niyongabo T, Henzel D, Durack-Bown I, Henri SC, Boulier A. Malnutrition and wasting, immunodepression, and chronic inflammation as independent predictors of survival in HIV-infected patients. Nutrition. 1999;15(11-12):865-9.
154. Rodolphe T, Denis M, Catherine M, Francois D, Grp Epidemiologie Clin Sida A. Anthropometric indices as predictors of survival in AIDS adults. Aquitaine Cohort, France, 1985-1997. European Journal of Epidemiology. 2000;16(7):633-9.
155. Smit E, Skolasky RL, Dobs AS, Calhoun BC, Visscher BR, Palella FJ, et al. Changes in the incidence and predictors of wasting syndrome related to human immunodeficiency virus infection, 1987-1999. American Journal of Epidemiology. 2002;156(3):211-8.
156. Tabi M, Vogel RL. Nutritional counselling: an intervention for HIV-positive patients. Journal of Advanced Nursing. 2006;54(6):676-82.
157. Crum-Cianflone NF, Roediger M, Eberly LE, Ganesan A, Weintrob A, Johnson E, et al. Impact of Weight on Immune Cell Counts among HIV-Infected Persons. Clinical and Vaccine Immunology. 2011;18(6):940-6.
158. Mueller I, Vounatsou P, Smith T, Allen B. Subsistence agriculture and child growth in Papua New Guinea. Ecology of food and nutrition. 2001;40(4):367-95.
159. Gelfand AE, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics. 2003;4(1):11.
160. World Health Organization. WHO Anthro 2005, Beta version Feb 17th, 2006: Software for assessing growth and development of the world's children. Geneva2006.
161. Beydoun MA, Kuczmarski MTF, Mason MA, Ling SM, Evans MK, Zonderman AB. Role of depressive symptoms in explaining socioeconomic status disparities in dietary quality and central adiposity among US adults: a structural equation modeling approach. The American journal of clinical nutrition. 2009;90(4):1084.
162. Casella G, George EI. Explaining the Gibbs sampler. American Statistician. 1992;46(3):167-74.
163. R-cran. R: A Language and Environment for Statistical Computing. R Development Core Team. 2.10.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2010.
164. Finley AO, Banerjee S, Carlin BP. spBayes overview Spatial models and Bayesian inference Illustration Future direction Summary. 2007.
165. Bailey RC, Kamenga MC, Nsuami MJ, Nieburg P, St Louis ME. Growth of children according to maternal and child HIV, immunological and disease characteristics: a prospective cohort study in Kinshasa, Democratic Republic of Congo. International Journal of Epidemiology. 1999;28(3):532-40.
166. Henderson RA, Miotti PG, Saavedra JM, Dallabetta G, Chiphangwi J, Liomba G, et al. Longitudinal growth during the first 2 years of life in children born to HIV-infected mothers in Malawi, Africa. Pediatric Aids and Hiv Infection-Fetus to Adolescent. 1996;7(2):91-7.

167. Hendricks MK, Eley B, Bourne LT. Nutrition and HIV/AIDS in infants and children in South Africa: implications for food-based dietary guidelines. Maternal and Child Nutrition. 2007;3(4):322-33.
168. Simpore J, Zongo F, Kabore F, Dansou D, Bere A, Nikiema JB, et al. Nutrition rehabilitation of HIV-infected and HIV-negative undernourished children utilizing spirulina. Annals of Nutrition and Metabolism. 2005;49(6):373-80.
169. Stambullian M, Feliu S, Slobodianik NH. Nutritional status in patients with HIV infection and AIDS. British Journal of Nutrition. 2007;98:S140-S3.
170. Lepage P, Msellati P, Hitimana DG, Bazubagira A, VanGoethem C, Simonon A, et al. Growth of human immunodeficiency type 1-infected and uninfected children: A prospective cohort study in Kigali, Rwanda, 1988 to 1993. Pediatric Infectious Disease Journal. 1996;15(6):479-85.
171. Sutcliffe CG, Scott S, Mugala N, Ndhlovu Z, Monze M, Quinn TC, et al. Survival from 9 months of age among HIV-infected and uninfected Zambian children prior to the availability of antiretroviral therapy. Clinical infectious diseases. 2008;47(6):837.
172. Duggan C, Fawzi W. Micronutrients and child health: Studies in international nutrition and HIV infection. Nutrition Reviews. 2001;59(11):358-69.
173. Kuhn L, Aldrovandi GM, Sinkala M, Kankasa C, Semrau K, Mwiya M, et al. Effects of early, abrupt weaning on HIV-free survival of children in Zambia. New England Journal of Medicine. 2008;359(2):130-41.
174. Papathakis PC, Rollins NC, Chantry CJ, Bennish ML, Brown KH. Micronutrient status during lactation in HIV-infected and HIV-uninfected South African women during the first 6 mo after delivery. American Journal of Clinical Nutrition. 2007;85(1):182-92.
175. Kabue MM, Kekitiinwa A, Maganda A, Risser JM, Chan W, Kline MW. Growth in HIV-infected children receiving antiretroviral therapy at a pediatric infectious diseases clinic in Uganda. Aids Patient Care and Stds. 2008;22(3):245-51.
176. Sutcliffe CG, van Dijk JH, Munsanje B, Hamangaba F, Sinywimaanzi P, Thuma PE, et al. Weight and height z-scores improve after initiating ART among HIV-infected children in rural Zambia: a cohort study. Bmc Infectious Diseases. 2011;11:7.
177. Naidoo R, Rennert W, Lung A, Naidoo K, McKerrow N. The Influence of Nutritional Status on the Response to HAART in HIV-Infected Children in South Africa. Pediatric Infectious Disease Journal. 2010;29(6):511-3.
178. Eley B, Davies MA, Apolles P, Cowburn C, Buys H, Zampoli M, et al. Antiretroviral treatment for children. Samj South African Medical Journal. 2006;96(9):988-93.
179. Weigel R, Phiri S, Chiputula F, Gumulira J, Brinkhof M, Gsponer T, et al. Growth response to antiretroviral treatment in HIV-infected children: a cohort study from Lilongwe, Malawi. Tropical Medicine & International Health. 2010;15(8):934-44.
180. Larrea C, Kawachi I. Does economic inequality affect child malnutrition? The case of Ecuador. Social Science & Medicine. 2005;60(1):165-78.

181. Downing J, Jones L, Bates G, Sumnall H, Bellis MA. A systematic review of parent and family-based intervention effectiveness on sexual outcomes in young people. Health Education Research. 2011;26(5):808-33.
182. Rotheram-Borus MJ, Li L, Liang L-J, Wen Y, Wu Z, Prevent NCHS. Attitudinal, behavioural and biological outcomes of a community popular opinion leader intervention in China. Aids Education and Prevention. 2011;23(5):448-56.
183. Simoni JM, Nelson KM, Franks JC, Yard SS, Lehavot K. Are Peer Interventions for HIV Efficacious? A Systematic Review. Aids and Behavior. 2011;15(8):1589-95.
184. Smith BD, Kalayil EJ, Patel-Larson A, Chen B, Vaughan M. Retaining clients in an outcome monitoring evaluation study: HIV prevention efforts in community settings. Evaluation and program planning. 2012;35(1):16-24.
185. UNAIDS. Report on the global AIDS epidemic 2011. Available from: http://www.avert.org/africa-hiv-aids-statistics.htm [cited 23 October 2012].
186. WHO. One million lives can be saved between now and the end of 2015 by preventing and treating tuberculosis among people living with HIV. Available from: http://www.who.int/tb/challenges/hiv/fastfactsbrochure.pdf [cited 23 October 2012].
187. Bradshaw D, Bourne D, Nannan N. What are the leading causes of death among South African children. MRC Policy Brief & Unicef. 2003;3:1-4.
188. Doherty T, Sanders D, Goga A, Jackson D. Implications of the new WHO guidelines on HIV and infant feeding for child survival in South Africa. Bulletin of the World Health Organization. 2011;89(1):62-7.
189. Rutstein SO. Factors associated with trends in infant and child mortality in developing countries in the 1990s.2000. Report No.: 78.
190. Guerrero L, La Valley A. Conflict, emotion, and communication. The Sage handbook of conflict communication: Integrating theory, research, and practice. 2006:69-96.
191. Hargreaves J, Collinson M, Kahn K, Clark S, Tollman S. Childhood mortality among former Mozambican refugees and their hosts in rural South Africa. Int J Epidemiol. 2004;33(6):1271 - 8.
192. Marshall RJ. A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease. Journal of the Royal Statistical Society Series A (Statistics in Society). 1991;154(3):421-41.
193. Moonan PK, Bayona M, Quitugua TN, Oppong J, Dunbar D, Jost KC, et al. Using GIS technology to identify areas of tuberculosis transmission and incidence. International Journal of Health Geographics. 2004;3(1):23.
194. Tanser F, LeSueur D, Solarsh G, Wilkinson D. HIV heterogeneity and proximity of homestead to roads in rural South Africa: an exploration using a geographical information system. Tropical Medicine & International Health. 2000;5(1):40-6.
195. Musenge E, Vounatsou P, Kahn K. Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spatial and Spatio-temporal Epidemiology. 2011;2(4):205-17.
196. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa:

Data and methods of the Agincourt Health and Demographic Surveillance System. Scandinavian Journal of Public Health. 2007;35(69 suppl):8-20.
197. Clark SJ, Collinson MA, Kahn K, Drullinger K, Tollman SM. Returning home to die: circular labour migration and mortality in South Africa. Scand J Public Health Suppl. 2007;69:35-44.
198. Kahn K, Tollman SM, Garenne M, Gear JS. Validation and application of verbal autopsies in a rural area of South Africa. Tropical Medicine and International Health. 2000;5(11):824-31.
199. Tollman SM, Kahn K, Sartorius B, Collinson MA, Clark SJ, Garenne ML. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. The Lancet. 2008;372(9642):893-901.
200. Lewis JJC, Donnelly CA, Mare P, Mupambireyi Z, Garnett GP, Gregson S. Evaluating the proximate determinants framework for HIV infection in rural Zimbabwe. Sexually transmitted infections. 2007;83(suppl 1):i61-i9.
201. Boerma J, Weir SS. Integrating demographic and epidemiological approaches to research on HIV/AIDS: the proximate-determinants framework. Journal of Infectious Diseases. 2005;191(Supplement 1):S61.
202. Lopman B, Nyamukapa C, Mushati P, Mupambireyi Z, Mason P, Garnett GP, et al. HIV incidence in 3 years of follow-up of a Zimbabwe cohort—1998–2000 to 2001–03: contributions of proximate and underlying determinants to transmission. International journal of epidemiology. 2008;37(1):88-105.
203. StataCorp. Stata Statistical Software Release 10. College Station. Texus,USA: StataCorp LP; 2007.
204. Coetzee D, Hilderbrand K, Boulle A, Draper B, Abdullah F, Goemaere E. Effectiveness of the first district-wide programme for the prevention of mother-to-child transmission of HIV in South Africa. Bulletin of the World Health Organization. 2005;83(7):489-94.
205. Kimani-Murage EW, Manderson L, Norris SA, Kahn K. 'You opened our eyes': care-giving after learning a child's positive HIV status in rural South Africa. Health & Social Care in the Community. 2010;18(3):264-71.
206. Ndirangu J, Newell M-L, Tanser F, Herbst AJ, Bland R. Decline in early life mortality in a high HIV prevalence rural area of South Africa: evidence of HIV prevention or treatment impact? AIDS. 2010;24(4):593-602.
207. Katapa R. A comparison of female-and male-headed households in Tanzania and poverty implications. Journal of Biosocial Science. 2006;38(3):327.
208. Wiegers E, Curry J, Garbero A, Hourihan J. Patterns of vulnerability to AIDS impacts in Zambian households. Development and change. 2006;37(5):1073-92.
209. Oni S, Obi C, Okorie A, Thabede D, Jordan A. The economic impact of HIV/AIDS on rural households in Limpopo Province. South African Journal of Economics. 2002;70(7):551-62.
210. Collinson MA. Striving against adversity. : the dynamics of migration, health and poverty in rural South Africa. Umeå: Umeå University medical dissertations; 2009.
211. Posel D, Casale D. What has been happening to internal labour migration in South Africa, 1993 - 1999? South African Journal of Economics. 2003;71(3):455-79.

212. Hahn MB, Riederer AM, Foster SO. The Livelihood Vulnerability Index: A pragmatic approach to assessing risks from climate variability and change--A case study in Mozambique. Global Environmental Change. 2009;19(1):74-88.
213. Beguy D, Bocquier P, Zulu EM. Circular migration patterns and determinants in Nairobi slum settlements. Demographic Research. 2010;23(20):549-86.
214. Kline RB. Principles and Practice of Structural Equation Modeling. 3rd ed. New York: Guilford Press; 2011.
215. Bollen KA. Structural equations with latent variables: Wiley New York; 1989.
216. Bollen KA, Curran PJ. Latent curve models: A structural equation perspective: Wiley-Interscience; 2006.
217. Sachs J, McArthur J. The Millennium Project: a plan for meeting the millennium development goals. Lancet. 2005;365(9456):347-53.
218. UNDG. The Millenium Development Goals Report 2008. Available from: http://www.un.org/millenniumgoals/pdf/MDG_Report_2009_ENG.pdf [cited 23 October 2012]
219. UNAIDS. Report on the global AIDS epidemic 2011. Available from: http://www.avert.org/africa-hiv-aids-statistics.htm [cited 23 October 2012]
220. UNAIDS. Report on the global AIDS epidemic 2008. Available from: http://www.unaids.org/en/dataanalysis/knowyourepidemic/epidemiologypublications/2008reportontheglobalaidsepidemic/ [cited 23 October].
221. UNESCO. South Africa Millennium Development Goals Mid term country report. 2007.Available from: http://www.europarl.europa.eu/meetdocs/2009_2014/documents/d-za/dv/5_m/5_mdg.pdf [cited 23 October 2012].
222. UNICEF. State of the World's Children 2009. Available from: http://www.unicef.org/publications/index_47127.html. [cited 23 October 2012]
223. Becher H, Muller O, Jahn A, Gbangou A, Kynast-Wolf G, Kouyate B. Risk factors of infant and child mortality in rural Burkina Faso.2004. Report No.: 82.
224. Kleinschmidt I, Pettifor A, Morris N, MacPhail C, Rees H. Geographic distribution of human immunodeficiency virus in South Africa. The American Journal of Tropical Medicine and Hygiene. 2007;77(6):1163-9.
225. Kahn K, Tollman SM, Garenne M, Gear JS, Clark SJ, Twine R, et al. Building a programme in health, population and social transitions research in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System. Scandanavian Journal of Public Health. 2007.
226. Greven S, Dominici F, Zeger SL. A Spatio-Temporal Approach for Estimating Chronic Effects of Air Pollution. Department of Biostatistics2009.
227. Burnett R, Ma RJ, Jerrett M, Goldberg MS, Cakmak S, Pope CA, et al. The spatial association between community air pollution and mortality: A new method of analyzing correlated geographic cohort data. Environmental Health Perspectives. 2001;109:375-80.
228. Greenland S. Modeling and variable selection in Epidemiologic analysis. American Journal of Public Health. 1989;79(3):340-9.
229. Tollman SM. The Agincourt field site – Evolution and current status. South African Medical Journal. 1999;89(8):853-8

230. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa: Data and methods of the agincourt health and demographic surveillance system. Scandinavian Journal of Public Health. 2007;35:8-20.
231. Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003;3:21.
232. Fahrmeir L, Echavarria LO. Structured additive regression for overdispersed and zero-inflated-count data. Applied Stochastic Models in Business and Industry. 2006;22(4):351-69.
233. Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. Biostat. 2003;4:11 - 25.
234. Kamman E, Wand M. Geoadditive models, Department of Biostatistics, School of Public Health, Havard University. 2001.
235. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. Operations Research. 1983;31(6):1109-44.
236. Geweke J, Minneapolis FRBo. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments: Federal Reserve Bank of Minneapolis, Research Dept.; 1991.
237. Raftery AE, Lewis S. How many iterations in the Gibbs sampler. Bayesian statistics. 1992;4(2):763-73.
238. Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2002;64(4):583-639.
239. Belitz C, Brezger A, Kneib T, Lang S, Fronk E, Heinzl F, et al. BayesX-software for bayesian inference in structured additive regression models. Version 2.01 2009.
240. Press WH. Numerical recipes in C++ : the art of scientific computing. Cambridge, UK; New York: Cambridge University Press; 2002.
241. Fleming T, Harrington D. Non parametric estimation of the survival distribution in censored data. Communications in statistics Theory and methods. 1984;13(20):2469-86.
242. Ranta J, Penttinen A. Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. Statistics in medicine. 2000;19(1718):2345-59.
243. Tollman S, Kahn K, Sartorius B, Collinson M, Clark S, Garenne M. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. The Lancet. 2008;372(9642):893-901.
244. Collinson M, White M, Short S, Lurie M, Byass P, Kahn K, et al. Child mortality,Migration and Parental Presence in rural South Africa near the border with Mozambique. 2009.
245. Argeseanu S. Risks, Amenities, and Child Mortality in Rural South Africa. African Population Studies. 2004;19(1):13-33.
246. Ndirangu J, Newell ML, Tanser F, Herbst AJ, Bland R. Decline in early life mortality in a high HIV prevalence rural area of South Africa: evidence of HIV prevention or treatment impact? AIDS. 2010;24(4):593.

247. Sachs JD. Millennium Development Goals at 10 A decade's worth of targeted accomplishments shows extreme poverty can be eliminated. Scientific American.302(6):30-.
248. Case A, Menendez A. Does money empower the elderly? Evidence from the Agincourt demographic surveillance site, South Africa. Scand J Public Health Suppl. 2007;69:157-64.
249. Zuma J. Address by President Jacob Zuma on the occasion of World AIDS Day, Pretoria Showgrounds. South African Government Information; 2009.
250. Bradshaw D, Johnson L, Schneider H, Bourne D, Dorrington R. Orphans of the HIV/AIDS epidemic - the time to act is now. Cape Town, South Africa: Medical Reasearch Council2002.
251. Ridout M, Hinde J, Demetrio CGB. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics. 2001;57(1):219-23.
252. Browne WJ, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Statistics. 2000;15(3):391-420.
253. Rose G. Sick individuals and sick populations. International journal of epidemiology. 2001;30(3):427.
254. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data: Chapman & Hall/CRC; 2004.
255. Banerjee S, Carlin B. Semiparametric spatio-temporal frailty modeling. Environmetrics. 2003;14(5):523-35.
256. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society,Series B. 2008;70,Part 4:825-48.
257. Johnson ME, Moore LM, Ylvisaker D. Minimax and maximin designs. Journal of Statistical Planning and Inference. 1990;26:131-48.
258. French JL, Kammann EE, Wand MP. Comment on paper by Ke and Wang. Journal of American Statistical Association. 2002;96:1285-8.
259. Lindgren F, Rue H, Lindstrom J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the royal statistical society: Series b (statistical methodology). 2011;73(4):423-98.
260. Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modelling of particulate matter concentration through the SPDE approach. AStA Adv Stat Anal, 2011.(In press)
261. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology). 2009;71(2):319-92.
262. Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. Journal of the Royal Statistical Society Series C Applied Statistics. 2001;50(Part 2):201-20.
263. Simpson D, Lindgren F, Rue H. Fast approximate inference with INLA: the past, the present and the future. Arxiv preprint arXiv:11052982. 2011.

264. Simpson D, Lindgren F, Rue H. Think continuous: Markovian Gaussian models in spatial statistics. Arxiv preprint arXiv:11106796. 2011.
265. Tollman SM, Herbst K, Garenne M, Gear JS, Kahn K. The Agincourt Demographic and Health Study: Site description, baseline findings and implications. . South African Medical Journal. 1999;89(8):858-64.
266. Agarwal D, Gelfand A, Citron-Pousty S. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics. 2002;9(4):341-55.
267. Ridout M, Hinde J, Demetrio C. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics. 2001;57(1):219-23.
268. Chopra M, Piwoz E, Sengwana J, Schaay N, Dunnett L, Sanders D. Effect of a mother-to-child HIV prevention programme on infant feeding and caring practices in South Africa. Samj South African Medical Journal. 2002;92(4):298-302.
269. Papathakis PC, Rollins NC, Chantry CJ, Bennish ML, Brown KH. Micronutrient status during lactation in HIV-infected and HIV-uninfected South African women during the first 6 months after delivery. American Journal of Clinical Nutrition. 2007;85(1):182-92.
270. Tohill BC, Heilig CM, Klein RS, Rompalo A, Cu-Uvin S, Piwoz EG, et al. Nutritional biomarkers associated with gynecological conditions among US women with or at risk of HIV infection. American Journal of Clinical Nutrition. 2007;85(5):1327-34.
271. Crum-Cianflone NF, Roediger M, Eberly LE, Ganesan A, Weintrob A, Johnson E, et al. Impact of Weight on Immune Cell Counts among HIV-Infected Persons. Clinical and Vaccine Immunology. 2011;18(6):940-6.
272. Brocklehurst P, French R. The association between maternal HIV infection and perinatal outcome: a systematic review of the literature and meta-analysis. British Journal of Obstetrics and Gynaecology. 1998;105(8):836-48.
273. Coley JL, Msamanga GI, Fawzi MCS, Kaaya S, Hertzmark E, Kapiga S, et al. The association between maternal HIV-1 infection and pregnancy outcomes in Dar es Salaam, Tanzania. British Journal of Obstetrics and Gynaecology. 2001;108(11):1125-33.
274. Bakaki P, Kayita J, Machado JEM, Coulter JBS, Tindyebwa D, Ndugwa CM, et al. Epidemiologic and clinical features of HIV-infected and HIV-uninfected Ugandan children younger than 18 months. Journal of Acquired Immune Deficiency Syndromes. 2001;28(1):35-42.
275. Newell ML, Borja MC, Peckham C, European Collaborative S. Height, weight, and growth in children born to mothers with HIV-1 infection in Europe. Pediatrics. 2003;111(1).
276. Jiamton S, Pepin J, Suttent R, Filteau S, Mahakkanukrauh B, Hanshaoworakul W, et al. A randomized trial of the impact of multiple micronutrient supplementation on mortality among HIV-infected individuals living in Bangkok. AIDS. 2003;17(17):2461-9.
277. Fawzi WW, Msamanga GI, Spiegelman D, Wei RL, Kapiga S, Villamor E, et al. A randomized trial of multivitamin supplements and HIV disease progression and mortality. New England Journal of Medicine. 2004;351(1):23-32.

278. Austin J, Singhal N, Voigt R, Smaill F, Gill MJ, Walmsley S, et al. A community randomized controlled clinical trial of mixed carotenoids and micronutrient supplementation of patients with acquired immunodeficiency syndrome. European Journal of Clinical Nutrition. 2006;60(11):1266-76.
279. Irlam JH, Visser MME, Rollins NN, Siegfried N. Micronutrient supplementation in children and adults with HIV infection. Cochrane Database of Systematic Reviews. 2010(12).
280. Range N, Changalucha J, Krarup H, Magnussen P, Andersen AB, Friis H. The effect of multi-vitamin/mineral supplementation on mortality during treatment of pulmonary tuberculosis: a randomised two-by-two factorial trial in Mwanza, Tanzania. British Journal of Nutrition. 2006;95(4):762-70.
281. Coovadia H. Current issues in prevention of mother-to-child transmission of HIV-1. Current Opinion in Hiv and Aids. 2009;4(4):319-24.
282. Taha TE, Hoover DR, Chen S, Kumwenda NI, Mipando L, Nkanaunena K, et al. Effects of Cessation of Breastfeeding in HIV-1-Exposed, Uninfected Children in Malawi. Clinical Infectious Diseases. 2011;53(4):388-95.
283. Abrams EJ, Matheson PB, Thomas PA, Thea DM, Krasinski K, Lambert G, et al. Neonatal predictors of infection status and early death among 332 infants at risk of HIV-1 infection monitored prospectively from birth. Pediatrics. 1995;96(3):451-8.
284. Magder LS, Mofenson L, Paul ME, Zorrilla CD, Blattner WA, Tuomala RE, et al. Risk factors for in utero and intrapartum transmission of HIV. Journal of Acquired Immune Deficiency Syndromes. 2005;38(1):87-95.
285. Lambert D. Zero-inflate Poisson regression, with application to defects in manufacturing. Technometrics. 1992;34(1):1-14.
286. Clark SJ, Collinson MA, Kahn K, Drullinger K, Tollman SM. Returning home to die: circular labour migration and mortality in South Africa. Scandinavain Journal of Public Health. 2007;69:S35-44.

9.0 Appendices: Published and in-press papers and ethics certificate

The following are the papers for Chapters 4 to 6 as published or in-press online; as such page numbers and other detail are according to the specific journals.

BUILDING NEW KNOWLEDGE SUPPLEMENT

COACTION PUBLISHING

The contribution of spatial analysis to understanding HIV/TB mortality in children: a structural equation modelling approach

Eustasius Musenge[1,2*], Penelope Vounatsou[3], Mark Collinson[1,4,5], Stephen Tollman[1,4,5] and Kathleen Kahn[1,4,5]

[1]MRC/Wits Rural Public Health & Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [2]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [3]Swiss Tropical and Public Health Institute, University of Basel, Basel, Switzerland; [4]Centre for Global Health Research, Umeå University, Umeå, Sweden; [5]INDEPTH Network, Accra, Ghana

Background: South Africa accounts for more than a sixth of the global population of people infected with HIV and TB, ranking her highest in HIV/TB co-infection worldwide. Remote areas often bear the greatest burden of morbidity and mortality, yet there are spatial differences within rural settings.

Objectives: The primary aim was to investigate HIV/TB mortality determinants and their spatial distribution in the rural Agincourt sub-district for children aged 1–5 years in 2004. Our secondary aim was to model how the associated factors were interrelated as either underlying or proximate factors of child mortality using pathway analysis based on a Mosley-Chen conceptual framework.

Design: We conducted a secondary data analysis based on cross-sectional data collected in 2004 from the Agincourt sub-district in rural northeast South Africa. Child HIV/TB death was the outcome measure derived from physician assessed verbal autopsy. Modelling used multiple logit regression models with and without spatial household random effects. Structural equation models were used in modelling the complex relationships between multiple exposures and the outcome (child HIV/TB mortality) as relayed on a conceptual framework.

Results: Fifty-four of 6,692 children aged 1–5 years died of HIV/TB, from a total of 5,084 households. Maternal death had the greatest effect on child HIV/TB mortality (adjusted odds ratio =4.00; 95% confidence interval = 1.01–15.80). A protective effect was found in households with better socio-economic status and when the child was older. Spatial models disclosed that the areas which experienced the greatest child HIV/TB mortality were those without any health facility.

Conclusion: Low socio-economic status and maternal deaths impacted indirectly and directly on child mortality, respectively. These factors are major concerns locally and should be used in formulating interventions to reduce child mortality. Spatial prediction maps can further guide policy makers to target interventions where they are most needed.

Keywords: *HIV/TB*; *child mortality*; *conceptual framework*; *spatial analysis*; *pathway analysis*; *South Africa*

Received: 2 August 2012; Revised: 25 October 2012; Accepted: 26 October 2012; Published: xxxx

Tuberculosis (TB) and the HIV epidemic are among the top public health research priorities especially in sub-Saharan Africa. The sub-Saharan African region, which constitutes 10% of the world's population, is home to more than 67% (22.9 million) of people living with HIV/AIDS (1). TB, which is preventable and curable, was responsible for 25% of HIV-related deaths in 2009 and infected 33% of those living with HIV. As a result, the World Health Organization (WHO) aims to reduce TB-related deaths by 1 million by 2015 (2). South Africa's population is about 7% [50.6 million (3)] of the sub-Saharan African population, yet accounts for

Glob Health Action 2013. © 2013 Eustasius Musenge et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

24.6% (5.63 million of the 22.9 million) of people living with HIV and an estimated 67% of these have TB. South Africa ranks as the country with the highest co-infection rates (1).

Children are affected in two ways: a very high proportion of child deaths (60% in the year 2000) are attributed to the HIV epidemic (4), and approximately 1.9 million children have been orphaned by the HIV epidemic in South Africa (1). The nation at the onset of the 21st century had an estimated 70,000 children [about 25% rate of mother-to-child-transmissions (MTCT)] newly infected, whereas 10 years later only 10,000 [3.5% rate of MTCT (5)], which is indicative of an improved and more effective mother-to-child prevention program (6). In public health, determining patterns of disease occurrence provides the first steps toward increased understanding of determinants and potentially greater control of the disease (7). Understanding the associated factors of child HIV/TB mortality and how these are geographically distributed is key for effective interventions.

Several determinants of child mortality have been reported over the past three decades. In the 1980s, researchers explored causes of child mortality including medical, socio-economic, and demographic (8). In the 1990s, fertility behaviour, nutritional status, breastfeeding and infant feeding, the use of health services by mothers and children, ecological variables, and socio-economic status were studied (9). In the 21st century, the focus has been on HIV exposure, poor maternal health, inadequate infant care, increased exposure to infections, deaths of parents or caregivers, immune system abnormalities, poor nutrition, reduced breastfeeding and antiretroviral treatment exposure (10). These determinants are interrelated and can be combined into a conceptual framework. This combined with spatial aspects forms social ecological models, which are a component of the fast developing systems science that examines multiple effects and interrelatedness of social elements in an environment (11). Analysis of such frameworks can now be handled with the aid of structural equation models also known as pathway analysis.

The motivation for spatial analysis in epidemiological modelling is based on the notion that people living in a household and those in close proximity share similar exposures, which impact on the observed outcomes. In HIV/TB epidemiological research, the focus has largely been on drug development, societal integration, personal nutrition, physical health, and care giving (8–10, 12). Country-wide or region-specific pooled results are often reported, disregarding geographical confounding that commonly exists. Little has been done to interlink factors associated with child HIV/TB mortality or model their impact on areas of residence (households) (13). The few studies implementing spatial analysis have looked across different infectious diseases (14–16), used different study designs (cross-sectional and longitudinal) (17, 18), and considered varying levels of analysis (village, household, and individual) (16, 18, 19). As this is an emerging field, few studies have focused on HIV/TB cause-specific mortality assessing geographical differentials in South Africa over time (19).

The primary aim of this paper was to investigate the cause-specific (HIV/TB) mortality determinants and their spatial (geographical) distribution in the Agincourt sub-district of rural northeast South Africa for children aged 1–5 years in 2004. Our secondary aim was to model how the associated factors were interrelated as either direct (underlying) or indirect (proximal) factors of child HIV/TB mortality using pathway analysis based on the Mosley-Chen conceptual framework that incorporates social, biological, and ecological determinants of child mortality (8).

Methods

Agincourt sub-district data and study design

The Agincourt health and socio-demographic surveillance system was established in 1992 in a remote rural sub-district in the Bushbuckridge Municipality of Mpumalanga Province, northeast South Africa near the Mozambique border (20). There are annual census updates that collect socio-demographic information including in-migration and out-migration, mortality, fertility, health, and behavioural data (20).

In this study, we extracted data for 2004 describing a cross-sectional study of households in the study site. In that year, the Agincourt health and socio-demographic surveillance system covered a sub-district population of over 70,000 persons living in approximately 12,000 households scattered throughout 21 neighbouring villages. There were several health facilities in the area as shown in Fig. 1.

Cause of death data were obtained through verbal autopsies conducted on every recorded death during the year (21). Trained lay field workers conducted interviews within 1–11 months after a death, with the closest caregiver of the deceased in his/her mother tongue. Two medical practitioners independently determined cause of death; a third diagnosis was sought to resolve discrepancies. Their consensus cause of death was classified according to the World Health Organization's International Classification of Diseases 10 (ICD10) (22). The reported signs and symptoms ascertained HIV/TB mortality in children, and in some instances this was substantiated by the mother's cause of death (22). Over 90% of the health and socio-demographic surveillance system (HDSS) households were geo-coded, thus enabling spatial analyses at household level.

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

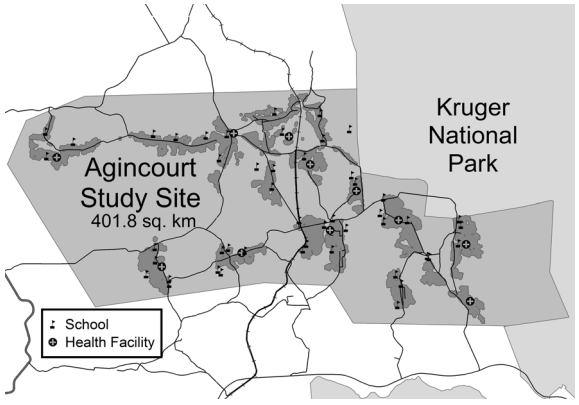


Fig. 1. Agincourt sub-district in 2004 showing villages (grey shading) schools (flags), and health facilities (circled grey crosses).

Dependent and independent variables

The study sample included all children aged between 1 and less than 5 years who were registered in the Agincourt HDSS between January and December 2004. The independent variables used were child's gender, nationality and age; mother's age, parity and death; gender of household head, cumulative deaths in household, number of household dwellers, minimum distance to health facility, and household socio-economic status. The latitude and longitude of each household were used to construct the variables for the spatial correlation (random effects), based on the straight line (Euclidean) distances between all pairs of households. The dependent variable was death due to HIV and tuberculosis (TB) determined by the ICD10 codes A16–A19[1] for TB and B20–B24[2] for HIV. The verbal autopsy cannot reliably distinguish HIV and TB due to low sensitivity and specificity; thus, these were combined in our analysis (23).

Conceptual framework: determinants of child mortality

As a way of interrelating child mortality determinants, we formulated a hybrid conceptual framework for child mortality determinants and disease ecology as proposed by Mosley-Chen and Meade (8, 24). This approach integrates ecological, social, and biological variables in the analysis, with the idea that these determinants operate through a common set of either 'proximate' (intermediate or indirect) or 'underlying' (direct) variables that impact on HIV/TB child mortality (8, 25–27). Variables are either endogenous (outcome variables) or exogenous (explanatory) or both, and these are modelled using structural equation models. Details of the structural equation model are given in Appendix 1.

We used the observed variables: child specific, mother specific, household related; and unobserved (latent) variables: socio-economic quintiles and spatial (ecological) variables. We tested the pathways shown in Fig. 2, to measure the direct and indirect determinants of child mortality for the Agincourt sub-district in 2004. Two approaches that complemented each other were used in our modelling. Firstly, we modelled the HIV/TB mortality determinants for children aged 1–5 years in 2004 and adjusted for their spatial random effects. In relation to our Fig. 2, we estimated the direct effects pathways (red arrows) from all categories of the explanatory (exogenous) variables controlling for confounders amongst them and their spatial or geographical effects (orange arrows). Results for these are presented in Table 1 and Fig. 3. Lastly, using the significant variables as outcomes (endogenous), we determined the indirect effects (proximal denoted by purple arrows) within and between categories to establish those factors impacting on child HIV/TB mortality.

Statistical methodology

We modelled the child HIV/TB mortality data to find associated factors and determine geographic disparities. Statistical techniques take into account spatial correlations by introducing spatially structured random effects into the model. These are usually based on independent observations. However, with multilevel correlated structures commonly used, maximum likelihood estimation (MLE) methods often underestimate the standard error; thus, the statistical significance is overestimated (28). Analysis was carried out using a geospatial logit regression model, which catered for individual- and household-level random effects (29). Inference was done using classical MLE and Bayesian inference techniques to fit the non-spatial and spatial models, respectively. Bayesian estimation of the many parameters was done using Markov Chain Monte Carlo (MCMC) simulations with the aid of Metropolis Hastings (MH) steps (29). Geospatial prediction maps showing the geographical coordinates, adjusting for child HIV/TB mortality determinants, were used in the analysis.

Data extraction was carried out using Structured Query Language (SQL), data management was done in STATA 10.0 (30), and structural equation modelling was performed in STATA 12.1 (Stata Corporation, College Station, Texas, USA). The Bayesian models were fit using software available in the public domain: BayesX version 2.1 for the spatial modelling and R-cran version 2.12.2 for geospatial mapping using the analysis and mapping

[1]A16 = Respiratory tuberculosis; not confirmed bacteriologically or histologically; A17 = Tuberculosis of nervous system; A18 = Tuberculosis of other organs; A19 = Miliary tuberculosis.
[2]B20 = Human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21 = Human immunodeficiency virus (HIV) disease resulting in malignant neoplasms; B22 = Human immunodeficiency virus (HIV) disease resulting in other specified diseases; B23 = Human immunodeficiency virus (HIV) disease resulting in other conditions; B24 = Unspecified human immunodeficiency virus (HIV) disease.

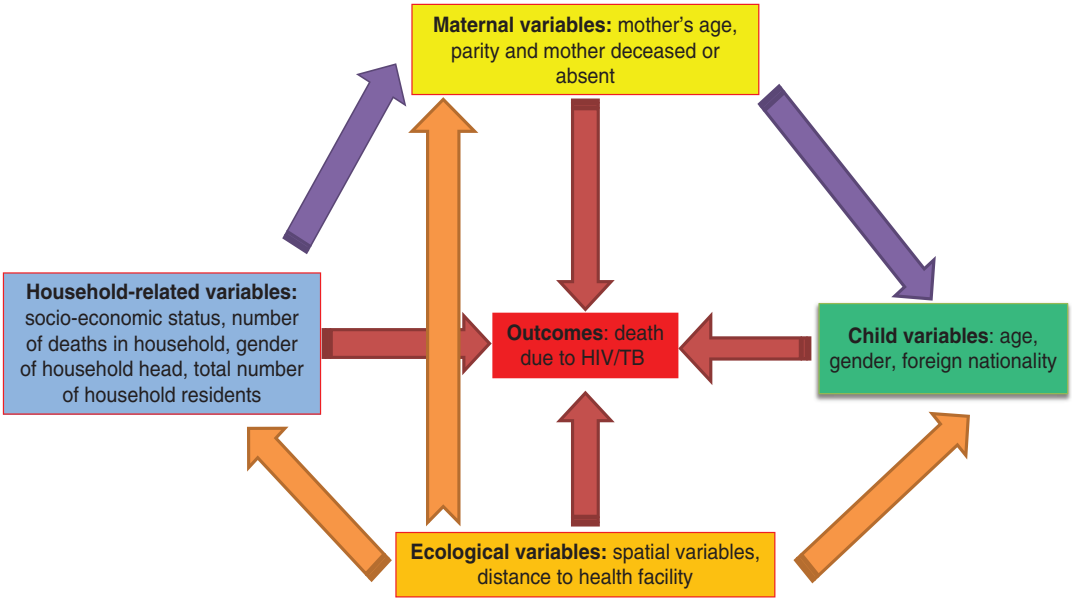


Fig. 2. Conceptual framework for modelling childhood HIV/TB-related mortality [adapted from Meade (24)]. The red arrows reflect a direct effect on mortality and the purple (explanatory variables) and orange (spatial random effects) arrows depict an indirect effect.

packages, maptools, maps, spBayes, and Multilevel B-spline Approximation (MBA) (31, 32).

Ethical clearance and informed consent

While this study used secondary data, the original study on which it was based was careful to consider ethical issues. Verbal informed consent was obtained when the Agincourt health and socio-demographic update rounds were conducted and also when verbal autopsy data were collected from a close relative of the deceased. To avoid emotionally stressing the interviewee, a culturally accepted period of 1 month was observed before the interview was conducted. Risks of stigmatisation were minimised through special training of fieldworkers on the importance of confidentiality. This work was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (No. 960720 and M081145).

Results

Of the 6,692 children between 1 and less than 5 years, with a mean age of 2.8 years, 54 HIV/TB-related deaths were recorded in the HDSS in 2004. There were 5,084 fully geo-coded households that recorded children either living or dead. Just over half the children were female (51%), and more than a third (36%) were born to former Mozambican refugees. The mean age of mothers was 29 years, and the mean distance to the nearest health facility was 2.33 km. Table 1 shows the results for the descriptive, univariate, and multivariable logit regression analyses. Maternal death had the greatest effect on child HIV/TB mortality [logit = 1.386 (0.707)], equivalent to an adjusted odds ratio (AOR) of 4.00 [95% Bayesian credible interval (BCI) = 1.01–15.80] controlling for spatial correlation and keeping all other variables constant.

Multivariable analysis showed three further significant predictors of child mortality in 2004, which were child, maternal, and household specific. Adjusting for spatial correlation and keeping all other variables constant, there were several significant determinants of child HIV/TB mortality. Child-specific differentials showed that boys were almost twice as likely to die compared to girls [logit = 0.641 (0.291); AOR = 1.89; 95% BCI = 1.07–3.36]. As the children got older, there was a protective effect against death; for every year the child aged, the likelihood of death decreased by about 50% [logit = −0.630 (0.137); AOR = 0.53; 95% BCI 0.40–0.70]. Male-headed households had a 60% lower risk of child HIV/TB mortality compared to female-headed households [logit = −0.895 (0.30); AOR = 0.41; 95% BCI 0.23–0.74].

Spatial disparities can be viewed on maps, which can show unexpected relationships that may have been overlooked in standard regression analysis. We obtained household-specific posterior odds ratios, which were adjusted for child-, maternal-, and household-related variables in the multivariable models. One main hotspot emerged with high mortality in the north, and there were two southerly areas with low mortality. The odds ratio posterior map in Fig. 3 shows the hotspot area between longitude 20° and 25° and latitude −2,750° and −2,745°. This area is characterised by the following: higher rate of maternal deaths, more male child mortality, lack of a public or private health facility, and being furthest from any health facility as shown in Fig. 1.

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

Table 1. Child HIV/TB mortality: descriptive statistics and results of two multiple logit regression analyses, Agincourt sub-district 2004

Variable	Summary number (%)	Univariate logit coefficient (standard error)	Non-spatial multiple logit coefficient (standard error)	Spatial Multiple logit coefficient (standard error)
Age of child in years[1]	2.77±1.20	−0.617 (0.131)**	−0.601 (0.146)**	−0.630 (0.137)**
Gender				
Female	3,379 (50.52)	0	0	0
Male	3,310 (49.48)	0.556 (0.283)*	0.698 (0.303)*	0.641 (0.291)*
Child's parents former refugee				
Yes	2,388 (35.71)	0		
No	4,299 (64.29)	−0.137 (0.280)		
Mother's age in years[1]	28.70±7.59	−0.003 (0.018)		
Cumulative household deaths[2]	0.68±0.96 (0–6)	0.684 (0.086)**		
Mother deceased by 2004				
No	6,634 (99.13)	0	0	0
Yes	58 (0.87)	1.508 (0.732)*	1.504 (0.867)*	1.386 (0.707)*
Minimum distance to health facility[1]	2.33±1.61	0.078 (0.083)		
Household size[1]	4.88±2.92	−0.011 (0.048)		
Gender of household-head				
Female	2,434 (36.37)	0	0	0
Male	4,258 (63.63)	−0.865 (0.418)**	−0.880 (0.302)**	−0.895 (0.30)**
Parity[2]	2.12±1.43 (1–10)	0.085 (0.090)		
Socio-economic status quintiles				
Most poor	1,823 (27.24)	0		
Very Poor	1,073 (16.03)	−0.112 (0.393)		
Moderately poor	1,120 (16.74)	−0.516 (0.444)		
Poor	1,316 (19.67)	−0.425 (0.406)		
Least poor	1,360 (20.32)	−0.458 (0.406)		
Non-spatial variance			0.10; 95% BCI (0.001; 0.68)	0.04; 95% BCI (0.001–0.40)
Spatial variance estimate				0.04; 95% CI (0.001–0.28)
Deviance information criteria			457.26±15.56	555.87±11.65

*$p < 0.05$; **$p < 0.01$.

[1]mean ± standard deviation, [2]mean ± standard deviation (range). 95% BCI = 95% Bayesian credible interval.

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

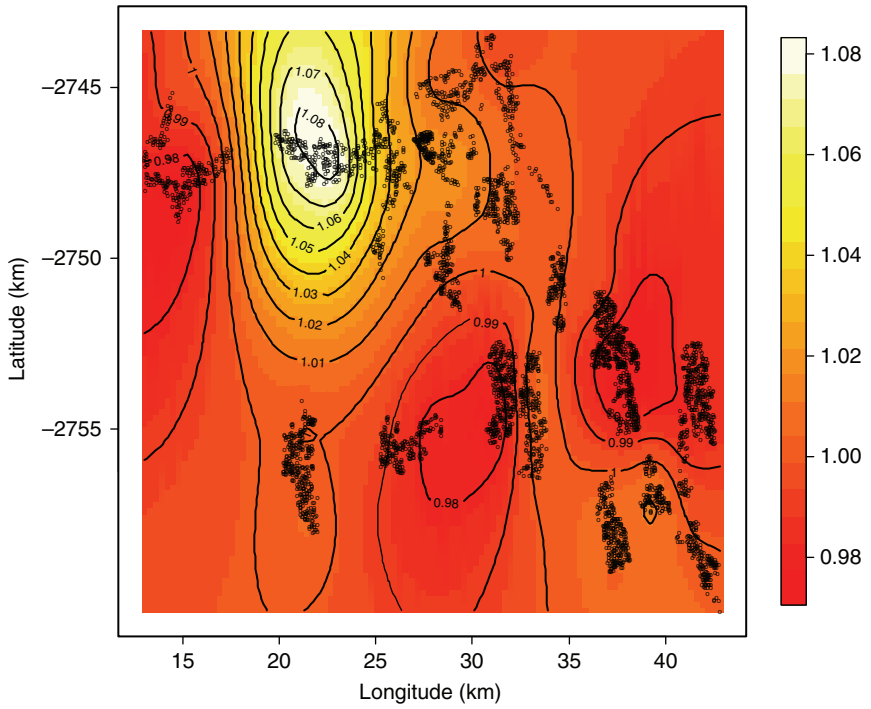


Fig. 3. Child HIV/TB mortality: posterior adjusted odds ratio map, Agincourt sub-district, 2004.

The results from the conceptual framework analysis of child HIV/TB mortality are shown in Fig. 4 and Table 2. All of the connected arrows show pathways that were statistically significant whether directly or indirectly. Being male (logit $=5.2\times10^{-3}$) and older child age (logit $=-4.6\times10^{-3}$) both impacted directly on HIV/TB mortality controlling for potential confounders related to the maternal and households variables. Upon treating male household headship as an exogenous variable, this directly impacted negatively on child HIV/TB deaths (logit $=-7.6\times10^{-3}$), adjusting for household variables. Moreover, when male headship was an endogenous variable, the following household variables indirectly impacted mortality: better socio-economic status and larger household size had a protective effect (logit $=7.7\times10^{-3}$) and increase in household deaths resulted in greater child HIV/TB mortality (logit $=-8.3\times10^{-2}$).

Table 2 details the direct pathways for child HIV/TB deaths (arrows directly linked to the red square) and the indirect pathways showing all possible routes for child HIV/TB deaths as shown in Fig. 4. In our analysis, poverty indirectly impacted on HIV/TB child mortality as conferred by two pathways: first, male headship of households (protective), and second, refugee status (a risk) controlling for child's age (as shown on Table 2 foot note). The greatest direct impact on child HIV/TB mortality was maternal death (logit $=2.7\times10^{-2}$), adjusting for the indirect effects of parity (logit $=-2.6\times10^{-3}$) and mother's age (logit $=6.8\times10^{-4}$).

Discussion

In this study, we showed that HIV/TB child mortality is influenced by a combination of child, maternal, household, and spatial determinants. Over the years 2002–2005, HIV/TB alone accounted for 34% of all-cause mortality among children living in the area (23). While being male was a risk factor for HIV/TB mortality as children got older, irrespective of sex they were less likely to die. A major finding in our study was that maternal death had the greatest direct impact on HIV/TB child deaths. A comparable study in the same area, adjusting for spatial random effects at village level, showed that infants whose mothers had died (presumptively of HIV) were at greater risk of death during the first 12 months of life (16).

The main source of TB transmission among children are adults, as shown by a study on a high-risk spatial cluster in South Africa with an incidence three times higher than the WHO levels (17). The most likely cause of HIV infection was mother-to-child transmission. Prevention of mother-to-child transmission (PMTCT), using a single-dose of nevirapine (sdNVP) administered to the child and mother, was found to reduce child mortality by about 9% in South Africa (33). Between 2004 and 2012 the guidelines for PMTCT and ART eligibility have changed. There is also a seven-fold lower (from 25 to 3.5%) rate of infection from mother to child over the period 2000–2010 (5). Kimani-Murage et al. (2009) found that knowledge of child HIV status influenced the manner of care-giving as well as acceptance and up-take

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

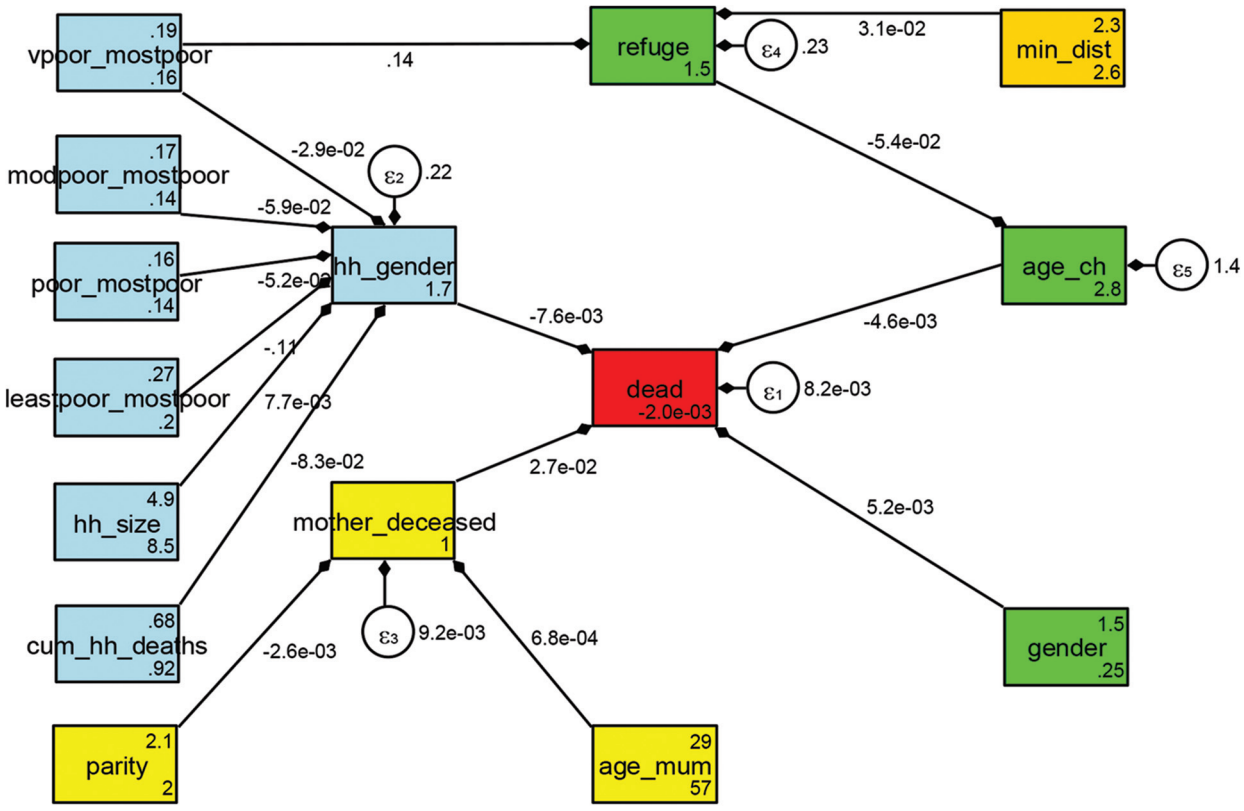


Fig. 4. Conceptual framework pathways results for Agincourt sub-district child mortality determinants in 2004. Key for variable names: very poor versus most poor (**vpoor_mostpoor**), moderately poor versus most poor (**modpoor_mostpoor**), least poor versus most poor (**leastpoor_mostpoor**), poor versus most poor (**poor_mostpoor**), age of child in years (**age_ch**), household size (**hh_size**), cumulative household deaths (**cum_hh_deaths**), gender of household head (**hh_gender**), mother age in years (**age_mum**), mother deceased by 2004 (**mother_deceased**), child born to former refugee parents (**refuge**), gender (**gender**), minimum distance to health facility (**min_dist**) parity (**parity**), child deceased in 2004 (**dead**). Other: the arrows pointing from the exogenous (explanatory) to endogenous (dependent) variables and the error terms (*ε*) placed on all five endogenous variables.

of community-wide interventions (31). The mother's survival significantly increases the likelihood of the child's survival, a finding that strengthens the current policy focus on maternal health. In South Africa, a comprehensive three stage process has been set up to improve maternal, foetal, infant, and ultimately child survival through antenatal care, labour and delivery, and postnatal care (34). Reduction of TB morbidity and mortality is the focus of the WHO 2015 goal of zero new TB infections and zero TB-related deaths (2). The impact of PMTCT aided by sdNVP, ziduvudine (AZT) and ART has also been demonstrated from evidence gathered from another site in rural South Africa (35).

In Agincourt, female-headed households had a higher child mortality compared to those headed by males. Studies in Tanzania, Zambia, and Limpopo province South Africa showed similar results (36–38). The pathway analysis (Fig. 4) demonstrates that male-headship does not independently lead to lower mortality but is a proxy for better socio-economic status of households. Main breadwinners are usually male with females having fewer opportunities in the labour market and consequently earning less. The absence from the household of a male due to death or permanent out-migration tended to impact negatively on household socio-economic status, leading to 90% greater child mortality odds (39). In other African countries, similar patterns were observed as men temporarily migrate within a country or to neighbouring countries due to either pull or push factors (40–42). HIV/TB deaths have lowered the life expectancy of working men from the area who often "return home to die" from their jobs (21). A recent qualitative study found that a father's unemployment, inability to work, and reluctance to provide support were common factors in HIV-infected households (31). From our pathway analysis (Table 2 and Fig. 4), we observed an indirect impact on child HIV/TB mortality: as the number of household deaths increased, the risk of child deaths was higher. This suggests that households with a greater burden of mortality are more vulnerable socio-economically through loss of

Table 2. Direct, indirect and total effects of child HIV/TB mortality in 2004, Agincourt sub-district

	Direct effects on the deceased child as shown in Fig. 2 conceptual framework						
	Gender of household-head	Child a former refugee	Child deceased	Mother deceased by 2004	Age of child in years	Indirect effects on child HIV/TB death	Total effects on child HIV/TB death
Very poor vs most poor	−0.029 (0.019)	0.140 (0.015)***				2.6×10^{-4} $(1.6 \times 10^{-4})^{ns!}$	2.6×10^{-4} $(1.6 \times 10^{-4})^{ns}$
Moderately poor vs most poor	−0.059 (0.020)***					4.5×10^{-4} $(2.1 \times 10^{-4})^{**}$	4.5×10^{-4} $(2.1 \times 10^{-4})^{**}$
Poor vs most poor	−0.052 (0.020)***					4.0×10^{-4} $(2.0 \times 10^{-4})^{**}$	4.0×10^{-4} $(2.0 \times 10^{-4})^{**}$
Least poor vs most poor	−0.111 (0.018)***					8.4×10^{-4} $(3.0 \times 10^{-4})^{***}$	8.4×10^{-4} $(3.0 \times 10^{-4})^{***}$
Household size vs most poor	0.008 (0.002)***					-5.9×10^{-5} $(2.5 \times 10^{-5})^{**}$	-5.9×10^{-5} $(2.5 \times 10^{-5})^{**}$
Cumulative household deaths	−0.083 (0.006)***					6.4×10^{-4} $(2.1 \times 10^{-4})^{***}$	6.4×10^{-4} $(2.1 \times 10^{-4})^{***}$
Minimum distance to health facility		0.031 (0.004)***				7.72×10^{-6} $(4.88 \times 10^{-6})^{ns}$	7.72×10^{-6} $(4.88 \times 10^{-6})^{ns}$
Male household-head			−0.0076 (0.002)***				−0.0076 (0.002)***
Mother deceased by 2004			0.027 (0.012)*				0.027 (0.012)**
Age of child in years			−0.0046 (0.001)***				−0.0046 (0.001)***
Gender			0.0052 (0.0023)**				0.0052 (0.0023)**
Mother's age (years)				0.001 (0.000)***		1.9×10^{-5} $(9.99 \times 10^{-6})^{*}$	1.9 ×10–05 $(9.99 \times 10^{-6})^{*}$
Parity				−0.003 (0.001)**		-7.0×10^{-5} $(4.0 \times 10^{-5})^{ns}$	−7.0 ×10–05 $(4.0 \times 10^{-5})^{ns}$
Parents former refugee					−0.054 (0.032)*	-2.5×10^{-4} $(1.5 \times 10^{-4})^{*}$	-2.5×10^{-4} $(1.5 \times 10^{-4})^{*}$
Intercept	1.719 (0.017)***	1.537 (0.011)***	−0.002 (0.014)	0.995 (0.005)***	2.814 (0.054)***		

*p <0.10; **p <0.05; ***p <0.01; ns: not significant; !: indirect effects computed as the product along the related pathways, i.e $-0.029 \times -0.0076 + 0.14 \times -0.054 \times -0.0046 = 2.55 \times 10^{-4} \approx 2.6 \times 10^{-4}$.

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

breadwinners and high health care and funeral costs. This compromises a family's ability to feed and care for all children, including those who are HIV positive. We also observed that former Mozambicans are at greater risk of child HIV/TB deaths, indicating the need for health care interventions that target these marginalised and vulnerable households.

Strengths and limitations of the study

The strengths and limitations of this paper are discussed in four categories: classification of exposures and outcomes; local or global relevance of results; confounding and effect modification; and analysis deviance.

A major strength in our observed variables is that the data are consistently updated, corrected, and checked for inconsistencies. Nevertheless, there are potentially two variable-related biases, namely selection and recall bias. Selection bias may occur if permanent out-migrants differ from residents or temporary migrants, which results in incomparable groups. Recall bias on the outcome variable mortality is unlikely as death is a major household event. However, it may affect the quality of the verbal autopsy interview, conducted at least 1 month but up to 11 months after death, and hence the ability of medical experts to assign a likely cause of death.

Our study focused on one sub-district only, but demonstrated geographical differentials in child HIV/TB mortality. There was one main hotspot with highest mortality and a visible downward trend in mortality moving southwards. These patterns are comparable to those found in other studies carried out in the area (16, 19). While many rural areas are not geo-coded to enable direct comparison of trends and patterns, our findings reflect similar patterns in other rural settlements at least in northeast South Africa.

To deal with confounding variables, we used multiple regression and added the spatial random effects in our modelling and also performed pathway analysis. Mediation was tested for in the pathway analysis as we determined factors directly and indirectly related to mortality and malnutrition. However, we were unable to model time varying covariates as this was a cross-sectional study design.

The robustness of our analysis procedures was demonstrated in that after catering for the spatial household random effects, we obtained narrower confidence bands compared to the non-spatial analogue. Therefore, catering for the spatial random effects further improves the model and parameter estimation compared to classical statistical models. Bayesian and pathway analysis combined were able to handle the complex structure of hierarchical data from the Agincourt health and socio-demographic surveillance system. Other advantages of Bayesian modelling include the ability to adjust for spatial confounding, and the ability to handle many parameters and inclusion of prior-known information on the distribution of data.

Conclusion

The growing field of spatio-temporal epidemiology brings in person, place, and time aspects in the risk factor analysis. Spatial Bayesian models provide better-fitting models compared to non-spatial models. This procedure also enables the drawing of risk maps, which can be used by policy makers to target and develop community-based interventions.

Going forward, further studies investigating trends over time using spatio-temporal Bayesian analysis will strengthen the evidence on associated factors and trends found in this study and demonstrate the value of this analytic approach. Studies comparing deaths due to other causes with those for HIV/TB in children would add policy value. Similarly, adopting a secondary analysis of a cohort study design would have enabled temporality to be investigated in greater detail and survival analysis to be undertaken. Moreover, a combination of qualitative and quantitative analysis methods is needed to fully understand the proximal and distal factors that influence child HIV/TB mortality. Findings that can be used for policy nationally would require sentinel sites located strategically across the country.

Acknowledgements

This work was made possible by doctoral fellowship support from the World Health Organization, Tropical Disease Research (WHO/TDR), and Swiss–South Africa Joint Research Programme. The MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt) is funded by the Wellcome Trust, UK, (grants 058893/Z/99/A, 069683/Z/02/Z and 085477/Z/08/Z), the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew W Mellon and William and Flora Hewlett Foundations, USA. Special thanks goes to Benn Sartorius for data extraction, David Rees for his valuable review of an early draft, community leaders, field workers, supervisors, and the Agincourt community.

Conflict of interest and funding

The authors declare that they have no competing interests.

References

1. UNAIDS. Report on the global AIDS epidemic 2011. Available from: http://www.avert.org/africa-hiv-aids-statistics.htm [cited 23 October 2012].
2. WHO. One million lives can be saved between now and the end of 2015 by preventing and treating tuberculosis among people living with HIV. Available from: http://www.who.int/tb/challenges/hiv/fastfactsbrochure.pdf [cited 23 October 2012].
3. Statistics South Africa (StatsSA). Statistical release P0302 2011. Available from: http://www.statssa.gov.za/publications/P0302/P03022011.pdf [cited 23 October 2012].

4. Bradshaw D, Bourne D, Nannan N. What are the leading causes of death among South African children. MRC Policy Brief & Unicef 2003; 3: 1–4.
5. Doherty T, Sanders D, Goga A, Jackson D. Implications of the new WHO guidelines on HIV and infant feeding for child survival in South Africa. Bull World Health Organ 2011; 89: 62–7.
6. UNICEF. State of the World's Children year 2009. Available from: http://www.unicef.org/publications/index_47127.html [cited 23 October 2012].
7. Waller LA, Gotway CA. Applied spatial statistics for public health data. Atlanta, GA: Wiley; 2004.
8. Mosley WH, Chen LC. An analytical framework for the study of child survival in developing countries. Popul Dev Rev 1984; 10: 25–45.
9. Rutstein SO. Factors associated with trends in infant and child mortality in developing countries in the 1990s. Bull World Health Organ 2000; 78: 1256–70.
10. Filteau S. The HIV-exposed, uninfected African child. Trop Med Int Health 2009; 14: 276–87.
11. Guerrero L, La Valley A. Conflict, emotion, and communication. In: Oetzel JG, Ting-Toomey S, eds. The SAGE handbook of conflict communication: integrating theory, research, and practice. Thousand Oaks, CA: SAGE; 2006 pp. 69–96.
12. Hargreaves J, Collinson M, Kahn K, Clark S, Tollman S. Childhood mortality among former Mozambican refugees and their hosts in rural South Africa. Int J Epidemiol 2004; 33: 1271–8.
13. Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. J Roy Stat Soc Series A (Statistics in Society) 1991; 154: 421–41.
14. Gosoniu L, Vounatsou P, Tami A, Nathan R, Grundmann H, Lengeler C. Spatial effects of mosquito bednets on child mortality. BMC Publ Health 2008; 8: 356.
15. Moonan PK, Bayona M, Quitugua TN, Oppong J, Dunbar D, Jost KC, et al. Using GIS technology to identify areas of tuberculosis transmission and incidence. Int J Health Geogr 2004; 3: 23.
16. Sartorius B, Kahn K, Vounatsou P, Collinson M, Tollman S. Young and vulnerable: spatial-temporal trends and risk factors for infant mortality in rural South Africa (Agincourt), 1992–2007. BMC Publ Health 2010; 10: 645.
17. Middelkoop K, Bekker LG, Morrow C, Zwane E, Wood R. Childhood tuberculosis infection and disease: a spatial and temporal transmission analysis in a South African township. S Afr Med J 2009; 99: 738–43.
18. Tanser F, LeSueur D, Solarsh G, Wilkinson D. HIV heterogeneity and proximity of homestead to roads in rural South Africa: an exploration using a geographical information system. Trop Med Int Health 2000; 5: 40–6.
19. Musenge E, Vounatsou P, Kahn K. Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spat Spatio-temporal Epidemiol 2011; 2: 205–17.
20. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt Health and Demographic Surveillance System. Scand J Publ Health 2007; 35: 8–20.
21. Clark SJ, Collinson MA, Kahn K, Drullinger K, Tollman SM. Returning home to die: circular labour migration and mortality in South Africa. Scand J Public Health Suppl 2007; 69: 35–44.
22. Kahn K, Tollman SM, Garenne M, Gear JS. Validation and application of verbal autopsies in a rural area of South Africa. Trop Med Int Health 2000; 5: 824–31.
23. Tollman SM, Kahn K, Sartorius B, Collinson MA, Clark SJ, Garenne ML. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. Lancet 2008; 372: 893–901.
24. Meade MS. Medical geography as human ecology: the dimension of population movement. Geogr Rev 1977; 67: 379–93.
25. Lewis JJC, Donnelly CA, Mare P, Mupambireyi Z, Garnett GP, Gregson S. Evaluating the proximate determinants framework for HIV infection in rural Zimbabwe. Sex Transm Infect 2007; 83: i61–9.
26. Boerma J, Weir SS. Integrating demographic and epidemiological approaches to research on HIV/AIDS: the proximate-determinants framework. J Infec Dis 2005; 191: S61.
27. Lopman B, Nyamukapa C, Mushati P, Mupambireyi Z, Mason P, Garnett GP, et al. HIV incidence in 3 years of follow-up of a Zimbabwe cohort—1998–2000 to 2001–03: contributions of proximate and underlying determinants to transmission. Int J Epidemiol 2008; 37: 88–105.
28. Cressie NAC. Statistics for spatial data. New York: Wiley; 1993.
29. Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. Boca Raton, FL: Chapman & Hall/CRC; 2004.
30. Stata Corporation. Stata statistical software release 10. College Station, TX: Stata Corporation LP; 2007.
31. Kimani-Murage EW, Manderson L, Norris SA, Kahn K. 'You opened our eyes': care-giving after learning a child's positive HIV status in rural South Africa. Health Soc Care Community 2010; 18: 264–71.
32. Belitz C, Brezger A, Kneib T, Lang S, Fronk E, Heinzl F, et al. Bayesx-software for bayesian inference in structured additive regression models. 2011. Version 2.01. Available from: http://www.stat.uni-muenchen.de/~bayesx [cited 23 October 2012].
33. Coetzee D, Hilderbrand K, Boulle A, Draper B, Abdullah F, Goemaere E. Effectiveness of the first district-wide programme for the prevention of mother-to-child transmission of HIV in South Africa. Bull World Health Organ 2005; 83: 489–94.
34. National Department of Health. Clinical guidelines: prevention of mother-to-child transmission (PMTCT) 2010. Available from: http://wwwfidssacoza/images/PMTCT_Guidelinespdf [cited 23 October 2012].
35. Ndirangu J, Newell M-L, Tanser F, Herbst AJ, Bland R. Decline in early life mortality in a high HIV prevalence rural area of South Africa: evidence of HIV prevention or treatment impact? AIDS 2010; 24: 593–602.
36. Katapa R. A comparison of female-and male-headed households in Tanzania and poverty implications. J Biosoc Sci 2006; 38: 327.
37. Wiegers E, Curry J, Garbero A, Hourihan J. Patterns of vulnerability to AIDS impacts in Zambian households. Dev Change 2006; 37: 1073–92.
38. Oni S, Obi C, Okorie A, Thabede D, Jordan A. The economic impact of HIV/AIDS on rural households in Limpopo Province. S Afr J Econ 2002; 70: 551–62.
39. Collinson MA. Striving against adversity: the dynamics of migration, health and poverty in rural South Africa. Medical Dissertations, Umeå University, Umeå; 2009.
40. Posel D, Casale D. What has been happening to internal labour migration in South Africa, 1993–1999? S Afr J Econ 2003; 71: 455–79.
41. Hahn MB, Riederer AM, Foster SO. The livelihood vulnerability index: a pragmatic approach to assessing risks from

Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

climate variability and change--a case study in Mozambique. Global Environ Change 2009; 19: 74–88.
42. Beguy D, Bocquier P, Zulu EM. Circular migration patterns and determinants in Nairobi slum settlements. Demographic Res 2010; 23: 549–86.
43. Stata Corporation. Stata statistical software release 12. College Station, TX: Stata Corporation LP; 2011.
44. Kline RB. Principles and practice of structural equation modeling. 3rd ed. New York: Guilford Press; 2011.
45. Bollen KA. Structural equations with latent variables. New York: Wiley; 1989.
46. Bollen KA, Curran PJ. Latent curve models: a structural equation perspective. New York: Wiley-Interscience; 2006.

***Eustasius Musenge**
Division of Epidemiology and Biostatistics
School of Public Health
Faculty of Health Sciences
University of the Witwatersrand
7 York Road, Parktown
Johannesburg 2193
South Africa
Tel: +27 11 717 2610
Email: Eustasius.Musenge@wits.ac.za

Appendix 1

Structural Equation Model for paths analyses

The structural equation model was fit in this paper for given individual-level outcome (endogenous) variables y_i (observed; child HIV/TB deaths) and/or η_i (for latent endogenous) which then form the vectors of exogenous variables $\boldsymbol{y}$ and $\boldsymbol{\eta}$. The individual level explanatory variables x_j (observed; maternal variables, child variables, and household-related variables) and/or ξ_j (latent; spatial and household related-variables such as socio-economic status) make up the vectors of endogenous variables $\boldsymbol{x}$ and $\boldsymbol{\xi}$. These can be expressed in a multivariate equation as follows:

$$\boldsymbol{Y} = \boldsymbol{BY} + \boldsymbol{\Gamma X} + \boldsymbol{\alpha} + \boldsymbol{\zeta}$$

where $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{\eta} \end{pmatrix}$ is the vector of all endogenous variables with observed and latent variables $\boldsymbol{y}$ and $\boldsymbol{\eta}$, respectively, the vector of exogenous variables $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\xi} \end{pmatrix}$ with observed and latent variables $\boldsymbol{x}$ and $\boldsymbol{\xi}$, respectively, and the vector of error terms $\boldsymbol{\zeta} = \begin{pmatrix} e.\boldsymbol{y} \\ e.\boldsymbol{\eta} \end{pmatrix}$. $\boldsymbol{B} = [\beta_{ij}]$ is the matrix of coefficients on endogenous variables predicting other endogenous variables which were gender of household head, mother's death, former Mozambican, and age of child in years. $\boldsymbol{\Gamma} = [\gamma_{ij}]$ is the matrix of coefficients on exogenous variables which were parity at child's birth, mother's age, socio-economic status, total deaths in household by 2004, household size, and minimum distance to health facility in kilometres. $\boldsymbol{\alpha} = [\alpha_i]$ is the vector of intercepts from the endogenous variables, whereas $\boldsymbol{\zeta}$ is assumed to have mean zero and covariance $\text{cov}(\mathbf{X},\boldsymbol{\zeta}) = \mathbf{0}$ such that $\boldsymbol{\zeta} \sim MVN(\boldsymbol{0},\boldsymbol{\Sigma})$,

where $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} = Var(\boldsymbol{X}) & \boldsymbol{\Sigma}_{X\zeta} = \text{cov}(\boldsymbol{X},\boldsymbol{\zeta}) \\ \boldsymbol{\Sigma}'_{X\zeta} = \text{cov}(\boldsymbol{X},\boldsymbol{\zeta}) & \boldsymbol{\Sigma}_{\zeta\zeta} = Var(\boldsymbol{\zeta}) \end{pmatrix} = \begin{pmatrix} [\phi_{ij}] = \boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & [\Psi_{ij}] = \boldsymbol{\Psi} \end{pmatrix}$. Let $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix}$ be the vector of all endogenous and exogenous variables, then $\boldsymbol{Z} \sim MVN(\boldsymbol{\mu}_Z,\boldsymbol{\Sigma}_Z)$,

where $\boldsymbol{\mu}_Z = \begin{pmatrix} E(\boldsymbol{Y}) = \boldsymbol{\mu}_Y \\ E(\boldsymbol{X}) = \boldsymbol{\kappa} \end{pmatrix}$ and $\boldsymbol{\Sigma}_Z = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}'_{YX} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}'_{YX} & \boldsymbol{\Phi} \end{pmatrix}$ with expectation vectors $\boldsymbol{\mu}_Y = E[(\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma X}+\boldsymbol{\alpha}+\boldsymbol{\zeta})] = (\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma\kappa}+\boldsymbol{\alpha})$ and $\boldsymbol{\kappa} = [\kappa_j]$, variance matrix with variables:

$\boldsymbol{\Sigma}_{YY} = Var(\boldsymbol{Y}) = (\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi})\{(\boldsymbol{I}-\boldsymbol{B})^{-1}\}'$ and $\boldsymbol{\Sigma}_{YX} = \text{cov}(\boldsymbol{Y},\boldsymbol{X}) = (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma\Phi}$ (43–46). The parameter estimation was done using maximum likelihood in STATA 12.1 (Stata Corporation, College Station, Texas USA).

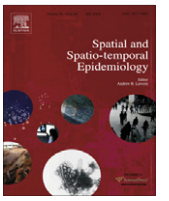
Citation: Glob Health Action 2013, **6**: 19266 - http://dx.doi.org/10.3402/gha.v6i0.19266

Spatial and Spatio-temporal Epidemiology 2 (2011) 205–217

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Review Article

Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa

Eustasius Musenge [a,b,*], Penelope Vounatsou [c], Kathleen Kahn [a,d,e]

[a] MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 7 York Road, Parktown 2193, Johannesburg, South Africa
[b] Biostatistics and Epidemiology Division, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 7 York Road, Parktown 2193, Johannesburg, South Africa
[c] Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland
[d] Centre for Global Health Research, Umeå University, Umeå, Sweden
[e] INDEPTH Network, Accra, Ghana

ARTICLE INFO

Article history:
Received 12 March 2011
Revised 9 June 2011
Accepted 6 July 2011
Available online 18 July 2011

Keywords:
Spatiotemporal confounding
HIV/TB
Zero inflated
Over-dispersed
Agincourt rural South Africa

ABSTRACT

South Africa is experiencing a major burden of HIV/TB. We used longitudinal data from the Agincourt sub-district in rural northeast South Africa over the years 2000 to 2005. A total of 187 HIV/TB deaths were observed among 16,844 children aged 1–5 years coming from 8,863 households. In this paper we used Bayesian models to assess risk factors for child HIV/TB mortality taking into account the presence of spatial correlation. Bayesian zero inflated spatiotemporal models were able to detect hidden patterns within the data. Our main finding was that maternal orphans experienced a threefold greater risk of HIV/TB death compared to those with living mothers (AHR = 2.93, 95% CI[1.29;6.93]). Risk factor analyses which adjust for person, place and time provide evidence for policy makers that includes a spatial distribution of risk. Child survival is dependent on the mother's survival; hence programs that promote maternal survival are critical.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The years 2000–2005 mark the middle of the United Nation's Millennium Development Goals (MDG) set out by the 192 member states and 23 international organizations (Sachs and McArthur, 2005). Goal 4.1 aims at reducing child (under five) mortality by two-thirds over the years 1990 and 2015. The sub-Saharan African region has the greatest burden of disease and highest mortality of HIV/TB. In 1990 the sub-Saharan Africa region had 184 child deaths per every 1000 live births, by the year 2006 27 of the countries in the region had made no progress in reducing child deaths (UNDG, 2008). Globally, 2.1 million children were living with HIV/TB in the year 2008 of which 90% were from the sub-Saharan region (UNAIDS, 2009). South Africa's HIV/TB orphans in 2008 were about 1.4 million, the highest absolute number globally (UNAIDS, 2008). In 1998, the year Demographic and Health Surveys were started in South Africa, there was an under five mortality rate (U5MR) of 59.4 deaths per 1000 live births, and aims to reduce this to 20 per 1000 by 2015 (UNESCO, 2007). The main cause of child deaths is HIV epidemic, contracted mainly through mother-to-child-transmissions (MTCT) with over 70,000 new infections yearly (UNICEF, 2009).

Public health priorities such as disease control intervention are made difficult in many developing countries due to lack of good data mainly due to poor resources, poor management and inadequate skills. The analyses of public health data often build from the statistical notion of each person having a risk or probability of contracting a disease.

* Corresponding author at: MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 7 York Road, Parktown 2193, Johannesburg, South Africa. Tel.: +27 11 717 2610; fax: +27 11 717 2084.
E-mail address: Eustasius.Musenge@wits.ac.za (E. Musenge).

1877-5845/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2011.07.001

The main goals in analyses are to identify and quantify any exposures, behaviours and characteristics that may modify an individuals' or a population's risk. However HIV/TB interventions are hampered by a complex interlink of social, behavioural, biological and spatial determinants.

Many articles have been published on HIV/TB mortality in Sub-Saharan Africa. The issues focused mainly on the impact of epidemics, assessing risk factors, measuring impact of interventions and identifying high risk clusters. A study done in rural Senegal showed a decreasing trend in child mortality over a 37 year period (Delaunay et al., 2001). Spatial analyses on 39 rural villages in Burkina Faso under the Health and socio-Demographic Surveillance System (HDSS) over a 9 year period, showed that village and distance from the health facility where among the risk factors for child mortality (Kynast-Wolf et al., 2002). A seasonal mortality trend was also disclosed by the same authors in another study of the same HDSS (Becher et al., 2004). A Tanzanian study used spatial modeling techniques and found that there is a need to adopt both a group-based and a place-based approach in identification of high-risk HIV groups for intervention (Msisha et al., 2008). In South Africa rural Kwazulu-Natal a study found that prevalence of HIV was associated with ethnicity, urban status and unemployment based on over 11,000 persons in 700 enumeration areas (Kleinschmidt et al., 2007). In north east South Africa the relative risk (RR) of HIV and TB has increased 10-fold over an 11 year period; reaching 32.3 from 3.2 for the period 2001–2003 compared to 1992–1994, (Kahn et al., 2007). Covariates that have mainly been used to model child mortality are: HIV exposure, poor maternal health, inadequate infant care, increased exposure to infections, increased mortality, immune system abnormalities, poor nutrition, reduced breastfeeding and antiretroviral exposure (Filteau, 2009). Methods applied in the analyses were mostly non-Bayesian based, which are limited in catering for large datasets, time and spatial random effects (Pacheco et al., 2008).

In epidemiological research, two study designs have been used for follow-up studies: time series studies, that estimate acute risk associated with short-term exposure and cohort studies, that estimate chronic effects associated with long term exposure. These approaches have been criticized for potential confounding by time varying covariates and individual risk factors or household level traits, respectively (Greven et al., 2009). Failure to account for extra-variation induced by temporal and spatial autocorrelation can lead to an understatement of the uncertainty of the risk of mortality (Burnett et al., 2001). Several approaches can be used to address confounding, at the design level – randomization, during the study – use of experts and validated instruments and during analyses – multivariable regression, standardization and stratification (Greenland, 1989). Bayesian and spatial statistical techniques can be used to evaluate differences in rates observed from different geographical locations, separate noise from patterns and identify disease clusters.

However, these methods have been rarely used in HIV/TB research to assess the diseases' impact on populations and their areas of residence. In this paper, we used Bayesian models to assess risk factors of child HIV/TB mortality taking into account the presence of spatial correlation. The modelling procedure adjusts for the household level spatial random effects for child HIV/TB mortality. Some related earlier work has looked at all-cause mortality adjusting for village (areal) level random effects (Sartorius et al., 2011, 2010). Our findings can aid in understanding geographically adjusted risk factors of child HIV/TB mortality in a rural South African cohort with sparse discrete outcomes.

2. Methods

2.1. Agincourt Health and Socio-demographic Surveillance System (HDSS) data

The Agincourt HDSS was set up in 1992 in a rural subdistrict in Bushbuckridge Municipality of Mpumalanga Province, northeast South Africa. The location was selected mainly due to availability of several clinics and its lack of infrastracture (Tollman, 1999). The Agincourt subdistrict had a population of over 70,000 persons, living in approximately 12,000 households scattered in 21 villages. The area covered 432 square kilometres, with 34.2 km as the maximum distance between two households (Fig. 1). Information at individual level (infants, children and adults) is collected once every year including vital events updates, that compile migration, health, demographic, mortality and fertility data (Kahn et al., 2007).

Special modules are run periodically such as socio-economic status and food security. Cause of death data are obtained through verbal autopsies conducted on every recorded death by trained lay fieldworkers (Kahn et al., 2000). The cause of death is independently determined by two or three medical doctors and classified according to the World Health Organization's International Classification of Diseases (ICD10). Every household is geo-coded thus enabling spatial analysis at household as well as at village level. Agincourt HDSS longitudinal data are correlated in time and space (spatiotemporal), since they are collected yearly on the same population. This makes it difficult to analyze the data due to their collinearity and also their large size.

2.2. Outcome and explanatory variables

The child mortality independent variables were selected in accordance with the approach proposed by Rutstein (2000) who mentions six main risk factors for child mortality: socio-economic status, nutritional status, use of health facilities, environmental health conditions, maternal factors and infant feeding (Rutstein, 2000). The persons included in the study were all the children aged between 1 and under 5 years who lived or had lived in the Agincourt HDSS between January 2000 and December 2005. The explanatory variables used were: child's gender, child's and mother's nationality, mother's age, parents' dead or alive, gender of household head, antenatal clinic visit, mother's parity at pregnancy and birth, cumulative deaths in household, number of household dwellers, number of deaths in household and mother's total live

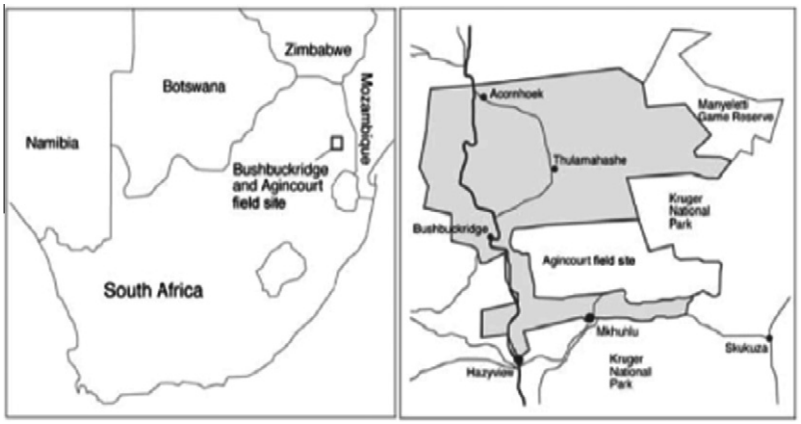


Fig. 1. Agincourt HDSS location North-East of South Africa. Source Kahn et al., 2000.

births. The individual child's age was treated as an offset variable.

The outcome variable was death due to HIV and or tuberculosis (TB) determined by the WHOs ICD10 verbal autopsy codes A16–A19[1] for TB and B20–B24[2] for HIV. The outcome death was treated as a discrete (count) variable since this provides epidemiologically better estimates of prevalence and relative risk compared to the odds ratio which results from treating it as binary outcome (Barros and Hirakata, 2003; Fekedulegn et al., 2010). Treating the outcome as such allowed us to employ the Poisson and Negative Binomial which yields more epidemiologically plausible relative risks for these cohort data. These models which are members of the exponential family have a log link function which has less identifiability issues in the generalized linear model compared to the logit link function which results from treating the outcome as binary and using logistic regression techniques (Agarwal et al., 2002).

2.3. Spatiotemporal zero inflated Poisson and Negative Binomial models

Agincourt child HIV deaths data had several problems that made it difficult to use standard statistical procedures: over dispersion, caused by unobserved heterogeneity or spatio-temporal correlation and data collected over large numbers of locations with zero deaths. Conventional statistical methods applied to spatiotemporal data often underestimate the standard error and thus the statistical significance is overestimated (Cressie, 1993). Bayesian analyses of multilevel models for complex geostatistical data can be done with the aid of Markov Chain Monte Carlo (MCMC) parameter estimation procedure (Banerjee et al., 2004). Bayesian geostatistical models for temporal count data with large number of zeros have been proposed by Fernandes et al. (2009). We employ the above methodology combined with semi-parametrically structured geo-additive predictors, to identify the determinants of HIV/TB related mortality in Agincourt HDSS data.

Geo-additive spatiotemporal zero inflated Poisson (ZIP) as well as Negative Binomial (ZINB) models (see Appendix A) were used in the analysis. The models have many parameters and are hierarchical (multilevel), thus we resort to full Bayesian inference with the computationally efficient MCMC techniques. Multiple variable models were used to assess the effect of different covariates in the presence and absence of geographical heterogeneity (Gosoniu et al., 2008). We investigated the presence of over-dispersion (when $\delta \to \infty$ it is absent) and zero inflation (when $\theta = 0$, it is absent) (Fahrmeir and Echavarria, 2006). Spatial dependence $f_{spat}(\mathbf{s})$ was modelled by assuming that the random effects follow a zero mean stationary multivariate Gaussian random field (GRF) with spatial variance σ^2_{spat} and isotropic Matern correlation functions. For the exponential correlation which is a special case of the Matern function, the multivariate GRF $f_{spat}(\mathbf{s}) \sim MVN(\mathbf{0}, \sum)$ has variance–covariance matrix $\sum = \sigma^2_{spat} \exp[-\widehat{\rho} d_{ij}]$, where d_{ij} is the Euclidean distance between households i and j, σ^2_{spat} is the geographic (spatial) variability known as the sill and $\widehat{\rho} = max_{ij}\|s_i - s_j\|/0.001$ is the empirically estimated rate of spatial correlation decay (Brezger et al., 2005). Multivariate Gaussian random field (GRF) priors $f_{spat}(\mathbf{s}) = p(\beta^{spat}|\sigma^2_{spat}) \propto |\sum|^{-\frac{1}{2}} \exp\left(\frac{-1}{\sigma^2_{spat}}(\beta^{spat})'\sum^{-1}\beta^{spat}\right)$, where β^{spat} is the vector of household specific structured spatial coefficients, whereas the non-structured utilize the independent and identically distributed (i.i.d) Gaussian priors with Inverse Gamma (IG) hyper-priors (Brezger and Lang, 2006).

Bayesian based model fitting requires the inversion of the variance–covariance matrix ($\sum$) with the same size as the number of geo-locations (households). Due to the large number of observations in our dataset, the estimation of model parameters becomes unstable and unfeasible; this problem is popularly known as the "big N". We

[1] A16 = respiratory tuberculosis; not confirmed bacteriologically or histologically; A17 = tuberculosis of nervous system; A18 = tuberculosis of other organs; A19 = miliary tuberculosis.

[2] B20 = human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21 = human immunodeficiency virus (HIV) disease resulting in malignant neoplasms; B22 = human immunodeficiency virus (HIV) disease resulting in other specified diseases; B23 = human immunodeficiency virus (HIV) disease resulting in other conditions; B24 = unspecified human immunodeficiency virus (HIV) disease.

addressed this problem by using low rank kriging to approximate the stationary Gaussian random field (Kamman and Wand, 2001). Firstly the spatial correlation decay $\widehat{\rho}$ is estimated from a "representative" subsample obtained using the mini-max space filling criteria, hence only parameter $\sigma^2_{spat}(1/\tau^2_{spat})$ needs to be estimated. This reduces the computational burden since the spatial decay is fixed and the large Euclidean distance matrix is also constant, resulting in estimation of the spatial variance and parameter vector of household spatial effects.

Full Bayesian inference was done via MCMC simulation based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data. Gibbs sampling was used for closed conditionals and for the numerically intractable a slightly modified form of the Metropolis-Hastings (MH) algorithm based on iteratively weighted least squares (IWLS) (Casella and George, 1992; Banerjee et al., 2004). A total of 55,000 iterations on a single chain were run with the first 5000 discarded as burn-in and to reduce auto-correlation every 25th value was taken to form a posterior sample of 2000. MCMC convergence was assessed using diagnostic procedures proposed by Raftery and Lewis (1992), Geweke and Minneapolis (1991) and Heidelberger and Welch (1983). These convergence diagnostics techniques were done in the R-package, "convergence diagnosis and output analysis" (CODA) (Plummer et al., 2006). Additionally a graphical assessment was done using trace-plots, histogram density plots and auto-correlations of parameters and the summary statistics assessing the closeness of the mean and median posterior estimates (Nylander et al., 2008).

2.4. Goodness of fit and best model selection

The deviance information criteria (DIC) was used in determining goodness of fit of the proposed models, this can be defined as "classical estimate of fit, plus twice the effective number of parameters" (Spiegelhalter et al., 2002). This is a technique for Bayesian model selection for MCMC derived posteriors. The procedure works on the assumption that the posterior distribution is multivariate normal and the un-standardised deviance is given as: $D(\Omega) = -2\log(p(y|\Omega))$, where y are the data, Ω are the unknown parameters of the model, $p(y|\Omega)$ is the likelihood. After addition of the standardising constant $C = 2\log f(y)$ we get the Bayesian or saturated deviance $D(\Omega) = -2\log(p(y|\Omega)) + C$. The expectation of this deviance $\bar{D} = E^{\Omega}[D(\bar{\Omega})]$ measures how well the model fits the data, the smaller it is the better. The measure of model complexity or parsimony is the effective number of parameters; $p_D = \bar{D} - D(\bar{\Omega})$, where $\bar{\Omega}$ is the expectation of Ω the smaller it is the more parsimonious. From these measures we get the deviance information criteria; $\mathrm{DIC} = p_D + \bar{D}$ and upon substitution of $\bar{D}$ we get $\mathrm{DIC} = 2p_D + D(\bar{\Omega})$, which penalises on complexity and lack of fit. Selection of a better model based on the DIC basically favours models with the smaller values.

The data extraction was done using Structured Query Language (SQL). Preliminary data analyses, trend analyses, Kaplan-Meier survival graphs and data management were done using STATA 10.0 (StataCorp, 2007). The Bayesian MCMC analysis was done using a software package called BayesX (Belitz et al., 2009). This package utilises computationally efficient C++ algorithms (Press, 2002). The software ran on a quad core processor with 4.0 gigabytes of random access memory (RAM) within a Linux environment. The routines used to estimate the parameters of the models via MCMC which perform Gibbs sampling and Metropolis Hastings within Gibbs sampling. Multivariable analyses were performed to determine patterns for HIV/TB mortality adjusting for possible household confounders. Posterior risk maps were produced using mapping packages in R (R-cran, 2010).

2.5. Ethical clearance and informed consent

The Agincourt HDSS site was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (No. 960720). This work was also granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (M081145). Verbal informed consent was obtained when the census rounds were conducted and also when verbal autopsy data were collected from a close relative of the deceased.

3. Results

3.1. Exploratory analysis and descriptive statistics

The data used were for children aged 1–5 years residing in Agincourt HDSS from 2000 to 2005 whose geo-location data were available. These totalled 16,844 children from 8863 households, with a range of 138–835 households across the 21 villages. A total of 187 deaths were HIV/AIDS and Tuberculosis related from 59,448.15 person years yielding a 1–5 years mortality rate (1–4MR) of 3.15 deaths per 1000 person years. There were slightly more females than males (51% vs 49%) and 37% were formerly Mozambican in origin. Mothers had an average age of 29 years, 65% were South African and about 1% had deceased compared to 2% of the males. The women in the area on average visited the antenatal facility four times during their pregnancy and 34% were heads of their households.

An investigation of the survival experience using the Kaplan–Meier (KM) curves (Fig. 2, left) showed differences in the survival experiences over the years 2000–2005. We assessed the differences in survival experiences over the years using the generalized Fleming–Harrington test weighted for later years failures ($p = 0.048$) (Fleming and Harrington, 1984). However there was no significant trend over the years ($p = 0.728$). This procedure of stratifying over the years was used to assess for the possibility of temporal random effects.

The year specific child 1-4MRs per 1000 person years from 2000 to 2005 were 3.57, 2.41, 3.33, 2.54, 4.44 and 1.55, respectively, consistent with the KM curves (Fig. 2, right) for the respective years. There was a statistically significant difference in the 2002 mortality rates compared to 2000 from the spatial and non-spatial analyses (see Table 2). The knots used in the low rank kriging were fit

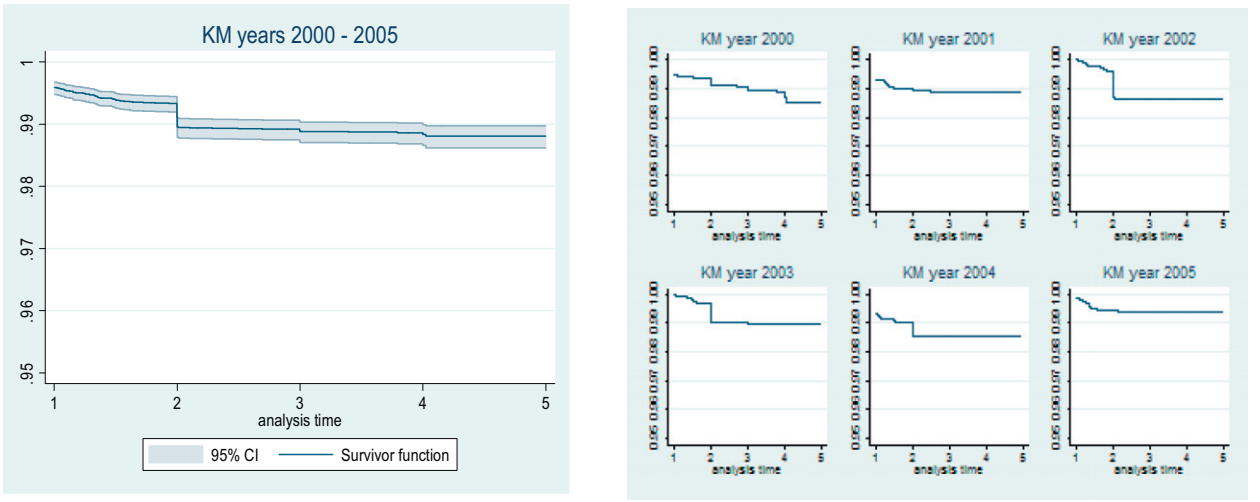


Fig. 2. Kaplan–Meier (KM) survival function over the years 2000–2005 (left) and year specific Kaplan–Meier (KM) curves (right).

using the mini-max space filling criteria, Fig. 3 shows distribution of 50 knots. The small faint (red)[3] dots are the locations and the squares are the selected 50 knots, since the $\widehat{\rho}$ (rate of spatial correlation decay) is fixed the number of knots only impacts on the computational speed and not on the accuracy of the estimates. The ZIP model was used to cater for zero inflation and to cater for both the excess zeros and over-dispersion, the ZINB model was used. The zero inflated Negative Binomial spatial temporal model showed an over-dispersion estimate $\delta = 1.79 \pm 1.87$[95% CI(0.041; 6.935)] and the zero inflation estimate was $\theta = 0.077 \pm 0.062$[95% CI(0.010; 0.217)]. These values show that there is some over-dispersion and zero inflation, thus ignoring these entirely in the modelling may cause model instability.

3.2. Spatiotemporal confounding adjusted determinants of child HIV/TB mortality

Multivariable analysis was done using non-spatial and spatial models for child specific variables, maternal variables and household variables. Bayesian semi-parametric geo-statistical models with location-specific random effects were fit to estimate the degree of zero inflation, over-dispersion, temporal and spatial correlation in the child HIV/TB mortality data. Spatial models showed more stable results with smaller standard errors in comparison to the non-spatial, the mean-based adjusted hazard ratios [AHR = exp(mean)] estimations were reported. The ages (interval start ages) were used in the analysis as the duration of follow-up (offset variables) until death, loss-to-follow up or survival between 1 and 5 years of age over the period 2000–2005. The outcome variable was treated as discrete count of child deaths due to HIV/TB (0 = Alive and 1 = Death). The outcome was treated as such to get more epidemiologically intuitive estimates for cohort data and to avoid identifiability problems associated with the logit link functions as highlighted earlier. Multivariable regression models were fit to account for spatial and temporal correlation; co-variables were added to the models to control for potential confounding in determining factors associated with child HIV/TB mortality (Pacheco et al., 2008). Poisson and Negative Binomial models were fit to the data as well as their zero inflated variants. The offset variable was fit using the piece wise exponential (p.e.m) with a first order autoregressive model temporal prior (see Fig. 4). Fig. 4 shows the log-baseline hazard function for selected models, showing a narrower credible interval for the spatial multivariable model in comparison to the non-spatial.

Four models were fit in the multivariable analyses these were the non-spatial zero inflated Poisson (ZIP) model, spatial zero inflated Poisson model, non-spatial zero inflated Negative Binomial (ZINB) model and spatial zero inflated Negative Binomial model. The first two were to cater for zero inflation and the last two were catering for zero inflation as well as over dispersion. The deviance information criteria (DIC) was used to measure model goodness of fit and the effective number of parameters (p_D) was used to measure model complexity (Spiegelhalter et al., 2002). Generally the spatial models had higher DIC because of the spatial parameters and also the Negative Binomial models had greater DICs compared to the Poisson; this made it difficult to judge the best model based on the DIC alone. The p_D which assesses model simplicity (parsimony) was then used to determine the most parsimonious model amongst these four. The zero inflated Negative Binomial spatial model had the least number of effective parameters and catered for over-dispersion, zero inflation and spatial random effects (see Table 1). We therefore focused on this model in the results and discussion sections.

The spatial multivariable analyses results showed several determinants of child HIV/TB mortality in the categories consistent with the non-spatial analyses. Several variables were statistically significant at the 5% level, these were either child, mother or household related. The results

[3] For interpretation of color in Fig. 3, the reader is referred to the web version of this article, these appear light grey in the black and white printed version.

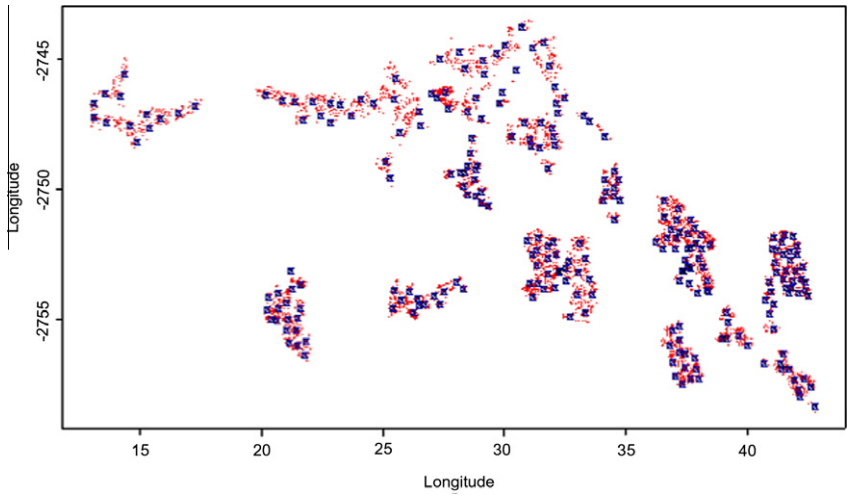


Fig. 3. Knots fitted using the mini–max space filling criteria from Agincourt households 2000–2005.

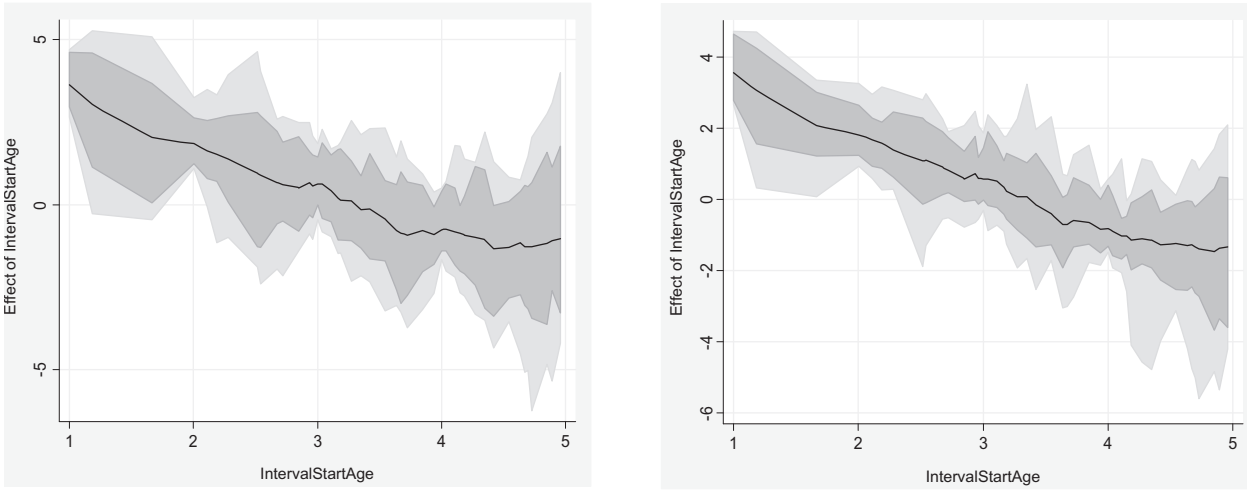


Fig. 4. Non-spatial ZINB with AR(1) baseline hazard (left) and multivariable spatial ZINB with AR(1) baseline hazard (right).

Table 1
Model goodness of fit non-spatial/spatial zero inflated Poisson and Negative Binomial models.

	Non-spatial zero inflated Poisson Un-standardized	nsZIP Saturated	Spatial zero inflated Poisson Un-standardized	sZIP Saturated	Non-spatial zero inflated Negative Binomial Un-standardized	nsZINB Saturated	Spatial zero inflated Negative Binomial Un-standardized	sZINB Saturated
$\overline{D}$	649.16	273.16	664.3	288.33	1674.77	1252.4	1865.41	1228.41
pD	246.51	246.51	247.7	247.7	83.86	104.72	47.95	36.23
DIC	1142.2	766.19	1159.79	783.79	1842.49	1461.86	1961.32	1300.88

from the analyses are shown in tables 2 and 3 for the ZIP and ZINB, respectively. The ZIP models had five and six out of the sixteen variables significant in the non-spatial and spatial analysis, respectively, whereas the ZINB had eight and nine variables for non-spatial and spatial models, respectively. The final spatiotemporal ZINB model was considered the "best" model which caters well for zero inflation, over-dispersion and spatiotemporal confounding.

The final zero inflated spatial Negative Binomial model results showed these nine predictors of child mortality: gender, nationality of child, death of mother, gender of head of household, year, antenatal clinic visits, socio-eco-

Table 2
Non-spatial and spatial multivariable zero inflated Poisson models.

Variable	Category	Descriptive, *n* (%) *N* = 16,844	Non-spatial zero inflated Poisson					nsZIP	Spatial zero inflated Poisson					sZIP
			Mean	SD	2.5%	Median	97.50%	AHR	Mean	SD	2.5%	Median	97.50%	AHR
Gender	Female (ref)	8506 (50.50)												
	Male	8838 (49.50)	0.415	0.25	−0.12	0.42	0.93	1.51	0.38	0.25	−0.11	0.39	0.88	1.46
Year	2000 (ref)	3360 (19.95)												
	2001	2826 (16.78)	0.49	0.30	−0.14	0.50	1.10	1.63	0.50	0.30	−0.10	0.52	1.02	1.65
	2002	2633 (15.63)	0.837*	0.37	0.16	0.86	0.54	2.31	0.81*	0.333	0.17	0.81	1.46	2.25
	2003	2496 (14.82)	0.56	0.43	−0.21	0.53	1.43	1.75	0.56	0.40	−0.32	0.58	1.27	1.75
	2004	3002 (17.82)	0.15	0.34	−0.50	0.15	0.80	1.16	0.17	0.32	−0.52	0.17	0.80	1.19
	2005	2527 (15.00)	−0.22	0.47	−1.14	−0.20	0.61	0.80	−0.21	0.45	−1.07	−0.20	0.65	0.81
Child's nationality	Mozambican (ref)	6242 (37.06)												
	South African	10,602 (62.94)	−0.82	0.65	−2.11	−0.86	0.34	0.44	−0.70	0.53	−1.65	−0.69	0.36	0.50
Mother's nationality	South African (ref)	10,883 (64.61)												
	Mozambican	5961 (35.39)	−0.11	0.67	−1.56	−0.11	1.25	0.90	−0.05	0.54	−1.07	−0.05	1.00	0.95
Mother's age at birth	Mean years ± SD	29.00 ± 7.60	0.01	0.02	−0.02	0.01	0.04	1.01	0.01	0.02	−0.03	0.01	0.04	1.01
Mother's status	Alive (ref)	16,736 (99.36)												
	Deceased	108 (0.64)	2.665*	0.81	0.90	2.74	3.90	14.37	2.456*	0.63	1.15	2.46	3.74	11.66
Father's status	Alive (ref)	16,531 (98.14)												
	Deceased	313 (1.86)	0.98	0.66	−0.33	1.00	2.33	2.66	1.081*	0.49	0.06	1.09	2.04	2.95
Household head gender	Female (ref)	3012 (33.98)												
	Male	5851 (66.02)	−0.49	0.26	−1.04	−0.50	0.01	0.61	−0.45	0.25	−0.96	0.44	0.02	0.64
Antenatal visits	Mean visits ± SD	4.04 ± 2.42	−0.185*	0.05	−0.29	−0.19	−0.09	0.83	−0.184*	0.05	−0.29	−0.18	−0.90	0.83
Pregnancy parity	Mean ± SD	2.104 ± 1.396	0.25	0.22	−0.14	0.24	0.67	1.28	0.25	0.21	−0.16	0.25	0.67	1.28
Household socio-economic status	Most poor (ref)	1566 (17.67)												
	Very poor	1701 (19.19)	−0.17	0.32	−0.83	−0.17	0.41	0.84	−0.22	0.30	−0.82	−0.22	0.36	0.80
	Poor	1833 (20.68)	−0.20	0.35	−0.84	−0.20	0.51	0.82	−0.28	0.29	−0.85	−0.29	0.27	0.76
	Moderately poor	1905 (21.49)	−0.26	0.33	−1.03	−0.27	0.36	0.77	−0.30	0.32	−0.95	−0.29	0.29	0.74
	Least poor	1858 (20.97)	−0.856*	0.42	−1.60	−0.84	−0.04	0.42	−0.941*	0.42	−1.83	−0.94	−0.13	0.39
Cumulative household deaths	Mean ± SD	0.252 ± 0.553	−0.16	0.25	−0.68	−0.17	0.29	0.85	−0.10	0.24	−0.62	−0.13	0.30	0.90
Parity at birth	Mean ± SD	1.569 ± 0.856	−0.369*	0.20	−0.77	−0.35	−0.03	0.69	−0.37*	0.18	−0.74	−0.37	−0.02	0.69
Total living household	Mean ± SD	1.569 ± 0.884	0.01	0.02	−0.04	0.01	0.06	1.01	0.00	0.02	−0.04	0.00	0.04	1.00
Number of deaths in household	Mean ± SD	0.047 ± 0.222	0.854	0.50	−0.11	0.86	1.80	2.35	0.89	0.48	−0.17	0.92	1.79	2.44
Total live births	Mean ± SD	2.030 ± 1.855	−0.21	0.18	−0.60	−0.21	0.13	0.81	−0.22	0.18	−0.61	−0.21	0.12	0.80

* Significant at 5% level of significance.

Table 3
Non-spatial and spatial multivariable zero inflated Negative Binomial models.

Variable	Category $N = 16,844$	Descriptive, n (%)	Non-spatial zero inflated Negative Binomial					nsZINB	Spatial zero inflated Negative Binomial					sZINB
			Mean	SD	2.5%	Median	97.50%	AHR	Mean	SD	2.5%	Median	97.50%	AHR
Gender	Female (ref)	8506 (50.50)												
	Male	8338 (49.50)	0.31*	0.14	0.07	0.28	0.62	1.32	0.326*	0.17	0.08	0.31	0.63	1.36
Year	2000 (ref)	3360 (19.95)												
	2001	2826 (16.78)	0.46	0.27	−0.05	0.48	1.05	1.62	0.55	0.23	−0.06	0.58	0.92	1.78
	2002	2633 (15.63)	0.74*	0.24	0.16	0.79	1.14	2.20	0.762*	0.31	0.12	0.72	1.38	2.05
	2003	2496 (14.82)	0.39	0.32	−0.19	0.40	1.01	1.48	0.45	0.31	−0.14	0.45	1.02	1.56
	2004	3002 (17.82)	0.00	0.25	−0.42	0.00	0.46	1.00	0.03	0.24	−0.40	0.03	0.49	1.03
	2005	2527 (15.00)	−0.54	0.41	−1.27	−0.53	0.26	0.59	−0.43	0.32	−1.08	−0.41	0.12	0.67
Child's nationality	Mozambican (ref)	6242 (37.06)												
	South African	10,602 (62.94)	−0.41	0.32	−1.02	−0.35	0.09	0.71	−0.643*	0.31	−1.43	−0.68	−0.04	0.51
Mother's nationality	South African (ref)	10,883 (64.61)												
	Mozambican	5961 (35.39)	0.05	0.32	−0.59	0.06	0.57	1.06	−0.12	0.33	−0.68	−0.19	0.57	0.83
Mother's age at birth	Mean years ± SD	2900 ± 7.60	0.01	0.01	−0.01	0.01	0.05	1.01	0.01	0.01	−0.01	0.01	0.03	1.01
Mother's status	Alive (ref)	16,531 (99.36)												
	Deceased	108 (0.64)	1.479*	0.36	0.63	1.52	2.15	4.55	1.09*	0.47	0.25	1.08	1.94	2.93
Father's status	Alive (ref)	16,531 (98.14)												
	Deceased	313 (1.86)	0.17	0.32	−0.55	0.24	0.70	1.27	0.11	0.44	−0.82	0.16	0.82	1.17
Household head gender	Female (ref)	3017 (33.98)												
	Male	5851 (66.02)	−0.449*	0.16	−0.70	−0.49	−0.15	0.62	−0.54*	0.14	−0.79	−0.55	−0.29	0.58
Antenatal Visits	Mean visits ± SD	4.04 ± 2.42	−0.156*	0.03	−0.21	−0.16	−0.09	0.85	−0.17*	0.03	−0.22	−0.17	−0.12	0.84
Pregnancy parity	Mean ± SD	2.104 ± 1.396	−0.04	0.11	−0.24	−0.03	0.18	0.97	0.08	0.11	−0.14	0.08	0.26	1.09
Household socio-economic status	Most poor (ref)	1566 (17.67)												
	Very poor	1701 (19.19)	−0.32	0.18	−0.65	−0.32	0.08	0.72	−0.36	0.24	−0.81	−0.31	0.09	0.73
	Poor	1833 (20.68)	−0.37	0.20	−0.77	−0.38	0.05	0.68	−0.41	0.21	−0.79	−0.43	0.02	0.65
	Moderately poor	1905 (21.49)	−0.23	0.24	−0.78	−0.22	0.18	0.80	−0.28	0.24	−0.75	−0.26	0.22	0.77
	Least poor	1858 (20.97)	−0.945*	0.33	−1.57	−0.97	−0.28	0.38	−0.99*	0.27	−1.44	−0.97	−0.49	0.38
Cumulative household deaths	Mean ± SD	0.252 ± 0.553	−0.10	0.14	−0.38	−0.09	0.16	0.92	−0.14	0.16	−0.48	−0.15	0.14	0.86
Parity at birth	Mean ± SD	1.569 ± 0.856	−0.291*	0.12	−0.54	−0.31	−0.04	0.74	−0.314*	0.11	−0.55	−0.31	−0.12	0.73
Total living household	Mean ± SD	1.569 ± 0.884	0.01	0.01	−0.01	0.01	0.04	1.01	0.01	0.01	−0.01	0.01	0.04	1.01
Number of deaths in household	Mean ± SD	0.047 ± 0.222	0.703*	0.27	0.14	0.71	1.16	2.03	0.681*	0.29	0.12	0.69	1.20	2.00
Total live births	Mean ± SD	2.030 ± 1.855	0.03	0.09	−0.11	0.02	0.23	1.02	−0.06	0.09	−0.26	−0.07	0.08	0.94

* Significant at 5% level of significance.

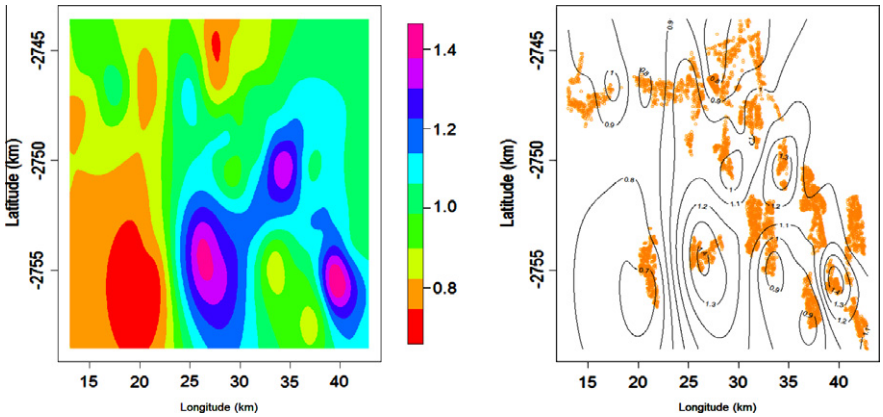


Fig. 5. Posterior adjusted hazard rates (AHRs) smoothed map (left) and contour AHRs map (right) for child HIV/TB mortality 2000–2005.

nomic status, birth order position and number of deaths in household. The boys were 36% more at risk of death compared to the girls adjusting for other variables and space-time confounders $(1.36 = e^{0.311}, 95\%\text{CI}[1.08, 1.88])$. There was a temporal effect on HIV/TB mortality risk for the year 2002 which was more than twice that of 2000 $(2.02 = e^{0.719}, 95\%\ \text{CI}[1.13, 4.00])$. South African children were almost twice better protected from HIV/TB death compared to the Mozambican children after adjusting for several variables $(0.51 = e^{-0.683}, 95\%\ \text{CI}[0.24, 0.96])$. Orphans with a deceased mother were almost thrice at risk of HIV/TB death compared to those whose mother's were alive keeping other variables constant $(2.93 = e^{1.075}$, $95\%\ \text{CI}[1.29; 6.93])$. Adjusting for other factors male headed households were 42% less likely to experience a child death in comparison to female headed households $(0.58 = e^{-0.548}, 95\%\ \text{CI}[0.45; 0.75])$.

For every increase by one visit to the antenatal clinic the adjusted risk of death reduced by 16% $(0.84 = e^{-0.173}$, $95\%\ \text{CI}[0.81; 0.89])$. The least economically poor households were 62% less likely to experience a child death compared to the most poor households $(0.38 = e^{-0.966}$ adjusted, $95\%\ \text{CI}[0.24; 0.61])$. The birth order position was protective such that as this increased by one level the adjusted risk of child death decreased by 27% $(0.73 = e^{-0.311}, 95\%\ \text{CI}[0.58; 0.89])$. Lastly as the number of household deaths increased by one the adjusted risk of child HIV/TB death increased two fold $(2.00 = e^{-0.694}$, $95\%\ \text{CI}[1.12; 3.33])$.

Posterior risk estimates were displayed on smoothed maps showing the smoothed areas where child deaths occurred. Maps of posterior mean adjusted hazard rates (AHRs) show point estimates and extreme high and low risk areas (Ranta and Penttinen, 2000). Risk factors for child HIV/TB mortality adjusting for potential spatiotemporal confounders were: being a boy, a former foreigner, losing your mother, having a female head of house, being socio-economically disadvantaged, being born earlier and coming from households were many deaths occurred. Protective of child HIV/TB mortality were increasing number of clinic visit and coming from households with more people. Fig. 5 shows the maps of the households and the mean AHRs smoothed posterior maps adjusting for the aforementioned risk factors, spatial and temporal potential confounders.

4. Discussion and conclusions

In our study we found that maternal orphans experienced a threefold greater risk of death due to HIV/TB than children whose mothers were still living (AHR = 2.93, 95% CI[1.26; 6.93]). This was adjusted for potential spatiotemporal confounding, number of household deaths, birth order position and parity, child and mother's nationality, gender of house head and gender of child, which were all significant risk factors. Increased antenatal clinic visits and cumulative total of adults in the house were protective of the child's death. These risk factors are interrelated and fall into either child-specific, mother-related, household socio-economic status, temporal or spatial. Risk map analysis showed that there were three hotspots central, south easterly and south-westerly over the years 2000–2005; similar hot spots were seen for the year 2004 alone and also in another study considering only the areal (village) level and temporal random effects (Sartorius et al., 2011). The trends seen in this paper are consistent with other studies in South Africa, in the same site and elsewhere. A study in Agincourt showed that HIV/TB accounted for about a third of the under-five deaths from 2002 to 2005 (Tollman et al., 2008). A similar non-spatial study on the impact of parental presence on child mortality in Agincourt 1997–2005, showed that death of a mother was the most predictive of the child's death (OR = 5.22, 95% CI[2.37; 11.54]) (Collinson et al., 2009). The study also showed that former Mozambican mothers' children were more than a third at greater risk of mortality than the South Africans' children (OR = 1.34, 95% CI[1.09; 1.65]). Similarly in rural Kwazulu-Natal child Province, mother's death increased the mortality risk fourfold adjusting for child, household and mother's variables (Argeseanu, 2004).

There are different schools of thoughts as to how the maternal orphan HIV/TB mortality trends can be explained, biological and socio-economical. Mothers who die due to

AIDS, may have transmitted the virus to their children, regardless of the stage at which this occured (in utero, intrapartum or postnatal). Support programs currently favour local South African mothers, making it more difficult for breast feeding foreign mother's to access available services and resources. Socially, if the mother (who will most likely be the remaining bread winner) dies, this leaves AIDS orphans with no caregiver or with grandparents who can no longer work. This contributes to the deaths of children due to poor nutrition, weakened immune system and lack of adequate support. Although we concentrated on HIV as the main cause of death for these under five children, TB may have played a significant role leading to death. However the verbal autopsy technique used to determine cause of death does not distinguish these very well for children. Although Agincourt is in rural South Africa, the use of available health facilities seems to be increasing; using antenatal clinic visits as a proxy.

The other risk factors were mainly related to poverty as direct proxies or indirect consequences of it. We found that poorer socio-economic households had higher child HIV/TB mortality rates, a finding which corresponds with a study in rural Kwazulu-Natal Province (Ndirangu et al., 2010). Poverty is a major contributor to child deaths in Africa, every 3.6 s one person dies of starvation, usually a child under the age of 5. One of the Millennium Development Goals (MDGs) is to: "Eradicate extreme poverty and hunger" (Eftim et al., 2008; Sachs, 2010). The child support grant provided by the South African government is an amount of ZAR260 p/m (US$38), is often all that some families in the study area have to live on. The elderly get an old-age pension of ZAR1140 p/m (US$163) which can be the main source of livelihood of families. A study showed that female old-age pensioners cared for children and improved the girl child's likelihood to go to school in Agincourt (Case and Menendez, 2007). These studies reveal gender differentials in access to services, and schooling.

Controlling for potential confounding in epidemiological analyses gives strength to the estimates derived from such modelling procedures. Our modelling approach caters for spatial and temporal confounders and also for large zero inflated spatiotemporal sparse data. Spatiotemporal Bayesian modelling provided good estimates and maps which can guide interventional programmes to be more effective in addressing HIV/TB. There are some modelling limitations which if addressed in future research will improve on the epidemiological modelling. The temporal component was discretized into whole years; this could have been modelled as continuous and allow for formulation of time varying covariates and spatiotemporal interactions. As a way of dealing with identifiability issues and getting more epidemiologically correct estimates; members of the exponential (log link) family were used, other less flexible family of distributions could also be used for sensitivity testing. Mostly non-informative priors have been used; experts opinion could be sought to build and apply more informative priors in our modelling. In dealing with the "big N" problem several approaches can be used. The package BayesX, whilst it cannot address all the modelling issues, provides very useful Bayesian spatiotemporal modelling algorithms which are applicable even in a resource scare environment.

Our results provide evidence useful for policy makers advocating for better access to health-care and child support grants especially for non-South Africans, increase and diversification of HIV mother to child transmission awareness programmes in rural areas. This study provides evidence to support the South African government's revised HIV policy effected on 1 April 2010. The policy allows HIV positive mothers to be on ante-retro-virals with a CD4 count of 350 (national cut off is 200) or less, symptomatic HIV regardless of CD4 and also pregnant mothers from 14 gestational weeks onwards. Infants can now be tested without the mother's consent to save the child (Zuma, 2009). Child survival is dependent on the survival of the mother, hence the need to strengthen anti-retro-viral treatment. South African policy supporting maternal survival need to be strengthened. In an address to the nation the South African president said: "One of the most important results of the roll-out of anti-retro-viral therapy among the general population will be the extension of the lives of AIDS sick parents leading to a dramatic decline in the number of orphans" (Bradshaw et al., 2002). The main challenge as with many other policies is implementation. We suggest targeting endeavours to reduce mother-to-child transmissions by effectively implementing interventions that reach rural villages and the poorest households which are often neglected.

Authors' contributions

E.M. drafted the manuscript, did the statistical analysis and wrote this paper. P.V. and K.K. gave their expert advice on methodology and epidemiology aspects of this paper, respectively.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was made possible by doctorial fellowship support from World Health Organisation, Tropical Disease Research (WHO/TDR) and Swiss South African Joint Research Program (SSAJP), Project Number JRP IZLSZ3_122926. The Agincourt HDSS is funded by the Wellcome Trust UK (Grant Nos. 058893/Z/99/A and 069683/Z/02/Z) and the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew Mellon and Hewlett Foundations, USA. Special thanks goes to Benn Sartorius for data extraction, community leaders, field workers, supervisors and the Agincourt community.

Appendix A. Zero inflated Poisson and Negative Binomial Hierarchical Bayesian models

Observational outcome data can be modelled by different known statistical distributions including the flexible

exponential family. Mortality outcomes of the count of deaths can be modelled using the basic log-linear Poisson model were the outcomes $y_i|X_i \sim Po(\lambda_i)$ are conditionally independent with mean and variance $\lambda_i = \exp(\eta_i)$ with linear predictor $\eta_i = X_i\beta$. The commonly applied logistic regression models which have logit link functions have been seen to cause identifiability problems, resulting in unstable model especially for spatial models (Agarwal et al., 2002). We therefore treated the outcome death as a discrete (count) variable since this provides better estimates of prevalence and relative risk compared to the odds ratio which results from treating it as binary outcome (Barros and Hirakata, 2003; Fekedulegn et al., 2010). The Poisson model can also incorporate the exposure time by introducing piece-wise exponential offsets o_i resulting in $y_i|X_i \sim Po(\mu_i), \mu_i = \exp(o_i + \eta_i)$ and in log-linear form $y_i|\eta_i \sim Po(\mu_i)$ (Brezger and Lang, 2006). In this paper, we extend this model in three dimensions to address different inherent problems in our hierarchical data structure. Firstly the problem of over-dispersion when the mean is less than the variance which can be addressed by use of conditionally independent Negative Binomial (NB) distribution (Ridout et al., 2001). We then extend our model incorporate the dispersion parameter $\delta > 0, y_i|\eta_i, \delta \sim NB(\mu_i, \delta)$. Secondly we have the problem of excess of zeros known as zero inflation, this is addressed by use of a latent process that inflates the number of zero points with a weighting θ for extra zeros and $1-\theta$ for the ZIP and ZINB where $0 < \theta < 1$. We then obtain the conditionally independent zero inflated Poisson $y_i|\eta_i, \theta \sim ZIP(\mu_i, \theta)$ and zero inflated Negative Binomial (ZINB) $y_i|\eta_i, \delta, \theta \sim ZINB(\mu_i, \delta, \theta)$ whose hierarchical model structures and likelihoods are shown below. Thirdly modelling which is flexible to include; time varying covariates, temporal random effects, spatial random effects, non-linear covariates and desired interactions. This was done by extending the linear predictor $\eta_i = X_i\beta$ to a structured additive predictor:

$$\eta_i = \sum_{j=1}^{p} f_j(z_{ij}) + X_i\beta + f_{spat_i}(s_k) + \varepsilon_{s_k}$$

where $f_i(z_{ij})$ are nonlinear effects of continuous covariate z_{ij}, X_i is the linear (fixed) effects design matrix and vector of parameters $\beta, f_{spat_i}(s_k)$ is the structured random spatial at household $s_k \in$s (all households with children in the area) and ε_{s_k} being the unstructured non-spatial random effect. Due to the complexity of the spatiotemporal data (correlated in time and space) and their large size. Statistical methods applied to spatial data which have spatial autocorrelation often underestimate the standard error and thus the statistical significance is overestimated (Cressie, 1993). We used the Bayesian estimation called the Markov Chain Monte Carlo (MCMC), which require Gibbs sampling and Metropolis Hastings within Gibbs (Banerjee et al., 2004). Modelling spatial longitudinal data which are highly autocorrelated has been made possible by introduction of the Hierarchical Bayesian (HB) modelling approach. This decomposes a complex problem into a series of simpler conditional levels such that at any given level in the hierarchy inference conditionally relies on hypotheses made at higher levels (Banerjee et al., 2004). The HB model constitutes of four levels: firstly the data level which specifies the conditional distribution of data given parameters i.e y_i; secondly a process level which specifies the conditional distributions of process given their parameters (these are usually logit and log functions of Generalised Linear Models); thirdly a parameter level with conditional parameter distributions given priors and lastly hyper-parameter level which specifies the prior distributions for hyper-parameters. The posterior distribution can then be used for full Bayesian inference of the ZIP and ZINB models given the following full conditional distributions. To estimate the parameters Gibbs sampling is used by sampling from the full conditional distributions for posteriors with known closed forms (Casella and George, 1992). For those without the closed form we use the Metropolis Hastings (MH) random walk algorithms (Browne and Draper, 2000).

A.1. Zero inflated Poisson model

A.1.1. Level 1: Data

The ZIP with a semi-parametric predictor offers a flexible way to model the zero inflation parameter (θ).

$y_i|\eta_i\theta \sim \text{ZIP}(\mu_i\theta)$ with probability density function

$$P(y_i|\eta_i, \theta) = \begin{cases} \theta + (1-\theta)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i^!} & \text{if } y_i = 0 \\ (1-\theta)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i^!} & \text{if } y_i \neq 0 \end{cases}$$

With expectation $E(y_i|\eta_i, \theta) = (1-\theta)\mu_i$ and variance $V(y_i\eta_i, \theta) = (1-\theta)(\mu_i) + \theta(1-\theta)\mu_i^2$.

A.1.2. Level 2: Processes

$$\eta_i = X_i'\beta + f_{spat_i}(s_k) + \varepsilon_{s_k};$$
$$\theta \sim Uniform[0, 1]$$

X_i is the linear (fixed) effects design matrix and vector of parameters $\beta, f_{spat_i}(s_k)$ is the structured random spatial effect and ε_{s_k} being the unstructured non-spatial random effect.

A.1.3. Level 3: Parameters

Priors for $P(\beta), P(f_{spat_i}|\tau^2_{spat}), P(\varepsilon_{s_k}|\tau^2_{error}), P(\theta)$. Uniform priors for $P(\beta)$ and $P(\theta)$. Multivariate Gaussian random filed for $P(f_{spat_i}|\tau^2_{spat})$ and $P(\varepsilon_{s_k}|\tau^2_{error})$ are Gaussian independent and identically distributed (i.i.d) random effects.

A.1.4. Level 4: Hyper-parameters and likelihood

Hyperpriors for $P(\tau^2_{spat})$ and $P(\tau^2_{error})$ are all Inverse Gamma $\sim IG(a^*, b^*)$; with hyper parameters $a^* = b^* = 0.001$ (Brezger and Lang, 2006). After some computations the likelihood of the ZIP models is:

$$L(y_i|\eta_i, \theta) = \prod_{i=1}^{n} P(y_i|\eta_i, \theta) \propto \exp\left\{\sum_{y_i=0} \log(\theta + (1-\theta)\exp(-\mu_i)) + W_o \log(1-\theta) + \sum_{y_i \neq 0}(-\mu_i + y_i\log(\mu_i))\right\}$$

where W_0 is the number of units with strictly positive responses.

A.2. Zero inflated Negative Binomial model

A.2.1. Level 1: Data

The ZINB with a semi-parametric predictor offers a flexible way to model the variability of the data catering for over-dispersion (δ), zero inflation (θ).

$y_i|\eta_i, \delta, \theta \sim ZINB(\mu_i, \delta, \theta)$ with probability density function

$$P(y_i|\eta_i,\delta,\eta) = \begin{cases} \theta + (1-\theta)\left(\frac{\delta}{\delta+\mu_i}\right)^{\delta} & \text{if } y_i = 0 \\ (1-\theta)\frac{\Gamma(y_i+\delta)}{\Gamma(y_i+1)\Gamma(\delta)}\left(\frac{\delta}{\delta+\mu_i}\right)^{\delta}\left(\frac{\mu_i}{\delta+\mu_i}\right)^{y_i} & \text{if } y_i \neq 0 \end{cases}$$

With expectation $E(y_i|\eta_i,\theta) = (1-\theta)\mu_i$ and variance $V(y_i|\eta_i,\delta,\theta) = (1-\theta)(\mu_i + \frac{\mu_i^2}{\delta}) + \theta(1-\theta)\mu_i^2$.

A.2.2. Level 2: Processes

$$\eta_i = X_i'\beta + f_{spat_i}(s_k) + \varepsilon_{s_k};$$
$$\delta|b \sim Gamma(a,b); \theta \sim Uniform[0,1]$$

X_i is the linear (fixed) effects design matrix and vector of parameters β, $f_{spat_i}(s_k)$ is the structured random spatial effect and ε_{s_k} being the unstructured non-spatial random effect. We set $a = 1$ and assume a Gamma hyper-prior for $b \sim Gamma(\alpha 1, \alpha 2)$.

A.2.3. Level 3: Parameters

Priors for $P(\beta), P(f_{spat_i}|\tau^2_{spat}), P(\varepsilon_{s_k}|\tau^2_{error}, P(\theta), P(\delta|b))$. Uniform priors for $P(\beta)$ and $P(\theta)$, Gamma priors for $P(\delta|b)$. Multivariate Gaussian random field (GRF) priors for $P(f_{spat_i}|\tau^2_{spat})$ and $P(\varepsilon_{s_k}|\tau^2_{error})$ are Gaussian independent and identically distributed (i.i.d) random effects.

A.2.4. Level 4: Hyper-parameters and likelihood

Hyperpriors for $P(\tau^2_{spat})$, $P(\tau^2_{error})$ and $P(b)$ are all Inverse Gamma $\sim IG(a^*, b^*)$; with hyper parameters $a^* = b^* = 0.001$ (Brezger and Lang, 2006). After some computations the likelihood of the ZINB models is:

$$\begin{aligned} L(y_i|\eta_i,\delta,\theta) &= \prod_{i=1}^{n} P(y_i|\eta_i,\delta,\theta) \propto \\ &\propto \exp\left\{\sum_{y_i=0} \log\left(\theta + (1-\theta)\left(\frac{\delta}{\delta+\mu_i}\right)^{\delta}\right)\right. \\ &\quad + W_o(\log(1-\theta) - \log(\Gamma(\delta)) + \delta\log(\delta)) \\ &\quad + \sum_{y_i \neq 0} \log(\Gamma(y_i+\delta)) + y_i\log(\mu_i) \\ &\quad \left. - (y_i+\delta)\log(\delta+\mu_i)\right\} \end{aligned}$$

where W_0 is the number of units with strictly positive responses.

References

Agarwal DK, Gelfand AE, Citron-Pousty S. Zero-inflated models with application to spatial count data. Environ Ecol Stat 2002;9:341–55.

Argeseanu S. Risks, amenities, and child mortality in rural South Africa. Afr Popul Stud 2004;19:13–33.

Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. Boca Raton, FL: Chapman & Hall/CRC; 2004.

Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol 2003;3:21.

Becher H, Muller O, Jahn A, Gbangou A, Kynast-Wolf G, Kouyate B. Risk factors of infant and child mortality in rural Burkina Faso, Report 82; 2004. p. 265–73.

Belitz C, Brezger A, Kneib T, Lang S, Fronk E, Heinzl, et al. BayesX-software for bayesian inference in structured additive regression models; 2009. Version 2.01.

Bradshaw D, Johnson L, Schneider H, Bourne D, Dorrington R. Orphans of the HIV/AIDS epidemic – the time to act is now, Report, Medical Reasearch Council, Cape Town, South Africa; 2002.

Brezger A, Kneib T, Lang S. BayesX-Reference Manual. Department of Statistics, University of Munich 2005.

Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. Comput Stat Data Anal 2006;50:967–91.

Browne WJ, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Comput Stat 2000;15:391–420.

Burnett R, Ma RJ, Jerrett M, Goldberg MS, Cakmak S, Pope CA, et al. The spatial association between community air pollution and mortality: a new method of analyzing correlated geographic cohort data. Environ Health Perspect 2001;109:375–80.

Case A, Menendez A. Does money empower the elderly? Evidence from the Agincourt demographic surveillance site, South Africa. Scand J Public Health Suppl 2007;69:157–64.

Casella G, George EI. Explaining the Gibbs sampler. Am Stat 1992;46:167–74.

Collinson M, White M, Short S, Lurie M, Byass P, Kahn K, et al. Child mortality, migration and parental presence in rural South Africa near the border with Mozambique; 2009.

Cressie NAC. Statistics for spatial data. New York: Wiley; 1993.

Delaunay V, Etard JF, Preziosi MP, Marra A, Simondon F. Decline of infant and child mortality rates in rural Senegal over a 37-year period (1963–1999). Int J Epidemiol 2001;30:1286–93.

Eftim SE, Samet JM, Janes H, McDermott A, Dominici F. Fine particulate matter and mortality – a comparison of the six cities and American Cancer Society cohorts with a medicare cohort. Epidemiology 2008;19:209–16.

Fahrmeir L, Echavarria LO. Structured additive regression for overdispersed and zero-inflated-count data. Appl Stoch Models Bus Ind 2006;22:351–69.

Fekedulegn D, Andrew M, Violanti J, Hartley T, Charles L, Burchfiel C. Comparison of statistical approaches to evaluate factors associated with metabolic syndrome. J Clin Hypertens 2010;12:365–73.

Fernandes MVM, Schmidt AM, Migon HS. Modelling zero-inflated spatio-temporal processes. Stat Model 2009;9:3–25.

Filteau S. The HIV-exposed, uninfected African child. Trop Med Int Health 2009;14:276–87.

Fleming T, Harrington D. Non parametric estimation of the survival distribution in censored data. Commun Stat Theory Methods 1984;13:2469–86.

Geweke J, Minneapolis FRB. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Federal Reserve Bank of Minneapolis, Research Departmant; 1991.

Gosoniu L, Vounatsou P, Tami A, Nathan R, Grundmann H, Lengeler C. Spatial effects of mosquito bednets on child mortality. BMC Public Health 2008;8:356.

Greenland S. Modeling and variable selection in Epidemiologic analysis. Am J Public Health 1989;79:340–9.

Greven S, Dominici F, Zeger SL. A spatio-temporal approach for estimating chronic effects of air pollution. John Hopkins University; 2009.

Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. Oper Res 1983;31:1109–44.

Kahn K, Tollman SM, Garenne M, Gear JS. Validation and application of verbal autopsies in a rural area of South Africa. Trop Med Int Health 2000;5:824–31.

Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa: data and methods of the agincourt health and demographic surveillance system. Scand J Public Health 2007;35: 8–20.

Kamman E, Wand M. Geoadditive models, Department of Biostatistics, School of Publicc Health, Havard University; 2001.

Kleinschmidt I, Pettifor A, Morris N, MacPhail C, Rees H. Geographic distribution of human immunodeficiency virus in South Africa. Am J Trop Med Hyg 2007;77:1163–9.

Kynast-Wolf G, Sankoh OA, Gbangoun A, Kouyate B, Becher H. Mortality Patterns, 1993–98, in a rural area of Burkina Faso, West Africa, based on the Nouna demographic surveillance system. Trop Med Int Health 2002;7:349–56.

Msisha WM, Kapinga SH, Earls FJ, Subramanian SV. Place matters: multilevel investigation of HIV distribution in Tanzania. AIDS 2008;22:741–8.

Ndirangu J, Newell ML, Tanser F, Herbst AJ, Bland R. Decline in early life mortality in a high HIV prevalence rural area of South Africa: evidence of HIV prevention or treatment impact? AIDS 2010;24:593.

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 2008;24:581.

Pacheco AG, Tuboi SH, Faulhaber JC, Harrison LH, Schechter M. Increase in non-AIDS related conditions as causes of death among HIV-infected individuals in the HAART Era in Brazil. PLoS One 2008;3:7.

Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News 2006;6:7–11.

Press WH. Numerical recipes in C++: the art of scientific computing. Cambridge, UK, New York: Cambridge University Press; 2002.

R-cran. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2010.

Raftery AE, Lewis S. How many iterations in the Gibbs sampler. Bayesian Stat 1992;4:763–73.

Ranta J, Penttinen A. Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. Stat Med 2000;19:2345–59.

Ridout M, Hinde J, Demetrio CGB. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 2001;57:219–23.

Rutstein SO. Factors associated with trends in infant and child mortality in developing countries during the 1990s. Bull World Health Organ 2000;78:1256–70.

Sachs J, McArthur J. The Millennium Project: a plan for meeting the millennium development goals. Lancet 2005;365:347–53.

Sachs J. Sustainable developments-millennium development goals at 10: a decade's worth of targeted accomplishments shows extreme poverty can be eliminated. Sci Amer Mag 2010;302:30.

Sartorius B, Kahn K, Vounatsou P, Collinson M, Tollman S. Young and vulnerable: spatial–temporal trends and risk factors for infant mortality in rural South Africa (Agincourt), 1992–2007. BMC Public Health 2010;10:645.

Sartorius B, Kahn K, Collinson MA, Vounatsou P, Tollman SM. Survived infancy but still vulnerable: spatial–temporal trends and risk factors for child mortality in rural South Africa (Agincourt), 1992–2007. Geospat Health 2011;5:285–95.

Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 2002;64:583–639.

StataCorp. Stata Statistical Software Release 10. StataCorp LP, Texas, USA; 2007.

Tollman S, Kahn K, Sartorius B, Collinson M, Clark S, Garenne M. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. Lancet 2008;372:893–901.

Tollman SM. The Agincourt field site – evolution and current status. South Afr Med J 1999;89:853–8.

UNAIDS. Report on the global AIDS epidemic, Report; 2008.

UNAIDS. Report on the global AIDS epidemic, Report; 2009.

UNDG. The Millenium Development Goals Report, Report; 2008.

UNESCO. South Africa Millennium Development Goals Mid term country report; 2007.

UNICEF. State of the World's Children; 2009.

Zuma J. Address by President Jacob Zuma on the occasion of World AIDS Day, Pretoria Showgrounds. South African Government Information; 2009.

International Journal of Applied Earth Observation and Geoinformation 22 (2013) 86–98

Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa

Eustasius Musenge[a,b,*], Tobias Freeman Chirwa[b], Kathleen Kahn[a,d,e], Penelope Vounatsou[c]

[a] MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[b] Biostatistics and Epidemiology Division, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[c] Swiss Tropical and Public Health Institute, Basel, Switzerland
[d] Centre for Global Health Research, Umeå University, Umeå, Sweden
[e] INDEPTH Network, Accra, Ghana

ARTICLE INFO

Article history:
Received 7 July 2011
Accepted 2 April 2012

Keywords:
GMRF
Big "N"
Zero inflated
INLA SPDE
HIV/TB mortality
Spatiotemporal
Agincourt South Africa

ABSTRACT

Longitudinal mortality data with few deaths usually have problems of zero-inflation. This paper presents and applies two Bayesian models which cater for zero-inflation, spatial and temporal random effects. To reduce the computational burden experienced when a large number of geo-locations are treated as a Gaussian field (GF) we transformed the field to a Gaussian Markov Random Fields (GMRF) by triangulation. We then modelled the spatial random effects using the Stochastic Partial Differential Equations (SPDEs). Inference was done using a computationally efficient alternative to Markov chain Monte Carlo (MCMC) called Integrated Nested Laplace Approximation (INLA) suited for GMRF. The models were applied to data from 71,057 children aged 0 to under 10 years from rural north-east South Africa living in 15,703 households over the years 1992–2010. We found protective effects on HIV/TB mortality due to greater birth weight, older age and more antenatal clinic visits during pregnancy (adjusted RR (95% CI)): 0.73(0.53;0.99), 0.18(0.14;0.22) and 0.96(0.94;0.97) respectively. Therefore childhood HIV/TB mortality could be reduced if mothers are better catered for during pregnancy as this can reduce mother-to-child transmissions and contribute to improved birth weights. The INLA and SPDE approaches are computationally good alternatives in modelling large multilevel spatiotemporal GMRF data structures.
© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Public Health data on mortality have been growing increasingly rich as more accurate information on "who", "where" and "when" becomes available. These form hierarchical (multilevel) data structures which are correlated such that person-level ("who") information can be repeated, geo-statistical ("where") data often has spatial correlation and temporal ("when") data can be auto-correlated. Classical statistical techniques are usually based on independent observations, but when applied to multilevel data structures they often underestimate the standard errors. As a result of this the statistical significance is overestimated leading to erroneous results and subsequent inferences (Cressie, 1993). This defeats the main goal in epidemiological analysis, which is to identify and quantify correctly any exposures, behaviours and characteristics that may modify a population's or individuals risk and use these to implement more appropriate interventions (Rose, 2001).

In modelling hierarchical data we can take into account spatial and temporal correlations by introducing in the model spatiotemporal random effects. Several other hurdles have to be overcome when modelling hierarchical mortality data such as: zero inflation when there is a greater proportion of non-occurrence for an outcome, handling large data structures, repeated measures and estimating many parameters rapidly and accurately. Bayesian techniques with the aid of the Markov chain Monte Carlo (MCMC) simulation methods have successfully overcome these hurdles and fit spatiotemporal random effects for reasonably sized geo-locations of Gaussian fields (GF) (Berliner et al., 2000; Gilks et al., 1996; Casella and Robert, 1999; Wikle, 2003; Wikle et al., 1998). However as the number of geo-locations increases, MCMC computations of a dense GF $m \times m$ spatial correlation matrix become

* Corresponding author at: MRC/Wits Rural Public Health & Health Transitions Research Unit, School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, 7 York road, Parktown 2193, Johannesburg, South Africa.
E-mail address: Eustasius.Musenge@wits.ac.za (E. Musenge).

0303-2434/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jag.2012.04.001

infeasible or extremely slow in the order of power three ($O(m^3)$), this problem is popularly known as the "big m" or "big N" (Banerjee et al., 2004). Several approaches have been used to resolve the "big m". Banerjee et al. (2004) give brief summaries of these: sub-sampling, spectral, lattice, dimension reduction and course fine coupling methods (Banerjee et al., 2004, 2008; Banerjee and Carlin, 2003; Kamman and Wand, 2001; Johnson et al., 1990; French et al., 2002). Generally these techniques attempt to reduce the dimension of the GF by selecting a "representative" sub-sample or fixing some parameters or changing the scale from continuous to discrete with the aim of reducing the computational burden in running MCMC simulations.

We addressed this problem using techniques proposed by Rue and Held (2005) who changed the continuous scale GF to a discrete scale Gaussian Markov Random Field (GMRF), for the Matérn family of covariance structures (Rue and Held, 2005). More recently Lindgren et al. (2011) provides the detail of how the GF and GMRF relate via Stochastic Partial Differential Equations (SPDE) using basis functions (Lindgren et al., 2011; Cameletti et al., 2012). Secondly we performed inference and prediction using Integrated Nested Laplace Approximation (INLA) well suited for GMRF as opposed to the commonly used MCMC (Rue et al., 2009). Hence we greatly reduced the computational burden and could do in hours what usually took days having reduced the computational operations for a spatiotemporal model from power 3 to power 1.5 ($O(m^3) \rightarrow O(m^{3/2})$).

The aim of this paper is to discuss and apply a Bayesian model that can handle large zero-inflated spatiotemporal observational data on mortality producing reliable estimates speedily. In Section 2 we explore the Bayesian methods and model fitting inference, prediction and goodness of fit. Section 3 we apply the discussed Bayesian spatiotemporal model to the data from Agincourt in rural South Africa which has 71,057 children aged 0–9 years living in 15,703 households over the years 1992–2010. In Section 4 we discuss the merits of our model and distil the Public Health implications of our results in interventional studies.

2. Methods

2.1. Spatiotemporal model structure

The outcome $y_i(s_j, t)$ was the observed HIV/TB related death of a child $i = 1, \ldots, N$ from a given household $j = 1, \ldots, m$ in a specific year $t = 1, \ldots, T$ which is a realisation of the spatio-temporal process $y(.,.) \in Y(.,.)$. Assuming the outcomes distribution belongs to the exponential family of distributions, we can fit flexible structural additive models belonging to the generalized linear mixed models (GLMM) (Brezger and Lang, 2006; Fahrmeir and Lang, 2001). Our data may be represented by the equation:

$$y_i(s_j, t) = X(s_j, t)\boldsymbol{\beta} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}, t) + \varepsilon_i(s_j, t) \quad (1)$$

where $X(s_j, t)$ is the design matrix with fixed p covariates, $\boldsymbol{\beta} = (\beta_0, \ldots \beta_p)'$ is the regression coefficients vector, $f(.)$ which is one of the $f^{(j)}$ used to relax the linear relationship or introduce random effects or both and $\varepsilon_i(s_j, t) \sim N(0, \sigma_\varepsilon^2)$ are the error terms which are neither temporally nor spatially correlated (Cameletti et al., 2012). As our data are spatially and temporally correlated we can introduce random effects $f(.) = f(s_j, t)$ a Gaussian random field with a first order autoregressive temporal effect $\xi(s_j, t-1)$ and coefficient ϕ and zero mean multivariate normal (temporally independent) spatial effects $\omega(s_j, t) \sim MVN(0, \Sigma = \sigma_\omega^2 C(||s_j - s_k|| = h); j \neq k)$ resulting in the equation:

$$f(s_j, t) = \phi\xi(s_j, t-1) + \omega(s_j, t) \quad (2)$$

where $|\phi|<1$ in case of stationarity, $\xi(s_j, 1) \sim N(0, \sigma_\omega^2/(1-\phi^2) = 1/\tau_\omega(1-\phi^2))$ and the spatial effect is second order stationary and isotropic. When the spatial correlation follows a Matérn covariance structure we obtain $C(h) = (1/(\Gamma(\nu)2^{\nu-1}))(\kappa h)^\nu K_\nu(\kappa h)$ for the Euclidean distance lags h. The parameter ν measures the degree of smoothness and also the order of the modified Bessel function (when $\nu > 0$) of second kind K_ν and finally $\kappa > 0$ is the scaling parameter with a range $\rho = (\sqrt{8\nu}/\kappa)$ where the spatial correlation is close to 0.1 for each ν (Lindgren et al., 2011; Rue et al., 2009).

2.2. Zero inflated Poisson and Binomial spatiotemporal models

Observational binary outcome data are commonly analysed using the logistic regression model, which has a logit linear predictor in the GLMMs canonical link structure. However this model has problems of instability especially with spatial random effects, which would be exacerbated due to zero inflation (Agarwal et al., 2002). In epidemiological cohort studies a relative risk is more preferred than an odds ratio as this caters for temporality and also a more intuitive measure of burden of morbidity or mortality (Barros and Hirakata, 2003; Fekedulegn et al., 2010). In light of this and in an endeavour to have better fitting models, two models that can handle zero inflation were employed. The two conditionally independent models fit were the zero inflated Poisson and zero inflated Binomial with a log and a logit canonical link functions respectively: $\boldsymbol{y}_t|\boldsymbol{\eta}_t, \theta_t \sim ZIP(\boldsymbol{\mu}_t, \theta_t)$ and $\boldsymbol{y}_t|\boldsymbol{\eta}_t, p_t, \theta_t \sim ZIBin(n_t, p_t, \theta_t)$.

The mortality outcome (count/binary) data $\boldsymbol{y}_t(s_j)$ observed at the households in the Agincourt area are zero inflated and assumed to follow either a Poisson (count) or Binomial distribution (binary). We will occasionally drop the s_j for notational convenience in the rest of the article. We therefore resorted to the zero inflated models to cater for the imbalance due to many zeros. The model that takes care of zero inflation (θ_t) can then be represented as:

$$\pi(y_t|\boldsymbol{\eta}_t, \theta_t) = \begin{cases} \theta_t + (1-\theta_t)\pi(0|\boldsymbol{\eta}_t, \theta_t) & \text{if } \boldsymbol{y}_t = 0 \\ (1-\theta_t)\pi(1|\boldsymbol{\eta}_t, \theta_t) & \text{if } \boldsymbol{y}_t \neq 0 \end{cases} \quad (3)$$

With canonical links of the expected means: $g(E(y_t(.)) = \log(\boldsymbol{\mu}_t) = \boldsymbol{\eta}_t$ and $g(E(y_t(.)) = \text{logit}(p_t) = \boldsymbol{\eta}_t$ with means are $\boldsymbol{\mu}_t = \exp(\boldsymbol{\eta}_t)$ and $p_t = (\exp(\boldsymbol{\eta}_t)/(1+\exp(\boldsymbol{\eta}_t)))$ for the Poisson and Binomial distributions, respectively. A spatiotemporal canonical link (linear predictor) model can be expressed as

$$\begin{aligned} \boldsymbol{\eta}_t(s_j) &= X_t(s_j)\boldsymbol{\beta} + f_t(s_j) + \boldsymbol{\varepsilon}_t(s_j) \\ f_t(s_j) &= \phi\xi_{t-1}(s_j) + \omega_t(s_j) \end{aligned} \quad (4)$$

where $\boldsymbol{\varepsilon}_t(s_j) \sim N(\mathbf{0}, \sigma_\varepsilon^2 I_m)$ with identity matrix I_m of dimension $m \times m$, $\omega_t \sim N(\mathbf{0}, \Sigma = \sigma_\omega^2 \tilde{\Sigma})$ with a stationary AR(1) process $\boldsymbol{\xi}_1 \sim N(\mathbf{0}, \Sigma/(1-\phi^2))$ (Cameletti et al., 2012). The Σ is a dense GF m by m dimensional matrix from a Matérn distribution with scale and smoothness parameters κ and ν (which is fixed in all our computations) respectively. As the size of m increases computations become increasingly more difficult due to the "big m" as previously highlighted.

2.3. Solving the "big m" using the SPDEs to estimate the spatial random effects

To resolve the computational burden associated with the GF Matérn covariance function we used a technique that changes this to a GMRF proposed by Rue and Held (2005). Briefly the locations are converted into areal triangulations firstly by making them the initial triangle vertices before adding more vertices for proper triangulation which extends the grid and very useful for prediction. Fig. 1 shows how we employed triangulation to our

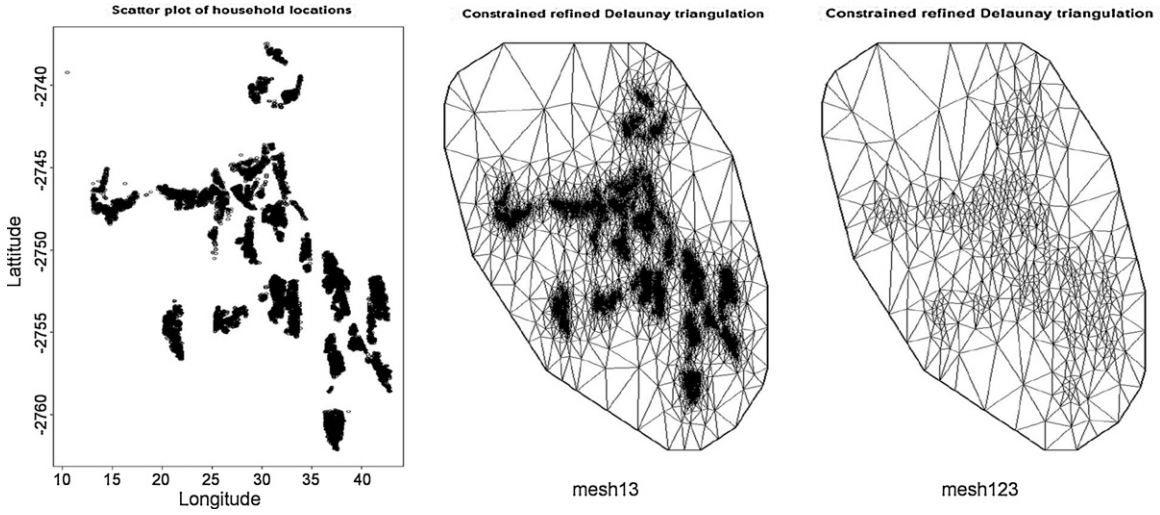


Fig. 1. Agincourt original household locations (left), triangulation of all household (centre) and triangulation of households within 500 m (right).

data, the diagram on the right was made assuming households within 500 m were similar we fit 488 vertices and 938 triangles.

The SPDE technique redefines the Matérn field as a finite "representative" linear combination of basis functions on a triangulation of the locations (Cameletti et al., 2012). Hence for our spatiotemporal random effects in GMRF representation we get $\tilde{\omega}_t \sim N(\mathbf{0}, \Sigma = Q_s^{-1})$ and $\boldsymbol{\xi}_1 \sim N(\mathbf{0}, (Q_s^{-1}/(1-\phi^2)) = Q_T^{-1})$ where Q_s is a sparse time-invariant precision matrix with dimension m^* vertices from the triangulations. A joint spatiotemporal GMRF $\tilde{f}_t \sim N(\mathbf{0}, Q^{-1} = (Q_T \otimes Q_s)^{-1})$ whose precision matrix is the Kronecker product of the temporal and spatial precision matrices; is such that $\tilde{f}_t \approx \boldsymbol{B}f_t$ where the basis $\boldsymbol{B}$ is a sparse matrix with unit elements for matching triangle vertices and zero's elsewhere (Lindgren et al., 2011). Therefore Eq. (4) becomes $\boldsymbol{\eta}_t = X_t\boldsymbol{\beta} + \boldsymbol{B}f_t + \boldsymbol{\varepsilon}_t$, where we let $\boldsymbol{x}_t = \{\boldsymbol{\beta}, \boldsymbol{B}, \theta_t\}$ be the vector of latent Gaussian fields and $\boldsymbol{\varphi}_t = \{\sigma_\omega^2, \phi, \kappa, \sigma_\varepsilon^2\}$ being a vector of unknown parameters. We can thus express our model into a hierarchical Gaussian latent variable fashion as follows, stage 1 – observational equation, stage 2 – latent Gaussian field and stage 3 – parameter model (Simpson et al., 2011):

$$\begin{aligned}
&\text{stage 1}: \quad \boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\varphi}_t \sim N(X_t\boldsymbol{\beta} + \boldsymbol{B}f_t; \boldsymbol{\sigma}_\varepsilon^2 I_m) = N(\boldsymbol{\mu}_{\boldsymbol{y}_t}(\boldsymbol{\varphi}_t) = A\boldsymbol{x}_t, Q_{\boldsymbol{y}_t}^{-1}(\boldsymbol{\varphi}_t)) \\
&\text{stage 2}: \quad x_t|\boldsymbol{\varphi}_t \sim N(\boldsymbol{\mu}_{\boldsymbol{x}_t}(\boldsymbol{\varphi}_t); Q_{\boldsymbol{x}_t}^{-1}(\boldsymbol{\varphi}_t)) \\
&\text{stage 3}: \quad \boldsymbol{\varphi}_t \sim \boldsymbol{\pi}(\boldsymbol{\varphi}_t)
\end{aligned} \tag{5}$$

where the precision matrices $Q_{\cdot}^{-1}(.)$ are either small enough (for easier multiple factorisation) or sparse (Simpson et al., 2011b). These models cover a wide range of models and are easily estimable using INLA as shown in the next subsection.

2.4. Bayesian inference using INLA

In accordance with the Bayesian paradigm we aim to find the posterior distribution of the processes and parameters updated by data (Wikle, 2003). This could be expressed as:

$$\begin{aligned}
&\textit{Probability}\ (\textit{process}, \ \textit{parameters}|\textit{data}) \propto \textit{Likelihood} \\
&\times(\textit{data}|\textit{process}, \ \textit{parameters}) \times \textit{Probability}\ (\textit{process}|\textit{parameters}) \\
&\times \textit{Probability}\ (\textit{parameters})
\end{aligned}$$

Applying this expression to our model, letting $\Theta = \{\boldsymbol{\beta}, \phi, \theta_t, \sigma_\omega^2, \kappa, \sigma_\varepsilon^2\}$ denote the vector of all parameters and dropping the subscripts to present in vector form, $\boldsymbol{\xi} = \{\xi_t\}$ and data $\boldsymbol{y} = \{y_t\}$ (Cameletti et al., 2012; Rue et al., 2009), their joint posterior distribution is thus:

$$\pi(\Theta, \boldsymbol{\xi}, \boldsymbol{\eta}|y) \propto \pi(y|\Theta, \boldsymbol{\xi}, \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\boldsymbol{\xi}, \Theta)\pi(\boldsymbol{\xi}, \Theta) \tag{6}$$

Fig. 2 gives a simplified pictographical view, where level 1 are the data and assumed distributions, level 2 is a process, level 3 and 4 are parameters and level 5 gives the default hyper-parameters used in the INLA package.

The posterior marginals are required, standard (i), nested approximation (ii) and numerical integrations (iii) for latent fields $\boldsymbol{x} = \{\beta, f, \theta\}$ and hyper-parameters $\varphi = \{\phi, \sigma_\omega^2, \kappa, \sigma_\varepsilon^2\}$ respectively (Rue and Held, 2005; Rue et al., 2009):

$$\begin{aligned}
&i)\pi(\boldsymbol{x}, \boldsymbol{\varphi}|\boldsymbol{y}) = \int \pi(\boldsymbol{x}_i|\boldsymbol{\varphi}, y)\pi(\boldsymbol{\varphi}|\boldsymbol{y})\,d\boldsymbol{\varphi} \qquad \pi(\varphi_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\varphi}|\boldsymbol{y})\,d\boldsymbol{\varphi}_{-j} \\
&ii)\tilde{\pi}(\boldsymbol{x}, \boldsymbol{\varphi}|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{x}_i|\boldsymbol{\varphi}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{y})\,d\boldsymbol{\varphi} \qquad \tilde{\pi}(\varphi_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{y})\,d\boldsymbol{\varphi}_{-j} \\
&iii)\tilde{\pi}(\boldsymbol{x}, \boldsymbol{\varphi}|\boldsymbol{y}) = \sum_k \tilde{\pi}(\boldsymbol{x}_i|\boldsymbol{\varphi}_k, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\varphi}_k|\boldsymbol{y})\Delta_k \quad \tilde{\pi}(\varphi_j|\boldsymbol{y}) \propto \int \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\varphi}|\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\varphi}, \boldsymbol{y})}\right|_{x=x^*(\varphi)} d\boldsymbol{\varphi}_{-j}
\end{aligned} \tag{7}$$

Using this technique we aim to initially get the terms "nested" inside the integrand in Eq. (7ii) left hand side without integration (Simpson et al., 2011b). To do so we firstly estimate the marginal $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\varphi}, \boldsymbol{y})$ which is a Gaussian approximation of $\boldsymbol{x}$ with mode $\boldsymbol{x}^*(\boldsymbol{\varphi})$, for a given $\boldsymbol{\varphi}$. Secondly we estimate $\tilde{\pi}(x_i|\boldsymbol{\varphi}, \boldsymbol{y}) = N(x_i; \mu_i(\boldsymbol{\varphi}), \sigma_i^2(\boldsymbol{\varphi}))$ using either Gaussian or Laplace or a simplified Laplace approximations (Rue et al., 2009). We computed these marginals using the R package Integrated Nested Laplace Approximation (INLA) which uses the simplified Laplace approximations (Rue and Held, 2005; Rue et al., 2009; R-cran, 2010). The INLA procedure also enables easier spatial prediction since it computes posterior conditionals for the spatial random effects on all triangulation vertices including the extensions as shown in Fig. 1.

2.5. Model goodness of fit and convergence diagnostics

We assessed the accuracy of $\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{y})$ using the effective number of parameters, which can be approximated as the difference between the dimension of the normalised integral $\{\tilde{\pi}(\boldsymbol{\varphi}|\boldsymbol{y})\}$ n and the trace of the product of the prior precision matrix and the posterior covariance (Spiegelhalter et al., 2002):

$$p_D(\boldsymbol{\varphi}) \approx n - tr\{Q(\boldsymbol{\varphi})Q*(\boldsymbol{\varphi})^{-1}\} \tag{8}$$

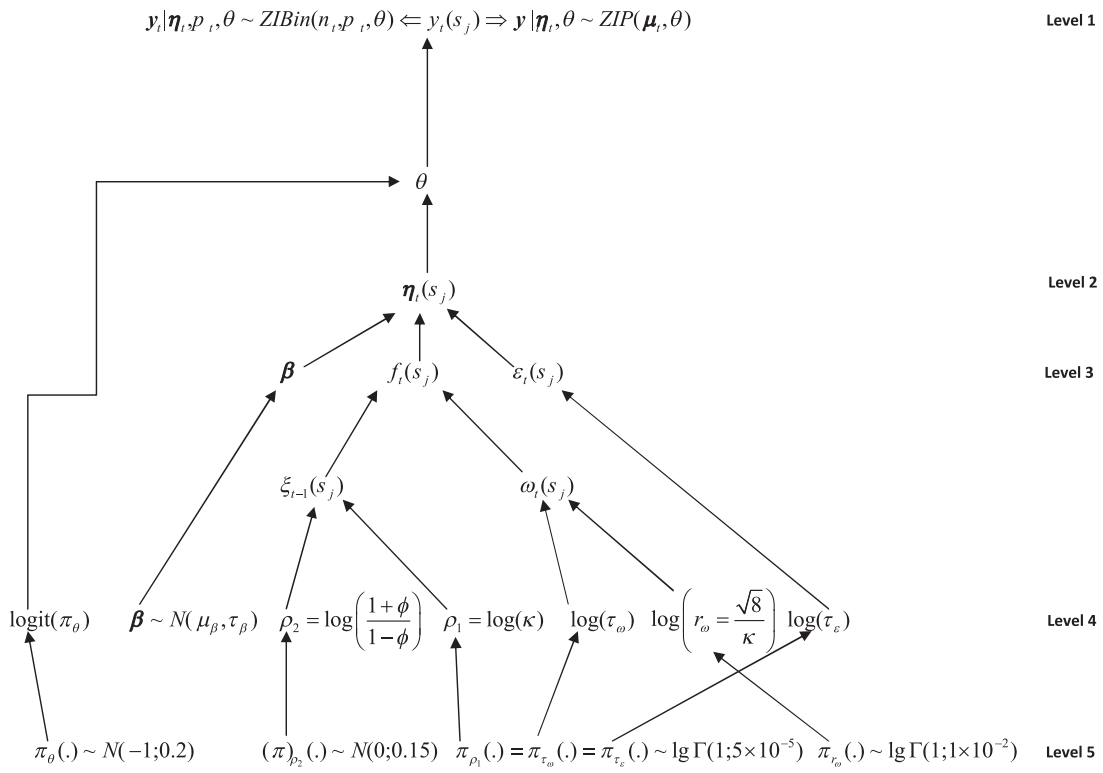


Fig. 2. Hierarchical structure of a zero inflated spatiotemporal model fit using INLA.

The deviance information criteria (DIC) was also used which is defined as the difference between twice the mean of the deviance and the deviance of the mean according to Spiegelhalter et al. (2002) and expressed as:

$$D(x, \boldsymbol{\varphi}) = -2\sum_i \log\{\tilde{\pi}(y_i|x_i, \boldsymbol{\varphi})\} + constant \quad (9)$$

Interpretations of these is quite straight forward the smaller the effective number of parameters the more parsimonious the model and the smaller the DIC the better the model fit, more-so most parsimonious is not always the best model.

3. Application

3.1. Rural South Africa Agincourt HDSS data, study design and ethics

The Agincourt health and demographic surveillance system (HDSS) site was set up in the Agincourt sub-district in 1992 due to its remote location, availability of several clinics and presence of Mozambican in-migrants (Tollman et al., 1999). By 2010, the Agincourt HDSS had a population of over 84,000 persons living in approximately 17,000 households scattered throughout 27 neighbouring villages. Cause of death data were obtained through verbal autopsies conducted on every recorded death (Clark et al., 2007). Interviews were conducted by trained field worker. This was done

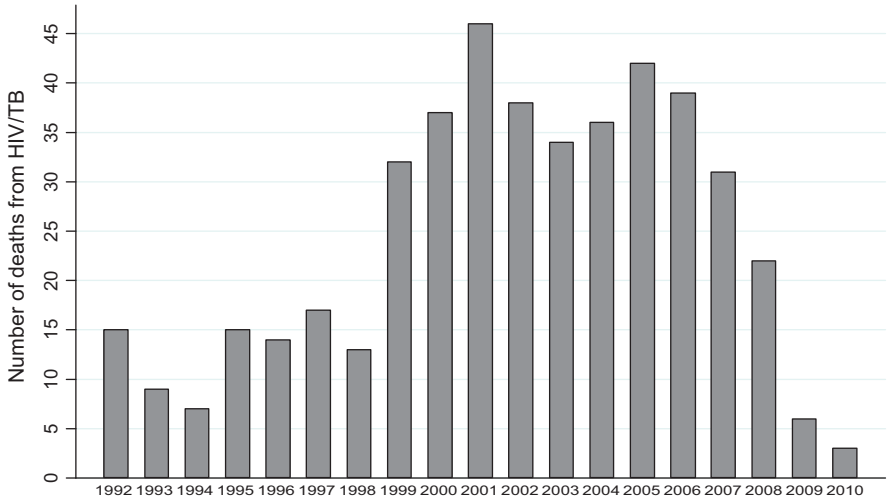


Fig. 3. Year specific child deaths due to HIV/TB from 1992 to 2010.

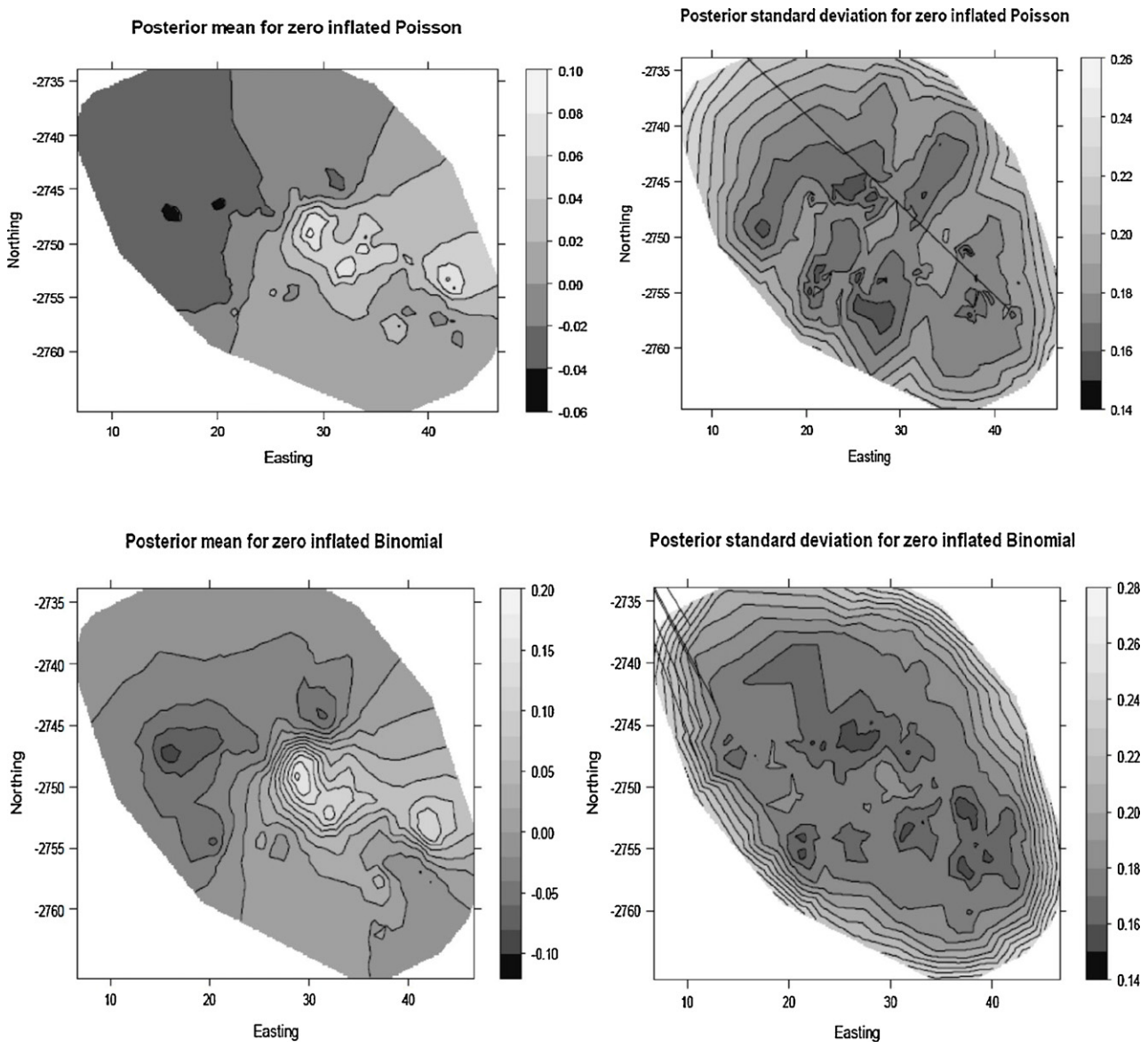


Fig. 4. Posterior point estimates for ZIP (top) and ZIB (bottom) models.

within 1 year after a death, with the closest caregiver of the deceased in their mother tongue. Cause of death was independently determined by two medical practitioners with the third as a tie breaker. Their consensus cause of death was classified according to the World Health Organization's International Classification of Diseases (ICD10) (Kahn et al., 2000). HIV/TB mortality in children was ascertained by the reported signs and symptoms, and in some instances this was verified through the mother's cause of death (Kahn et al., 2000). Over 90% of the HDSS households were geo-coded by 1992 and by 2010 all the households were geo-coded, thus enabling spatial analyses at household as well as village level. The study design was a retrospective cohort study covering households observed from the onset of the site to December 2010. The Agincourt HDSS site was granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (No. 960720). This work was also granted ethical clearance by the University of the Witwatersrand's Committee for Research on Human Subjects (M081145). Verbal informed consent was obtained when the census rounds were conducted and also when verbal autopsy data were collected from a close relative of the deceased.

3.2. Dependent and independent variables

The persons included in the study were all the children aged between 0 and under 10 years who lived or had lived in the Agincourt HDSS between January 1992 and December 2010. The independent variables used were: child's gender, birth-weight category, age category, slum (availability of water, electricity and toilet) and number of mother's antenatal clinic visits during pregnancy and year of observation. The household's latitude and longitude were used to construct the latent variables for the spatial random effects and year the AR(1) temporal effects. The dependent variable was death due to HIV and or tuberculosis (TB) determined by the WHOs ICD10 verbal autopsy codes A16-A19[1] for HIV and B20-B24[2] for TB. These data were extracted using Structured Query

[1] A16 = respiratory tuberculosis; not confirmed bacteriologically or histologically; A17 = tuberculosis of nervous system; A18 = tuberculosis of other organs; A19 = miliary tuberculosis.

[2] B20 = human immunodeficiency virus (HIV) disease resulting in infectious and parasitic diseases; B21 = human immunodeficiency virus (HIV) disease resulting in

Table 1
Univariate and multiple regression results models using zero inflated Poisson adjusting for spatiotemporal random effects.

Variable	Summary, *n* (%) or mean ± SD	Univariate results, RR (95% credible interval)	Non-spatial multiple variable model, adjusted RR (95% credible interval)	Temporal multiple variable model, adjusted RR (95% credible interval)	Spatial multiple variable model, adjusted RR (95% credible interval)	Spatiotemporal multiple variable model, adjusted RR (95% credible interval)
Sex						
Male	35,317(49.70)	1.00				
Female	35,740(50.30)	0.89(0.74;1.07)				
Birth weight category						
Low weight	6,320(8.89)	1.00	1.00	1.00	1.00	1.00
Moderate weight	15,241(21.45)	0.81(0.61;1.07)	0.70*(0.53;0.93)	0.85(0.63;1.16)	0.71*(0.53;0.93)	0.85(0.63;1.16)
High weight	49,496(69.66)	0.39*(0.30;0.51)	0.57*(0.42;0.76)	0.73*(0.53;0.99)	0.58*(0.43;0.77)	0.73*(0.53;0.99)
Age category						
0–1 years	8,580(12.07)	1.00	1.00	1.00	1.00	1.00
1–5 years	19,619(27.61)	0.21*(0.17;0.26)	0.20*(0.16;0.25)	0.18*(0.23;0.64)	0.20*(0.16;0.25)	0.18*(0.14;0.22)
5–9 years	42,858(60.31)	0.05*(0.04;0.07)	0.04*(0.03;0.05)	0.04*(0.05;0.54)	0.04*(0.03;0.05)	0.04*(0.03;0.05)
Slum (electricity, water and toilet)						
None of the three	10,051(14.14)	1.00	1.00		1.00	
At least one	27,329(38.46)	0.78(0.60;1.01)	0.79(0.61;1.02)		0.77(0.60;1.00)	
At least two	21,871(30.78)	0.74*(0.56;0.97)	0.73*(0.56;0.96)		0.71*(0.54;0.93)	
All three	5,748(8.09)	0.57*(0.37;0.86)	0.55*(0.35;0.83)		0.53*(0.34;0.81)	
Antenatal clinic visits	6.00(±1.50)	0.98*(0.97;0.99)	0.98*(0.97;0.99)	0.96*(0.94;0.97)	0.98*(0.96;0.99)	0.96*(0.94;0.97)
Zero inflation parameter			*0.06*(0.002;0.25)*	*0.05*(0.003;0.18)*	*0.07*(0.003;0.28)*	*0.04*(0.002;0.15)*
Precision for year				*0.91*(0.23;2.19)*		*0.86*(0.16;2.32)*
Rho for year				*0.88*(0.72;0.97)*		*0.89*(0.70;0.98)*
Kappa					1.98*(1.09;3.66)	*1.55*(1.41;1.63)*
Tau					1.37*(0.18;2.600	*2.31*(1.19;9.08)*
Moran's Indexes: Observed(Expected ± standard deviation)						*0.32*$(2.1 \pm 3.33 \times 10^{-3})$*
Effective number of parameters			*12.05(0.05)*	*25.61(1.10)*	*23.18(5.04)*	*29.16(6.33)*
DIC			*4855.31*	*4532.69*	*4846.13*	*4532.31*

* Statistical significance at the 5% level.

Language (SQL). Preliminary data analyses and data management were done using STATA 10.1 (StataCorp, 2007). The Bayesian analysis was done using an R software package called INLA (Rue et al., 2009; R-cran, 2010).

3.3. Descriptive statistics

There were similar proportions of boys and girls (50.3%) with an average age of 6.14 years (standard deviation of 3.38). Almost 9% of the births were low-weight, the majority (60.3%) of the children were only observed after their 5th birthday. Only 8.09% had access to all three: clean tap water, flush toilet inside house and electricity, which is expected in typical rural areas. Health check-ups during pregnancy averaged about 6 visits, with the majority (73%) still not going at all. A total of 456 deaths were HIV/AIDS (including HIV/tuberculosis) which was a small proportion of the population indicative of the presence of zero inflation. Mortality was moderate in the early to mid 1990s before gradually growing from 1999 and reaching a peak in 2001, remaining high for several years before gradually declining from 2007 as shown in Fig. 3.

3.4. Zero inflated Poisson and Binomial spatiotemporal modelling results

We performed univariate and multiple variable analyses using both the ZIP and ZIB models. Multiple variable models were fit systematically starting with one without random effects, spatial only, temporal only and finally spatiotemporal random effects. The R-codes using INLA are given in Appendix 1. From our results shown in Tables 1 and 2, our discussion will be centred on the ZIP spatiotemporal model which was the best fitting model since this had the lowest DIC (4532.31) and catered for spatiotemporal random effects, as later indicated in Section 3.5.

Two variables consistently showed significant associations with child HIV/TB mortality, firstly the age category of the child, which showed a decrease in mortality with increase in age (Chi-Square p-value < 0.001). Those aged 1 to under 5 (127/19,619) were 82% {0.18(0.14; 0.22)} less likely to experience a death due to HIV/TB compared to those under 1(262/8580) and those over 5 and under10 (67/42,858) were 96%{0.04(0.03;0.05)} less likely to die

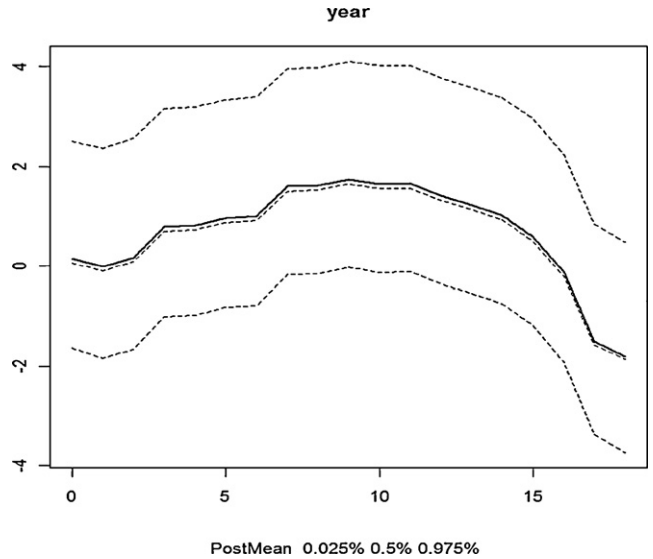


Fig. 5. Posterior means (bold), medians (0.5% = 50th percentiles, middle), lower credible limits (0.025% = 2.5th percentiles) and upper credible limits (0.975% = 97.5th percentiles).

malignant neoplasms; B22 = human immunodeficiency virus (HIV) disease resulting in other specified diseases; B23 = human immunodeficiency virus (HIV) disease resulting in other conditions; B24 = unspecified human immunodeficiency virus (HIV) disease.

Table 2
Multiple regression results of four models using zero inflated Binomial adjusting for spatiotemporal random effects.

Variable	Non-spatial multiple variable model, adjusted OR (95% credible interval)	Temporal multiple variable model, adjusted OR (95% credible interval)	Spatial multiple variable mode, adjusted OR (95% credible interval)	Spatiotemporal multiple variable model, adjusted OR (95% credible interval)
Sex				
Male				
Female				
Birth weight category				
Low weight	1.00	1.00	1.00	1.00
Moderate weight	0.70*(0.52;0.92)	0.86(0.62;1.20)	0.70*(0.52;0.93)	0.005*(0.003;0.07)
High weight	0.62*(0.46;0.83)	0.67*(0.48;0.95)	0.63*(0.47;0.85)	0.05*(0.007;0.04)
Age category				
0–1 years	1.00	1.00	1.00	1.00
1–5 years	0.07*(0.05;0.09)	0.05*(0.04;90.07)	0.07*(0.06;0.09)	0.18*(0.14;0.22)
5–9 years	0.006*(0.004;0.008)	0.005*(0.003;0.95)	0.006*(0.004;0.008)	0.04*(0.03;0.05)
Slum (electricity, water and toilet)				
None of the three	1.00		1.00	
At least one	0.77(0.60;1.01)		0.76*(0.58;0.99)	
At least two	0.71*(0.54;0.95)		0.69*(0.52;0.93)	
All three	0.52*(0.34;0.81)		0.51*(0.32;0.78)	
Antenatal clinic visits		0.94*(0.92;0.96)	0.97*(0.95;0.98)	0.94*(0.93;0.96)
Zero inflation parameter	*0.22*(0.042;0.66)*	*0.43*(0.19;0.79)*	*0.15*(0.031;0.45)*	*0.42*(0.18;0.67)*
Precision for year		*0.78*(0.17;1.96)*		*0.80*(0.20;1.93*
Rho for year		*0.87*(0.67;0.97)*		*0.87*(0.69;0.97)*
Kappa			1.98*(1.14;3.55)	*1.54*(1.41;1.62)*
Tau			1.37*(0.27;2.54)	*2.23*(1.22;6.91)*
Moran's Indexes: Observed(Expected ± standard deviation)				*0.30*($2.1 \pm 3.33 \times 10^{-3}$)*
Effective number of parameters	*11.10(0.080)*	*24.58(1.03)*	*22.97(5.09)*	*31.12(5.18)*
DIC	*5179.43*	*4754.60*	*5180.50*	*4803.93*

* Statistical significance at the 5% level.

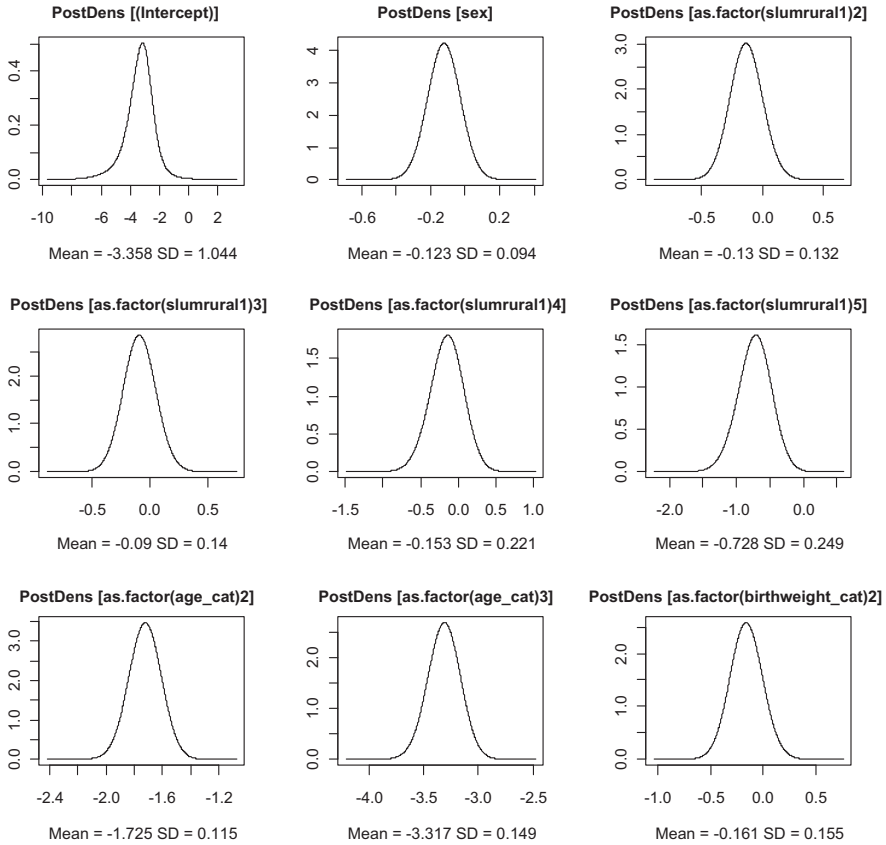


Fig. 6. Posterior density plots of log-means for fixed effects of the ZIP spatiotemporal model.

from HIV/TB relative to the under 1 s keeping all other variables constant. Secondly the category of the birth weight was significantly associated with deaths due to HIV/TB, showing a decreasing trend with greater birth weight (Chi-Square p-value < 0.001). Those who had a birth weight of 3.5 kg and above (232/49,496) were 27% {0.73(0.53;0.99)} less likely to die due to HIV/TB relative to those who were born with a low birth weight of less than 2.5 kg (76/6329) independent of other risk factors. The only other significant protective factor was the number of antenatal clinic visits made by the mother; for every additional visit the mother attended the risk of losing the child from HIV/TB decreased by 4% {0.96(0.94;0.97)} keeping other variables constant as shown by the spatial only and no-random effects model. Those who had access to all three electricity, water and flush toilet (who in this area are the more affluent) experienced fewer deaths due to the HIV/TB compared to those who had none. We show the posterior marginal point estimates from the ZIP and ZIB models in Fig. 4 top and bottom respectively.

As shown in Fig. 4, the region with the greatest risk of child HIV/TB mortality is the central region with a trend towards the south-westerly. In the context of our findings, this is the region whose households have lower birth weights, greater under 1 mortality and having lower visits to the health facility compared to households in other regions. The lowest risk region is the top north-easterly with the ZIP model showing few clusters than the ZIB model. The standard error maps indicate the greatest errors on the peripherals which are also the regions where the extensions of the triangulation shown in Fig. 1. Also consistent with our results from Tables 1 and 2 there is a greater margin of error with the ZIB model as compared with the ZIP model.

3.5. Model assumptions, goodness of fit and convergence diagnostics

Using the final spatiotemporal models in both ZIB and ZIP we investigated the presence of zero inflation, temporal and spatial correlation. The null hypothesis for zero inflation was that the inflation parameter θ_t is zero (no zero inflation) our results in Tables 1 and 2 indicate significant presence of zero inflation of 0.04 and 0.42 for ZIP and ZIB respectively. Investigation of temporality was done by testing the ϕ coefficient of the first order autocorrelation model with null hypothesis that the process was not stationary $|\phi|>1$. Both our models indicated significance presence of stationarity that is 0.86 and 0.87 for ZIP and ZIB respectively. Fig. 5 shows the posterior means plus 95% credible bands, time series plots from the modelling which resembles the observed data shown in Fig. 3. Lastly on checking model assumptions we investigated the presence of spatial effects by inspecting the significance of the components (κ and τ) of the Matérn covariance (see Tables 1 and 2). Also this was investigated with the classical Moran's indices on the spatial residuals testing the null hypothesis that there is zero spatial autocorrelation which from both tests yielded significant results, hence rejecting the null.

We compared the two models the ZIB and ZIP and found that the latter fitted better in all the variants from the no-random effects

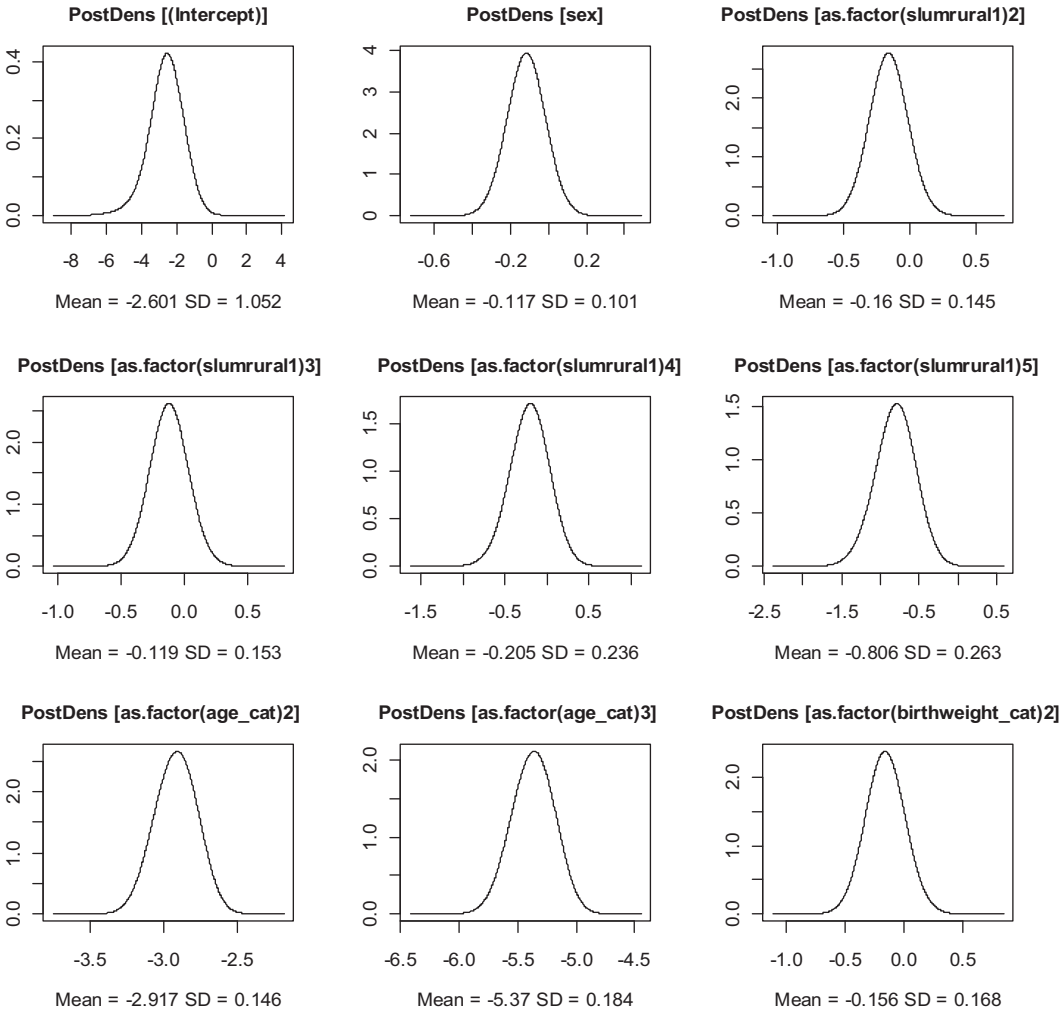


Fig. 7. Posterior density plots of log-means for fixed effects of the ZIB spatiotemporal model.

model to the spatiotemporal as this yielded smaller DICs and model parsimony. Some of the parameter estimates were similar, except for the zero inflated parameter, age categories and birth weights. The ZIP model also showed more stable estimates and consistent results in the adjusted relative risks (adjusted RRs) compared to the ZIB and narrower credible intervals. As a way of showing the convergence pictographically for some parameter estimates we show the fixed effects logarithmic form posterior means density plots in Figs. 6 and 7.

The plots in Figs. 6 and 7 show that the parameters fitted well as GMRF since they are symmetrical about thus the mean equals the mean and mode.

4. Conclusions

The Public Health rational of our results is centred on lack of health seeking behaviour in rural areas. Interventions could target two groups of mothers (also caregivers) those not yet infected and those already infected. For both groups maternal attendance of health facility can minimize the risk of under one mortality and possibly giving birth to low-weight children. For those mothers' already infected several options are there; before child birth (prevention of mother to child) and after birth (anti-retro-virals – ARV uptake) and also early treatment on ARVs for the parent to increase their longevity. A significant decline in mortality can already be observed in our cohort post 2007 the year when ARVs were started in the area. Poverty however still remains the greatest challenge in the developing countries and mostly interventions when available are very scarce. As such our results from spatiotemporal Bayesian modelling coupled with maps can be very handy in allocation of the limited resources of aid.

Modelling that controls for potential confounding is an add-on to epidemiological studies and gives strength to the estimates derived from such modelling procedures. Our modelling approach caters for potential spatial and temporal confounders. This approach was able also to cater for "large" zero inflated spatiotemporal data. The years 1992–2010 cover a span of 19 years (76%) of the United Nation's Millennium Development Goals (MDG) period of 25 years. Goal 4.1 aims at reducing child (under five) mortality by two thirds over that period (Black et al., 2003). Our results demonstrate such a significant decline in mortality due to HIV/TB.

A major limitation in developing nations is lack of reliable and user friendly analyses software, which was addressed by using INLA a package available on the public domain. Zero inflated Poisson (ZIP) and zero inflated Negative Binomial (ZINB) models were used for log link function linear predictors and also catered for spatial random effects (Lambart, 1992; Ridout et al., 2001). It has also been noted from other studies that the log link function linear predictor models were more stable than logit link function models when they contained a spatial component. This finding was also confirmed in our study as we observed that models from the exponential family with a log-linear predictor were more stable and able to converge better than the logit-link function model. This motivated us to treat the outcome as a discrete measure as opposed to binary mostly used for logistic regression (Agarwal et al., 2002). We demonstrated that the "big m" can be resolved with great ease for Public Health mortality hierarchical structures with the aid of SPDEs and INLA.

Contributions

EM drafted the manuscript and did the statistical analysis. KK provided critical intellectual content and TFC and PV the statistical content. KK was also responsible for the design and quality of the Agincourt verbal autopsy data. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgements

This work was made possible by a doctoral fellowship support from World Health Organisation's, Tropical Disease Research (WHO/TDR) from 2008 to 2010 the Swiss South African Joint Research Program (SSAJP), project number JRP IZLSZ3122926, from 2009 to 2010. The Agincourt HDSS is funded by the Wellcome Trust UK (Grant number 058893/Z/99/A, 069683/Z/02/Z and 085477/Z/08/Z) and the University of the Witwatersrand and Medical Research Council, South Africa, and the Andrew Mellon and Hewlett Foundations, U.S.A. Special thanks goes to Benn Sartorius for data extraction, community leaders, field workers, team supervisors and the Agincourt communities. A very special thanks to Finn Lindgren for answering all my questions promptly whilst I developed models using INLA.

Appendix A. Zero inflated Poisson R-code using INLA adapted from http://www.r-inla.org/ (Lindgren et al., 2011)

```
#1.#IMPORTING LIBRARIES
library(foreign)
library(INLA)
library(coda)
library(fields)

#2.# Read data:
decdata = read.dta("c:/jagnovredo/childleveldec.dta")
attach(decdata)
names(decdata)

#3.# Build triangular mesh of locations within 0.5 km see figure 1
mesh123 = (inla.mesh.create(cbind(decdata$coordlongx/1000,
decdata$coordlaty/1000),refine=TRUE,cutoff=0.5))
plot(mesh123)


#4.# Store the data-->vertex mapping:
## ( mesh$idx$loc[m] is the mesh vertex for data location number m )
decdata$hhid = mesh123$idx$loc

#5.# Create the SPDE/GMRF model, (kappa^2-Delta)(tau x) = W:
spde123 = inla.spde.create(mesh123, model="matern")

#6.# Functions of four models multivariate models fit
#6a.No random effects model with household and individual random effects
norand=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(birthwei
ght_cat)+antenatalvisits+f(hhid, model="iid")+f(idno, model="iid")

#6b.Temporal random effects model with household and individual random
effects
temponly=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(birthw
eight_cat)+antenatalvisits+f(year, model="ar1")+f(hhid, model="iid")+f(idno,
model="iid")

#6c.Spatial random effects model with individual random effects
spatonly=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(birthw
eight_cat)+antenatalvisits+f(idno,model="iid")+f(hhid,model=spde123)

#6d.Spatiotemporal random effects model with individual random effects
spatiotemp=hivtb2~sex+as.factor(slumrural1)+as.factor(age_cat)+as.factor(birt
hweight_cat)+antenatalvisits+f(year, model="ar1")+f(idno,
model="iid")+f(hhid,model=spde123)


#7.#  Running INLA models on the four stated in step 6
norand  = inla(norand, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals.pr
edictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(norand)

spatonly  = inla(spatonly, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals.pr
edictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(spatonly)

temponly  = inla(temponly, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals.pr
edictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(temponly)
```

```
spatiotemp  = inla(spatiotemp, family="zeroinflated.poisson1",data =
decdata,control.results=list(return.marginals.random=TRUE,return.marginals.pr
edictor=TRUE),control.compute=list(dic=TRUE,cpo=TRUE))
summary(spatiotemp)

#Saving output for model to store posterior estimates and other
save(spatiotemp, file="c:/resultsdec/spatiotemp.Rdata")

#-------------------------------------------------------------------------#

#8.#Loading data from previous analyses in new session
spatiotemp<-load(file="c:/jagnovredo/spatiotemp.Rdata")
names(spatiotemp)

#9.#Assessing for posteriors  Spatial auto-correlation (Moran's I)
library(ape)
D.dists <- as.matrix(dist(cbind(mesh123$loc[,1], mesh123$loc[,2])))
D.dists.inv <- 1/D.dists
diag(D.dists.inv) <- 0
Moran.I(spatiotemp$summary.random$hhid[,"mean"], D.dists.inv)


#10.# Diagnostic plots see figures 6 or 7
plot.inla(spatiotemp,
             plot.fixed.effects = TRUE,constant=FALSE,
             plot.lincomb = TRUE,
             plot.random.effects = TRUE,
             plot.hyperparameters = TRUE,
             plot.predictor = TRUE,
             plot.q = TRUE,
             plot.cpo = TRUE)


#11.# Extract the SPDE parameters, point and interval estimates
tauconf = exp(spatiotemp$summary.hyperpar[2,])
kappaconf = exp(spatiotemp$summary.hyperpar[3,]/2)

tau = exp(spatiotemp$summary.hyperpar[2,"mean"])
kappa = exp(spatiotemp$summary.hyperpar[3,"mean"]/2)

#12a.# Get the spatial precision matrix (Qs) see equation section 2.1 and 2.3
Q = inla.spde.query(spde123, precision=list(tau=tau,
kappa2=kappa^2))$precision


#12b.# Reference point for covariance/correlation comparisons:
ref.s = (which.min((mesh123$loc[,1]-mean(range(mesh123$loc[,1])))^2 +
                   (mesh123$loc[,2]-mean(range(mesh123$loc[,2])))^2))

#12c.# Calculate covariances (S) and correlations (SS):
S = solve(Q)
SS = diag(1/sqrt(diag(S))) %*% S %*% diag(1/sqrt(diag(S)))
D = as.matrix(dist(mesh123$loc))

#12d.# Theoretical Matérn correlations and covariances:
dd = (0:1000)/1000
SS.theory = (dd*kappa)*besselK(dd*kappa,1)
SS.theory[1] = 1
S.theory = SS.theory/(4*pi*kappa^2)/tau^2

#-------------------------------------------------------------------------#
```

```
#13.# Prediction, plotting posterior prediction maps figure 4

## Calculate mapping between triangulation vertices and grid points:
## Resolution for gridded output was dims=c(300,300)
proj = inla.mesh.projector(mesh123, dims=c(300,300))

## Construct greyscale palette function:
my.grey.palette = function (n,...) { return (grey.colors(n,0.05,0.95,...))}
## Use it:
my.palette = my.grey.palette

## Construct map data appropriate for easy plotting:
mm = calc.map(map)

## Plot results:

par(mfrow=c(1,2))
## Map resulting posterior mean field to a grid:
plotdata = inla.mesh.project(proj, spatiotemp$summary.random$hhid[,"mean"])
## Plot PM contours:

postmap = (levelplot(row.values=proj$x, column.values=proj$y, x=plotdata,
                  mm=mm, panel=levelplotmap,
                  col.regions=my.palette,
                  xlim=range(proj$x), ylim=range(proj$y), aspect="iso",
                  contour=TRUE, cuts=11, labels=FALSE, pretty=TRUE,
                  xlab="Easting",ylab="Northing", main="Posterior mean for
zero inflated Poisson"))
plot(postmap)


## Map resulting posterior standard deviation field to a grid:
plotdata = inla.mesh.project(proj, spatiotemp$summary.random$hhid[,"sd"])
## Plot std.dev. contours:
#dev.new()
postmap2 = (levelplot(row.values=proj$x, column.values=proj$y, x=plotdata,
                  mm=mm, panel=levelplotmap,
                  col.regions=my.palette,
                  xlim=range(proj$x), ylim=range(proj$y), aspect="iso",
                  contour=TRUE, cuts=11, labels=FALSE, pretty=TRUE,
                  xlab="Easting",ylab="Northing",main="Posterior standard
deviation for zero inflated Poisson"))

plot(postmap2)
```

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics 9, 341–355.

Banerjee, S., Carlin, B., 2003. Semiparametric spatio-temporal frailty modeling. Environmetrics 14, 523–535.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society, Series B, 70 (Part 4), 825–848.

Barros, A.J.D., Hirakata, V.N., 2003. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology 3, 21.

Berliner, L.M., Wikle, C.K., Cressie, N., 2000. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. Journal of Climate 13, 3953–3968.

Black, R.E., Morris, S.S., Bryce, J., 2003. Where and why are 10 million children dying every year? Lancet 361, 2226–2234.

Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis 50, 967–991.

Cameletti, M., Lindgren, F., Simpson, D., Rue, H., 2012. Spatio-temporal modelling of particulate matter concentration through the SPDE approach. AStA Advances in Statistical Analysis, accepted.

Casella, G., Robert, C.P., 1999. Monte Carlo Statistical Methods. Springer-Verlag, New York.

Clark, S.J., Collinson, M.A., Kahn, K., Drullinger, K., Tollman, S.M., 2007. Returning home to die: circular labour migration and mortality in South Africa. Scandinavian Journal of Public Health Supplement 69, 35–44.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.

Fahrmeir, L., Lang, S., 2001. Bayesian inference for generalized additive mixed models based on Markov Random Field priors. Journal of the Royal Statistical Society Series C-Applied Statistics 50, 201–220.

Fekedulegn, D., Andrew, M., Violanti, J., Hartley, T., Charles, L., Burchfiel, C., 2010. Comparison of statistical approaches to evaluate factors associated with metabolic syndrome. The Journal of Clinical Hypertension 12, 365–373.

French, J.L., Kammann, E.E., Wand, M.P., 2002. Comment on paper by Ke and Wang. Journal of American Statistical Association 96, 1285–1288.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov chain Monte Carlo in Practice. Chapman & Hall/CRC.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin designs. Journal of Statistical Planning and Inference 26, 131–148.

Kahn, K., Tollman, S.M., Garenne, M., Gear, J.S., 2000. Validation and application of verbal autopsies in a rural area of South Africa. Tropical Medicine and International Health 5, 824–831.

Kamman, E., Wand, M., 2001. Geoadditive Models. Department of Biostatistics, School of Public Health, Harvard University.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Lindgren, F., Rue, H., Lindstrom, J., 2011. An explicit link between Gaussian fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 423–498.

R-cran, 2010. R: A Language and Environment for Statistical Computing. R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria.

Ridout, M., Hinde, J., Demetrio, C., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative Binomial alternatives. Biometrics 57, 219–223.

Rose, G., 2001. Sick individuals and sick populations. International Journal of Epidemiology 30, 427.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 319–392.

Simpson, D., Lindgren, F., Rue, H., 2011. Fast Approximate Inference with INLA: The Past, The Present and The Future, Arxiv preprint arXiv:1105.2982.

Simpson, D., Lindgren, F., Rue, H., 2011b. Think Continuous: Markovian Gaussian Models in Spatial Statistics, Arxiv preprint arXiv:1110.6796.

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64, 583–639.

StataCorp, 2007. Stata Statistical Software Release 10. StataCorp LP, College Station, TX, USA.
Tollman, S.M., Herbst, K., Garenne, M., Gear, J.S., Kahn, K., 1999. The Agincourt Demographic and Health Study: site description, baseline findings and implications. South African Medical Journal 89, 858–864.
Wikle, C.K., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84, 1382–1394.
Wikle, C.K., Berliner, L.M., Cressie, N., 1998. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 5, 117–154.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Musenge

CLEARANCE CERTIFICATE **PROTOCOL NUMBER M081145**

PROJECT	Modelling Spatiotemporal Patterns of Childhood HIV/TB Related Morality and Malnutrition: Applications to Agincourt Data in Rural South Africa (revised title)
INVESTIGATORS	Mr E Musenge
DEPARTMENT	School of Public Health
DATE CONSIDERED	08.11.28
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 24/10/2012 **CHAIRPERSON** .....
(Professor P E Cleaton Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof K Kahn

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...