ANALYSIS OF ZERO ALTERED AND INFLATED DATA: APPLICATION TO HUMAN HELMINTHS

MSc. (BIOSTATISTICS) THESIS

MICHAEL GIVE CHIPETA

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

JULY, 2012

ANALYSIS OF ZERO ALTERED AND INFLATED DATA: APPLICATION TO HUMAN HELMINTHS

MSc. (Biostatistics)

By

MICHAEL GIVE CHIPETA

BSc. Mathematical Sciences(Statistics & Computing)

Thesis submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science in Biostatistics.

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

JULY, 2012

DECLARATION

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

Michael Give Chipeta

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

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

Signature: _____ Date: _____

Lawrence N. Kazembe, PhD (Associate Professor)

Main Supervisor

Signature: _____ Date: _____

Bagrey M. Ngwira, PhD (Lecturer)

Member, Supervisory Committee

DEDICATION

To Susan, Japhet, Vita and Tionge

ACKNOWLEDGEMENTS

My deepest and sincere gratitude is hereby expressed to the Lord, whom I believe gave me strength, courage and ability to complete this task. I thank God for the grace and love he abundantly, constantly and continuously shows on me.

The present thesis was undertaken under joint supervision of Associate Professor Lawrence N. Kazembe and Dr. Bagrey Ngwira. During my research, I benefited tremendously from the statistical/epidemiological knowledge and experience my supervisors were willing to share. Only thanks to this close cooperation, I was able to find results presented in this thesis. Assoc. Prof. Kazembe and Dr. Ngwira jointly guided me through my time as a master student by lending patiently a helping hand. My sincerest thanks are addressed to them for their personal, scientific and statistical contributions.

Gratitude is hereby expressed to National Commission of Science and Technology (NCST) through Health Research Capacity Strengthening Initiative (HRCSI) for making funds available for the study.

I also thank all my classmates for their support, and above all, for their knowledge which they offered in abundance. Well done guys **!** And to Amos Royle Bemeyani and Madalitso Rexy Tolani, I owe you so much guys, may God bless you.

My deepest gratitude goes to my parents (Japhet and Vita) who started me off on a brumous path that is now clearing off and bearing fruit. I am also grateful to my sister, friends, family and in a very special way to Susan McDowall for being there for me since my undergraduate studies up to now, providing support in every way possible. Many thanks Susan.

ABSTRACT

Background: Helminths infections (i.e., *S. haematobium*, *S. mansoni* and Hookworm) affect more than a quarter of world's population, with consequences for nutritional and educational development of infected individuals. They are a common cause for morbidity, especially among children in undeveloped countries. Control strategies require better understanding of helminths epidemiology. As such, appropriate statistical methods to model infection prevalence and intensity are crucial. This study aimed at joint modeling of helminths infection prevalence and intensity using robust methods in order to understand the epidemiology.

Methods: Zero altered models were fitted and applied to two datasets (one from Malawi and another from Zambia). Malawi data were collected in a cluster randomized study, in Chikhwawa district in 2004, with 18 villages randomised to intervention and control arms. Zambia data were collected from school children in a cross-sectional study in Lusaka province in 2004. A range of Zero Inflated (ZI) models (ZI-Poisson [ZIP] and ZI-Negative Binomial [ZINB]) and Hurdle models (Poisson Logit Hurdle [PLH] and Negative Binomial Logit Hurdle [NBLH]) were developed for infection analysis, adjusted for age, sex, education level, treatment arm, occupation, and polyparasitism, among others. Model estimation was based on maximum likelihood estimation (MLE) and model selection was based on Akaike Information Criteria (AIC). Exponential and Spherical variogram models were used to estimate residual spatial effects.

Results: Chikhwawa study had a total of 1, 642 participants. Overall, 55.4 % were female and mean age of study population was 32.4 with SD = 22.8. Prevalence was as follows: *S. haematobium* = 19.4 %, *S. mansoni* = 5.0 % and Hookworm = 22.9 %. Schistosomiasis and Hookworm infections were highly aggregated in a

relatively small and heavily infected population proportion. A large proportion of individuals were non-egg excretors (*S. haematobium* = 85.8 %, *S. mansoni* = 95.7 % and Hookworm = 80 %) for outcomes of interest hence data had a large number of zeros.

Data showed overdispersion evidence with p-value <0.0001. NBLH model offered the best fit to data with lowest AIC = 3, 482. Schistosomiasis infection was associated with age (RR = 0.96); with the highest intensity in school-age children. Fishing (RR = 0.73) and working in gardens (RR = 1.21) along the Shire River were also clear risk factors. Hookworm showed a high intensity in older people than in younger children and also high intensity in males than in females (RR = 1.19). Intervention reduced both infection intensity and prevalence in intervention arm as compared to control arm. Residual spatial effects showed some degree of spatial dependence across the study area.

The Zambia study had a total of 2040 participants. Overall, 50.4 % were female and mean age of study population was 9.98 with SD = 2.14. Urinary Schistosomiasis prevalence was 9.6 %. NBLH model offered the best fit with lowest AIC = 3, 230. Schistosomiasis prevalence was associated with age (AOR = 0.69), sex (AOR = 1.17), altitude (0.37), NDVI (AOR = 1.04) and temperature (AOR = 0.99). Infection intensity was associated with age (RR = 0.55), sex (RR = 1.28), altitude (RR = 0.11), temperature (RR = 0.84), and NDVI (RR = 1.07).

Conclusion: Helminths were highly localized, with small section of people harboring parasites; showing heterogeneous infection risk for both Malawi and Zambia settings. Joint modeling approach allowed identification of risk factors for infection presence and severity hence provide a platform to design combative control efforts. NBLH offered best-fit to data with capability to handle overdispersion, excess zeros and capture true zeros in the data. Its implementation and interpretation, ease of components, and its direct link with observed data make it a valuable alternative for analysing zero inflated count data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMNS AND ABBREVIATIONS

AIC	Akaike's Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
BIC	Bayesian Information Criterion
CAIC	Consistent Akaike's Information Criterion
CI	Confidence Interval
COMREC	College of Medicine Research Ethics Committee
DALYs	Disability-Adjusted Life Years
EPG	Eggs Per Gram
GLM	Generalised Linear Model
GPS	Geographical Positioning System
HIV	Human Immunodeficiency Virus
IUGR	Intrauterine Growth Retardation
LBW	Low Birth Weight
LF	Lymphatic Filariasis
MBG	Model Based Geo-statistics
MDA	Mass Drug Administration
MLE	Maximum Likelihood Estimation
NB	Negative Binomial
NBLH	Negative Binomial Logit Hurdle
NDVI	Normalised Difference Vegetation Index
PLH	Poisson Logit Hurdle
PSCL	Political Science Computational Laboratory
RPM	Revolutions Per Minute
SSA	Sub-Saharan Africa
STH	Soil Transmitted Helminths
UNICEF	United Nations Children's Fund
WHA	World Health Assembly
WHO	World Health Organization
ZI	Zero Inflated
ZINB	Zero Inflated Negative Binomial
ZIP	Zero Inflated Poisson

Chapter 1

INTRODUCTION

Human helminths infections (intestinal nematode infections such as *Ascaris lumbricodes, Trichuria trichura*, and Hookworm and schistosome infections like *Schistosoma haematobium* and *Schistosoma mansoni*) affect more than a quarter of the world's population, with consequences in health nutritional and educational development of infected individuals (Brooker, Hay, Tchuente, & Ratard, 2002). The burden of disease caused by infection with Schistosomiasis and soil-transmitted helminths (STH) remains enormous. According to WHO (2006), about 2 billion people are affected worldwide, of whom 300 million suffer associated severe morbidity. In 1999, WHO estimated that these infections represented more than 40 % of the disease burden caused by all tropical diseases, excluding malaria (WHO, 2001a).

Helminths are parasitic worms found in intestinal tract, urinary tract, blood and other tissues. Although helminths can infect all members of a population, it is clear that there are specific groups who are at greater risk of morbidity than others, and who are more vulnerable to harmful effects of chronic infections (Hotez et al., 2006). For schistosomes and STH, the most vulnerable groups are school-age children and women of child-bearing age, including adolescent girls (Brooker et al., 2009).

Schistosomiasis, or bilharzia, is caused by worms (flukes) that have a complex

life cycle involving freshwater snails as intermediate hosts. Several species exist, with *S. mansoni, S. japonicum,* and *S. haematobium* being the most prevalent. Chronic infection with *S. mansoni* and *S. japonicum* causes periportal liver fibrosis and portal hypertension with ascites and oesophageal varices. Long term infection with *S. haematobium* is associated with bladder scarring, renal obstruction, chronic urinary infection, and possibly bladder carcinoma (National Travel Health Network and Centre, 2008).

Hookworms are nematodes in the super family *Ancylostomatoidea*. In their normal hosts, they are parasites of intestinal tract (Center for Food Security and Public Health, 2005). In humans, some zoonotic Hookworms can cause cutaneous larva migrans. Most cases of classic Hookworm disease are caused by *Ancylostoma duodenale* or *Necator americanus*, species usually found only in humans (Center for Food Security and Public Health, 2005).

1.1 Disease burden and Endemicity

Hookworm is a leading cause of maternal and child morbidity in undeveloped countries of the tropics and subtropics (Larocque, Casapia, Gotuzzo, & Gyorkos, 2005). In susceptible children, Hookworms cause intellectual, cognitive and growth retardation, intrauterine growth retardation (IUGR), prematurity and low birth weight among new infants born to infected mothers (Bethony, Brooker, Albonico, & Hotez, 2006). In developed countries, Hookworm infection is rarely fatal, but anaemia can be significant in a heavily infected individual. Bethony et al. (2006) report that STHs are one of the world‘s most important causes of physical and intellectual growth retardation. Yet, despite their educational, economic, and public-health importance, they remain largely neglected by the medical and international community. The neglect stems from three features: first, those who are most affected are the world's most impoverished, particularly those who live on less than US$2 per day; second, the infections cause chronic ill health and have

insidious clinical presentation; and third, quantification of helminths infections effect on economic development and education is difficult.

Hookworms injure their human host by causing intestinal blood loss leading to iron deficiency and protein malnutrition (Pritchard & Hotez, 1995). The parasite induces blood loss directly through mechanical rupture of host capillaries and arterioles followed by release of a battery of pharmacologically active polypeptides including anticoagulants, antiplatelet agents, and antioxidants (Brooker, Hotez, Bethony, & Silva, 2003). Chronic blood loss and depletion of the body's iron stored in heavy Hookworm infections often lead to iron-deficiency anaemia, a condition better known as Hookworm anaemia (Lwambo, Bundy, & Medley, 1992). Some investigators believe that Hookworm anaemia is highly focal, and in some instances more common in coastal regions (Lwambo et al., 1992). Studies of anaemia associated with Hookworm blood loss indicate a disproportionate reduction in plasma haemoglobin concentration after some threshold worm burden is exceeded (Bundy, 1995).

Bethony et al. (2006) report evidence to support high disease-burden estimates from STH infections, and highlights Hookworm importance as a threat to maternal and child health. For example, cross-sectional evidence from Africa and Asia showed that 30 - 54% of moderate to severe anaemia in pregnant women was attributable to Hookworm. Intervention studies suggest that antenatal anthelmintics substantially increase maternal haemoglobin concentrations, birth weight and infant survival. In addition, Hookworm associated iron deficiency during childhood is partly responsible for physical and mental growth retardation effects (Brooker et al., 2003). Growth-stunting effects of Hookworm were well documented by the early part of the 20th Century (Brooker et al., 2003), as were some of the effects of Hookworm on intelligence quotient (Brooker et al., 2003). However, it is only within the last few years that Hookworm-induced iron deficiency was understood to also exert more subtle, yet profound, adverse effects on childhood memory,

reasoning ability and reading comprehension (Brooker et al., 2003).

Hookworm distribution of infection intensity and prevalence is strongly age dependent (Chan, Bradley, & Bundy, 1997). However, in contrast to other common intestinal helminths such as *Ascaris lumbricoides* (large roundworms) and *Trichuris trichiura* (whipworm) and also *Schistosomiasis*, where children are more heavily affected, Hookworm is more common in adults. Chan et al. (1997) report that it is generally thought that differences in levels of Hookworm infection in children and adults are due to exposure differences, as Hookworm is generally transmitted in the fields as opposed to near houses; as is the case with *Ascaris lumbricoides* and *Trichuris trichiura*.

Schistosomiasis exhibits a focal distribution, and symptoms are often difficult to recognize both by individuals infected and by health personnel who normally staff primary health care facilities in rural Africa (Fenwick & Hotez, 2009). Additionally, early stages of Schistosomiasis, when treatment is most beneficial, often show only mild yet debilitating symptoms, which lead to serious consequences later in life (Fenwick & Hotez, 2009). Through a full consideration of amount of end-organ pathologies to the liver (in case of *S. mansoni* and *S. japonicum* infections), and to the bladder and kidneys (in case of *S. haematobium* infection), together with chronic morbidities associated with impaired child growth and development, chronic inflammation, anaemia, and other nutritional deficiencies, some new disease burden assessments estimated that Schistosomiasis accounts for up to 70 million disability-adjusted life years (DALYs) lost annually (King & Dangerfield-Cha, 2008). This global burden estimate exceeds that of malaria or tuberculosis, and is almost equivalent to DALYs lost from HIV/AIDS (King & Dangerfield-Cha, 2008). Further, almost 300,000 people die annually from Schistosomiasis in Africa (Fenwick & Hotez, 2009), and there is evidence that female genital Schistosomiasis caused by *S. haematobium* may significantly increase likelihood of contracting HIV/AIDS (Kjetland, Ndhlovu, Gomo, & Mduluza, 2006;

Mbabazi, Andan1, Chitsulo, & Downs, 2011). Urogenital Schistosomiasis, caused by infection with *S. haematobium*, is widespread and causes substantial morbidity in Africa. Schistosomiasis infection has been suggested as an unrecognized risk factor for incident HIV infection (Mbabazi et al., 2011). In individuals of reproductive age, urogenital Schistosomiasis remains highly prevalent and, likely, under diagnosed (Mbabazi et al., 2011).

Schistosomiasis is a strictly regional endemic disease. It is dependent on *Oncomelania* snail distribution. There are three (3) types of endemic areas, namely, marshland and lake regions, hilly and mountainous regions, and plain regions with water-way networks. In plain regions, snails are distributed along river systems and Schistosomiasis spreads widely (Jiagang, 2003). The range of Schistosomiasis endemicity can change either through movements of infected persons to areas inhabited by host snails, or artificial creation of new habitat (that is, dams, canals, rice fields) for infected snails.

1.1.1 Geographical distribution

It is estimated that 85 % of people infected with Schistosomiasis are from Africa (Chitsulo, Engels, Montresori, & Savioli, 2000), and that 1.5 billion are infected with STHs worldwide (Hanzel, Karanja, Addiss, Hightower, & Rosen, 2003). Schistosomiasis is present worldwide, but it occurs most frequently in Sub-Saharan Africa (SSA), Brazil, southern China, and the Philippines (National Travel Health Network and Centre, 2008). Schistosomiasis is endemic in 74 countries and territories; and in particular, *S. haematobium* is endemic in 53 countries in Middle East and most of the African continent including the islands of Madagascar and Mauritius (Chitsulo et al., 2000), whereas *S. mansoni* is mostly endemic in SSA (Alemu et al., 2011). Roughly, 600 million cases of Hookworm are distributed predominantly in agricultural areas and among estimated 2.7 billion people who live on less than US$2 per day (P. Hotez, 2008). Environment and socioeconomic

status represent the two most important determinants for acquiring Hookworm. Hookworm infection is closely associated with poverty; inadequate sanitation, poor housing construction, and lack of access to essential medications are major factors in this relationship.

1.2 Epidemiology and control

Helminths infection is widely spread throughout tropical and subtropical areas (Bethony et al., 2006), with prevalence in some communities being as high as 90 %. Hookworm infection flourishes in rural communities with moist shaded soil and inadequate latrines and are linked to lack of sanitation . Agricultural laborers have traditionally been at high risk of Hookworm infection. Improper disposal of human faeces and the common habit of walking barefoot are important epidemiologic features. The predisposing factors for Schistosomiasis infection include: swimming, bathing, fetching and washing in infected freshwater habitats.

Several defined measures of helminth transmission are valuable to guide implementation of control programmes. The most common measure is infection prevalence (proportion of individuals infected). Prevalence is only an indirect measure of amount of disease transmission because infections may persist for varying lengths of time. Gemperli (2003) reports that a direct transmission measure is the incidence of disease; that is, the number of new cases of disease diagnosed per unit time and person. Incidence data can be biased when collected in health centers because it may reflect patients' access to those facilities. They also depend on accurate estimates of the population at risk. A second key measure is the intensity of infection (worm burden or severity) which is estimated based on quantitative egg counts or blood smears. The relative ease in collecting prevalence data means that decisions on where to implement control measures is typically based on whether prevalence of infection exceeds some species-specific threshold (Magalhes, Clements, Patil, Gething, & Brooke, 2011). For STH and Schistosomiasis, the goal of treatment is

morbidity control; mass treatment has been recommended where infection prevalence exceeds 20 % among school children (WHO, 2002, 2006). Regardless of treatment threshold, implementation of helminth control requires evidence-based maps of infection presence.

WHO recommends mass drug administration (MDA) with *praziquantel* (for schistosomes) and *albendazole or mebendazole* (for STH) wherever infection prevalence exceeds 10 %, and had a target of deworming at least 75 % of school-aged children and other high risk groups by 2010 (WHO, 2002). This goal has encouraged many countries to establish national action plans and programmes for controlling schistosomes and STH. However, implementation of such programmes requires reliable and up-to-date information on epidemiology and infection intensity in order to (i) guide control to areas in greatest need and (ii) estimate drug requirements (Brooker et al., 2009).

A number of international initiatives have supported mass school-based treatment for STH infection and Schistosomiasis as a control measure. These include **Deworm the World** (*www.dewormtheworld.org*) and **Children Without Worms** (*www.childrenwithoutworms.org*) for STH infection, and **Schistosomiasis Control Initiative** (*SCI: www.sci-ntds.org*), initially for Schistosomiasis and STH.

1.2.1 Morbidity and mortality

Helminths infections are important causes of morbidity and mortality in many undeveloped countries (Ezeamama, Friedman, Acosta, & Bellinger, 2005). In particular, Schistosomiasis and STHs are responsible for extensive morbidity and mortality in sub-Saharan Africa (SSA) (Hanzel et al., 2003). Among well-described morbidities associated with helminths infection in children are under-nutrition, anaemia, and failure to achieve genetic potential for growth (Ezeamama et al., 2005). Studies of morbidity caused by chronic Schistosomiasis have also confirmed

a general relationship between infection intensity and high morbidity in children (WHO, 1993).

Fortunately, much of morbidity associated with infection can be reversed with the use of effective anthelmintic drug treatments (Brooker et al., 2002, 2009). All of these parasites can be effectively treated with single dose oral therapies that are safe, inexpensive and required at periodic intervals. Treatment is typically implemented through mass chemotherapy whereby entire at-risk population is treated as part of either school or community-based campaigns.

In May 2001, World Health Assembly (WHA) passed resolution 54.19 endorsing regular treatment of high-risk groups, particularly school-age children, as the best means of reducing morbidity and mortality (WHO, 2001b; Hanzel et al., 2003). Under WHO guidelines, the decision to treat all persons (mass treatment) or only school children and other high risk groups (selective treatment) depends on prevalence of infection in a particular region (WHO, 1998b). Both schistosomes and STH infections tend to be highly aggregated in that a small percentage of infected persons have very high worm burdens (Hanzel et al., 2003; Vounatsou, Raso, Tanner, N'goran, & Utzinger, 2009).

A direct relationship exists between helminths infection intensity and anaemia (Larocque et al., 2005). Anaemia in pregnancy has been associated with poor birth outcome, such as low birth weight (Bethony et al., 2006) and increased maternal morbidity and mortality (Larocque et al., 2005). WHO (1998a) proposed a classification for intensities for each STH infection (based on quantitative counts obtained using Kato-Katz method). Especially in Hookworm infection, the degree of severity of morbidity varies not only according to number of worms present but also according to determinants of the host (that is, age), parasite (that is, species), and host-parasite interaction (that is, nutritional intake of iron). Increase in helminths infection intensity has been reported to significantly contribute to haemoglobin depletion in infected individuals (Kanzaria et al., 2005)

However, mortality from helminths has been poorly documented in most endemic countries, and death certificates and patients' records rarely identify helminths as an underlying cause of death. Therefore, there is no doubt that mortality due to helminths continues to be underestimated, and improved data collection in health services is needed (WHO, 1993). Even though mortality rate due to helminths is low (WHO, 1993), most of the people infected are children under 15 years with problems of faltering growth and/or decreased physical fitness (Kheir, Eltoum, Saad, Magdi M. Ali, & Homeida, 1999). Little is known about the pattern of mortality in areas of endemic Schistosomiasis (WHO, 1993). This is a manifestation of general deficiency in recording rates, causes and distribution of mortality in endemic countries (Kheir et al., 1999).

1.3 Justification

For an adequate medical, economic and public health appraisal of the importance of helminths infections, there are many factors to be considered. Among these are geographical distribution, prevalence, intensity of infection, transmission patterns, morbidity and mortality, which are influenced by demographic, environmental conditions, relative efficiency of molluscan intermediate hosts, agricultural practices, and human behavior. Understanding helminths infection burden and its determinants is needed and useful for designing combative interventions.

Of the many measures, prevalence (proportion of individuals infected) is the most commonly available measure of helminths disease burden. However, infection severity (worm burden), which is estimated based on quantitative egg counts (in faecal matter or urine) is very important in understanding medical, public health, and economic importance of helminths infections. Because of the easiness of data collection for prevalence, disease control studies mostly consider infection presence and ignore severity. If anything, those that consider severity or intensity, do it separately, that is, focusing on counts, ignoring the fact that infection prevalence and

severity can be driven by similar factors. There is need to come up with methods to jointly model helminths infection prevalence and severity. In so doing, we may help to learn risk factors associated with both prevalence and intensity and further discover important differences between the two outcomes, thus improve design of interventions. The current study considered a class of zero augmented models to accommodate excessive zeros in helminths infections. The study analysed not only the prevalence, but also helminths infection intensity with an emphasis on demographic, environment and social-economic covariates.

1.4 Study objectives

The study aimed at analysing prevalence and intensity of helminths infections with examples from Malawi and Zambia.

1.4.1 Specific objectives

The specific objectives of this study are:

1. Derive Zero adjusted models to estimate helminth epidemiology.
2. Assess demographic, environment and socio-economic covariates impacting on helminths infection prevalence and intensity.
3. Advance appropriate use of zero adjusted models in helminths disease control programmes.

1.5 Significance of the study

The study looked at helminths infection, emphasizing on factors that influence both intensity and prevalence. This is useful as it compliments existing knowledge of helminths and also ensures optimal control of infections in a cost-effective

way. These are important in public health policy making through improvement of national STH and Schistosomiasis control planning and also assist in designing of integrated intervention strategies to curb disease burden by:

- Predicting patterns of parasite prevalence and intensity for planners to target areas with highest potential risk.
- Allowing planners to determine intervention or combination of interventions that are cost effective in a particular area.

More importantly, the study contributes to zero adjusted models for count data applications in helminthology and disease mapping. Schistosomiasis and Hookworm infections have been chosen since both are influenced by demographic, environmental and socio-economic factors approximated by geographical location and depict spatial clustering.

1.6 Structure of thesis

The thesis is organised as follows: Chapter two (2) describes generation of helminths count data, models for count data, zero-altered/zero-inflated model specifications and their main properties. estimation parameters and model selection criteria. It also gives an overview of statistical modeling of helminths. Chapter three (3) outlines details for data, its sources, ethical clearance and methods used for data analysis. It also describes fitted models, model selection and details for residual spatial effects analysis. Chapter four (4) presents results and discussion for the applications. Chapter five (5) outlines main conclusions and recommendations from the study. A list of references used follows from chapter five (5). Some of R code snippets that were used in statistical analysis are in appendix A. Finally, a surrogate residual spatial effects figure follows in appendix B.

Chapter 2

METHODOLOGY REVIEW

2.1 Introduction

Statistical modeling gives mathematical descriptions of "factor-disease" relations (that is, demographic, environment and socio-economic factors e.t.c), identifies significant predictors of disease transmission and provides predictions of disease risk, among others. Various statistical models have been developed to model helminths disease burden. This chapter gives a review of count data and models/assumptions for count data, an overview of zero adjusted models (zero-inflated and Hurdle) and their properties, model selection standard, zero adjusted model applications, statistical methods in helminths modeling and residual spatial effects overview. The current review focuses on strengths and weaknesses of above-mentioned count data models and a justification for their adoption and subsequent use in modeling zero inflated count data.

2.2 Modeling zero inflated data

The encumbrance of helminths infection in a given community can be measured by either of these two indicators: infection intensity or infection prevalence. Infection intensity is a measure of number of eggs per gram of faeces (for STH and *S. mansoni*) or eggs per 10 ml of urine (for *S. haematobium*), and is a key deter-

minant of transmission dynamics within communities and morbidity risk among individuals (Brooker et al., 2009). Measuring intensity requires time-consuming, quantitative laboratory methods and consequently is not routinely assessed in field surveys. The more easily collected indicator is prevalence of infection: proportion of sampled individuals who have one or more eggs detected in their stool or urine sample. In view of relative simplicity of measuring prevalence, WHO recommends its use to determine the need for control, with mass treatment of whole populations recommended where prevalence exceeds 10 % (WHO, 2002).

In the above-mentioned infection intensity measures, count data are used; data in which observations can only take non-negative integer values: 0, 1, 2, ... (Winkelmann, 2008) and where these integers arise from counting rather than ranking, following a Poisson distribution. Transmission intensity of human helminths is a function of parasitic worm load within a group of individuals, which can be quantified by number of eggs that are excreted. Host heterogeneities in exposure and susceptibility to infection lead to an aggregated distribution of severity across individuals (Bradley, 1972). For this reason, a few individuals harbour large numbers of parasites whilst majority of individuals are uninfected or only carry a low parasite burden (Vounatsou et al., 2009). In addition, widely used diagnostic approaches for Schistosomiasis (that is, Kato-Katz technique for *S. mansoni* diagnosis) fail to detect some infected individuals particularly when only a single stool sample is examined and infection intensities are light (Utzinger et al., 2001). Due to these two issues, often a large proportion of individuals are considered as "zero egg excretors" (Vounatsou et al., 2009).

Poisson regression is traditionally conceived as the basic count model upon which a variety of other count models are based (Greene, 2005; Hilbe, 2011). Poisson distribution is characterized as:

$$f(y;\lambda) = \frac{e^{-\lambda_i}(\lambda_i^{y_i})}{y_i!}, \quad y_i = 0, 1, 2, ..., n_i; \ \lambda > 0 \tag{2.1}$$

where random variable y is the count response and parameter λ is the mean. Often, λ is also called the rate or intensity parameter referred to as μ (Hilbe, 2011). Unlike most other distributions, Poisson does not have a distinct scale parameter. Rather, the scale is assumed equal to 1. Poisson regression model derives from Poisson distribution. Relationship between μ, β, and x the fitted mean of the model, parameters, and model covariates or predictors, respectively is parameterized such that $\mu = \exp(x\beta)$. Here, $x\beta$ is the linear predictor, which is also symbolized as η within the context of generalised linear models (GLM). Exponentiating $x\beta$ guarantees that μ is positive for all values of η and for all parameter estimates (Hilbe, 2011).

The standard Poisson distribution, which assumes equal variance and mean, is not appropriate to fit the observed egg counts since variance of the counts is much larger than their mean. Violations of equidispersion indicate correlation in the data, which affects standard errors of the parameter estimates. Model fit is also affected. When such a situation arises, modifications are made to the Poisson model to account for discrepancies in the goodness of fit of the underlying distribution. Negative binomial (NB) is normally used to model overdispersed Poisson data. The NB model is employed as a functional form that relaxes the equidispersion restriction of the Poisson model (Greene, 2008). NB distribution is characterized as:

$$P(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k + 1)}\Big(\frac{1}{1 + \theta}\Big)^{\alpha}\Big(\frac{\theta}{1 + \theta}\Big)^{k} \quad k = 0, 1, 2, ... \tag{2.2}$$

where random variable X has a NB distribution with parameters $\alpha \geq 0$ and $\theta \geq 0$. Its mean and variance are given by:

$$E(X) = \alpha\theta \tag{2.3}$$

and

$$Var(X) = \alpha\theta(1+\theta) = E(X)(1+\theta) \tag{2.4}$$

Since $\theta \geq 0$, the variance of NB distribution generally exceeds its mean ("overdispersion") (Winkelmann, 2008). It has since been proposed to model excessive variation in helminth egg counts (Cohen, 1977). NB regression models have been widely used to analyse helminths infection intensity data (Ridout, Demetrio, & Hinde, 1998; Utzinger, Vounatsou, N'Goran, Tanner, & Booth, 2002; Brooker et al., 2006). However, distributional problems affect both models (Poisson and NB) such as overdispersion resulting from specification errors in the systematic part of the regression model, hence NB models themselves may be overdispersed (Hilbe, 2011). Nevertheless, both models can be extended to accommodate any extra correlation or dispersion in the data that result in a violation of distributional properties of each respective distribution. The enhanced Poisson or NB model can be regarded as a solution to a violation of distributional assumptions of the primary model (2.1). For a better fit, an overdispersed model that incorporates excess zeros should serve as an alternative (Famoye & Singh, 2006). Zero-adjusted mixture models such as Zero-Inflated (ZI) and Hurdle count models are capable of incorporating excess zeros. They are applied to count data when overdispersion exists and excess zeros are indicated (Flynn & Francis, 2009; Hilbe, 2011).

2.2.1 Assumption violations and diagnostics

1. No zeros in data

Poisson and NB distributions assume that zero counts are a possibility. When data to be modeled originate from a generating mechanism that structurally excludes zero counts, then Poisson or NB distribution must be adjusted to account for the missing zeros. Such model adjustment is not used when the data can have zero counts, but simply do not. Rather, an adjustment is made only when the data must be such that it is not possible to have zero counts. Zero-truncated Poisson

and zero-truncated NB models are normally used for such situations (Hilbe, 2011).

2. Excess zeros in data

Poisson and NB distributions define an expected number of zero counts for a given value of the mean. The greater the mean the fewer zero counts are expected. Some data, however, come with a high percentage of zero counts - far more than are accounted for by the Poisson or NB distribution. Zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) models handle this problem. Logistic or probit regression is typically used to model the structural zeros, and Poisson or negative binomial regression is used for the count outcomes (Hilbe, 2011).

3. Data separable into two or more distributions

When zero counts of a Poisson or NB model do not appear to be generated from their respective distributions, the model may be separated into two parts; somewhat like the ZIP and ZINB models mentioned above. In this case, hurdle models are used. However, in case of hurdle models (Poisson logit hurdle (PLH) and Negative binomial logit hurdle (NBLH)), the assumption is that a threshold must be crossed from zero counts to actually enter the counting process (Hilbe, 2011).

2.3 Zero adjusted distributions

Zero inflation in count data arises when one mechanism generates only zeros and the other process generates both zero and non-zero counts, hence they can be expressed as a two-component mixture model where one component has a degenerate distribution at zero and the other is a count model (Cameron & Trivedi, 1998). Zero adjusted models estimate two equations: one for the count model and another for the excess zeros. Zero inflated count data are common in a number of applications. Ridout et al. (1998) cite examples of data with too many zeros from various disciplines including agriculture, econometrics, patent applications, species abundance, medicine, and use of recreational facilities. Several models

have been proposed to handle count data with too many zeros than expected. ZI models include: zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) whereas Hurdle models include: Poisson logit hurdle (PLH) and negative binomial logit hurdle (NBLH) (Hilbe, 2011). Each of these models consists of an equation for "participation" and a model for the event count that is conditioned on the outcome of the first decision (Greene, 2005).

Both the Hurdles and ZI models allow for two sources of overdispersion (Cameron & Trivedi, 1998). One of these allows for extra (or too few) zeros; the second allows for overdispersion induced by individual heterogeneity in the positive set. The Hurdle model can also explain too few zeros (Cameron & Trivedi, 1998).

Zero adjusted models assume that a proportion of individuals have no chance to be infected, as they are not exposed. In case of helminths infection, there is a process which determines whether an individual is likely to be infected at all (infection probability), and a second process determining the number of excreted eggs among those who are at risk of infection (infection severity).

2.3.1 Zero Inflation (ZI) model

Model Description

A ZI count model is a special case of a finite mixture model that only permits mixing with respect to zeros. The assumption that mixing takes place with respect to zeros only is relatively more attractive if the population can be realistically divided into two components (Cameron & Trivedi, 1998). Members of one subpopulation are "never at risk" and hence never experience a positive number of events. Those of the second subpopulation are "at risk" and may experience a positive number of events (Cameron & Trivedi, 1998).

The latent class interpretation of the model suggests a two level decision process: the regime and the event count. ZIP models assume that the number of excreted

eggs follows a Poisson distribution. ZINB models assume that the number of eggs among those who are at risk of infection has a NB distribution (Vounatsou et al., 2009). Thus ZIP and ZINB models can be viewed as partial observation models or latent class models of a sort (Greene, 2005). A general distribution assumption for ZI models is that the number of observations having a value of zero (that is, no events experienced) on the dependent variable will generally exceed that which would be expected under the Poisson. The ZI model structure is given as follows:

$$d_i^* = w_i^{'}\delta + u_i \tag{2.5}$$

$$d_i = \mathbf{1}(d_i^* > 0) \tag{2.6}$$

$$Prob(d_i = 0|w_i) = \Pi_0(w_i^{'}\delta) \text{ (}\mathbf{regime\,selection\,equation}\text{)} \tag{2.7}$$

$$Prob(d_i = 1|w_i) = 1 - \Pi_0(w_i^{'}\delta) \text{ (}\mathbf{regime\,selection\,equation}\text{)} \tag{2.8}$$

$$y_i^*|X \sim P(y_i^*|X_i) \text{ (}\mathbf{latent\,Poisson\,or\,NBmodel}\text{)} \tag{2.9}$$

$$E[y_i^*|X_i] = exp(\alpha + X_i^{'}\beta) = \lambda_i \text{ (}\mathbf{conditional\,mean}\text{)} \tag{2.10}$$

$$y_i = d_i y_i^* \, and \, X_i \, \mathbf{are\,observed} \;\; \text{(}\mathbf{observation\,mechanism}\text{)} \tag{2.11}$$

From above, d_i is the participation decision or regime selection (binary variable) with $i = 0$ or 1, y_i is the count outcome variable 0,1,... with $i = 1$, ..., N. Thus, if d_i equals zero, then the observed y_i equals zero regardless of the latent value of

y_i^*. $w_i^{'}$ is a vector of covariates that appears in the participation equation (that is, in the regime selection equation) and δ is the standard deviation of a random heterogeneity term epsilon. If d_i equals one, Poisson or NB variable (which might then still equal zero) is observed (Greene, 2005). A common element throughout is the assumption that latent effects in the regime equation and the count outcome are uncorrelated.

Below are model specifications of ZI models adapted from Loeys, Moerkerke, Smet, and Buysse (2011) that have been used to analyse application data in the current study.

1. Zero Inflated Poisson (ZIP)

In ZIP regression, the counts Y_i equal 0 with probability p_i and follow a Poisson distribution with mean μ_i with probability 1 - p_i. ZIP model can thus be seen as a mixture of two component distributions:

$$Pr(Y_i = 0) = p_i + (1 - p_i)exp(-\mu_i) \tag{2.12}$$

$$Pr(Y_i = k) = (1 - p_i)exp(-\mu_i) - \mu_i^k/k!, \quad k = 1, 2, 3, ... \tag{2.13}$$

From (2.12), zero observations arise from both zero-component distribution and Poisson distribution. The zero-component distribution is therefore related to modeling 'excess' or 'inflated' zeros that are observed in addition to zeros that are expected to be observed under the assumed Poisson distribution. To assess impact of covariates on the count distribution in a ZIP model, p_i and μ_i can be explicitly expressed as a function of covariates. The most natural choice to model probability of excess zeros is to use a logistic regression model:

$$logit(p_i) = x_i^T \beta \tag{2.14}$$

where x_i represents a vector of covariates and β a vector of parameters. Impact of covariates on count data excluding excess zeros can be modeled through Poisson regression:

$$log(\mu_i) = x_i^T \gamma \tag{2.15}$$

Mean and variance of ZIP model are:

$$E(y_i|x_i, z_i) = \mu_i(1 - p_i) \tag{2.16}$$

$$V(y_i|x_i, z_i) = \mu_i(1 - p_i)(1 + \mu_i p_i) \tag{2.17}$$

2. Zero Inflated Negative Binomial (ZINB)

Count data often exhibit more variability than predicted by the mean of a Poisson distribution, even after accounting for excess zeros. A way of modeling over-dispersed zero-inflated count data is to assume a Zero-Inflated Negative Binomial (ZINB) distribution for Y_i:

$$Pr(Y_i = 0) = p_i + (1 - p_i)\frac{\theta^\theta}{(\mu_i + \theta)^\theta} \tag{2.18}$$

$$Pr(Y_i = k) = (1 - p_i)\frac{\Gamma(k + \theta)}{\Gamma(\theta)k!} \times \frac{\mu_i^k \theta^\theta}{(\mu_i + \theta)^{k+\theta}}, \quad k = 1, 2, 3, ... \tag{2.19}$$

with mean μ_i and shape parameter θ; Γ is the gamma function. The mean and variance of ZINB model are:

$$E(y_i|x_i, z_i) = \mu_i(1 - p_i) \tag{2.20}$$

$$V(y_i|x_i, z_i) = \mu_i(1 - p_i)(1 + \mu_i(p_i + \alpha)) \tag{2.21}$$

From the means and variances of ZIP and ZINB, it is noted that they both display over-dispersion such that $V(y_i|x_i, z_i) > E(y_i|x_i, z_i)$.

2.3.2 Hurdle model

Model description

The hurdle model is also a two part decision model similar to ZI model described above. The first part is a participation equation and second part is an event count, conditioned on participation (Greene, 2005). Hurdle models are based on an assumption that zero counts are generated from a different process than are positive counts in a given data situation. A Hurdle model partitions the model into two parts: a binary process generating positive counts (1) versus zero counts (0); and a process generating only positive counts (Hilbe, 2011). The binary process is generally estimated using a binary model; the positive count process is estimated using a zero-truncated count model. The binary process can also be estimated using a 'right censored at one' count model: Poisson, geometric, or negative binomial. Censoring takes place at a count value of one such that, for example, a Poisson count of zero is given the binary process value of zero and counts of one or greater are given the value of one (1) (Hilbe, 2011).

The above described partitioning permits the interpretation that positive observations arise from crossing a zero hurdle or zero threshold. In principle, the threshold need not be at zero; it could be any value. Furthermore, it need not be treated as known. The zero value has special appeal because in many situations it partitions the population into sub-populations in a meaningful way (Cameron & Trivedi, 1998). In contrast to ZI model, zero and non-zero counts are separated in hurdle models (Loeys et al., 2011) which makes them very useful in inferential studies.

Formally, the model can be constructed as follows:

$$d_i^* = w_i^{'}\delta + u_i \tag{2.22}$$

$$d_i = \mathbf{1}(d_i^* > 0) \tag{2.23}$$

$$Prob(d_i = 0|w_i) = \Pi_0(w_i^{'}\delta) \ \ (\textbf{hurdle equation}) \tag{2.24}$$

$$Prob(d_i = 1|w_i) = \Phi(w_i^{'}\delta) \ \ (\textbf{probit hurdle model}) \tag{2.25}$$

$$y_i|X_i, (d_i = 0) = \textbf{unobserved} \ \ (\textbf{non} - \textbf{participation}) \tag{2.26}$$

$$y_i|X_i, (d_i = 1) \sim P{+}(y_i|X_i) \ \ (\textbf{truncated Poisson or NB model given participation}). \tag{2.27}$$

An interpretation in this model is that the zero outcome is governed by a separate process; the zero outcome is a decision not to participate. The central feature of the model is the effect of hurdle decision on event count equation, which is denoted $P + (y_i|x_i)$. If $d_i = 1$, then by construction, $y_i > 0$. Thus, the resulting count model has truncated form (Greene, 2005). The underlying motivation is similar to the latent class interpretation in the ZI model above.

Below are model specifications of Hurdle models adapted from Loeys et al. (2011) that have been used to analyse application data in the current study.

1. Poisson Logit Hurdle (PLH)

Like the ZIP model, PLH model is a two-component model: a hurdle component

models zero versus non-zero counts, and a truncated Poisson count component is employed for the non-zero counts:

$$Pr(Y_i = 0) = p_i^* \tag{2.28}$$

$$Pr(Y_i = k) = (1 - p_i^*)\frac{exp(-\mu_i^*)(\mu_i^*)^k/K!}{1 - exp(-\mu_i^*)}, \quad k = 1, 2, 3, ... \tag{2.29}$$

p_i models all zeros. For PLH model, the most natural choice to model probability of zeros is to use a logistic regression model:

$$\boldsymbol{logit(}p_i^*\boldsymbol{)} = x_i^T\beta^* \tag{2.30}$$

while impact of covariates x_i on strictly positive (that is, censored) count data are modeled through Poisson regression:

$$\boldsymbol{logit(}\mu_i^*\boldsymbol{)} = x_i^T\gamma^* \tag{2.31}$$

2. Negative Binomial Logit Hurdle (NBLH)

Similarly, for the hurdle models, the NBLH can be used instead of Poisson distribution above in case of over-dispersion:

$$Pr(Y_i = 0) = P_i^* \tag{2.32}$$

$$Pr(Y_i = k) = (1 - p_i^*)\frac{\Gamma(k + \theta)}{\Gamma(\theta)k!} \times \frac{(\mu_i^*)^k\theta^\theta}{(\mu_i^* + \theta)^{k+\theta}} \times \frac{1}{1 - \theta^\theta/(\mu_i^* + \theta)^\theta}, \tag{2.33}$$

where k = 1, 2, 3, ...

The most natural choice to model probability of excess zeros is to use a logistic

regression model:

$$logit(p_i^*) = x_i^T \beta^* \tag{2.34}$$

Impact of covariates on count data modeled through Poisson regression:

$$log(\mu_i) = x_i^T \gamma \tag{2.35}$$

The expected value and the corresponding variance are given by:

$$\boldsymbol{E}(y_i \mid Z_i, X_i, y_i > 0) = \sum_{y_i=1}^{\infty} y_i f_2(y_i)\theta_i \tag{2.36}$$

$$\boldsymbol{V}(y_i \mid Z_i, X_i, y_i > 0) = \theta_i \sum_{y_i=1}^{\infty} y_i^2 f_2(y_i) - \theta_i^2 \Big[\sum_{y_i=1}^{\infty} y_i f_2(y_i) \Big]^2 \tag{2.37}$$

2.4 Parameter Estimation for Zero-inflation and Hurdle models

To implement the Zero inflated (equations 2.7 and 2.8) and Hurdle models (equations 2.24 and 2.25), the following indicator variables can be created: $\omega 1$ and $\omega 2$ where $\omega 1$ equals 1 when observed count is zero and zero elsewhere; while $\omega 2$ equals 1 when observed counts are ≥ 1 and zero elsewhere. The use of these indicators ensures that the maximization of the log-likelihood functions are uniform across the entire sample (Lawal, 2010) . The indicator variables are defined as:

$$\omega 1 = \begin{cases} 1 & \text{if } y_i = 0 \\ 0 & \text{elsewhere} \end{cases} \tag{2.38}$$

$$\omega 2 = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{elsewhere} \end{cases} \tag{2.39}$$

Consequently, the log-likelihood functions for a given observation y_i are estimated as follows for the zero inflation models (for ZIP model, see equation 2.40, ZINB model, see equation 2.41 and for NBLH model, see equation 2.43):

$$\begin{aligned} \boldsymbol{L} = (\omega 1 \times [log(\phi + (1-\phi)exp(-\mu_i))] \\ + \omega 2 \times [log(1-\phi) + y_i log(\mu_i) - log(y_i!) - \mu_i]) \end{aligned} \tag{2.40}$$

$$\begin{aligned} \boldsymbol{L} = \Big(\omega 1 \times [log(\phi + (1-\phi)(1+k\mu_i)^{-k^{-1}})] \\ + \omega 2 \times [log(1-\phi) + y_i log\mu_i + y_i logk - logy_i! \\ -(y_i + k^{-1})log(1+k\mu_i) + log\Gamma(y_i + k^{-1}) - log\Gamma(k^{-1})]\Big) \end{aligned} \tag{2.41}$$

The likelihood function for Hurdle models is given as:

$$\boldsymbol{L} = \prod_{y_i=0} f_1(0) \prod_{y_i>0} \Big[1 - f_1(0)\Big] \prod_{y_i>0} \left[\frac{f_2(y_i)}{[1-f_2(0)]}\right] \tag{2.42}$$

and its corresponding log-likelihood function is:

$$\boldsymbol{Ln\ L} = \sum_i 1(y_i = 0) ln\Big[P_1(y_i = 0 \mid Z_i, X_i)\Big] + \Big(1 - 1(y_i = 0)\Big) ln\Big[1 - P_1(y_i = 0 \mid Z_i, X_i)\Big] \tag{2.43}$$

$$+ \sum \Big[(1 - 1(y_i = 0)) ln\Big[P(y_i = j \mid Z_i, X_i)\Big]\Big]$$

Where L_1 can be regarded as a log-likelihood function for the binary (zero/positive) outcome, e.g., logit model and L_2 as a log-likelihood function for a truncated-at-zero (positive number of helminth parasites). The first hurdle follows the decision in which an individual does or does not have helminths parasites in the underlying binary probability distribution, logit. The second hurdle truncates non-zero counts in the underlying negative binomial distribution. Thus, MLEs of β_1 and

β_2 can be obtained separately from L_1 and L_2. Poisson model is obtained when the overdispersion parameter alpha, α_2, equals to zero. For computational simplicity, we use $\alpha_1 = 1$, which corresponds to using logit model at first stage in the double-hurdle negative binomial count data model.

While both ZI and hurdle models need distributional assumptions for their count component, both classes differ with respect to their dependencies of estimation of parameters of "zero" component on these assumptions (Loeys et al., 2011). Unlike ZI model, estimation of parameters β^* related to p_i^* in the hurdle model is not dependent on estimation of parameters γ^* related to μ_i^*. Hence, if assumptions about the (truncated) Poisson/negative binomial model are violated (for example due to extreme outlying observations), the hurdle model will, in contrast to ZI model, still yield consistent estimators for parameters in the logit part of the model (if correctly specified) (Loeys et al., 2011).

2.5 Zero adjusted models' applications

Lambert (1992) described Zero-Inflated Poisson (ZIP) regression models with an application to defects in manufacturing. Hall (2000) described Zero-Inflated Binomial (ZIB) regression model and incorporated random effects into ZIP and ZIB models. Cragg (1971) developed the idea for a hurdle model - a modified count model in which the processes generating zeros and positives are not constrained to be the same.

Ridout et al. (1998) considered various ZIP regression models for an Apple shoot propagation data. They concluded that ZIP models were inadequate for the data as there was still evidence of over dispersion. The first application of a ZINB model within a model based geostatistics (MBG) framework for *S. mansoni* infection was applied in Cote d'Ivoire (Vounatsou et al., 2009). This study showed that geostatistical ZI models produce more accurate maps of helminths infection

intensity than their spatial NB counterparts. Gurmu and Trivedi (1996) found out that Negative Binomial Hurdle model (NBLH), which allows for overdispersion and also accommodates presence of excess zeros, was more appropriate among all zero-adjusted models they considered. Recent applications have suggested that hurdle model is more plausible and, at the same time, a more manageable specification for explaining the preponderance of zeros that one typically finds in observed data (Greene, 2005).

A Bayesian analysis of ZIP models is given in Rodrigues (2003) and of ZINB models in Denwood et al. (2008). Zeileis, Kleiber, and Jackman (2008) compared generalised linear models (GLM) **:** Poisson and NB and zero adjusted models **:** Hurdle and ZI. From these, NBLH model presented a better fit to count data with overdispersion and excess zeros. Loeys et al. (2011) distinguished between zero-inflated models and hurdle models. They concluded that the choice between Hurdle and Zero-inflated models should be based on the aim and endpoints of the study. If the goal is prediction, it is not important which modeling framework is used, because predictions are (almost) identical. However, if the goal is inference, model choice is related to the study goal. From literature and empirical evidence available, Hurdle models (such as NBLH) therefore are more appropriate for modeling zero inflated count data as they allow for overdispersion and accomodate excess zeros.

2.6 Model selection

For comparison of non-nested models based on maximum likelihood to choose the best fitting model, Akaike's information criterion (AIC) has been proposed for model selection criteria based on the fitted log-likelihood function (Akaike, 1973; Cameron & Trivedi, 1998; Pan, 2001). A model with lowest AIC is preferred (Akaike, 1973). Several modifications of AIC also exist, viz Bayesian information criteria (BIC) and Consistent Akaike's information criterion (CAIC) (Cameron

& Trivedi, 1998). AIC is asymptotically optimal in selecting the model with the least mean squared error while BIC is not asymptotically optimal (Burnham & Anderson, 2004).

As a measure of the relative goodness of fit of a statistical model, AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. Since the log-likelihood is expected to increase as parameters are added to a model, the AIC criteria penalize models with larger k, the number of parameters in the model. This penalty function may also be a function of n, the number of observations (Cameron & Trivedi, 1998). This penalty discourages over-fitting. The AIC is specified as:

$$AIC = -2log(L) + 2k \tag{2.44}$$

where L is the maximized value of the likelihood function for the estimated model and $2k$ is the variance, with k being equal to number of parameters in the model. Bayesian information criterion is given as:

$$BIC = -2log(L) + (log(n))k \tag{2.45}$$

Consistent Akaike information criterion is given as:

$$CAIC = -2log(L) + (1 + log(n))k \tag{2.46}$$

Similar to AIC model (equation 2.44), k is the number of parameters in equations 2.45 and 2.46, whereas n is the number of observations.

2.7 Statistical methods in helminths modeling

Much work has been done on distribution and prevalence of Schistosomiasis and Hookworm infections. However mapping of disease distribution is complicated by limitations of available data. Most widely available data are from prevalence surveys (Gemperli, 2003). However, these surveys are generally carried out at arbitrary locations and include non-standardized and overlapping age groups (Gemperli, Vounatsou, Sogoba, & Smith, 2006). To augment approaches to rapid mapping and also address absence of suitable data in many settings, spatial prediction methods based on statistical relationships between individual and environmental predictors and infection risk, are increasingly being used (Magalhes et al., 2011).

Several methods have been used to estimate prevalence and distribution of Hookworm, Schistosomiasis and other tropical diseases rather than infection intensity, although the latter is particularly important for morbidity control. Most recently, predictive approaches to disease mapping have employed Bayesian model-based geostatistics (MBG) which embeds classical geostatistics in a GLM framework (Magalhes et al., 2011).

Raso et al. (2006) and Brooker et al. (2009) implemented multinomial spatial models for predicting risk of co-infection with multiple parasitic worms (helminths). Such models depend on observed co-infection data arising from a single survey on the same individuals. However, there is lack of these types of data since most surveys consider single infections.

In Man region, Cote d'Ivoire, the ability of models to capture co-infection risk was assessed on simulated data sets based on multinomial distributions assuming light- and heavy-dependent diseases, and a real dataset of *S. mansoni*-Hookworm co-infection (Schur, Gosoniu, Raso, Utzinger, & Vounatsou, 2011). In Mali, Niger and Burkina Faso, Clements et al. (2010) used a multinomial formulation to identify

areas with highest prevalence of high-intensity of *S. haematobium* infection and estimated the number of school-age children with high and low intensity infections.

Multinomial approach is the most straightforward and involves predicting prevalence of low and moderate/heavy intensity infections which can be useful tools for estimating burden of helminth diseases (Clements et al., 2010). However, the main limitation of this approach is that it involves stratifying egg counts, leading to a loss of information, whereas the NB-based approaches make full use of intensity data on a continuous scale (Magalhes et al., 2011). Therefore, an alternative approach is to model individual level egg counts.

Vounatsou et al. (2009) reported the first application of a ZINB model within an MBG framework for *S. mansoni* infection in Cote d'Ivoire to map Schistosomiasis transmission and predict infection intensity. This study showed that geostatistical zero-inflated models produce more accurate maps of helminths infection intensity than their spatial NB counterparts.

2.8 Residual spatial effects

In geostatistical studies, residual spatial effects have been investigated to check whether there are spatial patterns (spatial dependence). The variogram technique (Cressie, 1993) is used. In spatial statistics the theoretical variogram $2\gamma(x,y)$ is a function describing the degree of spatial dependence of a spatial random field or stochastic process $Z(x)$. It is defined as variance of the difference between field values at two locations across realizations of the field (Cressie, 1993):

$$2\gamma(x,y) = var(Z(x) - Z(y)) = E(|(Z(x) - \mu(x)) - (Z(y) - \mu(y))|^2) \qquad (2.47)$$

For a spatial random field with constant mean μ, the expectation reduces to equation 2.48 below where $\gamma(x,y)$ itself is called the *semivariogram*:

$$2\gamma(x,y) = E(|(Z(x) - (Z(y)|^2) \tag{2.48}$$

The empirical variogram is used in geostatistics as a first estimate of the (theoretical) variogram needed for spatial interpolation by kriging (Cressie, 1993). For observations Z_i, i = 1,...,k at locations x_1,...,x_k the empirical variogram $\hat{\gamma}$(h) is defined as equation (2.49). In this equation, *N(h)* denotes the set of pairs of observations *i,j* such that $|x_i$ - $x_j| = h$, and $|N(h)|$ is the number of pairs in the set. Generally an "approximate distance" h is used, implemented using a certain tolerance (Cressie, 1993).

$$\hat{\gamma}(h) := \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} |Z_i - Z_j|^2 \tag{2.49}$$

As an empirical variogram cannot be computed at every lag distance h, and due to variation in the estimation, it is not ensured that it is a valid variogram, as defined above. However, some geostatistical methods such as kriging need valid semivariograms. In applied geostatistics, empirical variograms are thus often approximated by model function ensuring validity (Cressie, 1993). Some of the common variogram models from the empirical variogram include *exponential* and *spherical* models, these are respectively given by:

$$\gamma(h) = (s-n)(1 - exp(-h/(ra))) + n1_{0,\infty}(h) \tag{2.50}$$

$$\gamma(h) = (s-n)\left(\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right)1_{0,r}(h) + 1_{r,\infty}(h)\right) \tag{2.51}$$

Bohling (2005) reports the following characteristics for variograms:

- **Nugget** - the height of jump of the semivariogram at the discontinuity of the origin.

- **Sill** - the limit of the variogram tending to infinity lag distances.
- **Range** - the distance in which the difference of the variogram from sill becomes negligible. Autocorrelation is essentially zero beyond the range.

For the study region's shapefile, all sampled villages had geo-coordinates determined by a portable Geographical Positioning System (GPS- Garmin eTrex®) machine.

In summary, the chapter has given an overview of zero inflated data generation and a review of basic methods of analysis with their properties/limitations. It also gives a review of zero inflated and altered methods for handling zero inflated count data and also methods for analysing residual spatial effects. The chapter has also introduced model selection techniques. The above reviewed methods have been used in analysing data for the current study in the applications section that follows.

Chapter 3

APPLICATIONS

3.1 Introduction

The need to update existing models in the context of renewed efforts to address disease burden in different ecological settings using different interventions, calls for new and updated models. With respect to this, various statistical models and methods have been developed to model helminths disease burden. For purposes of this study, selected zero adjusted models as covered in chapter two (2) have been applied to analyse count data with excess zeros to predict helminths infection intensity and prevalence in Chikhwawa district, Malawi and Lusaka province, Zambia. The two datasets were used in order to validate/fortify the method that offered a better fit to data. The methods used are Maximum Likelihood Estimation (MLE) based. This chapter gives a description of data, data sources and ethical clearance, the models fitted and model fitting, model comparison and selection criteria and residual spatial effects analysis.

3.2 Data sources

Two data sets were used in the analysis of zero inflated and altered data on human helminths. The first data set came from data collected in 2004 in a cluster randomised study that was conducted in Chikhwawa district in the Lower Shire

Valley - southern Malawi. This is a rural area whose population is mainly engaged in subsistence farming. This area lies between 100 and 300m above sea level. The rainy season extends from December to March. Temperatures can rise up to 50 °C in months preceding rainy season. Malaria is known to be holoendemic (Verhoeff, 2000).

Data were collected in eighteen villages, purposively selected from control and intervention arms of a cluster randomized study design. Ten percent of the households were randomly selected from the villages for baseline survey using random number tables (Ngwira, 2005). Further details are provided in Ngwira (2005). Briefly, in three of these villages, a survey mapping distribution of lymphatic filariasis had recently been completed. Four of these villages were taking part in malaria related entomological studies aimed at mapping genetic diversity of *Plasmodium falciprum* as well as a parallel lymphatic filariasis vector incrimination studies.

A two-stage sample selection was used. In the first stage, villages were selected, then at second stage, sample of households was listed and chosen. In the selected households all members aged one year and above were invited to participate. Consenting individuals had their demographic details completed and were given a full body clinical examinations (except genitals for females) for chronic manifestations of human helminths. In addition, they had anthropometric measurements taken and were asked to provide a single fresh stool and urine sample. All individuals (aged >1 year) were requested to provide a finger prick blood sample (Ngwira, 2005).

Fresh stool samples were transported in a cooler box to the laboratory and processed within four hours of collection. A single Kato-Katz thick smear was prepared from each sample and immediately examined under a light microscope for parasite eggs. Standardized and quality controlled procedures were followed . Briefly 41.7 mg of sieved stool was placed on a microscope slide through a punched

plastic template. Ova for each parasite observed were counted and expressed as eggs per gram (epg) of stool. Five percent of the slides were randomly selected for re-examination for quality control purposes (Ngwira, 2005).

Urine samples were processed on the day of collection. A measured volume (maximum 10 ml) was centrifuged at 300 revolutions per minute (RPM) for five minutes. The sediment was then examined under a light microscope. Number of eggs seen was counted and infection intensity per 10 ml of urine accordingly determined. All those infected were treated with *praziquantel* at 40 mg/kg (Ngwira, 2005).

The second data set came from data collected in a cross sectional study that was carried out in Lusaka province of Kafue and Luangwa districts, Zambia in 2004. The two districts were selected on the basis of their ecological representativeness of the country in general. In each of these districts 10 primary schools were selected. Approximately 100 school children, aged 6 to 15 years, were recruited from every school. The altitude and geographical location (longitude, latitude) of the surveyed schools were obtained from the archives of the Survey Department (2003). Further details of the study design are given elsewhere in Simoonga et al. (2008).

Briefly, data on *S. haematobium* prevalence and intensity were obtained using a Quantitative Filtration technique (Mott, 1984) to process duplicate urine samples collected about mid-morning. Two laboratory technicians were trained to prepare and read specimen filters. Both technicians read each specimen independently. This was useful in increasing sensitivity of the technique, particularly where egg intensity was low. All pupils found infected were treated with *praziquantel* (40 mg/kg body weight). Individual data sheets were used to collect ancillary information on each child examined.

In addition, data on intermediate host snails were also obtained through field collections and laboratory-based species identification. Sampling of potential Schistosomiasis transmission sites was done based on water body proximity to respective

primary school, that is, the nearest likely infection source. These water points were also qualified by relevant local people as the most frequented water contact points for both domestic and/or livestock. Climatic data were downloaded in 1-km image files from the website http://edcdaac.usgs.gov/1km-homepage.html. These images were captured by the Advanced Very High Resolution Radiometer (AVHRR) on board National Oceanic and Atmospheric Administration (NOAA) polar-orbiting meteorological satellites. They were then calibrated into Normalized Difference Vegetation Index (NDVI) and mid-day earth surface temperature (T_{max}) values using the ERDAS Imagine 8.5 (ERDAS, Atlanta, GA) software for each decadal (10-day) interval between April 1992 - September 1993 and February 1995-January 1996 (Simoonga et al., 2008).

3.3 Ethical approval

The study that collected data from Malawi received ethical clearance from Malawi College of Medicine Research Ethics Committee (COMREC) and the Ethics Committee of London School of Hygiene and Tropical Medicine (LSHTM) (Ngwira, 2005). Individual consent was obtained from each participant or (if they were aged <16) from one of their parents or guardian. Ethical approval for a study that collected data on urinary Schistosomiasis in school children in Zambia received ethical clearance from University of Zambia Ethics Committee (Simoonga et al., 2008).

3.4 Analysis

3.4.1 Descriptives analysis

Data were entered into *STATA 11.2 Inter Cooled (IC) Edition, StataCorp* which was also used to carry out exploratory data analysis and come up with descriptive statistics for summarizing and information presentation.

Descriptive data analysis that is, histograms/graphs and table summaries were done to present individual and group effects on the count outcomes. For categorical variables, proportions were generated as summaries whereas means and standard deviations were generated for continuous variables. P-values were generated for each of the variables to show whether it was significant or not. Polyparasitism analysis was performed for males and females and also for different age groups.

3.4.2 Statistical modeling

In order to see how well zero adjusted methods described in chapter two (2) fitted and captured the observed helminths count data, six models were fitted with main effects for predictor variables. They were assessed and compared for best fit to data. These are briefly described below:

- **Poisson model (P)** as the benchmark model for count data.
- **Negative Binomial Model (NB)**, a model derived from a Poisson - gamma mixture distribution. Its the standard parametric model to account for overdispersion in Poisson distributions.
- **Zero-Inflated Poisson model (ZIP)**, a mixture model in which the complete distribution of the outcome is represented by two separate components, a first part modeling the probability of excess zeros and a second part accounting for the non excess zeros and non-zero counts. It assumes a Poisson distribution.
- **Zero-Inflated Negative Binomial model (ZINB)**, similar to a ZIP model above but assuming a Negative binomial distribution.
- **Poisson Logit Hurdle model (PLH)**, a two part model assuming Poisson distribution. The first part is a binary outcome model, and the second part is a truncated count model. Such a partition permits the interpretation

that positive observations arise from crossing the zero hurdle or the zero threshold.

- **Negative Binomial Logit Hurdle model (NBLH)**, a two part model similar to PLH model above but assuming a Negative binomial distribution.

3.5 Model fitting

Statistical model fitting and likelihood ratio test (LRT) to compare the models' fit to data were carried out in *R version 2.14.0* (The R Foundation for Statistical Computing).

R's Political Science Computational Laboratory (PSCL) package developed to fit maximum likelihood estimation of ZI and Hurdle models for count data and goodness-of-fit measures for GLMs among others, was used. The PSCL package has several model-fitting functions and the one used in this analysis, for the Hurdle models is called *Hurdle ( )*. The function fits hurdle regression models for count data via maximum likelihood, which accepts as main arguments "a formula for regression fit" and a "character specification of the zero hurdle model family that is, NB or Poisson". Another function called *zeroinfl ( )* from the same package was used to fit ZI models. Both the fitting function interface and the returned model objects of class *zeroinfl ( )* are almost identical to the corresponding *Hurdle ( )* functionality and again modeled after the corresponding GLM functionality in R.

3.6 Model assessment

Model assessment and subsequent selection are important aspects in practical modeling. Choice between nested models (e.g., P versus NB) was made using a LRT. For a choice between non-nested mixture models (e.g., NBLH versus ZINB),

Akaike's Information Criterion (AIC) (Akaike, 1973) was employed, preferring the model with smallest AIC value. However, it turns out that in practice there is no or little difference in AIC between Hurdle model and Zero-Inflated model (Cameron & Trivedi, 1998). For a single binary predictor, ZINB model can be seen as reparametrization of NBLH model, and vice versa (Loeys et al., 2011). Models' ability to capture zero counts (Zeileis et al., 2008) was also assessed and compared among the models that offered a best fit to data based on AIC value.

3.7 Residual spatial effects

Distribution of residuals across the area was analysed to see if there were any spatial patterns in data from Chikhwawa, Malawi. This lead to subsequent mapping of residual spatial effects. From the empirical variogram, two models were suggested and subsequently fitted, viz *exponential* and *spherical*, as described in Chapter two (2) have been applied in the current residual spatial analysis.

For exponential variogram (see model 2.50), the nugget = 0.12, nugget is the height of jump of the semivariogram at the discontinuity of the origin. The sill = 0.20, it is the limit of the variogram tending to infinity lag distances. The range = 12, the distance in which the difference of the variogram from sill becomes negligible. For spherical variogram (see model 2.51), the nugget = 0.1, the sill = 0.20 and the range = 12. For residual spatial effects analysis, geoR package in **R** (Ribeiro & Diggle, 2011) was used to implement variogram analysis and kriging.

Chapter 4

RESULTS AND DISCUSSION

In this chapter, study results are presented in form of tables, graphs and figures for descriptive statistics and tables for statistical modeling. In each of the two applications, the first part gives exploratory results of data analysis and the second part gives results for statistical modeling. This chapter also gives a discussion of findings on each of the applications.

4.1 Results of analysis for Helminths infection prevalence and intensity in Chikhwawa, Malawi

4.1.1 Descriptive statistics results

Characteristics for study participants are summarized in Table 4.1. In total, 1642 individuals participated. There was a female excess (55.4 %) amongst study participants (more marked amongst those aged $\geq$ 11, see Figure 4.1). Of the total study population, 22.9 % had *Hookworm*, 5.0 % had *S. Mansoni* and 19.4 % had *S. haematobium*. Study participants had a mean age of 32.4 with a standard deviation of 22.79. Age and sex distribution of study participants is shown in Figure 4.1.

Table 4.1: Characteristics for individuals who had *S. haematobium*, *S. Mansoni* and *Hookworm* (N = 1642)

				S. haematobium ***233 (19.4 %)***	**S. Mansoni** ***71 (5.0 %)***	**Hookworm** ***324 (22.9 %)***
Variable	**Mean (Std. Dev)**	**Number(%)**	***t***	χ^2 **(P-value)**	χ^2 **(P-value)**	χ^2 **(P-value)**
Age (years)	32.36 (22.79)		47.73	(<0.001)	(<0.001)	(<0.001)
Sex						
Male		733 (44.6)		0.85 (0.358)	3.25 (0.073)	0.56 (0.454)
Female		909 (55.4)				
Education						
None		745 (45.4)		0.96 (0.618)	1.03 (0.599)	2.92 (0.233)
Primary		850 (51.8)				
Secondary		47 (2.9)				
Treatment arm						
Control		811 (49.4)		13.09 (<0.001)	3.27 (0.071)	10.89 (<0.001)
Intervention		831 (50.6)				
Fishing						
No		221 (13.5)		14.76 (<0.001)	0.17 (0.681)	22.89 (<0.001)
Yes		1,421 (86.5)				
Garden						
No		682 (41.5)		26.33 (<0.001)	11.02 (<0.001)	9.12 (0.003)
Yes		960 (58.5)				
Occupation						
Other		909 (55.4)		1.85 (0.174)	1.23 (0.268)	83.23 (<0.001)
Farmer		733 (44.6)				
Polyparasitism						
None		807 (49.2)		326.31 (<0.001)	289.92 (<0.001)	463.68 (<0.001)
One		594 (36.2)				
Two		200 (12.2)				
Three		38 (2.3)				
Four		3 (0.2)				
Deworm						
No		1,220 (74.3)		2.69 (0.100)	0.004 (0.948)	8.77 (0.003)
Yes		422 (25.7)				

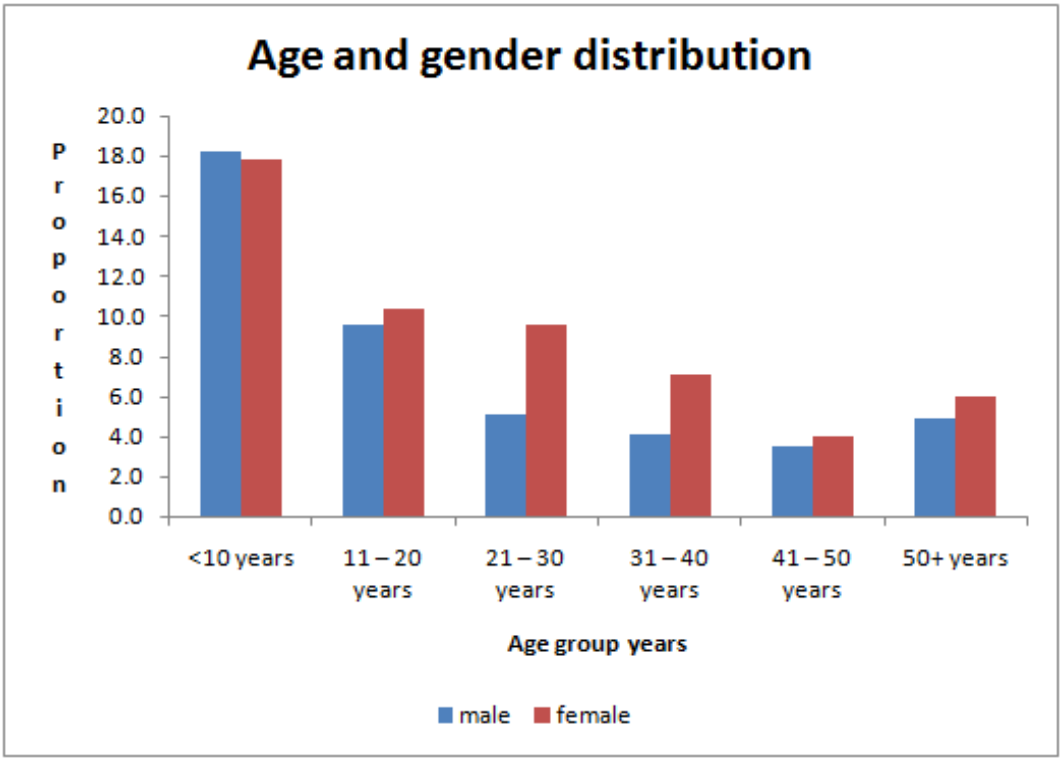

Figure 4.1: Age-gender distribution

From Table 4.1, age showed to be significant across all the outcomes, with a t - value = 47.73 and p-value <0.0001. All the outcomes showed to be insensitive to gender differences as well as education levels (see Table 4.1), the χ^2 values showed insignificant p - values. Working in gardens around the Shire river showed significant differences across all the three outcomes, with p - values <0.001 for various χ^2 values (see Table 4.1). Again, treatment showed to bring significant differences between the two arms and for outcomes: *S. haematobium* $\chi^2 = 13.09$, p - value <0.001 and *Hookworm* $\chi^2 = 10.89$, p - value <0.001. The number of parasites an individual hosted showed significant differences across the three outcomes with p - values <0.001 and χ^2 value = 326.31 for *S. haematobium*, $\chi^2 = 289.92$ for *S. mansoni* and $\chi^2 = 463.68$ for *Hookworm*.

The key outcome of interest was count data to quantify infection intensity and prevalence. For Schistosomiasis and Hookworm infections, often a large proportion of individuals are considered as "zero egg excretors" (Vounatsou et al., 2009) hence

the data is inflated with zeros. For Hookworm, 80 % were zero egg excretors, 95.7 % for *S. mansoni*, and 85.8 % for *S. haematobium* (see Figure 4.2)

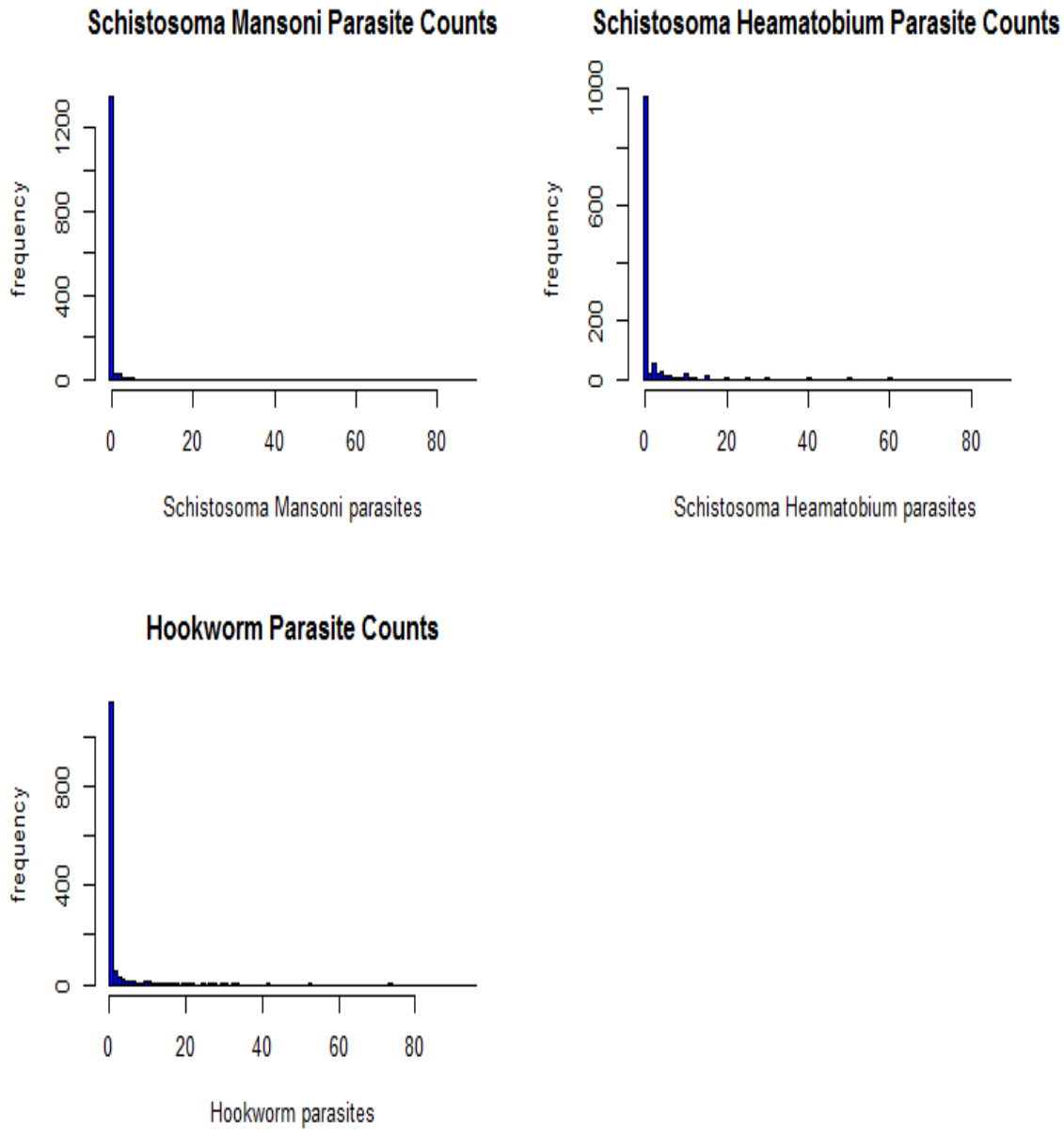

Figure 4.2: Zero inflated outcomes' counts for *S. mansoni*, *S. haematobium* and *Hookworm*

4.1.2 Polyparasitism

In tropical environments, Schistosomiasis, geohelminths, malaria and lymphatic filariasis, among others, are widespread parasitic infections posing an enormous toll on socio-economic development of the infected individuals as well as that of the general population (Ngwira, 2005). Schistosomiasis and Hookworm infections tend to be highly aggregated in a relatively small, heavily infected proportion of the population. Data were further analysed to investigate the epidemiology of multiple species parasite infections in the lower Shire valley. Parasites investigated in this analysis included *Hookworm*, *S. haematobium*, *S. mansoni*, and *lymphatic filariasis*. There were 993 individuals with a complete data set for this analysis. Of these, 544 (56 %) were female. Number of species and gender distribution is shown in Figure 4.3.

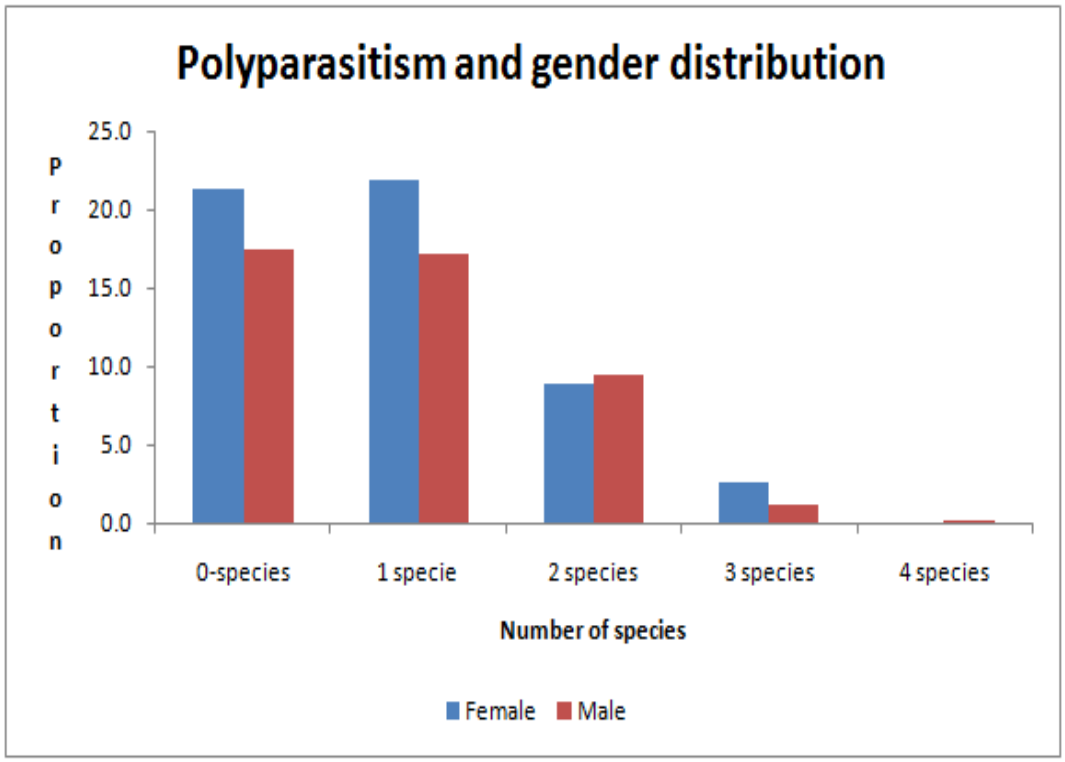

Figure 4.3: Number of species and gender distribution

There was female excess in all categories and following from Figure 4.1 above, it was shown that there was female excess in all age groups except in those ≤ 10.

This was probably a result of males being absent from homes during the study period as most of them were working on the sugar cane estate where they obtain temporary employment. Results of this investigation further showed that number of species reduced with an increase in age as shown in Figure 4.4.

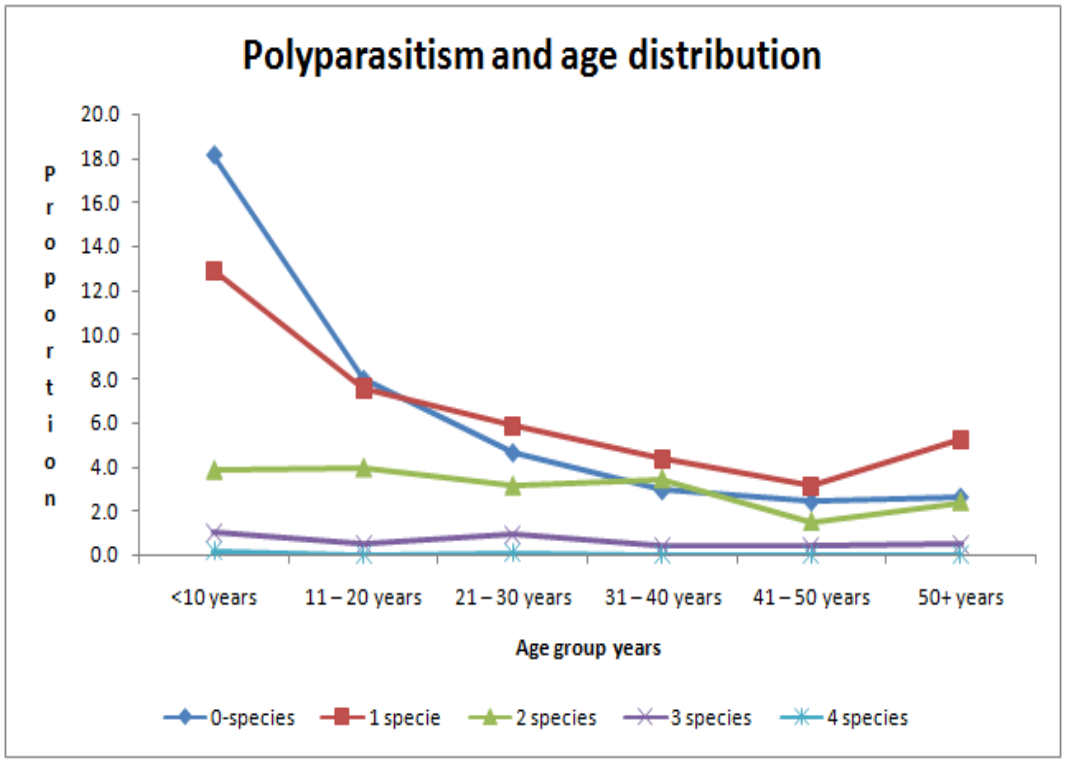

Figure 4.4: Multiple species and age graph

4.1.3 Statistical modeling results

As noted in chapter two (literature review), while the count nature of dependent variable implies use of a Poisson model or its variants, the methods used in this study suggest that the process generating the count outcomes was governed by a two-step structure also called hurdle process. For purposes of comparison, the following count models were estimated: Poisson, NB, Zero-Inflated (ZIP and ZINB) and Hurdle (Poisson and Negative Binomial) regressions, again including all independent variables in both transition and event stages in the two-component models. The LRT for overdispersion between Poisson and NB at $\alpha = 0.05$ showed a critical value test statistic = 2.71 with a χ^2 test statistic = 10606.51, p-value

<0.0001. Since p-value was <0.0001, there was overwhelming evidence of overdispersion. This therefore means that Poisson model was not appropriate for modeling the data hence NB - based models were the alternative. Table 4.2 gives a summary of Akaike Information Criterion (AIC) from the fitted models.

Table 4.2: Akaike information criterion (AIC)

	Poisson	**Neg. Bin.**	**ZIP**	**ZINB**	**PLH**	**NBLH**
AIC	14,182	3, 576	6, 854	3, 484	6, 854	**3, 482**

From Table 4.2, Negative Binomial Logit Hurdle (NBLH) showed to have the lowest AIC = 3,482, proving to be the model offering the best fit to data, however it had similar performance with ZINB model (AIC = 3,484). The Poisson model was inferior to all other fits with AIC = 14, 182 whereas NB based models dramatically improved the fit that is, for Neg. Bin model, AIC = 3, 576. This also reflects that overdispersion data was captured better by NB-based models than Poisson. Table 4.3 compares models in terms of zero count capturing ability. The Poisson model was again not appropriate as it could only capture 515 of the zeros whereas the NB-Zero adjusted based models were much better in capturing the zero counts. The NBLH model captured 971 zeros counts but was very comparable with ZIP and PLH models which each captured 970 zero counts. This therefore means that NB logit hurdle offered the best fit to zero inflated data for *S. Haematobium*, *S. Mansoni* and *Hookworm*.

Table 4.3: Zero count capturing

Observed	**Poisson**	**Neg. Bin**	**ZIP**	**ZINB**	**PLH**	**NBLH**
971	515	968	970	969	970	**971**

Since NB Logit Hurdle model offered the best fit to zero inflated helminth data in terms of the AIC (minimum value for all the models fitted) as well as true zero count capturing, it therefore became a natural choice for fitting a final model

to model helminths infection intensity and determination of factors that foster infections.

4.1.4 Fixed effects of infection probability

Table 4.4 provides estimates for fixed effects. The probability of *S. haematobium* infection was found to be associated with age (Adjusted Odds Ratio (AOR) = 0.97, 95 % Confidence Interval (CI): 0.96–0.99), with the risk of *S. haematobium* infection decreasing with age. The risk of infection was low in males than in females (AOR = 0.61, 95 % CI: 0.41–0.89). There was no association between risk of *S. haematobium* infection with education at both primary level (AOR = 1.18, 95 % CI: 0.81–1.71) and secondary level (AOR = 1.37, 95 % CI: 0.41–4.60) relative to no education. *S. haematobium* infection risk was also found to be associated with treatmment arm (AOR = 0.38, 95 % CI: 0.26–0.54) with those in intervention arm at a reduced risk of infection relative to control arm. There was a positive though insignificant association between *S. haematobium* infection probability and fishing (AOR = 0.73, 95 % CI: 0.44–1.20). Working in gardens along Shire river was observed not to be associated to *S. haematobium* presence (AOR = 1.34, 95 % CI: 0.90–1.99). Again, occupation (farmer/other) showed an association with *S. haematobium* infection probability though not signficant (OR = 0.61, 95 % CI: 0.35–1.06). Risk of infection increased with number of parasites an individual was hosting (Table 4.4) (AOR = 7.30, 95 % CI: 5.56–9.59).

For *S. Mansoni*, infection probability was associated with age, in that as age was increasing, the risk of infection decreased (AOR = 0.98, 95 % CI: 0.96–1.00). Similar to *S. haematobium*, *S. Mansoni* infection risk was low in males than in females (AOR = 0.55, 95 % CI: 0.31–0.97). There was an association between risk of *S. Mansoni* infection and treatment arm, that is, whether the village recieved intervention (MDA) or not, (AOR = 1.41, 95 % CI: 0.77–2.57), however from the 95 % CI, the association was not significant. Similar with *S. haematobium*, *S.*

Mansoni infection risk showed insignificant association with working in the garden (AOR = 1.30, 95 % CI: 0.66–2.54). Deworming and bathing in the river showed positive but insignicant associations as well, see Table 4.4. Again, similar to *S. haematobium*, having one parasite increased the risk of getting infected with other parasites (AOR = 5.78, 95 % CI: 4.15–8.06).

Hookworm infection risk was positively associated with age, with high risk as age was increasing (AOR = 1.02, 95 % CI: 1.01–1.03). There was a positive association between *Hookworm* infection and gender, with males showing higher risk of infection than females though this result was not significant (AOR = 1.04, 95 % CI: 0.78–1.38). Education and fishing showed positive but insignficant association with *Hookworm* infection, see Table 4.4. Those working in gardens showed a strong positive association with *Hookworm* infection as compared to those that did not (AOR = 1.32, 95 % CI: 0.99–1.75). Farmers were also at a greater risk of *Hookworm* infection as compared to non-farmers (AOR = 2.11, 95 % CI: 1.42–3.13).

Table 4.4: Fixed effects estimates for NBLH model (infection probability)

	S. haematobium	**S. Mansoni**	**Hookworm**
Variable	AOR **(95 % CI)**	AOR **(95 % CI)**	AOR **(95 % CI)**
Intercept	0.13 **(0.06, 0.29)**	0.00 **(0.00, 0.01)**	0.14 **(0.08, 0.25)**
age	0.97 **(0.96, 0.99)**	0.98 **(0.96, 1.00)**	1.02 **(1.01, 1.03)**
gender:			
Female	1.00	1.00	1.00
Male	0.61 **(0.41, 0.89)**	0.55 **(0.31, 0.97)**	1.04 **(0.78, 1.38)**
Education:			
None	1.00		1.00
Primary	1.18 **(0.81, 1.71**)		0.97 **(0.74, 1.27)**
Secondary	1.37 **(0.41, 4.60)**		0.35 **(0.12, 1.04)**
Treatment arm:			
Control	1.00	1.00	1.00
Intervention	0.38 **(0.26, 0.54)**	1.41 **(0.77, 2.57)**	1.57 **(1.21, 2.05)**
Fishing:			
No	1.00		1.00
Yes	0.73 **(0.44, 1.20)**		0.58 **(0.41, 0.83)**
Garden:			
No	1.00	1.00	1.00
Yes	1.34 **(0.90, 1.99)**	1.30 **(0.66, 2.54)**	1.32 **(0.99, 1.75)**
Occupation:			
Other	1.00		1.00
Farmer	0.61 **(0.35, 1.06)**		2.11 **(1.42, 3.13)**
Deworm:			
No		1.00	
Yes		1.19 **(0.61, 2.32)**	
Bath:			
No		1.00	
Yes		1.65 **(0.64, 4.24)**	
Polyparasitism	7.30 **(5.56, 9.59)**	5.78 **(4.15, 8.06)**	

4.1.5 Fixed effects of infection intensity

From Table 4.5, it was observed that *S. haematobium* infection intensity reduced with age (Relative Risk (RR) = 0.96, 95 % CI: 0.95–0.98). There was a marginal difference of infection intensity between males and females (RR = 1.03, 95 % CI: 0.72–1.47). Primary school children showed a high *S. haematobium* infection intensity relative to those that were in pre-school level (RR = 1.54, 95 % CI:

1.08–2.19) whereas those in secondary level showed a reduced infection intensity (RR = 0.34, 95 % CI: 0.11–1.06). There was a reduced *S. haematobium* infection intensity for those in intervention arm relative to those in the control arm, though, not significant (RR = 0.81, 95 % CI: 0.58–1.13). A positive association was also observed between those who did fishing in Shire river and infection intesity (Table 4.5) relative to those who did not do fishing. An increased risk of infection intensity was observed in those working in gardens relative to those that did not, (RR = 1.21, 95 % CI: 0.82–1.81) and also increased *S. haematobium* infection intensity for farmers compared to non-farmers (RR = 1.83, 95 % CI: 1.16–2.91).

For *S. Mansoni*, there was a positive association with age and intensity showed to increase with age (RR = 1.01, 95 % CI: 1.00–1.02). There was a reduced *S. mansoni* infection intensity for males as compared to females though the difference was marginal (RR = 0.61, 95 % CI: 0.35–1.08). Village level intevention efforts reduced infection intenisty with (RR = 0.77, 95 % CI: 0.46–1.28) for those in intervention arm compared to those in the control arm; this though showed no real significant association with infection intensity. Infection intenisty for *S. mansoni* was higher in those working in gardens as compared to those that did not, however this association was not very significant (RR = 1.29, 95 % CI: 0.64–2.60). Equally, there was increased infection intensity for those bathing in Shire river as compared to those that did not, (RR = 1.46, 95 % CI: 0.45–4.67) though the association was not significant. Deworming showed to have significantly reduced *S. mansoni* infection intensity among the dewormed group as compared to control group, (RR = 0.52, 95 % CI: 0.27–0.99).

There was a positive association between age and *Hookworm* infection intensity, thus as age increased, infection intensity also increased (RR = 1.01, 95 % CI: 1.00–1.03). Intensity of infection was higher in males as compared to females (RR = 1.19) though not significant (95 % CI: 0.80–1.76). Similar to age, an increase in education showed to be positively associated with Hookworm infection

intensity: RR = 1.18 for primary education level and RR = 2.18 for secondary education level. These however were not signficant (95 % CI: 0.80–1.72) (for primary education level) and (95 % CI: 0.45–10.55) (for secondary education level). Fishing (yes/no), working in the garden (yes/no) and occupation (farmer/other) all showed association with Hookworm infection (see Table 4.5)

Table 4.5: Fixed effects estimates for NBLH model (infection intensity)

	S. haematobium	**S. Mansoni**	**Hookworm**
Variable	RR **(95 % CI)**	RR **(95 % CI)**	RR **(95 % CI)**
Intercept	11.72 **(5.70, 24.08)**	1.48 **(0.44, 4.97)**	3.65 **(1.72, 7.71)**
age	0.96 **(0.95, 0.98)**	1.01 **(1.00, 1.02)**	1.01 **(1.00, 1.03)**
gender:			
Female	1.00	1.00	1.00
Male	1.03 **(0.72, 1.47)**	0.61 **(0.35, 1.08)**	1.19 **(0.80, 1.76)**
Education:			
None	1.00		1.00
Primary	1.54 **(1.08, 2.19)**		1.18 **(0.80, 1.72)**
Secondary	0.34 **(0.11, 1.06)**		2.18 **(0.45, 10.55)**
Treatment arm:			
Control	1.00	1.00	1.00
Intervention	0.81 **(0.58, 1.13)**	0.77 **(0.46, 1.28)**	1.25 **(0.87, 1.80)**
Fishing:			
No	1.00		1.00
Yes	0.68 **(0.45, 1.03)**		0.90 **(0.58, 1.40)**
Garden:			
No	1.00	1.00	1.00
Yes	1.21 **(0.82, 1.81)**	1.29 **(0.64, 2.60)**	1.12 **(0.75, 1.67)**
Occupation:			
Other	1.00		1.00
Farmer	1.83 **(1.16, 2.91)**		0.86 **(0.49, 1.53)**
Deworm:			
No		1.00	
Yes		0.52 **(0.27, 0.99)**	
Bath:			
No		1.00	
Yes		1.46 **(0.45, 4.67)**	
Polyparasitism	0.87 **(0.70, 1.08)**	0.91 **(0.69, 1.18)**	

4.1.6 Residual spatial effects

Distribution of residuals across the study area was analysed to check for spatial patterns. Estimating the continuous surface using interpolation function, spatial patterns in the residuals were observed and subsequently mapped, (Figure 4.5).

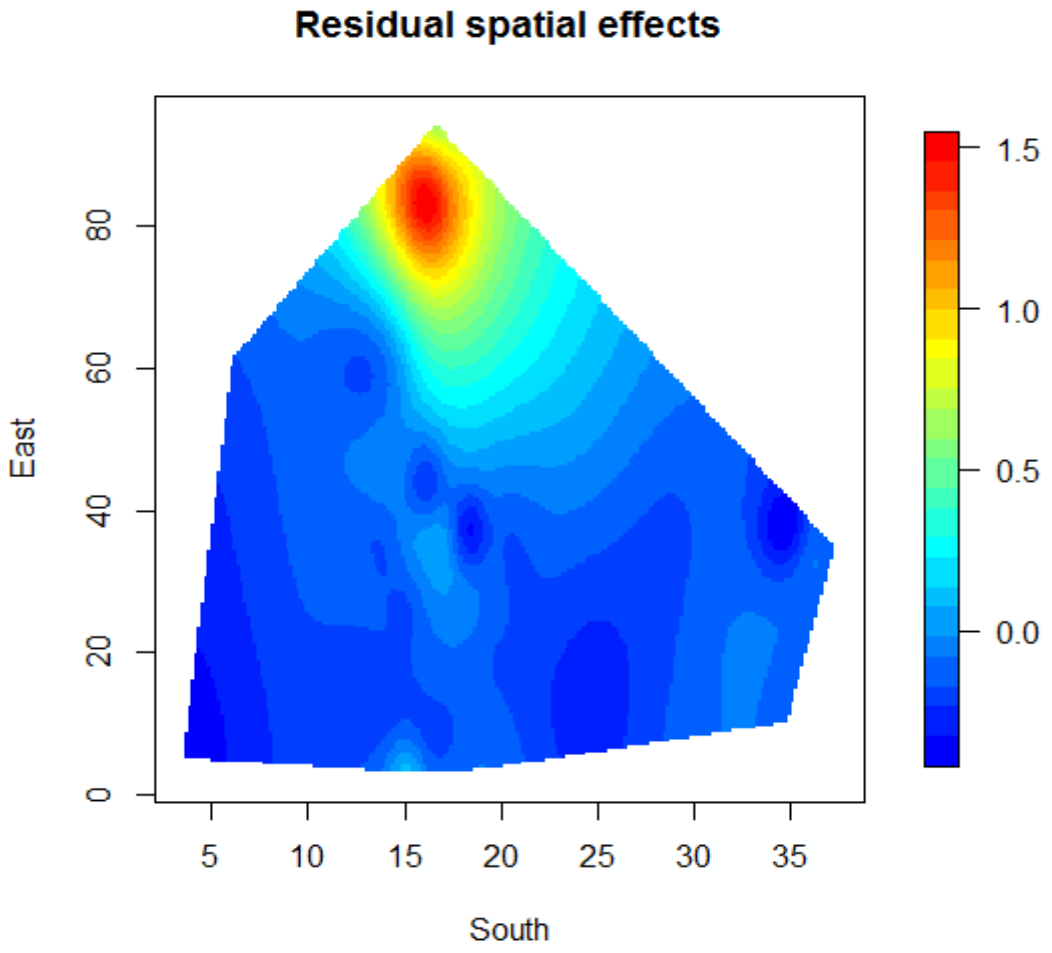

Figure 4.5: Estimated residual spatial effects, given in odds ratios

From the visual inspection of Figure 4.5, there was some degree of spatial dependence in residuals distribution across the study area. This encouraged further exploration using variogram analysis to quantify the range of spatial dependence and apportioning of variance into spatial and non-spatial components. Common descriptive statistics and histograms fail to identify, and quantify textural differ-

ences in the data, and they do not incorporate spatial locations of data into their defining computations. Variogram is a quantitative descriptive statistic that can be graphically represented in a manner which characterizes the spatial continuity (that is, roughness) of a data set. Empirical variogram suggested two models that have been fit, namely: exponential and spherical variogram models (Figure 4.6).

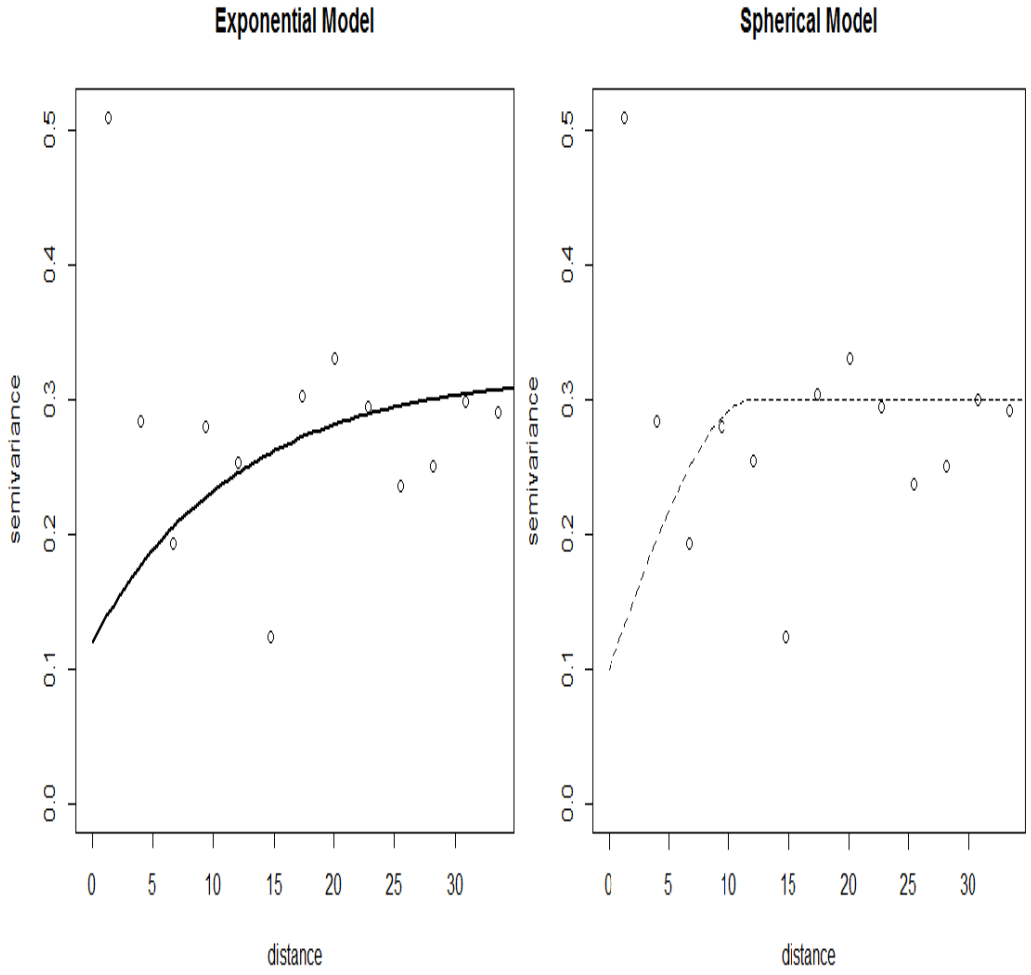

Figure 4.6: Exponential and spherical variogram plots based on the deviance residuals.

From the variogram plots (Figure 4.6), geometric anisotropy behaviour was exhibited by the residuals. The magnitude of spatial correlation decreased with separation distance until a distance at which no spatial correlation existed.

4.1.7 Discussion

The current study found a prevalence of 19.4 % for *S. haematobium* in Chikhwawa district. The finding highlights the fact that *S. haematobium* infections are highly localised (see Figure 4.5). The overall prevalence of *S. haematobium* was well below expectations. Based on a previous study in Malawi, overall prevalence of Schistosomiasis was thought to be between 40 % and 50 % in the population. These were surveys of selected populations, perhaps undertaken in the season of high transmission (Bowie, Purcell, Shaba, Makaula, & Perez, 2004). Certainly, the diseases are domiciled and localised in Malawi. The finding serves to highlight the fact that helminths infections were highly localised (see Figure 4.5), and that national wide surveys tend to overstep focus of heterogeneity of infection. In a study conducted in the northern lakeshore area by Randal et al. (2002), school children from four (4) schools were screened and there was a wide range of prevalence: 5 % - 57 % on *S. haematobium*. A national survey, representative of all school children in the country, and undertaken just before the rainy season, suggested far lower levels of 7 % for *S. haematobium* (Bowie et al., 2004). The prevalence for *S. mansoni* in Chikhwawa district was 5.0 %. The findings highlights the fact that *S. mansoni* was not common. A previous study showed a prevalence of 0.4 % for *S. mansoni* in standard 3 pupils (modal age 10 years of age) (Bowie et al., 2004). The prevalence of Hookworm was 22.9 %, well above previous studies' findings. Bowie et al. (2004) reported a prevalence of 1.3 % (95 % CI : 0.4-2.3 %) for Hookworm.

Robust and contemporary statistical methods in a two part application were used to analyse risk factors for *S. haematobium*, *S. mansoni* and Hookworm infection intensity and prevalence. This resulted in estimates of parasitic infection intensity and prevalence that could be used in control programmes planning. In this study, intensity and prevalence of helminths were examined in relation to factors such as age, sex, education level, treatment arm, fishing in Shire river, working in gardens,

occupation and polyparasitism. The study confirmed the critical importance of ascertaining infection intensity. The higher the intensity of helminths infection, the higher the burden.

S. haematobium and *S. mansoni* infection intensity showed a reduction with an increase in age. This confirmed what previous studies found. Chan et al. (1997) report that in common intestinal helminths such as *Ascaris lumbricoides* (large roundworms) and *Trichuris trichiura* (whipworm) and Schistosomiasis, children are more heavily affected and infected than adults. Several other studies have reported that school-aged children show high infection intensity and prevalence (Bowie et al., 2004; Saathoff et al., 2004; Midzi et al., 2011). Unlike Schistosomiasis, Hookworm infection intensity showed an increase with age. Chan et al. (1997) report that Hookworm is more common in adults which means that child-targeted chemotherapy programmes advocated for the treatment of other species may be less appropriate in the mass control of Hookworms. Exposure differences for Hookworm are responsible for differences between children and adults; as Hookworm is generally transmitted in the fields (where adults work) as opposed to near homes (where children play).

Fishing in Shire river and working in gardens along the river were clear risk factors for exposure to helminths and subsequent infection because transmission requires contact with the aquatic habitat of intermediate host snails as well as with soil. Clements et al. (2010) report for a study that was conducted in western Africa that contact with water bodies that are a habitat for intermediate host snails is one of the main risk factors. Results showed low probability of infection for males compared to females. This could be explained by a number of factors including that Malawi being an agriculture based economy, with females mainly carrying out agricultural activities,they are more exposed to risk factors such as working in gardens and farming.

Individuals who had received chemotherapy cure for helminth showed reduced

risk of infection as well as infection intensity as compared to those in the control area. Evidence has shown that, following chemotherapeutic cure of *S. mansoni* or *S. haematobium* infection, older individuals display a resistance to reinfection in comparison to younger children (Roberts et al., 1993). This shows that there is need to direct control and interventions for helminths to areas with diseases burden in order to reduce and/or eradicate infections - more especially in children.

Several studies have shown that having one infection is a risk factor for getting other infections. It is conceivable that the first parasite that establishes an infection may modulate body's immune response in such a way that it makes it easier for the next parasite to infect the body (Ngwira, 2005). Results from the current analysis have shown that polyparasitism was common in the population of Chikhwawa especially among females. Results also suggested that morbidity may be associated with number of parasite species an individual was carrying.

Worth noting are differences that existed in the associations between infection probability and infection intensity. For gender, males had a reduced risk of infection as compared to females (negative association) but high infection intensity (positive association). This could possibly be explained by the fact that women were mostly involved in agricultural activities there by being more exposed to risk factors. Also, for those infected (males), many studies find that men visit public health care facilities much less frequently than do women (Kuwane et al., 2009; Myburgh, 2011) hence the high intensity.

Polyparasitism was positively associated with infection probability but had a negative association with infection intensity. This could be explained by the fact that having other parasites increases the chance of the body being susceptible to new parasite infections (Cassia et al., 2007). Secondary level of education had a positive association with infection probability but showed a negative association with infection intensity. This finding could be explained by the fact that education may correspond to increased awareness and access to treatment (Spear et al., 2004) by

this group hence reduced intensity. A study by Spear et al. (2004) found out that those with highest level of education, that is, through high school, showed the lowest mean infection intensity.

Being a farmer had a negative association with probability of infection and a positive association with infection intensity. The finding was in line with what Spear et al. (2004) accounted. In their study, farmers showed highest levels of Schistosomiasis infection among occupational groups. Spear et al. (2004) reported that both education and occupation were proxies for nature and intensity of water contact. Amhanyunonsen (2009) reported that individuals become infected by prolonged contact (like farm irrigation, bathing, washing or swimming) with fresh water containing free-swimming *cercariae*.

4.2 Results of analysis for urinary Schistosomiasis in school children in Lusaka Province, Zambia

4.2.1 Descriptive statistics results

Table 4.6 gives characteristics of study population. A total of 2040 school children aged 6 to 15 years were enrolled into the study from 20 selected primary schools in the two districts, Kafue and Luangwa, 1909 (93.5 %) provided urine samples for parasitological examination. The remaining children 131 (6.5 %) did not provide urine samples for examination. Overall *S. haematobium* prevalence rate for two districts was 9.6 % (range: 0 – 36.1 %). Infection intensity had a mean of 31.4 eggs/10ml (range: 0 – 120 eggs/10ml). However there was a significant difference in the mean intensity of infection, with 40.2 (range: 3 – 53.1 eggs/10ml) observed in Kafue district and 22.6 (range: 0 – 116.0 eggs/10ml) in Luangwa district. For Schistosomiasis, a large proportion of individuals were "zero egg excretors" (84.6 %) - see Figure 4.7.

Table 4.6: Characteristics and intensity of infection with *S. haematobium* in 2040 children from 20 schools in Lusaka Province, Zambia, 2004

Variable	**Mean(Std. Dev)**	**Number(%)**	**t**	χ^2	**P-value**
Intensity of infection					
No infection (0 eggs/ml: epm)		1726 (84.6)			
Light infection (1-100 epm)		139 (6.8)			
Mod/heavy infection (>100 epm)		44 (2.2)			
Age (years)	9.98 (2.14)				
6-9 years		1130 (55.4)		4.0	0.059
10-15 years		900 (44.1)			
Sex					
Female		1027 (50.4)		2.4	0.124
Male		1000 (49.0)			
Altitude					
Plateau		723 (35.4)		29.5	<0.0001
Valley		1316 (64.5)			
NDVI	138.2 (5.1)		1280.47		<0.0001
TMAX	19.6 (2.9)		282.26		<0.0001
Snail abundance (B. globosus)	25.3 (29.9)		30.82		<0.0001

Similar to results from Malawi study, gender showed to be insensitive to urinary Schistosomiasis prevalence and intensity with a $\chi^2 = 2.4$ and a p-value = 0.124. Age showed marginal differences between the 6-9 years and 10-15 years age groups, $\chi^2 = 4.0$ with p-value = 0.059 (see Table 4.6). With a $\chi^2 = 29.5$, altitude showed significant difference between valleys and plateaus in influencing infection prevalence and intensity. From Table 4.6, Normalised difference vegetation index (NDVI) (t = 1280.47, p-value <0.001) showed a significant impact on urinary Schistosomiasis. Again, Tmax (t = 282.26) and snail abundance (t = 30.82) both with p-values <0.001 also showed significant impact on urinary Schistosomiasis.

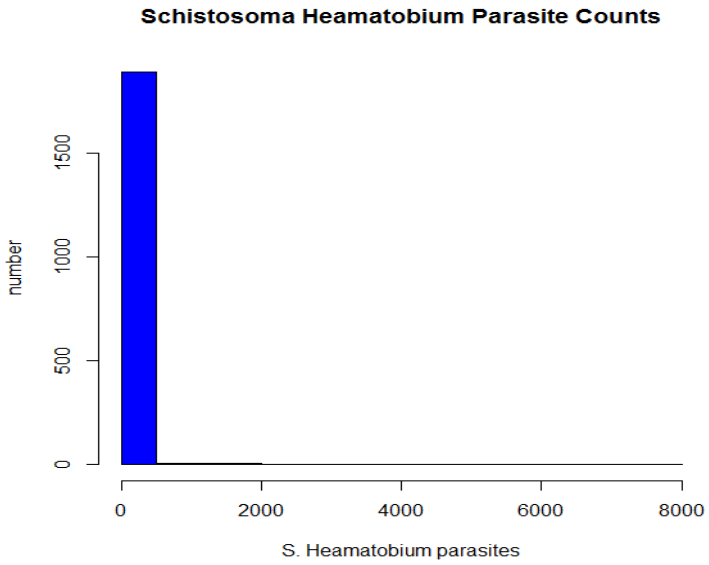

Figure 4.7: Zero inflated outcome count for urinary Schistosomiasis

4.2.2 Statistical modeling results

Zero Inflated (ZIP/ZINB) and hurdle (NBLH) models were compared using Akaike Information Criteria (AIC) (Akaike, 1973). Table 4.7 shows AIC values for the three estimated models. NBLH model had a lowest AIC and therefore the best fitting (AIC = 3,230 versus AIC = 3,232 in ZINB model). From NBLH model (Table 4.7), probability of urinary Schistosomiasis infection was shown to have a significant association with age - thus infection probability was lower in 6 - 9 years age group (AOR = 0.69, 95 % CI: 0.50–0.94). NBLH results showed lower risk in younger children as compared to older children. Infection probability showed a positive association with sex (AOR = 1.17, 95 % CI: 0.86–1.60) though there was no significant difference between females and males. However, significant differences in infection probability were observed between valley and plateau (AOR = 0.37, 95 % CI: 0.25–0.53) with those in the valley being at reduced risk of infection compared to those in the plateau. A positive relationship was observed between snail abundance and risk of infection, though marginally significant at 5 % (AOR=1.00, 95 % CI: 1.00-1.01). Marginal positive associations were ob-

served between urinary Schistosomiasis and NDVI (the mean Dec-Nov biannual composites of NDVI) with AOR=1.04 (95 % CI: 1.00-1.07), as well as with Tmax (maximum temperature) (AOR=0.99, 95 % CI: 0.94-1.04).

Table 4.7: Estimated adjusted odds ratios (AORs) of factors associated with Urinary Schistosomiasis infection obtained from the zero adjusted models

	ZIP Model	ZINB Model	NBLH Model
	AOR **(95 % CI)**	AOR **(95% CI)**	AOR **(95 % CI)**
Age			
6 - 9 years	1.46 **(1.06, 2.00)**	1.43 **(0.96, 2.11)**	0.69 **(0.50, 0.94)**
10 - 15 years	1.00	1.00	1.00
Gender:			
Male	0.85 **(0.62, 1.16)**	0.87 **(0.59, 1.28)**	1.17 **(0.86,1.60)**
Female	1.00	1.00	1.00
Altitude			
Valley	2.73 **(1.89, 3.96)**	2.38 **(1.46, 3.89)**	0.37 **(0.25, 0.53)**
Plateau	1.00	1.00	1.00
Tmax	1.01 **(0.96, 1.07)**	0.99 **(0.92, 1.05)**	0.99 **(0.94, 1.04)**
NDVI	0.96 **(0.93, 1.00)**	0.97 **(0.92, 1.01)**	1.04 **(1.00, 1.07)**
Snail abundance	1.00 **(0.99, 1.00)**	1.00 **(0.99, 1.00)**	1.00 **(1.00, 1.01)**
Model Selection			
AIC	148, 734	3, 232	**3, 230**

From Table 4.8, infection intensity was observed to be marginally associated with age (RR = 0.55, 95% CI: 0.25–1.19). Similar to infection probability pattern, intensity was lower in younger age group relative to older age group. Sex was positively associated with infection intensity, with increased intensity in males relative to females (RR = 1.28, 95 % CI: 0.57–2.87) albeit insiginficant. It was observed that children in the valley had low urinary infection intensity relative to those in plateaus (RR = 0.11, 95 % CI: 0.04–0.28). Temperature was negatively associated with urinary Schistosomiasis intensity (RR = 0.75, 95 % CI: 0.75–0.94). Snail abundance was marginally associated with infection intensity (RR = 1.00, 95 % CI: 0.99–1.01). A positive association was observed between infection intensity and NDVI (mean Dec-Nov biannual composites of NDVI) (RR = 1.07, 95 % CI: 0.99–1.16) .

Table 4.8: Estimated relative risk (RRs) factors associated with Urinary Schistosomiasis infection obtained from the zero adjusted models

	NBLH Model	
	RR	**(95%CI)**
Age		
6 - 9 years	0.55	**(0.25, 1.19)**
10 - 15 years	1.00	
Gender:		
Male	1.28	**(0.57, 2.87)**
Female	1.00	
Altitude		
Valley	0.11	**(0.04, 0.28)**
Plateau	1.00	
Tmax	0.84	**(0.75, 0.94)**
NDVI	1.07	**(0.99, 1.16)**
Snail abundance	1.00	**(0.99, 1.01)**

4.2.3 Discussion

Factors associated with *S. haematobium* infection among school children in Lusaka province were quantified using Negative Binomial Logit Hurdle (NBLH). The proportion with light infection (6.81 %) and moderate to high infection (2.2 %) were very low (Table 4.6 and Figure 4.7) compared to no infection (84.6 %). This finding related well with results obtained from helminths study (described above) carried out in Chikhwawa district, Malawi. Previous studies on helminths have highlighted that infection prevalence is usually low (Bowie et al., 2004) but highly focalised in a locality - with few individuals hosting large numbers of parasites.

Both infection prevalence and intensity showed differences in relation to the two age groups that were considered (Tables 4.6 and 4.7). The younger age group showed reduced infection risk and intensity as compared to older children. The differences were as a result of increased risk-behaviour of older school children who frequently contacted schistosome-infested water for both domestic and livestock purposes relative to younger children (Simoonga et al., 2008). Schistosomiasis is water dependent disease and incidence is usually more amongst people who

constantly interact with schistosome infected waters through activities such as farming, fishing, swimming and laundry (Amhanyunonsen, 2009).

In terms of gender, males showed increased infection risk and intensity albeit not significantly different from females, implying that infection prevalence and intensity were not gender selective. This could be explained by the fact that interaction with schistosome infested water was gender independent as both groups were involved with behavioral activities such as swimming, washing and farming. The results obtained were similar to a study that was done in Etsako east LGA, Edo state, Nigeria. In the study, Amhanyunonsen (2009) reported that disease impact was common to both males and females, although more males than females were found to be infected.

Results indicated that both prevalence and intensity of infection were low for school children in the valleys than those in plateaus. In relation to this finding, Simoonga et al. (2008) accounted that during dry season, school children on plateaus usually had a higher degree of water-contact unlike those in the valley. The reason was that water sources for domestic and livestock purposes remained relatively unlimited due to perennial rivers flowing through such as the Zambezi River. Temperature was shown to be a clear factor for Schistosomiasis infection. Warmer temperatures have been reported as being optimal for the development and maintenances of a *B. globosus – S. haematobium* system (Simoonga et al., 2008).

A study by Zhou et al. (2008) found a temperature threshold of 15.4 °C for development of *S. japonicum* within the intermediate host snail (that is, *Oncomelania hupensis*), and a temperature of 5.8 °C at which half the snail sample investigated was in hibernation.

As expected and observed, NDVI and abundance of intermediate host snails (*Bulinus globosus*) had influence on the transmission of urinary Schistosomiasis. One of the significant factors for Schistosomiasis transmission in an area is the successful

development of intermediate host snail-parasite system.

Chapter 5

RECOMMENDATIONS AND CONCLUSION

The apparent dominance of agricultural, environmental, socio-economic and demographic factors in determining human helminths infection risk in the communities carries important implications for disease surveillance and control strategies. Prevalence of helminths was highly associated with age of an individual as well as occupational/behavioral activities (such contact with infected water) and also number of parasites an individual hosted. Furthermore, helminths infection intensity was associated with gender, education level, garden, abundance of intermediate host snails, occupation and treatment (intervention). Results presented therefore highlight the need to understand environmental and human behavior patterns with respect to contact and contamination in human helminth epidemiology.

Cercariae control through environmental modifications and strategies involving socio-economic status improvement and MDA may be more promising approaches to disease control in the infected settings. Thus, it is important that an integrated approach within a primary health care system is adopted by ensuring infection source reduction through control of intermediate host snails combined with chemotherapy for morbidity control. Behavioral factors, such as water and soil contact activities have been shown to be particularly important for risk mapping

in both applications describe above. Therefore, helminths control programmes should consider provision of safer alternative water sources for both domestic and livestock purposes during dry season when the only available water sources are collected pools that are often infested with *S. haematobium*-infected snails. For Hookworm, awareness and ensuring the use of protective wear when working in the fields could also be promising. Again, treatment to adults, perhaps targeted to particular risk groups, may be advisable in addition to the usually practiced child-based chemotherapy.

In both applications for data from Chikhwawa district, Malawi and Lusaka province, Zambia, results have indicated that the risk of infection with helminths is heterogeneous. This therefore calls for the need to undertake further focalised studies to ascertain further exposure risk factors. By applying the methodology to two different applications, the study managed to use schools as spatial points in one application and also used households as spatial points in another application in assessing high infection risk areas.

The study has constructed a set of statistical models (ZI and Hurdle) which can be used together with planning interventions in human helminths and other helminth parasite systems. Zero adjusted methods represent a key advance in the analysis of helminth disease data inflated with zeros. There is an increasing number of examples in published literature where two part methods are being used for zero inflated data for disease control planning and implementation programmes. The joint modeling approach has allowed identification of risk factors for both infection prevalence and severity and provide a platform to evaluate progress of control efforts (that is, comparison of intervention (MDA) and control arms arm). Use of such robust methods also allowed discovery of important differences between the two outcomes (infection presence and severity). This has an implication in that it necessitates improvement of interventions designing.

The choice between Hurdle and Zero-Inflated models should be based on the aim and study endpoints. If the goal is prediction, it is not important which modeling framework is used, because predictions are (almost) identical. However, if the goal is inference, model choice is related to the study goal. Statistical modeling results presented have shown that NBLH offered the best fit to zero inflated data with a capability to handle overdispersion and excess zeros as well as capturing true zeros in the data.

The NBLH approach allows the decomposition of effects of factors into participation and consumption decisions: participation decision (helminths infection presence/absence) and consumption/frequency decision (helminths parasite counts or intensity). The NBLH also allows accounting for both unobserved heterogeneity and excess zeros in zero inflated data. The ease of implementation and straightforward interpretation of the components and its direct link with observed data makes the NBLH model a valuable alternative for researchers analysing zero-inflated count data.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267-281). Budapest: Akademiai Kiado.

Alemu, A., Atnafu, A., Addis, Z., Shiferaw, Y., Teklu, T., Mathewos, B., ... Gelaw, B. (2011). Soil transmitted helminths and *schistosoma mansoni* infections among school children in Zarima town, Northwest Ethiopia. *BMC Infectious Diseases*, *11*(1), 189.

Amhanyunonsen, J. A. (2009). *Ecological Correlates of schistosomiasis incidence and access to PHC delivery in Etsako East LGA, Edo State, Nigeria* (Tech. Rep.). Nigeria: Colloque International.

Bethony, J., Brooker, S., Albonico, M., & Hotez, P. J. (2006). Soil-transmitted helminth infections: Ascariasis, Trichuriasis, and Hookworm. *Lancet*, *367*, 1521-1532.

Bohling, G. (2005). *Introduction to geostatistics and variogram analysis*. Kansas: Kansas Geological Survey.

Bowie, C., Purcell, B., Shaba, B., Makaula, P., & Perez, M. (2004). A national survey of the prevalence of Schistosomiasis and soil transmitted helminths in Malawi. *BMC Infectious Diseases*, *4*(49), 49.

Bradley, D. J. (1972). Regulation of parasite populations: A general theory of the epidemiology and control of parasitic infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *66*, 697-708.

Brooker, S., Alexander, N., Geiger, S., Moyeed, R. A., Standerd, J., Fleming, F.,

... Bethony, J. (2006). Contrasting patterns in the small-scale heterogeneity of human helminth infections in urban and rural environments in Brazil. *International Journal of Parasitology*, *36*(10-11), 1143–1151.

Brooker, S., Hay, S., Tchuente, L., & Ratard, R. (2002, February). Using NOAA-AVHRR data to model human helminth distributions in planning disease control in Cameroon, West Africa. *Photogametric Engineering and Remote Sensing*, *68*(2), 175-179.

Brooker, S., Hotez, P., Bethony, J., & Silva, N. (2003, March). *Soil transmitted helminths: The nature, causes and burden of the condition.* Working Paper No. 3, Disease Control Priorities Project. Bethesda, Maryland: Fogarty International Center, National Institutes of Health.

Brooker, S., Kabatereine, N. B., Smith, J. L., Mupfasoni, D., Mwanje, M. T., & Ndayishimiye, O. (2009, July). An updated atlas of human helminth infections: the example of East Africa. *International Journal of Health Geographics*, *8*, 42.

Bundy, D. A. P. (1995). Epidemiology and transmission of intestinal helminths. In M. Farthing, G. Keusch, & D. Wakelin (Eds.), *Enteric infection 2,intestinal helminths.* London: Chapman & Hall Medical.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, *33*, 261–304.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data.* New York: Cambridge University Press.

Cassia, R., Silva, R., Barreto, M. L., Assis, A. M. O., de Santana, M. L. P., Parraga, I. M., ... Blanton, R. E. (2007). The relative influence of polyparasitism, environment, and host factors on schistosome infection. *Am. J. Trop. Med. Hyg.*, *77*(4), 672–675.

Center for Food Security and Public Health. (2005). *Hookworms.* Retrieved November 13, 2011 from http://www.cfsph.iastate.edu.

Chan, M. S., Bradley, M., & Bundy, D. A. P. (1997). Transmission patterns and the

epidemiology of Hookworm infection. *International Journal of Epidemiology*, *26*(6), 1392 – 1400.

Chitsulo, L., Engels, D., Montresori, A., & Savioli, L. (2000). The global status of Schistosomiasis and its control. *Acta Tropica*, *77*, 41–51.

Clements, A., Firth, S., Dembele, R., Garba, A., Toure, S., Sacko, M., ... Fenwick, A. (2010). Use of Bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in Western Africa. *Bull World Health Organisation*, *87*, 921–929.

Cohen, J. E. (1977). Mathematical models of Schistosomiasis. *Annual Review of Ecology and Systematics*, *8*, 209–233.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*, 829–844.

Cressie, N. (1993). *Statistics for spatial data*. United States of America: Wiley Interscience.

Denwood, M. J., Stear, M. J., Matthews, L., Reid, S. W. J., Toft, N., & Innocent, G. T. (2008). The distribution of the pathogenic nematode Nematodirus battus in Lambs is zero-inflated. *Parasitology*, *135*, 1225-1235.

Ezeamama, A. E., Friedman, J. F., Acosta, L. P., & Bellinger, D. C. (2005). Helminth infection and cognitive impairment among Filipino children. *Am. J. Trop. Med. Hyg.*, *72*(5), 540–548.

Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, *4*, 117–130.

Fenwick, A., & Hotez, P. (2009, September). Schistosomiasis in Africa: An Emerging Tragedy in Our New Global Health Decade. *PLoS Negl Trop Dis*, *3*(9), 485.

Flynn, M., & Francis, L. A. (2009). More flexible glms zero-inflated models and hybrid models. *Casualty Actuarial Society*, *Winter 2009*, 148–224.

Gemperli, A. (2003). *Development of spatial statistical methods for modelling point-referenced spatial data in malaria epidemiology*. Unpublished doctoral

dissertation, University of Basel, Basel.

Gemperli, A., Vounatsou, P., Sogoba, N., & Smith, T. (2006). Malaria Mapping Using Transmission Models: Application to Survey Data from Mali. *American Journal of Epidemiology*, *163*(3), 289 – 297.

Greene, W. (2005). Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*, *1*(2), 113–218.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, *99*, 585-590.

Gurmu, S., & Trivedi, P. K. (1996, October). Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics*, *14*(4), 469–477.

Hall, D. B. (2000, December). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, *56*(4), 1030-1039.

Hanzel, T., Karanja, D. M. S., Addiss, D. G., Hightower, A. W., & Rosen, D. H. (2003). Geographic distribution of schistosmiasis and soil transmitted helminths in Western Kenya: Implications for antihelminthic soil mass treatment. *Am. J. Trop. Med. Hyg.*, *69*(3), 318-323.

Hilbe, J. M. (2011). *Negative binomial regression*. United States of America: Cambridge University Press.

Hotez, Bundy, D. A. P., Beegle, K., Brooker, S., Drake, L., de Silva, N., ... Savioli, L. (2006). Helminth Infections: Soil-transmitted Helminth Infections and Schistosomiasis. In D. T. Jamison et al. (Eds.), *Disease control priorities in developing countries* (2nd ed.).

Hotez, P. (2008). Hookworm and poverty. *New York Academy of Sciences*, *1136*, 38-44.

Jiagang, G. (2003). *Schistosomiasis Control in China: Strategy of Control and Rapid Assessment of Schistosomiasis Risk by Remote Sensing (RS) and Geographic Information System (GIS)*. Unpublished doctoral dissertation, University of Basel, Basel.

Kanzaria, H. K., Acosta, L. P., Langdon, G. C., Manalo, D. L., Olveda, R. M.,

McGarvey, S. T., ... Friedman, J. F. (2005). *Schistosoma Japonicum* and occult blood loss in endemic villages in *LEYTE, THE PHILIPPINES*. *Am. J. Trop. Med. Hyg.*, *72*(2), 115-118.

Kheir, M. M., Eltoum, I. A., Saad, A. M., Magdi M. Ali, O. Z. B., & Homeida, M. M. A. (1999). Mortality due to schistosomiasis mansoni: A field study in sudan. *Am. J. Trop. Med. Hyg.*, *60*(2), 307-310.

King, C. H., & Dangerfield-Cha, M. (2008). The unacknowledged impact of chronic Schistosomiasis. *Chronic Illn*, *4*, 65–79.

Kjetland, E. F., Ndhlovu, P. D., Gomo, E., & Mduluza, T. (2006). Association between genital schistosomiasis and HIV in rural Zimbabwean women. *AIDS*, *20*, 593-600.

Kuwane, B., Appiah, K., Felix, M., Grant, A., Address, E., & Churchyard, G. (2009). Expanding HIV care in Africa: Making men matter in Johannesburg. *Lancet*, *374*(9698), 1329.

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

Larocque, R., Casapia, M., Gotuzzo, E., & Gyorkos, T. W. (2005). Relationship between intensity of soil-transmitted helminth infections and anaemia during pregnancy. *Am. J. Trop. Med. Hyg.*, *73*(4), 783–789.

Lawal, B. H. (2010). Zero-inflated count regression models with applications to some examples. *Qual Quant*, *46*(2012), 19-38.

Loeys, T., Moerkerke, B., Smet, O., & Buysse, A. (2011). The analysis of zero inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical psychology*, *65*(1), 163 – 180.

Lwambo, N., Bundy, D. A. P., & Medley, G. F. H. (1992, June). A new approach to morbidity risk assessment in Hookworm endemic communities. *Epidemiol. Infect.*, *108*(3), 469-481.

Magalhes, R. J. S., Clements, A. C., Patil, A. P., Gething, P. W., & Brooke, S. (2011). The applications of model-based geostatistics in helminth epidemiology and control. *Adv. Parasitol.*, *74*, 267–296.

Mbabazi, P. S., Andan1, O., Chitsulo, L., & Downs, J. A. (2011). Examining the relationship between urogenital Schistosomiasis and HIV Infection. *PLoS Negl Trop Dis*, *5*(12), 1396.

Midzi, N., Mtapuri-Zinyowera, S., Mapingure, M. P., Paul, N. H., Sangweme, D., Hlerema, G., ... Mduluza, T. (2011). Knowledge attitudes and practices of grade three primary school children in relation to Schistosomiasis, soil transmitted helminthiasis and malaria in Zimbabwe. *BMC Infectious Diseases*, *11*(1), 169.

Mott, K. E. (1984). *Schistosomiasis: new goal* (Tech. Rep.). Geneva, Switzerland: World Health Organization.

Myburgh, H. (2011). The clinic as a gendered space: Masculinities, health seeking behaviour and HIV & AIDS. Retrieved on July 12, 2012 from http://www.consultancyafrica.com.

National Travel Health Network and Centre. (2008, March). *Schistosomiasis.*

Ngwira, B. (2005). *The epidemiology and control of lymphatic filariasis and intestinal helminths in the lower shire valley- Chikwawa district southern Malawi.* Unpublished doctoral dissertation, University of Liverpool, Liverpool.

Pan, W. (2001, March). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120-125.

Pritchard, D. I., & Hotez, P. (1995). Hookworm infection. *Scientific American*, *272*, 68–74.

Randal, A. E., Perez, M. A., Floyd, S., Black, G. F., Crampin, A. C., Ngwira, B., ... Fine, P. E. M. (2002). Patterns of helminth infection and relationship to BCG vaccination in Karonga district northern Malawi. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *96*, 29–33.

Raso, G., Vounatsou, P., Singer, B. H., N'Goran, E., Tanner, M., & Utzinger, J. (2006, May). An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*-hookworm co-infection. *Proceedings of the National Academy of Sciences (PNAS)*, *103*(18), 6934–6939.

Ribeiro, P. J., & Diggle, P. (2011). Analysis of geostatistical data: Package 'geor'.

Retrieved on March 19, 2012 from http://www.leg.ufpr.br/geoR.

Ridout, M., Demetrio, C. G., & Hinde, J. (1998, December). Models for count data with many zeros. *Proceedings from the International Biometric Conference*. Cape Town.

Roberts, M., Buttterworth, A. E., Kiman, G., Kamau, T., Fulford, A. J. C., Dunne, D. W., ... Sturrock, R. F. (1993, Deceember). Immunity after Treatment of Human Schistosomiasis: Association between Cellular Responses and Resistance to Reinfection. *Infection and Immunity*, *61*(12), 4984–4993.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. *Communication in Statistics-Theory and Methods*, *32*, 281-289.

Saathoff, E., Olsen, A., Magnussen, P., Kvalsvig, J. D., Becker, W., & Appleton, C. C. (2004). Patterns of Schistosoma Haematobium infection, impact of praziquantel treatment and re-infection after treatment in a cohort of school children from rural KwaZulu-Natal / South Africa. *BMC Infectious Diseases*, *4*, 40.

Schur, N., Gosoniu, L., Raso, G., Utzinger, J., & Vounatsou, P. (2011). Modelling the geographical distribution of co-infection risk from single-disease surveys. *Statistics in Medicine*, *30*, 1761-1776.

Simoonga, C., Kazembe, L. N., Kristensen, T. K., Olsen, A., Appleton, C. C., Mubita, P., & Mubila, L. (2008). The epidemiology and small-scale spatial heterogeneity of urinary schistosomiasis in Lusaka province, Zambia. *Geospatial Health*, *3*(1), 57–67.

Spear, R. C., Seto, E., Liang, S., Birkner, M., Hubbard, A., Qiu, D., ... Davis, G. M. (2004). Factors influencing the transmission of *Schistosoma Japonicum* in the mountains of Sichuan province of China. *Am. J. Trop. Med. Hyg.*, *70*(1), 48-56.

Utzinger, J., Booth, M., NGoran, E. K., Muller, I., Tanner, M., & Lengeler, C. (2001). Relative contribution of day-to-day and intra-specimen variation in faecal egg counts of Schistosoma mansoni before and after treatment with praziquantel. *Parasitology*, *122*, 537-544.

Utzinger, J., Vounatsou, P., N'Goran, E. K., Tanner, M., & Booth, M. (2002). Reduction in the prevalence and intensity of Hookworm infections after praziquantel treatment for Schistosomiasis infection. *International Journal for Parasitology*, *32*, 759–765.

Verhoeff, F. (2000). *Malaria in pregnancy and its consequences for the infants in rural Malawi.* Unpublished doctoral dissertation, University of Leiden, Leiden.

Vounatsou, P., Raso, G., Tanner, M., N'goran, E., & Utzinger, J. (2009). Bayesian geostatistical modelling for mapping schistosomiasis transmission. *Parasitology*, *136*, 1695–1705.

WHO. (1993). Public health impact of schistosomiasis: disease and mortality. *WHO Expert Committee on the Control of Schistosomiasis. Bull World Health Organisation*, *71*, 657-662.

WHO. (1998a). *Guidelines for the evaluation of soil-transmitted helminthiasis and schistosomiasis at community level. a guide for managers of control programmes* (Tech. Rep.). Geneva: World Health Organization. (WHO/CTD/SIP/98.1).

WHO. (1998b). *Report of the who informal consultation on schistosomiasis control* (Tech. Rep.). Geneva: World Health Organization. (WHO/CDS/CPC/SIP/99.2).

WHO. (2001a, March). Control of schistosomiasis and soil-transmitted helminth infections. *Secretariat report, 54th World Health Assembly*.

WHO. (2001b, May). *World health assembly endorses whos strategic priorities.* Retrieved December 13, 2011 from http://www.who.int/inf-pr-2001/en/pr2001WHA-6.html.

WHO. (2002). Prevention and control of schistosomiasis and soil-transmitted helminthiasis: report of a who expert committee. *WHO Technical Report Series*, *912*, 1-57.

WHO. (2006). Schistosomiasis and soiltransmitted helminth infections preliminary estimates of the number of children treated with albendazole or mebendazole.

Weekly epidemiological record, *81*(16), 145–164.

Winkelmann, R. (2008). *Econometric analysis of count data,* (5th ed.). Berlin Heidelberg: Springer-Verlag.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, *27*(8), 1–25.

Zhou, X.-N., Yang, G.-J., Yang, K., Wang, X.-H., Hong, Q.-B., Sun, L.-P., ... Utzinger, J. (2008). Potential Impact of Climate Change on Schistosomiasis Transmission in China. *Am. J. Trop. Med. Hyg.*, *78*(2), 188-194.

Appendix A

R Code snippets

Some of packages used

library(pscl)

library(geoR)

library(MBA)

Histogram plotting

hist(shegml, col = "blue", plot = TRUE, xlab = "S. haematobium parasites", ylab = "number", main = "Schistosoma Haematobium Parasite Counts")

Poisson model fitting

poissh <- glm(sh ~ age + gender +education + intervention +fishing + garden + occupation + polyparasitism, data=mydata, family = poisson)

summary(poissh)

confint(poissh)

Zero inflated negative binomial model fitting

zinbsh <- zeroinfl(sh ~ age + gender +education + intervention +fishing + garden + occupation + polyparasitism, data=mydata, dist= "negbin")

summary(zinbsh)

confint(zinbsh)

Negative binomial logit Hurdle model fitting

hurdlenbsh <- hurdle(sh ∼ age + gender +education + intervention +fishing + garden + occupation + polyparasitism, data=mydata, dist="negbin", zero.dist = "binomial", link = "logit", control = hurdle.control(method = "BFGS", maxit = 10000, trace = TRUE, separate = TRUE))

summary(hurdlenbsh)

confint(hurdlenbsh)

Comparing Models' Full Likelihood

rbind(logLik = sapply(comparesh, function(x) round(logLik(x), digits = 0)), Df = sapply(comparesh, function(x) attr(logLik(x), "df")))

Zero count capturing

round(c("Obs" = sum(mydata$sh ¡ 1), "ML-Pois" = sum(dpois(0, fitted(poissh))), "NB" = sum(dnbinom(0, mu = fitted(nbsh), size = nbsh$theta)),"ZIP" = sum (predict(zipsh, type = "prob")[,1]), "ZINB" = sum(predict(zinbsh, type = "prob") [,1]), "Poison-Hurdle" = sum(predict(hurdlepsh, type = "prob")[,1]), "NB-Hurdle" = sum (predict (hurdlenbsh, type = "prob")[,1])))

Visual residual spatial effects inspection

x.res <- 200

y.res <- 200

surf <- mba.surf(cbind(coords, rex), no.X = x.res, no.Y = y.res, h = 5, m = 2, extend = FALSE)$xyz.est

image.plot(surf, xaxs = "r", yaxs = "r", xlab = "South", ylab = "East", main = "Residual spatial effects", col = col.br(25))

drape.plot(surf[[1]], surf[[2]], surf[[3]], col = col.br(150), theta = 225, phi = 50, border = FALSE, add.legend = FALSE, xlab = "South", ylab = "East", zlab = "Residuals")

image.plot(zlim = range(surf[[3]], na.rm = TRUE), legend.only = TRUE, horizontal = FALSE)

Variogram plotting

res.vario <- variog(res.geo1, max.dist=35)

plot(res.vario, main = "Spherical Model")

res.model1 <- list(cov.model = "sph", cov.pars = c(0.20, 12), nugget=0.1, max.dist = 35)

lines.variomodel(res.model1, lty = 2)

res.vario <- variog(res.geo1, max.dist=35)

plot(res.vario, main = "Exponential Model")

res.model <- list(cov.model = "exp", cov.pars = c(0.20, 12), nugget=0.12, max.dist = 35)

lines.variomodel(res.model, lwd = 2)

Appendix B

Residual spatial effects figure

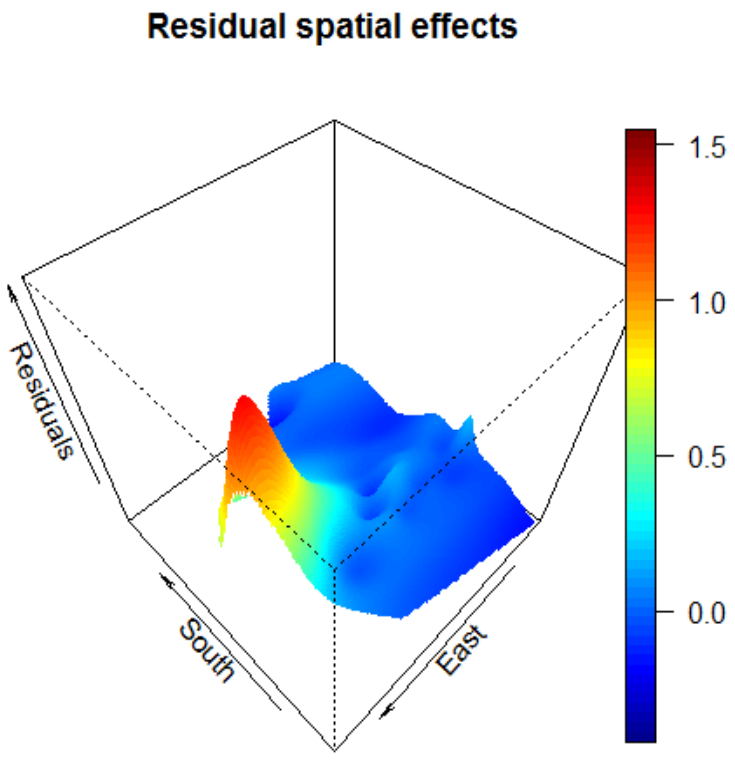

Figure B.1: Estimated residual spatial effects, given in odds ratios