The Modelling of African Animal Trypanosomiasis in Kwazulu-Natal, South Africa

Nada A. Abdelatif

A thesis presented in partial fulfilment of the requirements for the degree of
Master of Science in Statistics

School of Mathematics, Statistics and Computer Science
The University of Kwazulu-Natal
South Africa
2015

The candidate's supervisor(s) have approved this dissertation for submission:

Supervisor: _____ Date: _____

Co-supervisor: _____ Date: _____

Declarations

I, Nada A. Abdelatif, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) Their words have been re-written but the general information attributed to them has been referenced
   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: _____

ABSTRACT

African animal trypanosomiasis (AAT), restricted to parts of the KwaZulu-Natal Province, is a disease which contributes significantly to the disease burden of cattle. Drug resistance is a constraint and dipping of cattle using insecticides has proved to be unsustainable. Even though the incidence of AAT has increased, little is known about the epidemiology of the disease in the region. To better understand the dynamics of AAT, mathematical modelling was done to investigate the interactions between the cattle, tsetse flies and buffaloes which are considered to be the reservoir host. In addition, a statistical analysis of the data collected from three sites around the Hluhluwe-iMfolozi Game Park was done to assess the interactions between the variables.

A susceptible-infected (SI) model was constructed for the different classes of the population i.e. susceptible and infected cattle and tsetse flies and infected buffaloes. The basic reproduction number R_0, a threshold determining whether the disease will die out or persist in the population, was derived using the next-generation matrix since we had two-hosts and one vector. R_0 was used to assess which elements contribute to R_0 (i.e. transmission of AAT from the buffaloes and cattle to the tsetse flies or tsetse flies to the cattle and buffaloes). The important element was found to be the transmission of AAT from buffaloes to flies. Sensitivity analysis was done using the partial rank correlation coefficients (PRCC) measure. PRCC values can show which parameters to target when looking at intervention measures and determine how to efficiently reduce AAT. The mortality rate of tsetse flies and their biting rate were determined to be the most important parameters.

Generalized linear models (GLMs) were used to analyse the data since we had binary and count data. The AAT prevalence data was modelled using a binomial GLM, using the packed cell volume (PCV), which is an indicator of whether a cow has AAT or not, region i.e. whether the cow is located near or further away from the game park and month as the explanatory variables. PCV and region were found to be significant, so where the cattle are located seems to be important. The tsetse abundance data was modelled using Poisson

GLMs, however the problem of overdispersion was evident and so alternative models were considered. Since there were excess zeroes for *G. austeni*, zero-inflated models were used and the best fit was found to be the zero-inflated negative binomial, whereas the negative binomial model was used for *G. brevipalpis* to account for the overdispersion. Months 7 and 8 and year were found to be statistically significant for *G.austeni*. This could be because month 7 has the lowest minimum and maximum temperatures during the year and at lower temperatures, tsetse flies become less active and the pupal stage lengthens to around 50 days and the reproductive rate decreases. For *G.brevipalpis* only year was found to be statistically significant.

The AAT prevalence data was fit to the mathematical model using least squares, and the input parameters were estimated and used to calculate R_0 again so that it is more site-specific. Climate change was also briefly addressed, since it is predicted to affect the geographical distribution of tsetse flies. Higher temperatures could have a big impact on the AAT situation because tsetse flies might modify their behaviour and shift their geographical range to regions that are cooler, which might put cattle populations in other regions at risk of AAT outbreaks.

Acknowledgements

I would firstly like to thank my supervisors, Professor Henry Mwambi and Dr. Faraimunashe Chirove for their invaluable assistance and experience throughout this degree, and their support and motivation.

I would also like to thank Professor Latif for opening me to a new world in vector-borne diseases, for allowing me to use his data set and for his invaluable knowledge in the epidemiology of trypanosomiasis. I would also like to thank him for his endless support as a father.

Without the financial support of SACEMA, this project would have been difficult to complete, and also for their input into my work, which has certainly helped me think about things in a different way, and a particular thanks to Professor Hargrove, for the world of tsetse flies is certainly his world, Dr. Rachid Ouifki for his input and Dr. Gavin Hitchcock for always checking up on the progress that is being made.

Last and definitely not least, thank you to my mother, sister and brother, whom I look up to so much.

Contents

List of Figures

List of Tables

List of Abbreviations

AAT	African animal trypanosomiasis
BB	Beta-binomial
DFE	Disease free equilibrium
ELISA	Enzyme-linked immuno-sorbent assay
GIS	Geographic information systems
GLM	Generalized linear model
HAP	Herd average prevalence
HA-PCV	Herd average packed cell volume
KZN	KwaZulu-Natal
LHS	Latin hypercube sampling
NB	Negative binomial
NGM	Next-generation matrix
ODE	Ordinary differential equations
PCV	Packed cell volume
PDF	Probability density function
PMF	Probability mass function
PRCC	Partial rank correlation coefficient
SI	Susceptible-infected
SIT	Sterile insect technique

VBD	Vector-borne disease
ZIB	Zero-inflated binomial
ZINB	Zero-inflated negative binomial
ZIP	Zero-inflated Poisson

Chapter 1

Introduction

1.1 Background

Trypanosomiasis is a parasitic, vector-borne disease, caused by protozoa of the genus *Trypanosoma*, and affects humans, domestic and wildlife animals in tropical and sub-tropical countries in the world. It is transmitted by blood-sucking insects such as tsetse flies and biting flies. In South and Central America, *Trypanosoma cruzi* is transmitted by blood-sucking insects and causes Chagas disease in humans, whereas in Asia and North Africa *Trypanosoma evansi* is transmitted by biting flies and affects camels and horses (known as *Surra*). In Africa, the disease is transmitted mostly by tsetse flies of the genus *Glossina*, and affects both humans and animals.

Human African trypanosomiasis, or sleeping sickness, is endemic in East, West and Central Africa, with approximately 70 million people said to be at risk, one third of which are living in areas of high to moderate risk (Simarro et al, 2012). It is caused by *T. brucei gambiense*, which is found in West and Central Africa and is more of a chronic form of the disease, and *T. brucei rhodesiense*, found in East Africa and is acute.

Animal trypanosomiasis is a disease affecting livestock in tropical and sub-tropical countries; transmitted mechanically by biting flies in areas such as in Central and South America, and cyclically by the tsetse fly in Africa. African animal trypanosomiasis (AAT), or nagana (a Zulu word meaning powerless/useless), has a major impact on livestock production and economic development in Africa. It is a cause of poverty and food shortages affecting the livelihood of about 500 million farmers in rural villages (Deveze, 2010). Animals in Africa have great economic and social significance, as they provide milk, meat,

hides and skins and are a means of accumulating and distributing wealth (Perry et al, 2005). Sick animals eventually become unfit to work or produce meat and/or milk and will eventually die if they are not treated. AAT causes about 3 million deaths in cattle and estimated economic losses in production of about US$1-2 billion each year (FAO, 2004).

1.2 Epidemiology of AAT

1.2.1 Etiologic Agent of Trypanosomiasis

Tsetse-transmitted trypanosomiasis is caused by parasites of the genus *Trypanosoma*, which affects all domestic and wildlife animals. The major species are *Trypanosoma congolense*, *Trypanosoma vivax*, *Trypanosoma brucei brucei* and *Trypanosoma simiae*. The two main cyclically transmitted trypanosomes in cattle are *T. congolense*, *T. vivax* and *T. brucei brucei* to a smaller degree, while *T. brucei rhodensei* and *T brucei gambense* are considerably more important in causing sleeping sickness in humans (Clair, 1988). *T. simiae* is rare but very pathogenic to pigs. *T. vivax* can be transmitted mechanically in non-tsetse infested areas such as in Central and South America.

1.2.2 Natural History of the Disease

1.2.2.1 Symptoms

Trypanosomiasis is fatal if left untreated. In livestock, it leads to considerable weight loss and anaemia, and the various symptoms exhibited include fever, emaciation, oedema, anaemia and paralysis (Steverding, 2008). The advancement of trypanosomiasis varies according to the parasite strain and the species and breed of animal infected (Connor and Van den Bossche, 2004).

1.2.2.2 Pathogenesis and Virulence of the Disease

The disease varies from acute to chronic, where acute infection is characterised by high levels of parasitemia and a fall in packed cell volume (the volume percentage of red blood

cells in the blood), which leads to fever. Death occurs within 10 days (Connor and Van den Bossche, 2004). The chronic form of the disease can persist for months, eventually leading to anaemia and tiredness in the animal. The animal may go from the chronic form to the acute form if they are in stressful conditions or may recover spontaneously.

The virulence of the trypanosomes is influenced by certain factors such as those affecting the parasite and the host (Motloang, 2012). Parasite isolates from distinct geographic regions such as *T. vivax* strains from West African are more pathogenic in cattle than East African strains. Variations between sub-species of trypanosomes such as savanna type *T. congolense* are more pathogenic than the riverine-forest type. Similarly with hosts, wildlife are more resistant than domestic animals, and so are indigenous breeds compared to exotic breeds.

1.3 Methods of Diagnosis

A number of diagnostic tests are available to detect trypanosomiasis, such as serological, parasitological and molecular diagnostic tests (Marcotty et al, 2008). Microscopy identifies trypanosomes in stained blood smears. This method has a low sensitivity when low levels of parasitaemia are present, which is often the case in chronic cases of the disease and also in large animals (Motloang, 2012). Another parasitological test is the buffy coat method, where blood is transferred to micro-capillary tubes which are spun in a microhaematocrit centrifuge, after which the packed cell volume (PCV) can be determined and the presence of motile trypanosomes can be examined under the microscope (Marcotty et al, 2008).
PCV is a measure of anaemia, which is a known and inevitable consequence of infection with trypanosomes (Marcotty et al, 2008) and so, is a good indicator of disease. By using PCV with parasitological tests, detection of trypanosome-infected animals will be improved.

Molecular diagnostic tests detect the presence of pathogens and are more sensitive and specific than parasitological tests but they are expensive and require advanced infrastructure, and like parasitological tests, in chronic cases when parasitaemia levels are low, they are less sensitive (Marcotty et al, 2008).

The ELISA (enzyme-linked immuno-sorbent assay) test is a serological test which looks at antibodies produced during a trypanosomal infection. They are relatively cheaper than

molecular tests, but they require laboratory equipment, and are not an indication of current infections as antibodies can remain weeks after infection. ELISA tests, also, cannot be used in early stages of infection because the antibodies will not have been produced (Marcotty et al, 2008).

1.4 Disease Control Methods

Trypanocidal drugs are used to treat infected animals, which reduces the losses due to the disease and eliminates trypanosome reservoirs. However, the continuous use of trypanocides requires close monitoring of the herd health condition while treatment becomes expensive and unaffordable to communal farmers. The problem of development of drug resistance in trypanosomes is the threat to the sustainability of the strategy (Connor and Van den Bossche, 2004).
Prophylactic trypanocidal drugs can be administered to protect the animal for several months (Ntantiso, 2012), or as in the case of cattle, trypanotolerant livestock (cattle resistant to trypanosomiasis) can be reared. It has been shown that trypanotolerant cattle are able to keep their productivity under tsetse and trypanosomiasis infected areas as evidenced in West Africa (Murray, Morrison and Whitelaw, 1982).

Insecticides, applied from the ground or the air, can be used to control or eradicate the flies. Alternatively the use of sterile insect technique (SIT) can be used, where males are made sterile through gamma irradiation, and is effective when the tsetse fly population density is low (Kappmeier-Green, Potgieter and Vreysen, 2007).

1.5 The vector: The Tsetse Fly

Mechanical transmission of AAT can occur when blood infected with trypanosomes is transferred from one sick animal to another without the parasite undergoing developmental change within the fly. Biting flies of the genus *Tabanus* are considered important mechanical transmitters. On the other hand, tsetse flies transmit trypanosomiasis cyclically. This biological transmission happens when blood from a trypanosome-infected animal is ingested by the fly and the trypanosome multiplies within the fly and reach the infective stage

(Connor and Van den Bossche, 2004) (See Figure 1.1). Therefore tsetse flies are a crucial part of the transmission of AAT as without them the disease cannot be maintained.

Tsetse flies fall in the genus *Glossina* with three sub-genera, each one representing a group of species (Motloang, 2012). Among the three groups are the fusca group which are forest species and feed very little on domestic livestock, the palpalis group which are mainly riverine species and are important vectors in the transmission of human and animal trypanosomes, and lastly the morsitans group which are savanna species and play the predominant role in transmitting AAT as they feed on cattle, sheep and goats (Jordan, 1988). Up to date, over 30 species of *Glossina* have been identified, but it is believed that only 8 to 10 of them are of economic importance (Vreysen et al, 2013).

The distinguishable features of tsetse flies are their long and pointed mouthparts and their wings (See Figure 1.2), which overlap and are about 6-15mm. Both males and females feed on blood and the feeding preferences differ between species. They travel at low flying heights and are attracted to visual stimuli such as size and colours. Tsetse flies have low reproductive rates and long life cycles and lifespans, of up to 3 to 4 months (Leak, 1999). An adult female produces a single egg which hatches to a first-stage larva in 1-2 days while still in the fly uterus.The third-stage larva is deposited into the ground after 2-5 days, immediately changes to a pupa after which an adult fly will emerge after 30-40 days. About 10 offspring or less can be produced in a female's reproductive life (Leak, 1999).

1.5.1 Tsetse Fly Traps

There are a number of reasons for trapping of tsetse flies, such as to measure the biodiversity of the flies, to determine seasonal peaks of abundance, to determine AAT risk and it may also be used as a form of vector control (Leak, 1999). There are a number of trap designs available such as the biconical trap, pyramidal trap, vavoua trap and H-trap, to name a few. There are advantages and disadvantages to each trap and each has its own characteristics that make it suitable for certain environments and to attract specific species of flies. The biconical and Vavoua traps, for example, are adapted for the capture of riverine species along rivers. Various features make traps efficient and attractive to flies like the colour, odour and design. Blue and black are the two colours most used in traps and screens (Leak, 1999). The location and season are crucial factors to consider when placing the traps as

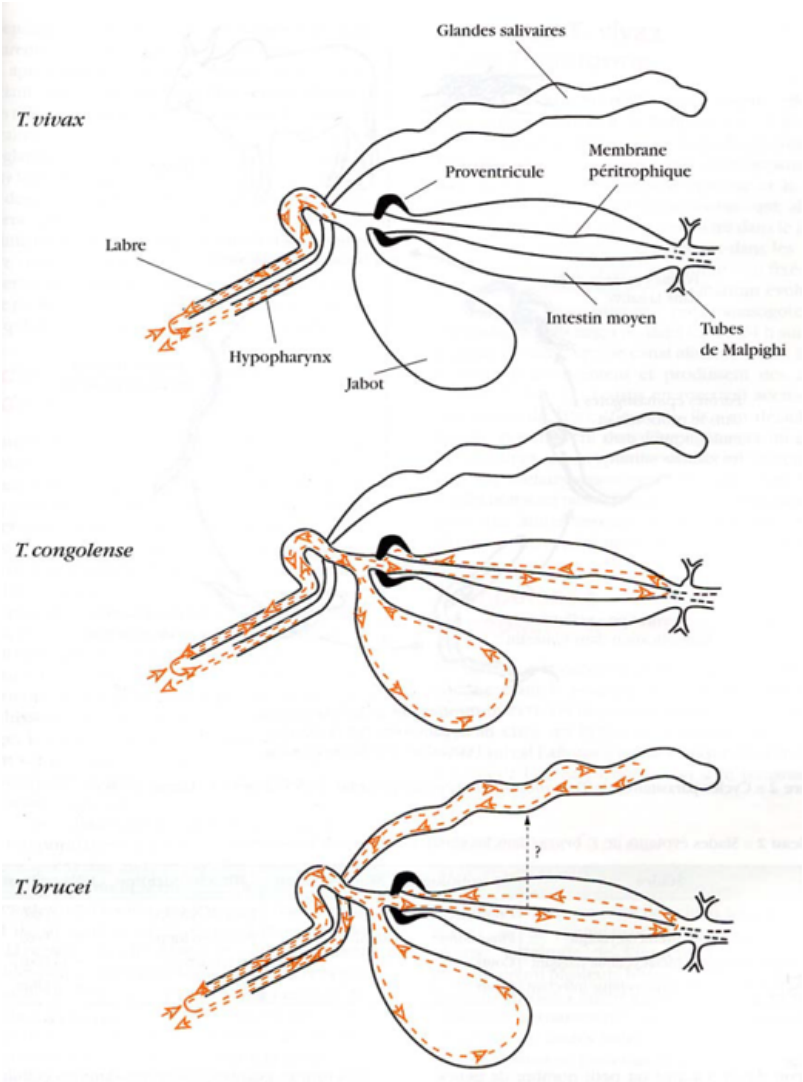


Figure 1.1: Cyclical development of trypanosomes within the tsetse fly (*T.vivax* develops in the mouthparts, *T.congolense* develops in the midgut and mouthparts, *T.brucei* develops in the salivary glands)(*Source: Leak, 1999*)

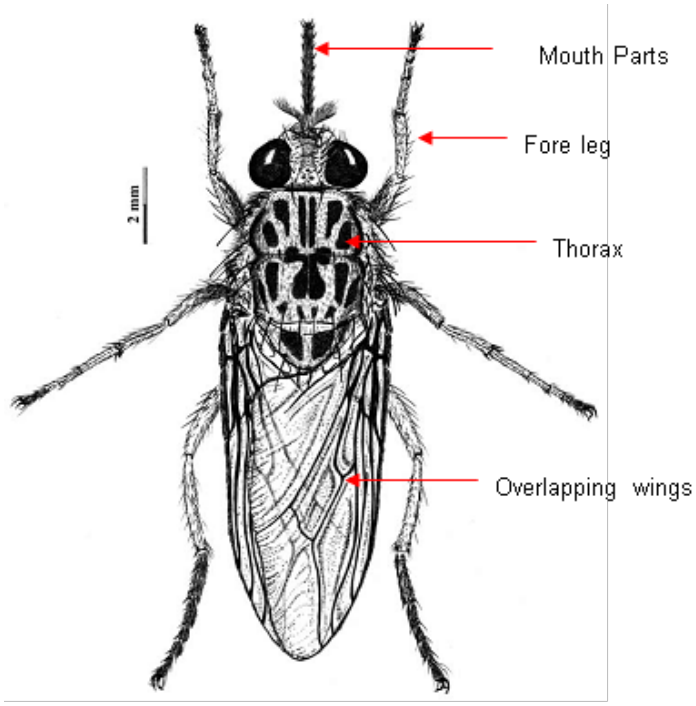


Figure 1.2: The Tsetse fly

they determine the efficiency with which tsetse flies are captured.

1.5.2 Tsetse Distribution in Africa

The tsetse fly is restricted to Africa, although some species were found in Saudi Arabia in 1990 but this seemed not to be as an established population. It inhabits about 10 million km^2 (latitude 150N-290S), which affects 37 countries in the Sub-Saharan region (Connor and Van den Bossche, 2004). The southernmost tsetse fly belt distribution goes from Mozambique to KwaZulu-Natal (KZN) in South Africa.

1.6 Trypanotolerance and the Role of Reservoir Hosts

Certain breeds of cattle such as the taurine N'Dama and the West African Shorthorn breeds are known to have a certain level of resistance to trypanosomiasis, that other breeds, such as the Zebu and European cattle, do not possess (Murray et al, 1982). This is referred to as trypanotolerance, which is defined as the "ability to survive and to be productive in tsetse infested areas without the aid of treatment where other breeds succumb to disease" (Murray, Trail and d'Ieteren, 1990). Even though this feature only applies to about one third of cattle in tsetse infested areas in Africa and 10% of cattle south of the Sahara, the production potential of these breeds is great, and possible breeding of these cattle may result in a solution to the AAT problem.

Wild animals are reservoir hosts of animal trypanosomiasis, and play an important role in disease transmission. Reservoir hosts are able to carry pathogens indefinitely and display no clinical symptoms. This complicates the epizootiology of AAT greatly, as elimination of the disease now requires consideration of and control measures for the wild animals as well (Clair, 1988). This introduces a multi-host problem in finding control measures which can be a big challenge.

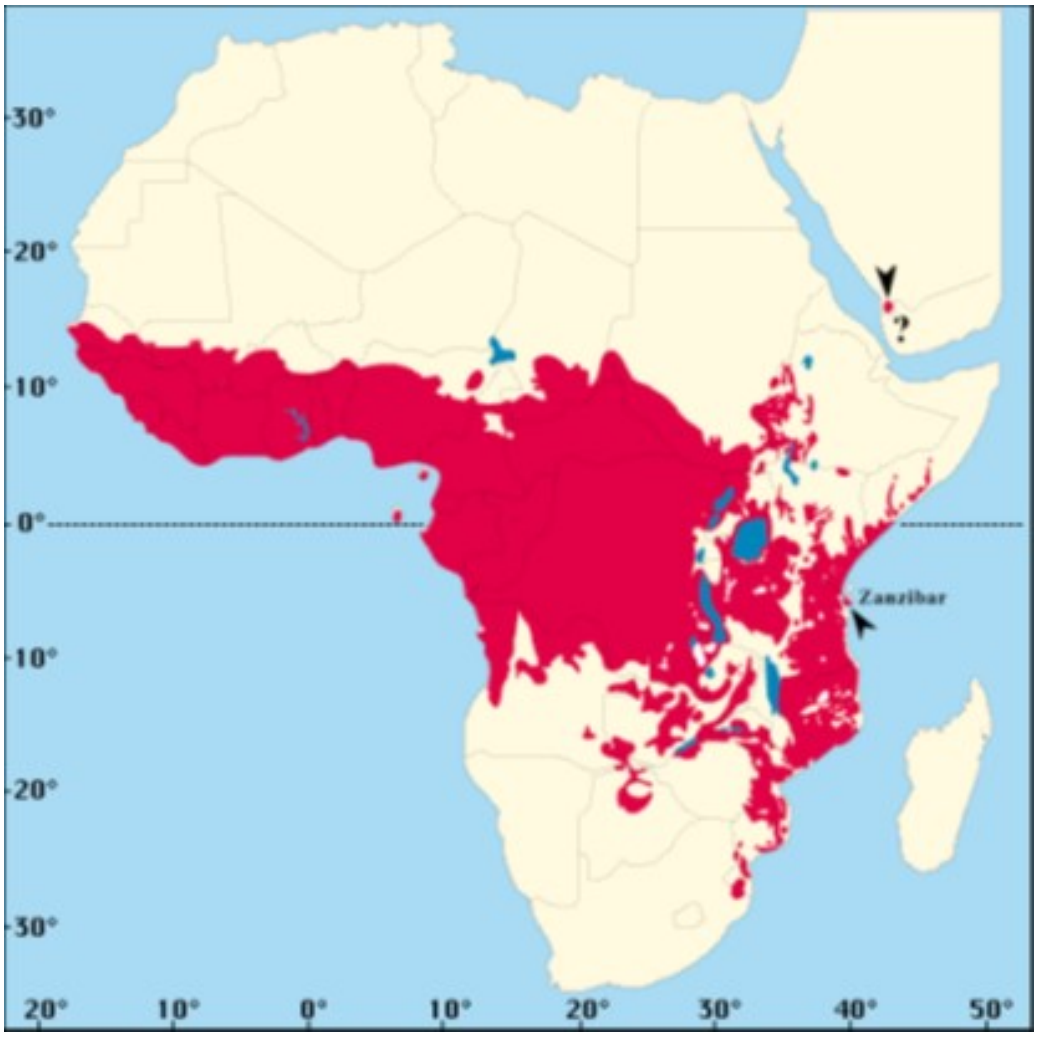


Figure 1.3: Tsetse distribution in Africa (*Source: Leak, 1999*)

1.7 Bovine Trypanosomiasis in South Africa

1.7.1 History

Four species of tsetse flies have been recorded in South Africa, namely *Glossina morsitans morsitans*, *G. pallidipes*, *G. brevipalpis* and *G. austeni* (Kappmeier, Nevill and Bagnall, 1998). These species were confined to the KwaZulu-Natal province (except *G. morsitans morsitans* which was in the northerly parts of South Africa). During a rinderpest epizootic during 1896-1897, *G. m. morsitans* completely disappeared. The three remaining species were the vectors of AAT, with *G.pallidipes* considered as the vector of the most pathogenic trypanosomes, until a massive campaign led to its eradication in the early 1950s (Du Toit, 1954).
Human trypanosomiasis (sleeping sickness) has not been known to occur in South Africa, whereas a lot of literature on the nagana situation is available from as early as 1923. Between 1955 and 1990, only sporadic cases of AAT were diagnosed, with localised outbreaks occurring in certain regions. In 1990, a widespread outbreak of AAT occurred in Zululand, killing 10,000 cattle and where 116,000 were treated (Kappmeier et al, 1998). The animals were diagnosed as being infected with *T. congolense* and *T. vivax*. The outbreak was controlled by introducing dipping of cattle and treatment of ill animals. Following the outbreak, tsetse distribution surveys of *G. austeni* and *G. brevipalpis* were conducted and greater research into the epidemiology of the disease in South Africa was and still is being done.

1.7.2 Tsetse Distribution in KwaZulu-Natal Province

The tsetse fly belt in Zululand, in KZN covers an area of approximately 18,000 km^2 (Eesterhuizen et al, 2005). The two tsetse species, *G. austeni* and *G. brevipalpis* are found in KZN and mainly reside in shaded areas along rivers and in forests and thickets (Van den Bossche et al, 2006). *G. brevipalpis* is generally found in high indigenous forests and open grasslands such as in the Hluhluwe-iMfolozi Game Park and the southern parts of the St. Lucia Wetlands Park, whilst *G. austeni* is found throughout central Zululand along communal areas and lake shores (see Figure 1.4).

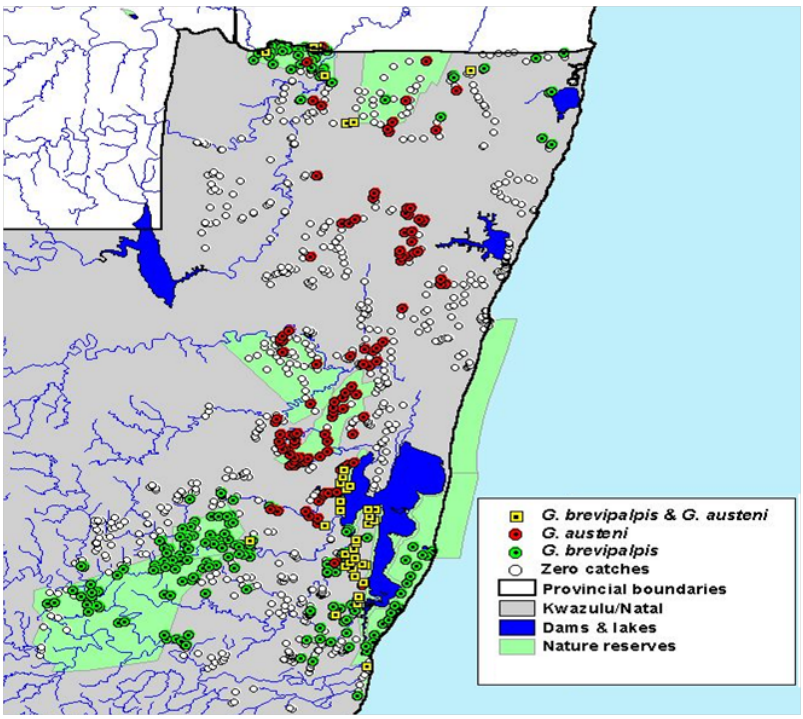


Figure 1.4: Tsetse distribution in KZN, South Africa (*Source: Hendrickx et al, 2003*)

1.7.3 AAT Prevalence in South Africa

Animal trypanosomiasis (nagana) or bovine trypanosomiasis is restricted to parts of northern KZN province and covers an area of 18,000 km^2 (Ntantiso et al, 2014). Little information is available on the prevalence of disease in cattle in KZN, since the 1990 outbreak. Van den Bossche et al (2006) conducted a once-off sample of 76 animals in an area near the Hluhluwe-iMfolozi game park to assess the contribution nagana had to the disease burden of cattle. The study showed that incidence of nagana caused mostly by *T. congolense*, had increased. Baseline data collected by Ntantiso et al (2014) aimed to unravel the epidemiology of nagana in the region where tsetse flies occur, as incidence of the disease in KZN is unknown. Nagana was seen to be prevalent in 10 of the diptanks surveyed and was the cause of anaemia in more than 60% of infected cattle.
The buffaloes in the Hluhluwe-iMfolozi Game Park are reservoir hosts of the disease and harbour the highly virulent strains of *T. congolense* which is considered as a risk at the game – livestock interface, as acute forms of AAT are observed in livestock near the game park (Motloang, 2012; Van den Bossche et al, 2006).

1.7.4 Current Control Policies and Measures

No national control policy has been in place for the last 20 years in South Africa (Ntantiso et al, 2014). Treatment is available but is not accessible to resource-poor farmers, which are the ones whom nagana affects the most. Drug resistance to trypanosomiasis is also a problem, and it becomes too costly to produce new drugs every so often. Dipping of cattle using insecticides which also kills ticks is done when the disease progresses in a herd but as it is not enforced, it has proved to be unsustainable. Migration of tsetse flies between conservation-protected game parks and the regions outside means that an integrated control policy with treatment is also not viable, and so eradication of the tsetse flies is seen as the only possible solution (Kappmeier-Green et al, 2007).

1.8 AAT and Climate Change

Vector-borne diseases (VBDs) such as malaria, dengue fever, schistosomiasis, African trypanosomiasis etc. are considerable public health problems in Africa (TDR, 2012). In-

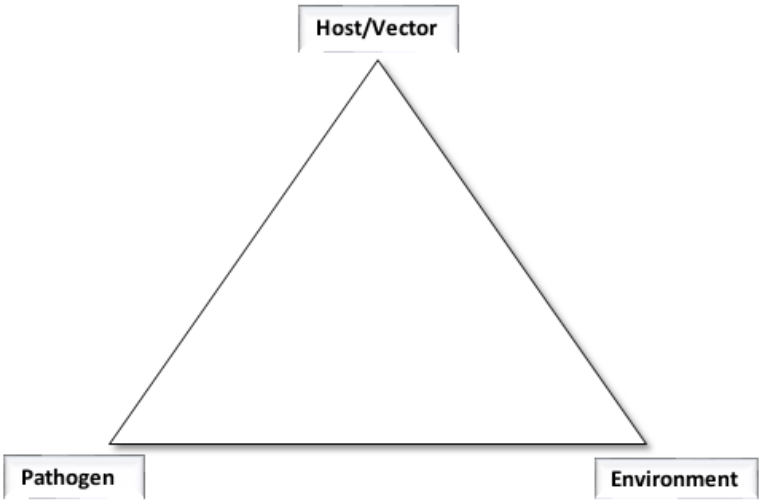


Figure 1.5: Disease triangle with host-vector-pathogen-environment interactions

creases in minimum temperatures compared to average and maximum temperature may allow dengue and other climate-sensitive VBDs to extend into regions that previously have been free of disease or exacerbate transmission in endemic parts of the world (IPCC, 2014). Insect vectors' distribution, abundance and intensity is greatly determined by environmental factors such as temperature and relative humidity (Pollock, 1982), and global climate change can potentially lengthen transmission seasons and shift the geographic range of these vectors. Climate change will also affect the development of pathogens in the vectors and population dynamics of reservoir hosts (Moore et al, 2012). Sutherst (2004) believes that variability of climate change and extreme weather events are of more importance than overall global warming which increases average temperature, and so, Africa especially is at risk.

African trypanosomiasis is one of the VBDs that is likely to be affected by climate change, and the effects of which will be seen in an increase of incidence and either an expansion or a shift of the current tsetse fly geographical range (Moore et al, 2012). Reproductive rates, pupal and adult stages, and mortality rates of the tsetse fly are dependent on factors such as temperature, relative humidity deficit and vegetation (Rogers, 1979, 2000; Pollock, 1982; Hargrove, 2004). Many studies have looked at these abiotic factors in relation to

the various life stages of the tsetse fly, especially the temperature-dependent parameters (Brightwell et al, 1992; Rogers, 1979, 1991, 2000; Hargrove, 2004; Artzrouni and Gouteux, 2006) and all show the importance of changes in average temperature to the ecology of tsetse flies. The study by Moore et al (2012) suggests that tsetse distributions may shift to more favourable conditions such as East and southern Africa, as temperatures become too hot in their current geographical range, which will put these densely populated areas at risk of HAT outbreaks. The effects of climate change on tsetse distributions therefore, need to be closely looked at, as changes could make naive human and animal populations very vulnerable to infection, and could lead to serious epidemics if not analysed.

1.9 Epidemiological and Statistical Modelling

In vector-borne diseases analytical methods are mostly used to describe the dynamics of the interactions between parasite/vector/host behaviour (Gettinby, Revie and Forsyth, 1994), through the use of ordinary differential equations (ODEs). Deterministic models are useful for reaching general principles about the disease transmission, through the use of Monte Carlo simulation methods (Morris and Marsh, 1994). Disease transmission dynamics require information about the determinants of risk to the hosts and challenge by the vector and their relationship to each other (Rogers, 1994). Other approaches used are statistical distribution models which describe spatial distribution and predictive mapping of a vectors' distribution by using logistic regression or discriminant analysis (Gettinby, 1989; Rogers, 2006), and geographic information systems (GIS) that use data to predict distribution and abundance of tsetse and trypanosomiasis in order to aid in the planning and control of trypanosomiasis and other VBDs (Robinson, 1998). Trypanosomiasis is a unique disease within VBDs due to its complexity and diversity, from the ranging habitats of the tsetse flies and their seasonal variability to the pathogenicity of the trypanosomes and the reservoir hosts involved (Gettinby et al, 1994; Gettinby, 1989). Differential models that divide the animal and fly population into classes according to their disease status, help in understanding the biology of the disease, while statistical models are used to establish relationships between the disease variables from the data collected in the field (Gettinby, 1989).

1.10 Problem Statement and Objectives of Study

AAT, or nagana, has been a serious problem for cattle farmers, leading to significant losses in production and livestock. There have been efforts to collect data in the surrounding farm areas of the Hluhluwe-iMfolozi Game Park, in a step towards enhancing knowledge on the spread of disease, as well as gather information about the tsetse fly population. A study done by Ntantiso et al (2014) showed tsetse population and disease in cattle in three sites in KwaZulu-Natal near the Hluhluwe-iMfolozi Game Park. The study also demonstrated a high prevalence rate of trypanosomiasis in two of the sites and a low tsetse population and low prevalence rate in the third site. We shall adopt the approach of mathematical models and statistical techniques to the specific case of Zululand, and the three regions mentioned, and use the data to predict the basic reproduction number (R_0) generated by the mathematical model.

The objectives of the study are as follows:

- To use mathematical models to investigate the interactions between the tsetse fly, cattle and buffalo, and analyse it to define the critical threshold, R_0 as well as the next generation matrix.
- To establish possible feedback mechanisms driving the interactions between the tsetse fly, cattle and buffalo and how these mechanisms influence changes in R_0.
- To estimate critical parameters by carrying out statistical analysis using generalized linear models (GLMs) on the 3 sites where data was collected.
- To fit the designed mathematical models to the data available to check the prognosis of AAT in the Zululand area and come up with recommendations of possible intervention strategies using results from fitted data.

1.11 Significance of Study

Mathematical modelling is central to infectious disease epidemiology, as it simplifies complex biological systems and leads to an understanding of the disease process, whilst statistical analysis is the first step in assessing relationships between variables and is used to

fit models to estimate site-specific parameters. The purpose of this study is to model the disease transmission of African animal trypanosomiasis in the KwaZulu-Natal Province, by integrating both the mathematical model and the statistical analysis of the data collected from the epidemiological study done in the region so that critical processes and patterns can be outlined and perhaps used to plan and evaluate future control policies.

Chapter 2

Mathematical Modelling of AAT

2.1 Introduction

The mathematical model developed by Rogers (1988) for African trypanosomiasis used the Ross-Macdonald model for malaria, as a starting point, and considers a two-host, one-vector species system. The model incorporated most of the parameters involved in the transmission process of African trypanosomiasis, for two hosts, which are humans and domestic animals for human African trypanosomiasis and domestic and wild animals for AAT, and the different trypanonomes, *T. congolense*, *T. vivax* and *T. brucei*. Rogers also looks at the basic reproduction number R_0, by calculating the effects of one infected fly biting uninfected hosts which in turn will infect other uninfected flies, and assesses the importance of each parameter in relation to disease prevalence.

Another model developed for cattle trypanosomiasis by Milligan and Baker (1988), looked at the epizootiology of the disease and the effects of different control methods, by considering the transmission from one species of parasite (trypanosome) at a time. A model for the disease when using control methods (chemotherapy) was developed and then extended to include density-dependent fly movement inside and out an insecticide controlled area. Changes in R_0 were assessed by looking at the different parameters and sensitivity analysis was done using Monte Carlo methods.

The basic reproduction number, denoted by R_0, is defined as the expected number of secondary cases produced, in a completely susceptible population by an infected individual over their infectious lifetime (Diekmann and Heesterbeek, 1990). R_0 is a threshold determining whether a disease will persist in the population (when $R_0 > 1$), or will die out

(when $R_0 < 1$). So if $R_0 < 1$ then the disease-free equilibrium (DFE) is locally asymptotically stable (van den Driessche and Watmough, 2002), and $R_0 > 1$ means the DFE is unstable. R_0 is useful when determining control measures or the need of, for established infections (Diekmann, Heesterbeek and Roberts, 2010). The next-generation matrix (NGM) is a method of arriving at R_0 when we have a population that can be divided into different discrete categories which are epidemiologically important, such as disease stage, age etc. (van den Driessche and Watmough, 2002). It was introduced by Diekmann and Heesterbeek (1990), and considers the infected states of the ODE system, and is referred to as the infected subsystem in Diekmann et al (2010). The NGM, denoted by $\boldsymbol{K}$, is made up of two parts i.e. the transmission part $\boldsymbol{F}$ which shows the production of new infections, and $\boldsymbol{V}$ which is the transition part or changes in state that come about as a result of death or immunity for example. R_0 is then the dominant eigenvalue, or spectral radius of the NGM $\boldsymbol{K} = \boldsymbol{F}\boldsymbol{V}^{-1}$. The elements of $\boldsymbol{K}$, K_{ij} are the expected number of new cases with state-at-infection i arising from an individual in the jth state-at-infection. The advantage of calculating R_0 from the NGM is that even when multiple hosts and vectors are considered, the R_0 value still has the same properties and interpretation as the one from a single-host species case (Davis, Aksoy and Galvani, 2011).

2.2 Model Formulation

The dynamics of AAT in KZN, is one of a mixed sylvatic/domestic cycle (Ntantiso, 2012), meaning that the domestic animals located near the Hluluwe-iMfolozi Game Park, and the wild animals residing within the game park are both being fed on by the tsetse fly, and therefore the transmission goes from wild animals to domestic animals, through the fly and vice versa. So the model included three populations, that is cattle, which is the domestic host, the buffalo which is considered the reservoir host and the tsetse fly which is the vector. There are two species of tsetse fly in KZN, *G. austeni* and *G. brevipalpis*, but for the sake of the mathematical model, we assume that only one species transmits *T.congolense*, since *G. austeni* is considered to be the main vector of AAT. We also assume constant population sizes, the transmission probabilities do not change over time and we do not distinguish between teneral and non-teneral flies for this study. We adopt an SI (susceptible-infected) model because without treatment the cattle remain infected for the entire period of study, tsetse flies remain infected for their entire lifetime and infected

buffalo can harbour the trypanosomes for a long time. We therefore have susceptible and infected compartments for cattle and the tsetse flies, and an infected compartment for the buffalo. The susceptible buffaloes are considered to be constant. Both the tsetse fly and buffalo do not die from trypanosomiasis.

Cattle, buffalo and tsetse fly are denoted by C, B and F respectively, and the compartments are denoted by S_i and I_i (where $i = C, B, F$) for the different populations. μ_i are the natural birth and death rates of the respective populations and α is the cattle mortality rate arising from AAT. β is the tsetse fly biting rate, and $\hat{f}$ and $\hat{f}_1$ are the probabilities that an infected fly will lead to an infected cow and buffalo, and $\hat{f}_2$ is the probability that an infected host will lead to an infected fly. η is a weight of infectivity for the cattle from the buffalo, so for cattle further away from the game park, the weight will be less and for cattle closer to the game park, the weight will be more.
The basic model is therefore formulated with the following system of differential equations:

$$\frac{dS_C}{dt} = \mu_C - \mu_C S_C - \beta \hat{f}(\frac{I_F}{N_C} + \eta \frac{I_F}{N_B}) S_C \tag{2.1}$$

$$\frac{dI_C}{dt} = \beta \hat{f}(\frac{I_F}{N_C} + \eta \frac{I_F}{N_B}) S_C - (\mu_C + \alpha) I_C \tag{2.2}$$

$$\frac{dI_B}{dt} = \beta \hat{f}_1(\frac{I_F}{N_C} + \eta \frac{I_F}{N_B}) S_B^0 - \mu_B I_B \tag{2.3}$$

$$\frac{dS_F}{dt} = \mu_F - \mu_F S_F - \beta \hat{f}_2(\frac{I_C}{N_C} + \frac{I_B}{N_B}) S_F \tag{2.4}$$

$$\frac{dI_F}{dt} = \beta \hat{f}_2(\frac{I_C}{N_C} + \frac{I_B}{N_B}) S_F - \mu_F I_F \tag{2.5}$$

$\frac{I_i}{N_i}$ is the proportion of the infected population, given

$$N_C = S_C + I_C$$
$$N_B = S_B^0 + I_B$$
$$N_F = S_F + I_F$$

where N_i is the total of each population and S_B^0 is the constant susceptible buffalo population.

2.3 The Next-Generation Matrix and R_0

To derive the basic reproduction number (R_0) we use the next-generation matrix (NGM), which consists of two parts, the matrices $\boldsymbol{F}$ and $\boldsymbol{V}$, which show the rate of appearances of new infections in compartment i and the transfer of individuals out of compartment i by all other means.
We have state variables $(S_C, I_C, I_B, S_F, I_F)$ and for the DFE we get $(1, 0, 0, 1, 0)$, so from that and the infectious subsystem from our system of ODEs, given by

$$\frac{dI_C}{dt} = \beta\hat{f}(\frac{I_F}{N_C} + \eta\frac{I_F}{N_B})S_C - (\mu_C + \alpha)I_C \tag{2.6}$$

$$\frac{dI_B}{dt} = \beta\hat{f}_1(\frac{I_F}{N_C} + \eta\frac{I_F}{N_B})S_B^0 - \mu_B I_B \tag{2.7}$$

$$\frac{dI_F}{dt} = \beta\hat{f}_2(\frac{I_C}{N_C} + \frac{I_B}{N_B})S_F - \mu_F I_F \tag{2.8}$$

we arrive at the matrices

$$\boldsymbol{F} = \begin{bmatrix} 0 & 0 & \frac{\beta\hat{f}}{\mu_F}(\frac{S_B^0+\eta}{S_B^0}) \\ 0 & 0 & \beta\hat{f}_1(S_B^0+\eta) \\ \beta\hat{f}_2 & \frac{\beta\hat{f}_2}{S_B^0} & 0 \end{bmatrix}$$

and

$$\boldsymbol{V} = \begin{bmatrix} \mu_C + \alpha & 0 & 0 \\ 0 & \mu_B & 0 \\ 0 & 0 & \mu_F \end{bmatrix}$$

To get the NGM $\boldsymbol{K}$, we multiply the matrices $\boldsymbol{F}$ and $\boldsymbol{V^{-1}}$, so that

$$\boldsymbol{K} = \boldsymbol{F}\boldsymbol{V^{-1}} = \begin{bmatrix} 0 & 0 & \frac{\beta\hat{f}}{\mu_F}(\frac{S_B^0+\eta}{S_B^0}) \\ 0 & 0 & \frac{\beta\hat{f}_1}{\mu_F}(S_B^0+\eta) \\ \frac{\beta\hat{f}_2}{\mu_C+\alpha} & \frac{\beta\hat{f}_2}{S_B^0\mu_B} & 0 \end{bmatrix}$$

where

$$\begin{aligned} R_{CF} &= \frac{\beta\hat{f}}{\mu_F S_B^0}(S_B^0+\eta) \\ R_{BF} &= \frac{\beta\hat{f}_1}{\mu_F}(S_B^0+\eta) \\ R_{FC} &= \frac{\beta\hat{f}_2}{\mu_C+\alpha} \\ R_{FB} &= \frac{\beta\hat{f}_2}{S_B^0\mu_B} \end{aligned}$$

R_0 is then the dominant eigenvalue of $\boldsymbol{K}$. To simplify, we re-write $\boldsymbol{K}$ as

$$\boldsymbol{K} = \begin{bmatrix} 0 & 0 & R_{CF} \\ 0 & 0 & R_{BF} \\ R_{FC} & R_{FB} & 0 \end{bmatrix}$$

which gives the basic reproduction number of

$$R_0 = \sqrt{R_{BF}R_{FB} + R_{CF}R_{FC}}. \tag{2.9}$$

R_{ij} is the average number of cases of type i caused by a single type j or transmission from i to j (Davis et al, 2011).

2.4 Analysis of Model

2.4.1 Equilibrium Points

To find the disease-free and endemic equilibrium points we can write the system of equations 2.1-2.5 in terms of the forces of infection λ_1^* and λ_2^* such that

$$S_C^* = \frac{\mu_C}{\mu_C + \lambda_1^*} \tag{2.10}$$

$$I_C^* = \frac{\lambda_1 *}{\mu_C + \lambda_1^*} A_1 \tag{2.11}$$

$$I_B^* = A_2 \lambda_1^* \tag{2.12}$$

$$S_F^* = \frac{\mu_F}{\mu_F + \lambda_2^*} \tag{2.13}$$

$$I_F^* = \frac{\lambda_2^*}{\mu_F + \lambda_2^*} \tag{2.14}$$

where

$$\lambda_1^* = \frac{\beta \hat{f} \lambda_2^* [(\mu_C + \lambda_1^*)(S_B^0 + A_2 \lambda_1^*) + \eta(\mu_C + A_1 \lambda_1^*)]}{(\mu_C + \lambda_2^*)(\mu_C + A_1 \lambda_1^*)(S_B^0 + A_2 \lambda_1^*)} \tag{2.15}$$

$$\lambda_2^* = \frac{\beta \hat{f}_1 [A_1 \lambda_1^* (S_B^0 + A_2 \lambda_1^*) + A_2 \lambda_1^* (\mu_C + A_1 \lambda_1^*)]}{(\mu_C + A_1 \lambda_1^*)(S_B^0 + A_2 \lambda_1^*)} \tag{2.16}$$

and $A_1 = \frac{\mu_C}{\mu_C + \alpha}$, $A_2 = \frac{S_B^0}{\mu_B}$.

It is clear that $(\lambda_1^*, \lambda_2^*) = (0, 0)$ is a solution to equations 2.15 and 2.16 which corresponds to the DFE $E_0 = (1, 0, 0, 1, 0)$ when substituting them in equations 2.14. This DFE is the point when the population remains free from the disease.

Next we consider what happens when only one or neither of the forces of infection (λ's) are zero:

1. $\lambda_1^* \neq 0, \lambda_2^* = 0$. When substituted into 2.15 we get $\lambda_1^* = 0$ which is contradictory since we said that $\lambda_1 \neq 0$.

2. $\lambda_1^* = 0, \lambda_2^* \neq 0$. When substitued into 2.16 we get $\lambda_2^* = 0$ which is contradictory

because we said $\lambda_2^* \neq 0$.

3. $\lambda_1^* \neq 0, \lambda_2^* \neq 0$. It follows that this is the only case which gives an endemic equilibrium.

Next we show that there exists a unique fixed point $(\lambda_1^*, \lambda_2^*)$ with $\lambda_1^* > 0$ and $\lambda_2^* > 0$ satisfying

$$\psi(\lambda_1^*, \lambda_2^*) = \begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix}$$

where

$$\psi_1^{\lambda_1}(\lambda_2) = \frac{\beta \hat{f} \lambda_2 [(\mu_C + \lambda_1)(S_B^0 + A_2\lambda_1) + \eta(\mu_C + A_1\lambda_1)]}{(\mu_C + \lambda_2)(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)} \tag{2.17}$$

$$\psi_2^{\lambda_2}(\lambda_1) = \frac{\beta \hat{f}_1 [A_1\lambda_1(S_B^0 + A_2\lambda_1) + A_2\lambda_1(\mu_C + A_1)]}{(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)} \tag{2.18}$$

that corresponds to the endemic equilibrium point E^*. Note that $\psi_1^{\lambda_1}$ and $\psi_2^{\lambda_2}$ are the two components of $\psi(\lambda_1^*, \lambda_2^*)$.
For each fixed $\lambda_1 > 0$ we have the following real valued function that depends on λ_2

$$\psi_1^{\lambda_1}(\lambda_2) = \frac{\beta \hat{f} \lambda_2 [(\mu_C + \lambda_1)(S_B^0 + A_2\lambda_1) + \eta(\mu_C + A_1\lambda_1)]}{(\mu_C + \lambda_2)(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)}$$

from which

$$\psi_1^{\lambda_1}(0) = 0$$

and

$$\lim_{\lambda_2 \to \infty} \psi_1^{\lambda_1}(\lambda_2) = \frac{\beta \hat{f} [(\mu_C + \lambda_1)(S_B^0 + A_2\lambda_1) + \eta(\mu_C + A_1\lambda_1)]}{(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)} < \infty$$

The first and second derivatives of $\psi_1^{\lambda_1}(\lambda_2)$ are given by

$$\frac{\partial \psi_1^{\lambda_1}(\lambda_2)}{\partial \lambda_2} = \frac{\beta \hat{f} [(\mu_C + \lambda_1)(S_B^0 + A_2\lambda_1) + \eta(\mu_C + A_1\lambda_1)\mu_C]}{[(\mu_C + \lambda_2)(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^2} > 0$$

$$\frac{\partial^2 \psi_1^{\lambda_1}(\lambda_2)}{\partial \lambda_2^2} = \frac{(\mu_c + \lambda_2) - 2\beta \hat{f} \mu_C [(\mu_C + \lambda_1)(S_B^0 + A_2\lambda_1) + \eta(\mu_C + A_1\lambda_1)]}{(\mu_C + \lambda_2)^3} < 0$$

so we can therefore see that $\psi_1^{\lambda_1}(\lambda_2)$ is an increasing function that is finite and concave down with no change of convexity $\forall t \geq 0$ and there exists $\lambda_2^* > 0$ such that $\psi_1^{\lambda_1}(\lambda_2^*) = \lambda_2^*$.

Similarly, for $\lambda_2^* > 0$ we have a function that depends on λ_1

$$\psi_2^{\lambda_2^*}(\lambda_1) = \frac{\beta \hat{f}_1[A_1\lambda_1(S_B^0 + A_2\lambda_1) + A_2\lambda_1(\mu_C + A_1)]}{(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)}$$

from which

$$\psi_2^{\lambda_2^*}(0) = 0$$

and

$$\lim_{\lambda_1 \to \infty} \psi_2^{\lambda_2^*}(\lambda_1) = 2\beta\hat{f}_1 < \infty$$

The first and second derivatives for $\psi_2^{\lambda_2^*}(\lambda_1)$ are

$$\begin{aligned}
\frac{\partial \psi_2^{\lambda_2^*}(\lambda_1)}{\partial \lambda_1} =& \frac{\beta\hat{f}_1\{[(A_1S_B^0 + A_1A_2\lambda_1 + A_2\mu_C)(\mu_C(S_B^0 + A_2\lambda_1) + A_1\lambda_1(S_B^0 + A_2\lambda_1))]\}}{[(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^2} \\
& - \frac{\beta\hat{f}_1\{[(A_1S_B^0 + A_2\mu_C)\lambda_1 + 2A_1A_2\lambda_1^2][\mu_C A_2 + A_1S_B^0 + 2A_1A_2\lambda_1]\}}{[(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^2} \quad > 0 \\
\text{if} \quad & \frac{\{[(A_1S_B^0 + A_1A_2\lambda_1 + A_2\mu_C)(\mu_C(S_B^0 + A_2\lambda_1) + A_1\lambda_1(S_B^0 + A_2\lambda_1))]\}}{-\{[(A_1S_B^0 + A_2\mu_C)\lambda_1 + 2A_1A_2\lambda_1^2][\mu_C A_2 + A_1S_B^0 + 2A_1A_2\lambda_1]\}} \quad > 1
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \psi_2^{\lambda_2^*}(\lambda_1)}{\partial \lambda_1^2} =& \frac{-2\beta\hat{f}_1A_1A_2[2A_1A_2\lambda_1 - 2(\mu_C A_2 + A_1S_B^0)][A_1S_B^0 + 4A_1A_2\lambda_1 + \mu_C A_2]}{[(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^3} \\
& \times \frac{[\mu_C S_B^0 + (\mu_C A_2 + A_1S_B^0)\lambda_1 + A_1A_2\lambda_1^2]}{[(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^3} \\
& + \frac{2\beta\hat{f}_1A_1A_2[(A_1S_B^0 + A_2\mu_C)\lambda_1 + 2A_1A_2\lambda_1^2][\mu_C A_2 + A_1S_B^0 + 2A_1A_2\lambda_1]}{[(\mu_C + A_1\lambda_1)(S_B^0 + A_2\lambda_1)]^3} \quad < 0 \\
\text{if} \quad & \frac{-[2A_1A_2\lambda_1 - 2(\mu_C A_2 + A_1S_B^0)][A_1S_B^0 + 4A_1A_2\lambda_1 + \mu_C A_2]}{[(A_1S_B^0 + A_2\mu_C)\lambda_1 + 2A_1A_2\lambda_1^2][\mu_C A_2 + A_1S_B^0 + 2A_1A_2\lambda_1]} \\
& \times \frac{[\mu_C S_B^0 + (\mu_C A_2 + A_1S_B^0)\lambda_1 + A_1A_2\lambda_1^2]}{[(A_1S_B^0 + A_2\mu_C)\lambda_1 + 2A_1A_2\lambda_1^2][\mu_C A_2 + A_1S_B^0 + 2A_1A_2\lambda_1]} \quad < 1
\end{aligned}$$

so $\psi_2^{\lambda_2^*}(\lambda_1)$ is an increasing concave down function (if the conditions for $\frac{\partial \psi_2^{\lambda_2^*}(\lambda_1)}{\partial \lambda_1}$ and $\frac{\partial^2 \psi_2^{\lambda_2^*}(\lambda_1)}{\partial \lambda_1^2}$ hold) with no change of convexity $\forall t \geq 0$ and there exists $\lambda_1^* > 0$ such that $\psi_2^{\lambda_2^*}(\lambda_1^*) = \lambda_1^*$.

Theorem. *There exists a nonzero fixed point $(\lambda_1^*, \lambda_2^*)$ of the system 2.15 and 2.16 corresponding to the endemic equilibrium point of the system 2.1-2.5.*

2.4.2 Stability of Equilibrium Points

Now we shall look at the global stability of our fixed points $\psi(\lambda_1^*, \lambda_2^*)$, by looking at the Jacobian matrix given by

$$J = \begin{bmatrix} \frac{\partial \psi_1}{\partial \lambda_1} & \frac{\partial \psi_1}{\partial \lambda_2} \\ \frac{\partial \psi_2}{\partial \lambda_1} & \frac{\partial \psi_2}{\partial \lambda_2} \end{bmatrix}$$

The first partial derivatives of $\psi(\lambda_1^*, \lambda_2^*)$ at fixed point $(0, 0)$ are given by

$$\begin{aligned} \frac{\partial \psi_1(0,0)}{\partial \lambda_1} &= 0 \\ \frac{\partial \psi_1(0,0)}{\partial \lambda_2} &= \frac{\beta \hat{f}(\mu_C S_B^0 + \eta \mu_C^2)}{\mu_C^2 S_B^0} \\ \frac{\partial \psi_2(0,0)}{\partial \lambda_1} &= \frac{\beta \hat{f}_1[(A_1 S_B^0 + A_2 \mu_C)\mu_C S_B^0]}{(\mu_C S_B^0)^2} \\ \frac{\partial \psi_2(0,0)}{\partial \lambda_2} &= 0 \end{aligned}$$

The characteristic equation obtained from the Jacobian at the fixed point $(0, 0)$ is

$$\lambda^2 - (\frac{\partial \psi_1(0,0)}{\partial \lambda_2})(\frac{\partial \psi_2(0,0)}{\partial \lambda_1}) = 0.$$

For stability, we need for $max\{|\lambda_a|, |\lambda_b|\} < 1$, where λ_a and λ_b are given by

$$\lambda_a = \sqrt{\frac{\beta}{\mu_C^2}(\frac{\hat{f}\hat{f}_1(\mu_C S_B^0 + \eta\mu_C^2)(A_1 S_B^0 + A_2\mu_C)\mu_C S_B^0}{{S_B^0}^3})}$$

$$\lambda_b = -\sqrt{\frac{\beta}{\mu_C^2}(\frac{\hat{f}\hat{f}_1(\mu_C S_B^0 + \eta\mu_C^2)(A_1 S_B^0 + A_2\mu_C)\mu_C S_B^0}{{S_B^0}^3})}$$

The fixed point $(0,0)$ is therefore stable when the dominant eigenvalue $\lambda_a < 1$. The Jacobian matrix of $\psi(\lambda_1, \lambda_2)$ at the fixed point $(\lambda_1^*, \lambda_2^*)$ is

$$J(\lambda_1^*, \lambda_2^*) \begin{bmatrix} \frac{\partial\psi_1(\lambda_1,\lambda_2)}{\partial\lambda_1}\Big|_{(\lambda_1^*,\lambda_{2*})} & \frac{\partial\psi_1(\lambda_1,\lambda_2)}{\partial\lambda_2}\Big|_{(\lambda_1^*,\lambda_{2*})} \\ \frac{\partial\psi_2(\lambda_1,\lambda_2)}{\partial\lambda_1}\Big|_{(\lambda_1^*,\lambda_{2*})} & \frac{\partial\psi_2(\lambda_1,\lambda_2)}{\partial\lambda_2}\Big|_{(\lambda_1^*,\lambda_{2*})} \end{bmatrix}$$

where

$$\begin{aligned}
\frac{\partial\psi_1(\lambda_1,\lambda_2)}{\partial\lambda_1}\Big|_{(\lambda_1^*,\lambda_{2*})} &= \frac{\beta\hat{f}\lambda_2^*[(\mu_C A_2 + S_B^0 + \eta A_1 + 2A_2\lambda_1^*)(\mu_C + \lambda_2^*)(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]}{[(\mu_C + \lambda_2^*)(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]^2} \\
&\quad \frac{-\beta\hat{f}\lambda_2^*[(\mu_C + \lambda_1^*)(S_B^0 + A_2\lambda_1^*) + \eta(\mu_C + A_1\lambda_1^*)(\mu_C A_2 + A_1 S_B^0 + 2A_1 A - 2\lambda_1^*)]}{[(\mu_C + \lambda_2^*)(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]^2} \\
\frac{\partial\psi_1(\lambda_1,\lambda_2)}{\partial\lambda_2}\Big|_{(\lambda_1^*,\lambda_{2*})} &= \frac{\beta\hat{f}[(\mu_C + \lambda_1^*)(S_B^0 + A_2\lambda_1^*) + \eta(\mu_C + A_1\lambda_1^*)\mu_C]}{[(\mu_C + \lambda_2^*)(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\psi_2(\lambda_1,\lambda_2)}{\partial\lambda_1}\Big|_{(\lambda_1^*,\lambda_{2*})} &= \frac{\beta\hat{f}_1\{[(A_1 S_B^0 + 4A_1 A_2\lambda_1^* + A_2\mu_C)(\mu_C(S_B^0 + A_2\lambda_1^*) + A_1\lambda_1^*(S_B^0 + A_2\lambda_1^*))]\}}{[(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]^2} \\
&\quad - \frac{\beta\hat{f}_1\{[(A_1 S_B^0 + A_2\mu_C)\lambda_1^* + 2A_1 A_2{\lambda_1^*}^2][\mu_C A_2 + A_1 S_B^0 + 2A_1 A_2\lambda_1^*]\}}{[(\mu_C + A_1\lambda_1^*)(S_B^0 + A_2\lambda_1^*)]^2} \\
\frac{\partial\psi_2(\lambda_1,\lambda_2)}{\partial\lambda_2}\Big|_{(\lambda_1^*,\lambda_{2*})} &= 0
\end{aligned}$$

and the characteristic equation is given by

$$\lambda^2 - \lambda\,(\frac{\partial\psi_1}{\partial\lambda_1})\Big|_{\lambda_1^*,\lambda_2^*} + (\frac{\partial\psi_1}{\partial\lambda_1} - \frac{\partial\psi_1}{\partial\lambda_2}\frac{\partial\psi_2}{\partial\lambda_1})\Big|_{\lambda_1^*,\lambda_2^*} = 0.$$

The roots of the above characteristic equation are as follows

$$
\begin{aligned}
D_1 &= \frac{1}{2}(P_0 + \sqrt{P_0^2 - 4P_1}) \\
D_2 &= \frac{1}{2}(P_0 - \sqrt{P_0^2 - 4P_1}) \\
\text{where} \quad P_0 &= (\frac{\partial \psi_1}{\partial \lambda_1})\Big|_{(\lambda_1^*, \lambda_2^*)} \\
P_1 &= (\frac{\partial \psi_1}{\partial \lambda_1} - \frac{\partial \psi_1}{\partial \lambda_2}\frac{\partial \psi_2}{\partial \lambda_1})\Big|_{(\lambda_1^*, \lambda_2^*)}.
\end{aligned}
$$

The fixed point $(\lambda_1^*, \lambda_2^*)$ is therefore stable when $max\{|D_1|, |D_2|\} < 1$.

2.5 Model Simulations

The initial variable values (Table 2.1) are either estimated from our sample data (see Chapter 4) or assumed, and the initial parameter values given in Table 2.2, are from various literature sources. They were used to assess how the different populations $(S_C, I_C, I_B, S_F, I_F)$ changed over time.

Table 2.1: Population variable description and initial values

Variable Description and initial values		
Variable	**Description**	**Value**
N_C	Number of cattle	1462
N_B	Number of buffaloes	500
S_B^0	Number of susceptible buffaloes	378
N_F	Number of tsetse flies	687

The model was run for a year (365 days) and R_0 was calculated. We first want to see whether infection can be maintained just within the cattle population without any infection from the buffaloes. For infection from only the cattle (Figure 2.1), infected cattle, buffalo and flies rise at a much slower rate and similarly, susceptible cattle and flies fall at slower rates. When infection comes from buffaloes (Figure 2.2) then the number of infected cattle rises very steeply in a shorter amount of time, and the number of susceptible flies falls at a faster pace (number of S_F is more than 100 when at 50 days when we look at infection from the cattle compared to less than 100 S_F when the infection is coming from the buffaloes).

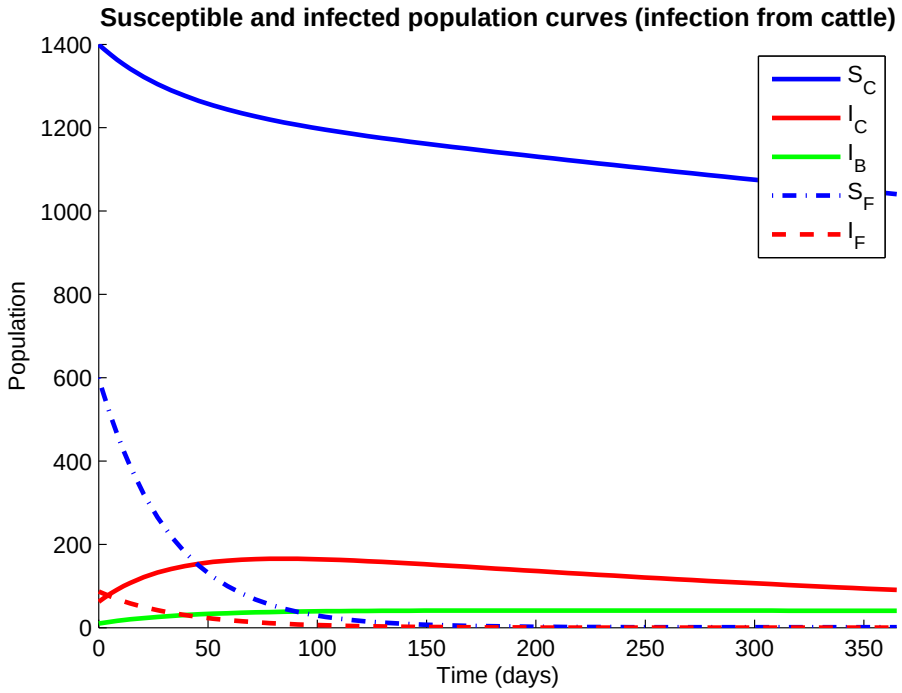


Figure 2.1: Population curves when infection is from the cattle only

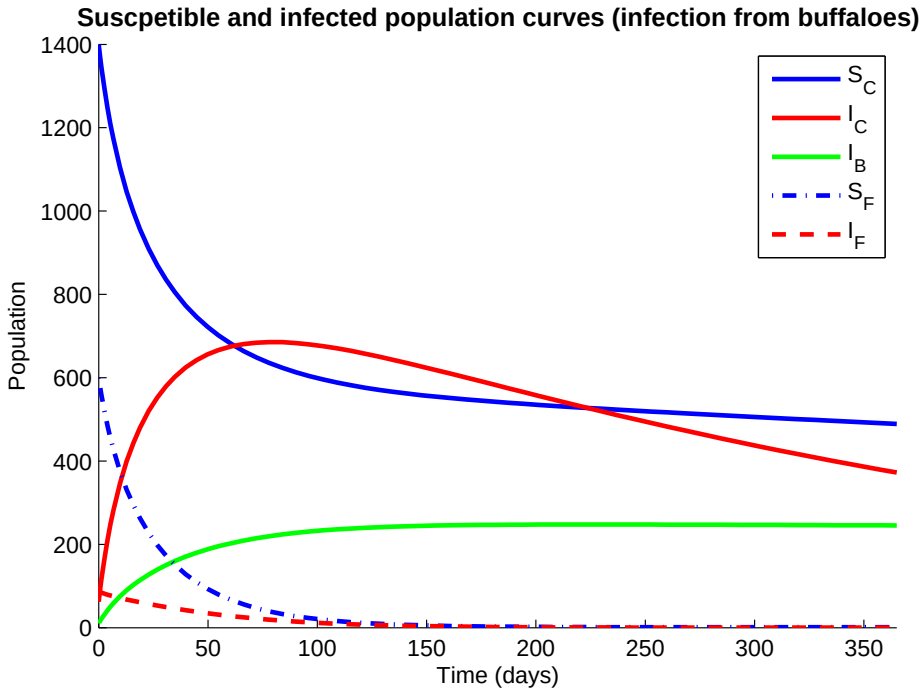


Figure 2.2: Population curves when infection is from the buffaloes only

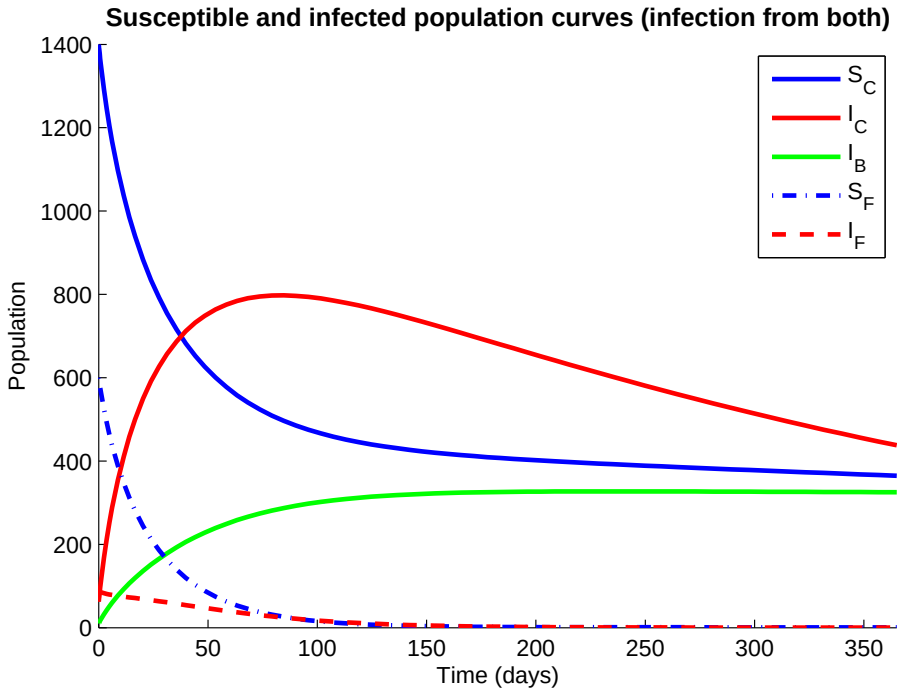


Figure 2.3: Population curves when the infection is from both the cattle and the buffaloes

Table 2.2: Initial parameter values obtained from different literature sources

Parameter description and values			
Parameter	**Description**	**Value**	**Reference**
μ_C	Mortality rate of cattle (excluding disease-induced)($days^{-1}$)	0.0005	Milligan and Baker, 1988
μ_B	Mortality rate of buffaloes ($days^{-1}$)	0.0001053	Hassan, Garba, Gumel and Lubuma, 2014
μ_F	Mortality rate of tsetse fly ($daily$)	0.03	Rogers, 1988
α	Disease-induced mortality rate of cattle ($days^{-1}$)	0.002	Milligan and Baker, 1988
β	Biting rate ($days^{-1}$)	0.25	Milligan and Baker, 1988
η	Weight (distance from buffaloes)	0-3	estimated
$\hat{f}$	Probability of infected fly producing infection in cow	0.2	Milligan and Baker, 1988
$\hat{f}_1$	Probability of infected fly producing infection in buffalo	0.46	Rogers, 1988
$\hat{f}_2$	Probability of infected blood meal from host producing infection in fly	0.025, 0.05	Rogers, 1988; Milligan and Baker, 1988

When we look at infection from both cattle and buffaloes then we notice the sharp rises of both infected cattle and flies in a shorter period of time (Figure 2.3). These figures show that buffaloes contribute to the number of infected cattle quite significantly, as opposed to when infection just occurs between the cattle population i.e. without the contribution of the reservoir host, the buffalo. To note, the curves for I_B in Figures 2.1-2.3 rise and do not dip down again and that is because the buffalo has a long lifespan, of approximately 15 years, and as stated before do not lose infectiousness throughout their lifetime. Next we look at the NGM and the value of R_0:

$$\boldsymbol{K} = \begin{bmatrix} 0 & 0 & 1.67 \\ 0 & 0 & 1453.98 \\ 2.5 & 0.157 & 0 \end{bmatrix}$$

$$R_0 = 15.25.$$

R_{CF}, R_{FC} and R_{BF} are 1.67, 2.5 and 1453.98 respectively. These values show that the transmission from cattle and buffaloes to flies and transmission from flies to cattle are important in maintaining AAT within their populations (since they are greater than 1) which is in agreement with what is known about the epidemiology of AAT in KZN (Motloang et al, 2014). When we just consider transmission between cattle and flies (R_{CF} and R_{FC}), we see that AAT can be maintained between them but the important element in our NGM $\boldsymbol{K}$ is the transmission from buffaloes to flies which contributes the most to the R_0 value. Transmission from flies to buffaloes ($R_{FB} = 0.314$) is not important as buffaloes do not display any clinical symptoms from the trypanosomal infection. Finally, the value of R_0 of 15.25, which is greater than 1, means AAT will persist in the population, so interventions or control measures should be looked at.

2.6 Sensitivity Analysis

2.6.1 Methods

Uncertainty and sensitivity analyses are performed in order to explore the behaviour of complex models and to assess the uncertainty from the input parameters that result in uncertainties in the model outcome variables and how the variations in those model outputs can be apportioned (Marino et al, 2008). These uncertainties in the input parameters are a result of natural variation, measurement error or the inability to measure them. Uncertainty analysis is used to assess the variability in the outcome variables that arises from the uncertainty of estimating the input parameter values (Blower and Dowlatabadi, 1994). An efficient statistical analysis technique that can be used to perform uncertainty analysis is the Latin hypercube sampling (LHS) method, which is a type of stratified Monte Carlo sampling (Mckay, Beckman and Conover, 2000). Sensitivity analysis then extends the uncertainty analysis by identifying the critical input parameters that are important in contributing to the variability of the outcome variables (Blower and Dowlatabadi, 1994). Global sensitivity analysis is better than local sensitivity analysis due to the range of values that are explored, as opposed to using fixed-point estimates (Davis et al, 2011). Differ-

ent sensitivity analysis measures can be used, such as the Pearson correlation coefficient, Spearman rank correlation coefficient and the partial rank correlation coefficient (PRCC). The latter two are for non-linear, monotonic relationships, whereas the former is best for linear relationships (Marino et al, 2008).

LHS (introduced by Mckay et al (1979)) generates samples of each parameter from a multidimensional distribution by randomly selecting without replacement values from each of the parameters' probability distribution function so that the entire range is explored (Marino et al, 2008). So, in other words, LHS allows for simultaneous variation of the input parameter values and each parameter is treated as a random variable with a defined probability distribution function for each one (Blower and Dowlatabadi, 1994). Since it uses a probabilistic selection technique, LHS enables results from a deterministic model to be interpreted within a statistical framework. If we have X_i $(i = 1, 2, \ldots, k)$ input parameters where i represents the first input parameter and k is the last input parameter, in a model and a sample size of N, that is the number of simulations run, we will have an LHS matrix with N rows and X_i columns, representing the number of simulations run and the number of varied parameters respectively. The range of each parameter is divided into N equiprobable intervals of equal marginal probability $\frac{1}{N}$ and sampled once (Mckay et al, 2000). Each interval is assigned a sampling index from $1, \ldots, N$, so say an input parameter X_1 is being sampled, we will have N sample values of $x_1, x_2, \ldots, x_N$.
The probability distribution function of X_i will be $f(X_i)$, where $f(X_i)$ can be from a uniform distribution, Weibull, Normal distribution etc. The resultant model will have N observations of the outcome variable of interest e.g. N values of R_0 using each combination of parameter values. The sample size (number of simulations run) N should at least be equal to $X + 1$ but larger sample sizes guarantee reduced variability in the estimates of the outcome variable and greater sampling coverage of the actual probability distribution functions from which the input parameters were sampled (Sanchez and Blower, 1997).

Since the input parameters of disease transmission models are rarely normally distributed and the outcome variables are non-linear functions, non-parametric tests of ranked data are used, therefore the sensitivity analysis measure that we will use is the PRCC (Blower and Dowlatabadi, 1994). PRCC, introduced earlier, performs a partial correlation on rank-transformed data and is used to evaluate the statistical relationship between each input parameter and each outcome variable (generated by the LHS scheme) while keeping the other input parameters constant. So if X_i and Y are the input parameters and outcome

variable respectively, then the partial correlation coefficient is the correlation coefficient between the residuals $(X_i - \hat{X}_i)$ and $(Y - \hat{Y})$ where X_i and Y are first rank transformed. PRCC is a robust measure of sensitivity when there is little or no correlation between the inputs (Marino et al, 2008).

MATLAB R2013a was used to generate the LHS matrix and do the sensitivity analysis using the PRCC measure. The parameters' were assumed to follow a uniform distribution with a minimum and maximum value for each, since their statistical distributions are not known, and 1000 simulations were run ($N = 1000$) for each simultaneously.

2.6.2 Results

The sensitivity analysis was done on the parameters affecting the infected classes (our outcome variables) i.e. infected cattle, buffaloes and flies. To assess the monotonicity between the input parameters and the outcome variables, scatter plots were done, given in Appendix A. The partial rank correlation coefficients (PRCCs) for each parameter in relation to I_C, I_B and I_F (Table 2.3) were determined as well as their significance based on a 5% level of significance (shown in Figures 2.4, 2.5 and 2.6). Five time points were used to assess how the parameters' PRCCs varied, namely at 10, 30, 50, 70 and 90 days. The sign of the PRCC shows the relationship between the parameters and the outcome variables, and the magnitude of the PRCC shows how important that parameter is in contributing to the prediction imprecision of the outcome variable (Blower and Dowlatabadi, 1994; Marino et al, 2008). When we look at all our outcome variables (I_C, I_B and I_F), the parameters that are consistently shown as important are μ_F and β. This means that the uncertainty in estimating the values of these parameters contributes the most to the prediction imprecision of our outcome variables. The other parameters that are important in relation to I_C are μ_C and $\hat{f}$ at time point 10 only. They both have negative relationships with I_C meaning as the natural mortality rate and the probability of an infected fly producing infection in the cow increases then the number of future infected cattle will decrease. For μ_C that makes sense, but we would expect that as $\hat{f}$ increases then so will the number of future infected cattle, but as it is only important at time point 10, we can say that $\hat{f}$ is not important in the prediction imprecision of future infected cattle cases. The other values for I_B and I_F that are in bold can be interpreted in a similar way.

The values of the PRCCs are given in Table 2.3.

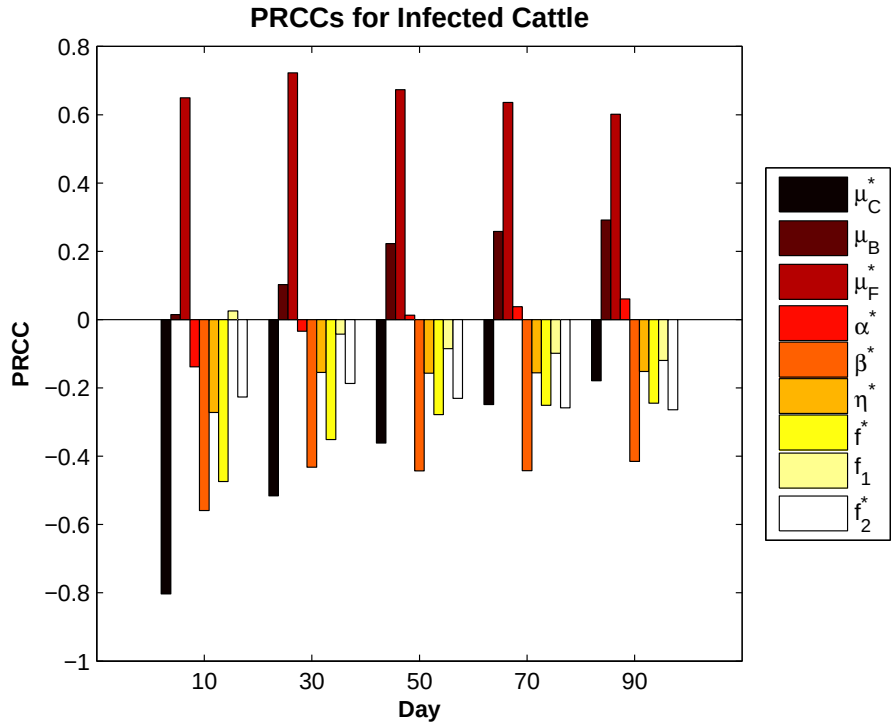


Figure 2.4: The partial rank correlation coefficients for infected cattle where the significant parameters (marked with *) are μ_C, μ_F, $\alpha, \beta, \eta, \hat{f}$ and $\hat{f}_2$.

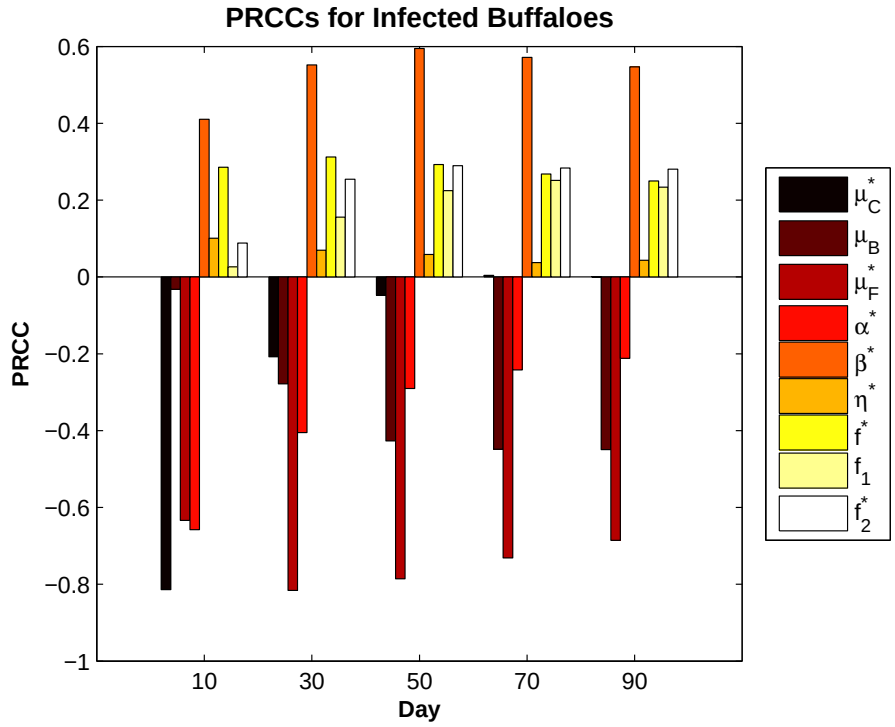


Figure 2.5: The partial rank correlation coefficients for infected buffaloes where the significant parameters (marked with *) are $\mu_C, \mu_F, \alpha, \beta, \eta, \hat{f}$ and $\hat{f}_2$.

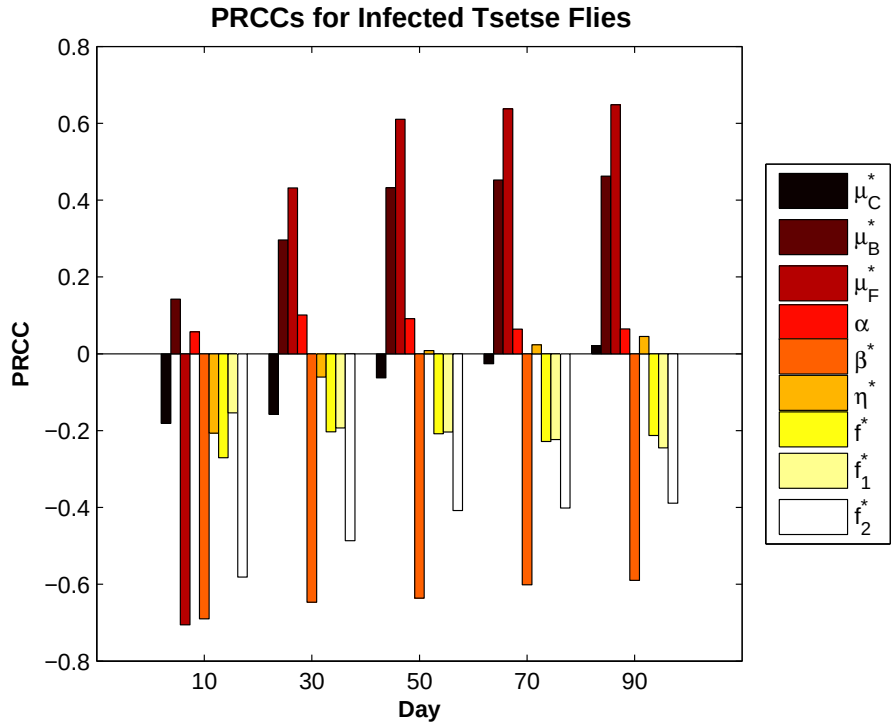


Figure 2.6: The partial rank correlation coefficients for infected flies where the significant parameters (marked with *) are $\mu_C, \mu_B, \mu_F, \beta, \eta, \hat{f}, \hat{f}_1$ and $\hat{f}_2$.

Table 2.3: The PRCC values for infected cattle, buffaloes and flies for the different time points with bold values ≤ -0.5 or ≥ 0.5.

Day	$\boldsymbol{\mu_C}$	$\boldsymbol{\mu_F}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\hat{\boldsymbol{f}}$	$\hat{\boldsymbol{f}}_{\boldsymbol{2}}$
Partial Rank correlation coefficients for Infected Cattle						
10	**-0.7693**	**0.6603**	-0.1685	**-0.6107**	**-0.5961**	-0.1514
30	-0.4505	**0.7204**	-0.0205	-0.4615	-0.3771	-0.1390
50	-0.2865	**0.6869**	-0.0012	-0.4442	-0.2885	-0.2102
70	-0.1687	**0.6516**	0.0153	-0.4704	-0.2660	-0.2598
90	-0.0641	**0.5853**	0.0319	-0.4514	-0.2355	-0.2422
Partial Rank correlation coefficients for Infected Buffaloes						
10	**-0.8062**	**-0.6539**	**-0.7023**	0.3426	0.2804	0.0116
30	-0.2203	**-0.7839**	-0.3974	**0.5485**	0.3235	0.2516
50	-0.1275	**-0.7657**	-0.2589	**0.6132**	0.3360	0.3457
70	-0.1025	**-0.7147**	-0.1899	**0.6022**	0.2855	0.3547
90	-0.0774	**-0.6561**	-0.1717	**0.5717**	0.2643	0.3222
Partial Rank correlation coefficients for Infected Tsetse flies						
10	-0.1879	**-0.6829**	0.0715	**-0.709**	-0.2648	**-0.6618**
30	-0.1168	0.3594	0.0919	**-0.6321**	-0.1464	**-0.5262**
50	-0.0728	**0.5759**	0.1238	**-0.6449**	-0.1581	-0.4637
70	-0.0228	**0.5858**	0.1197	**-0.6088**	-0.1361	-0.423
90	-0.0221	**0.5779**	0.1217	**-0.5960**	-0.1033	-0.3992

2.7 Parameters Affecting the Basic Reproduction Number

To see which parameters affect R_0 clearly, contour plots of the important ones were done to determine how the value of R_0 changes as they increase (Figures 2.7, 2.8 and 2.9). The parameters that increase R_0 the most are β, μ_F, $\hat{f}_1$ and $\hat{f}_2$. From Figures 2.7 and 2.8, we can see that high values of β and $\hat{f}_1$ and $\hat{f}_2$ increase R_0, more so for β and $\hat{f}_2$ than $\hat{f}_1$. Figure 2.9 shows that even for low values of β and high fly mortality rates (μ_F), R_0 is quite high and becomes a lot higher as β increases and μ_F decreases. This shows that R_0 is most sensitive to μ_F and β.

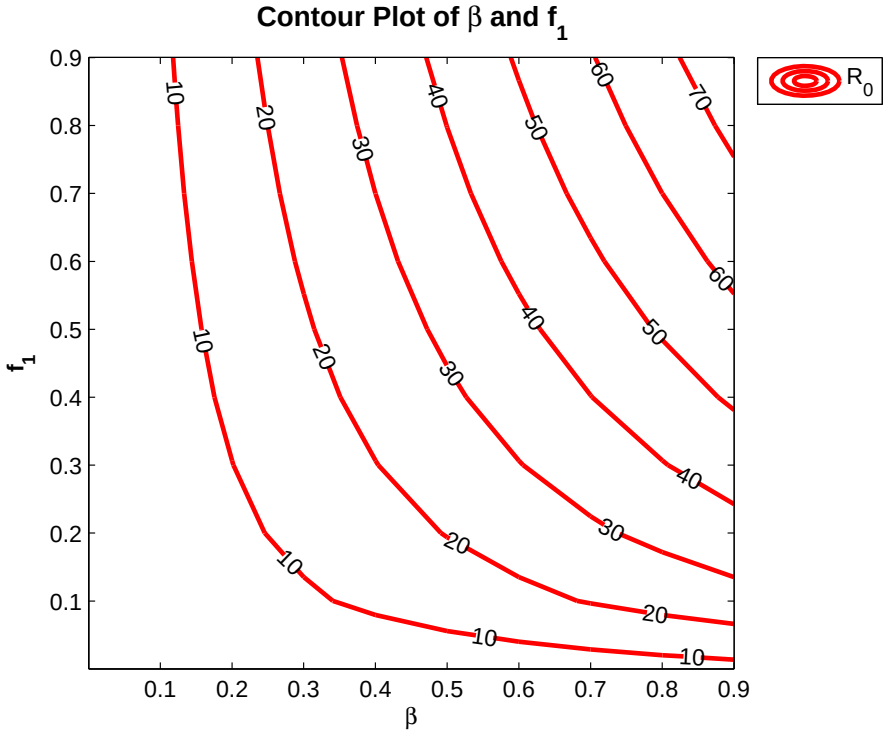


Figure 2.7: Contour plot of how changes in β and $\hat{f}_1$ affect R_0

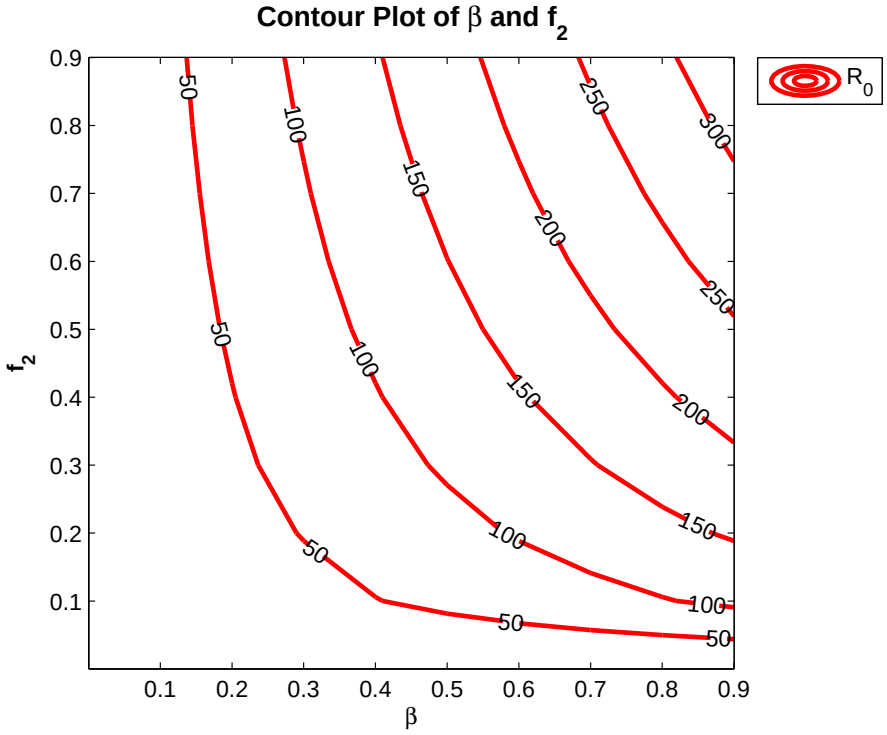


Figure 2.8: Contour plot of how changes in β and $\hat{f}_2$ affect R_0

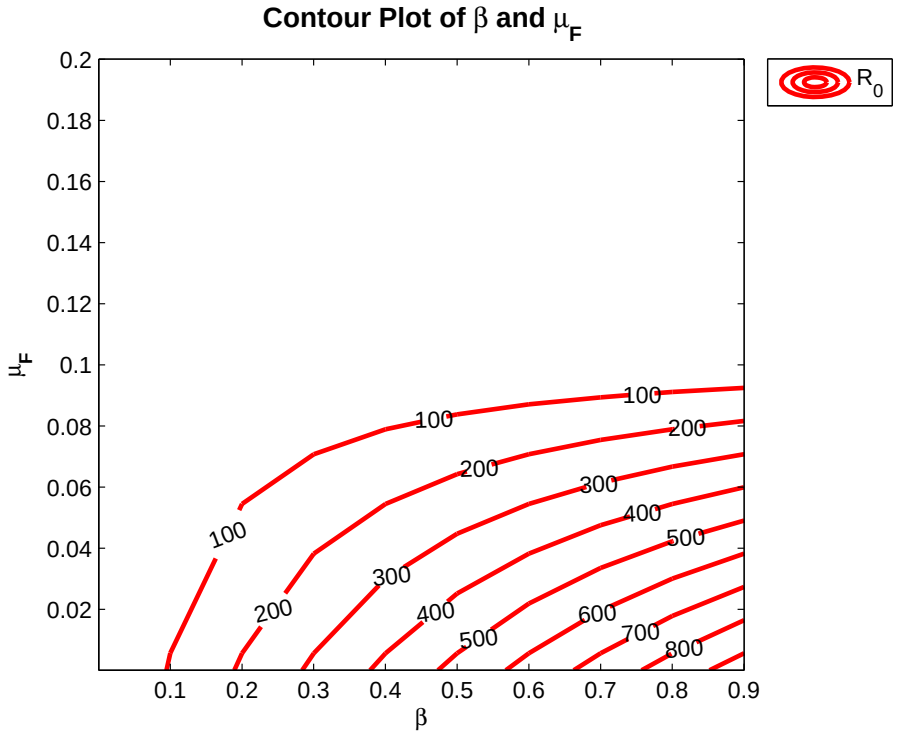


Figure 2.9: Contour plot of how changes in β and μ_F affect R_0

2.8 Summary

We have constructed a basic SI model (susceptible and infected) depicting the dynamics of AAT in KZN, where there is interaction between three populations, namely cattle which are the hosts, buffaloes which are considered the reservoir hosts and tsetse flies which are the vectors. We have used the next-generation matrix (NGM) for the two-host, one-vector species and used it to derive R_0. To assess how the populations vary when infection is just from the cattle or just from the buffalo or from both, we saw that the when infection was coming from only the cattle then the infected cattle, buffaloes and flies rose slower than when the infection was coming from only the buffalo. When we calculated R_0 this confirmed that transmission from the buffalo to the fly was greater than when the transmission is between the cattle and the fly.

A sensitivity analysis was done to assess which parameters are important in affecting the model outputs. This showed that μ_F and β were the most critical parameters in affecting the output variability, therefore getting better estimates of these parameters will result in better prediction precision of future AAT cases in cattle. This was also seen by the contour plots which showed how certain parameters affect the value of R_0.

Chapter 3

Generalized Linear Models

3.1 Introduction

Generalized linear models (GLMs) are often used to model biological data, which often have values that are binary (0 and 1) such as when looking at the distribution of a disease and whether it is present or absent in an area (Rogers, 2006), or count data which take on non-negative integer values. GLMs were introduced in 1972 by Nelder and Wedderburn, and are an extension of the classic linear models, to allow for data with non-normally distributed outcomes and non-linear relationships.

GLMs consist of three components (McCullagh and Nelder, 1989):

i. A *random component*, specifying the conditional distribution of the $n \times 1$ vector of response variables $\boldsymbol{Y}$, given the values of the $n \times p$ matrix of explanatory variables $\boldsymbol{X}$ in the model, where the distribution of $\boldsymbol{Y}$ comes from the exponential family such as the Binomial, Poisson and Gamma distributions, among others.

ii. A linear predictor (*systematic component*) η_i which is a linear function of predictor variables $\boldsymbol{x_i}' = (x_1, x_2, \ldots, x_p)$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad \text{or in matrix form}$$

$$\boldsymbol{\eta} = \boldsymbol{x_i}'\boldsymbol{\beta}.$$

So unlike in classic linear models, where the mean of Y_i has a linear relationship to $\boldsymbol{x}_i'$, GLMs use $\boldsymbol{\eta}$ in the same way.

iii. A *link function* $g(.)$, which transforms the means $\mu_i = E(Y_i)$ to the linear predictor, (i.e. a link between the random and systematic components), given by:

$$g(\mu_i) = \eta_i$$

$g(.)$ can take various forms depending on the distribution of the response variable of interest. For Gaussian outcomes, for example, the identity link is used, for Poisson counts, it is the log link, for binomially distributed data it is the logit link etc.

3.2 The Model

Since GLMs are extended from classical linear models, we shall first consider the form of general linear models. Let Y_i be a random variable which follows the Normal distribution, with mean μ_i and variance σ^2 that is

$$Y_i \sim N(\mu_i, \sigma^2)$$

where the expected value μ_i $(i = 1, 2, \ldots, n)$ is a linear function of p predictors, with values $X_i' = (x_i, \ldots, x_p)$ for the ith case and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ be a $p \times 1$ vector of unknown parameters, so that $\mu_i = X_i'\boldsymbol{\beta}$. The classic linear model assumes the error terms are Normally distributed with a constant variance σ^2, whereas GLMs allow flexibility in the error structure of the regression model. Using the components written earlier, if we have a random variable Y_i and a vector of independent observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, we can say that Y is an independent Normal variable with a constant variance σ^2 and mean $E(Y_i) = \mu_i$, with the identity link function and the linear predictor is equal to the mean. For GLMs, Y_i may come from an exponential family (or extended to non-exponential families such as the negative binomial distribution) and the link function may be any monotonic differentiable function (McCullagh and Nelder, 1989). So for example, if we assume that Y_i comes from a Gamma distribution, then the inverse link function is used and $\eta_i = g(\mu_i) = \mu_i^{-1}$. In the case of classic linear models, the parameters $\boldsymbol{\beta}$ are fit using least squares, but for GLMs they are fit using maximum likelihood techniques. Note that for Normally distributed observations, the two methods give the same results.

3.2.1 Exponential Family of Distributions

Given that our response variable Y_i comes from a distribution in the exponential family i.e. Gaussian, binomial, Poisson, gamma and inverse-Gaussian, we may use the following form to express the probability density function:

$$f(y_i;\theta,\phi) = \exp(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)) \tag{3.1}$$

where θ is a canonical parameter (or location parameter), and the canonical link function is such that $\theta_i = g(\mu_i) = \eta_i$ for the assumed distribution. ϕ is a dispersion (or scale) parameter with $\phi > 0$, and $a(.)$, $b(.)$ and $c(.)$ are some known functions that are related to the exponential family. The canonical link may be the identity function, log, logit etc. and there exists a sufficient statistic for $\boldsymbol{\beta}$ which is given by $\boldsymbol{X}^T\boldsymbol{Y}$ and ϕ may be known or otherwise estimated, which is something we shall consider later on.

The general expression given by (3.1) can be used to derive the mean and variance of the specific distribution, where the mean $E(Y_i)$ and variance $var(Y_i)$ are given by

$$E(Y_i) = \mu_i = \frac{db(\theta)}{d\theta} = b'(\theta) \quad \text{and} \tag{3.2}$$

$$var(Y_i) = \sigma_i^2 = a(\phi)\frac{d^2b(\theta)}{d\theta^2} = a(\phi)b''(\theta) \tag{3.3}$$

where $a(\phi) = \frac{\phi}{w}$, w being a 'prior weight' which varies across observations. This shows that the variance depends on both the canonical and dispersion parameters θ and ϕ. For the binomial and Poisson distributions ϕ is generally assumed to be equal to 1.

3.2.2 The Log-likelihood and Maximum Likelihood Estimation

The log-likelihood for n independent observations $\boldsymbol{y}$ can be written as

$$L(\boldsymbol{\theta},\phi;\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$.
If we let l_i be ith component for the log-likelihood (i.e. for a single observation) and $l =$

$\sum_i l_i$ then the estimating equations for the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$ are obtained by differentiating l with respect to each coefficient using the chain rule so that

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

for $j = 0, 1, \ldots, p$. Given $b'(\theta) = \mu$ and say $b''(\theta) = V$ it follows that $\frac{d\mu}{d\theta} = V$, and the derivative of $\eta_i = \sum \beta_j x_i$ with respect to β_j is x_i, therefore we get

$$\frac{\partial l}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)}\frac{1}{V}\frac{d\mu_i}{d\eta_i}x_i.$$

If the dispersion is constant then $a(\phi) = \phi$ and the above equation becomes

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V}\frac{d\mu_i}{d\eta_i}x_i = 0$$

which is the maximum-likelihood estimating equations for β_j.

The estimating equations are then solved using a form of iteratively re-weighted least squares for their solutions as they are nonlinear functions of the regression parameters (Nelder and Wedderburn, 1972). Given a trial estimate of parameters $\hat{\beta}$, we start with initial values of the estimated linear predictor $\hat{\eta_i}^0 = X_i'\hat{\beta}$ and the fitted values $\hat{\mu_i}^0 = g^{-1}(\hat{\eta_i})$ to calculate a working dependent variable:

$$z_i^0 = \hat{\eta_i}^0 + (y_i - \hat{\mu_i}^0)(\frac{d\eta_i}{d\mu_i})^0$$

with iterative weights

$$w_i^0 = \frac{1}{b''(\theta_i)(\frac{d\eta_i^0}{d\mu_i^0})^2}$$

where $b''(\theta_i)$ is the second derivative of $b(\theta_i)$. z_i^0 is then regressed on the covariates x_i with weight w_i^0 to obtain new estimates of $\hat{\boldsymbol{\beta}}^1$ from which a new $\hat{\eta_i}^1$ is formed. The weighted least-squares estimates are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wz}$$

where $\boldsymbol{X}$ is the $n \times p+1$ model matrix, W is a diagonal matrix of weights with entries w_i and $\boldsymbol{z}$ is a $n \times 1$ response vector with z_i entries. The procedure is repeated until the estimates $\hat{\boldsymbol{\beta}}$ converge to the maximum-likelihood estimates of $\boldsymbol{\beta}$. This algorithm uses the Newton-Raphson method and gives the same results as the Fisher scoring method.

3.2.3 Goodness of Fit Measures

3.2.3.1 Likelihood Ratio Criterion and the Deviance

When fitting the model to the data, we would like to know how well the fitted values $\hat{\boldsymbol{\mu}} = (\mu_1, \ldots, \mu_n)$ estimate the observed values $\boldsymbol{y}$. One goodness of fit measure is the Wald test (Bewick, Cheek and Ball, 2005). The deviance is another, which is formed from the logarithm of a ratio of likelihoods, which can be used to compare any two nested models say $\omega_1 \subset \omega_2$.

First we consider the full model, or saturated model, which has one parameter for each observation and does not consign any variation for the random component (McCullagh and Nelder, 1989). So if we have $\hat{\mu_i}$ and $\hat{\theta}_i$ which is the estimate of θ under our model of interest, and $\tilde{\theta}_i$ is the estimate of θ for the saturated model, where $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{y})$, then the likelihood ratio criterion to compare these two models (which are in the exponential family) is given by

$$-2\log\lambda = 2\sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)} \tag{3.4}$$

where $a_i(\phi) = \frac{\phi}{w_i}$. The deviance is then the numerator of (3.4) which can be re-written as

$$-2\log\lambda = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi}.$$

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)].$$

This shows that the deviance does not depend on unknown parameters i.e. it is a function of the data only, and for the Normal distribution reduces to the residual sum of squares.

Given $l(\mu; y) = \log f(y; \theta)$, the scaled deviance is

$$D^*(\boldsymbol{y}; \boldsymbol{\mu}) = 2l(\boldsymbol{y}, \boldsymbol{y}) - 2l(\boldsymbol{\mu}, \boldsymbol{y}) \tag{3.5}$$

$$= \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi} \tag{3.6}$$

where $l(\boldsymbol{y}, \boldsymbol{y})$ is the "maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data" (McCullagh and Nelder, 1989).

3.2.3.2 Pearson X^2 Statistic

Another important measure of fit is the generalized Pearson X^2 statistic given by

$$X^2 = \sum \frac{(y_i - \hat{\mu_i})^2}{var(\hat{\mu_i})}$$

where $var(\hat{\mu_i})$ is the estimated variance function for the assumed distribution of Y_i (given its from the exponential family). For the Normal distribution the Pearson X^2 also reduces to the residual sum of squares and has an exact χ^2 distribution, while for the Poisson and binomial distributions the original X^2 statistic is used.

3.3 GLMs for Binary Data

3.3.1 Binary Responses

When the data is in a form of binary responses i.e. the response variable Y_i has two possible values, for example in epidemiological studies where death status is of interest, Y_i can take the value of either 0 when an event is not observed and 1 if the event is observed. So the probabilities of 'not observed' and 'observed' can be written as $P(Y_i = 0) = 1 - \pi_i$ and $P(Y_i = 1) = \pi_i$ respectively. For an individual unit, there is a vector of explanatory variables or covariates $\boldsymbol{x_i}' = (x_1, \ldots, x_p)$ that are thought to influence the probability of a positive response, $\pi = \pi(x)$. The Bernoulli distribution may be used to model binary data when we have $n = 1$, and for $n > 1$ we may use the binomial distribution.

In this case we want to look at the relationship between the response probability π_i and

$\boldsymbol{x_i}'$, which will occur through the link function

$$g(\pi_i) = \eta_i = \sum_{j=1}^{p} x_i \beta_j; \quad \text{i=1,...,n.}$$

The different link functions that can be used for binary responses are the logit link function, probit, complementary log-log etc. The logit link function is the one that is commonly used, because of its simple interpretation as the logarithm of the odds and because it can be used on data that is sampled either prospectively or retrospectively (McCullagh and Nelder, 1989). It is also the canonical link that is directly obtained from the exponential family representation of the Bernoulli distributions.
Given we have covariates $x_1, x_2, \ldots, x_p$ and parameters $\beta_0, \beta_1, \ldots, \beta_p$, the linear logistic model can be written as

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

for the log odds of a positive response, or alternatively as the odds of a positive response

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p).$$

The inverse of the logistic function $g(\pi_) = \log(\frac{\pi}{1-\pi})$ gives the probability of a positive response which is written as

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} = \frac{e^n}{1 + e^n}$$

where $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

3.3.2 The Binomial Distribution

The binomial distribution is a discrete probability and occurs when the observations Y_i are non-negative counts or $Y_i \sim$ independent $B(n_i, \pi_i)$ where n_i are the number of trials and π_i is the probability of a 'success'. The probability density function (PDF) is given by

$$f(y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}.$$

By taking the logs of the PDF we can arrive at the general expression given by 3.1:

$$\begin{aligned}\log f(y_i) &= y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i) + \log \binom{n_i}{y_i} \\ &= y_i \log(\frac{\pi_i}{1-\pi_i}) + n_i \log(1-\pi_i) + \log \binom{n_i}{y_i} \\ &= \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\end{aligned}$$

where θ_i and $b(\theta_i)$ are a function of π_i. This means $\log(\frac{\pi_i}{1-\pi_i}) = \theta_i$. If we make π_i the function, we get

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$$

and

$$1 - \pi_i = \frac{1}{1+e^{\theta_i}}$$

so that $b(\theta_i) = n_i \log(1+e^{\theta_i})$, $a(\phi) = \phi = 1$ for the binomial distribution and $c(y_i, \phi) = \log \binom{n_i}{y_i}$.

Similarly to get the mean and variance we can use the expressions 3.3

$$E(Y_i) = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i\pi_i$$

$$var(Y_i) = a(\phi)b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i\pi_i(1-\pi_i).$$

The canonical link $\theta_i = g(\mu_i) = \eta_i$ for the binomial distribution is the logit function.

3.3.3 Estimation of the Parameters for Logistic Regression

To estimate the parameters for logistic regression we use the Fisher scoring method as described by Section 3.2.2. Given the n-vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$ and the observed value

$\boldsymbol{y}$, the log-likelihood can be written as

$$l(\boldsymbol{\pi}, \boldsymbol{y}) = \sum_{i=1}^{n} [y_i \log(\frac{\pi_i}{1-\pi_i}) + n_i \log(1-\pi_i)].$$

Since the systematic part of GLMs is the relationship between η_i and the covariates x_i, we can use the form $g(\pi_i) = \eta_i = \sum x_i \beta_j$. so that

$$g(\pi_i) = \eta_i = \log(\frac{\pi_i}{1-\pi_i}) \tag{3.7}$$

for linear logistic models. The log-likelihood can therefore be written as a function of the unknown parameters $\boldsymbol{\beta}$

$$l(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_i \sum_j y_i x_i \beta_j - \sum_i n_i \log(1 + \exp \sum_j x_i \beta_j) \tag{3.8}$$

from which the maximum likelihood estimating equations for β_j is derived as:

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{y_i - n_i \pi_i}{\pi_i (1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_i \tag{3.9}$$

$\hat{\boldsymbol{\beta}}$ can then be obtained, given the initial estimates of $\hat{\boldsymbol{\beta}^0}$, $\hat{\boldsymbol{\pi}^0}$ and $\hat{\boldsymbol{\eta}^0}$, and with the working dependent variable and iterative weights as

$$z_i = \hat{\eta_i} + \frac{y_i - n_i \hat{\pi}_i}{n_i} \frac{d\eta_i}{d\pi_i}$$

and

$$w_i = n_i \pi_i (1 - \pi_i)$$

The binomial deviance is given by

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i [y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{\mu}_i}]$$

where $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i for our model of interest.

3.4 GLMs for Count Data

3.4.1 The Poisson Model

Many outcomes, such as those in clinical medicine, epidemiology or in behavioural studies, are counts of events in a Poisson or Poisson-like process (McCullagh and Nelder, 1989; Byers et al, 2003). The basic Poisson model can be used when successive events occur at the same rate and independently of each other. The Poisson distribution is assumed to have an equal mean and variance of μ_i, and the events observed are non-negative integer values, with no finite upper limit. For log-linear models, the common choice for the canonical link is the log function where $\theta = \eta_i = \log(\mu_i)$.
The probability mass function (PMF) is given by

$$f(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad \text{with } y_i = 0, 1, 2, \ldots \tag{3.10}$$

To get the general expression of the form of (3.1), we take the logs to get

$$\log f(y_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!)$$

and can see that $\theta_i = \log(\mu_i)$ which confirms that the canonical link is the log function. If we write θ_i in terms of μ_i we have $b(\theta) = e^{\theta_i}$, $c(y_i, \phi) = -\log(y_i!)$ and again we assume that the dispersion parameter $\phi = 1$ so that $a(\phi) = 1$.
The mean and variance will then be

$$E(Y_i) = b'(\theta) = e^{\theta_i} \quad \text{and}$$
$$var(Y_i) = a(\phi)b''(\theta_i) = e^{\theta_i}$$

which verifies that the mean and variance of the Poisson distribution are equal. As $\mu_i \to \infty$ for each i, the Poisson distribution tends to the Normal distribution.

3.4.2 Estimation of the Parameters for Log-linear Models

The log-likelihood for the Poisson regression model is

$$l(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i)$$

for a vector of independent observations $\boldsymbol{y}$. The canonical link is $\eta_i = \log(\mu_i)$ for which the derivative is $\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i}$. The maximum likelihood estimating equations are the same as for the logisitic regression (equation 3.9).
The working dependent variable and iterative weights are given respectively as

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i} w_i = \mu_i$$

from which the Fisher scoring algortihm can be used to obtain the maximum likelihood estimates.
Since the log link function is used, the exponentiated coefficients e^{β_j} can be interpreted as the multiplicative effects on the expected response.

3.4.3 The Poisson Deviance

If we let $\hat{\mu_i}$ be the maximum likelihood estimate of μ_i for our model of interest ω and $\tilde{\mu_i} = y_i$ for the saturated model, then the deviance function can be written as

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = 2 \sum [y_i \log(\frac{y_i}{\hat{\mu_i}}) - (y_i - \hat{\mu_i})].$$

As $n \to \infty$, with fixed number of parameters p, the Poisson deviance has an asymptotic chi-squared distribution, which can be used as a goodness of fit test.

3.5 Overdispersion

Overdisperion occurs when there is greater variability in the data than would be expected by a regression model i.e. the variances of Y_i are greater than their expected values (Gard-

ner, Mulvey and Shaw, 1995). Dispersion is determined by the parameter ϕ, which is typically set to $\phi = 1$ for the binomial and Poisson models. In practice however, overdispersion ($\phi > 1$) is common, such as in epidemiological studies where the incidence of disease is geographically varied. For count data, overdispersion is almost always present, since the standard Poisson regression is often used to model it, and the requirement that the mean and variance are equal is hardly ever met. This means that ϕ should be estimated to determine whether overdispersion is present or absent, because failure to take it into account results in misleading inferences (Coxe, West and Aiken, 2009). Underdispersion ($\phi < 1$) can also occur in some cases when there is this less variation in the data than would be expected.

If we want to see whether overdispersion is present or not i.e. see whether $\phi = 1$ or $\phi > 1$, we estimate ϕ separately without using maximum likelihood methods such as the method of moments estimator to get

$$\hat{\phi} = \frac{1}{n-p-1} \sum \frac{(y_i - \hat{\mu_i})^2}{\hat{var(y_i)}}.$$

The estimated asymptotic covariance matrix of coefficients $\hat{\boldsymbol{\beta}}$ is given by

$$cov(\hat{\boldsymbol{\beta}}) = \hat{\phi}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$$

where W is a diagonal matrix of weights w_i.

If overdispersion is present, then alternative regression models may be used to account for them. For data that is binomially distributed we may use the beta-binomial model or quassi-binomial model, and for overdispersed count data, we can use overdispersed Poisson regression (quassi-Poisson) models, which includes the dispersion parameter ϕ in the Poisson model or the negative binomial regression model (Gardner et al, 1995; Coxe et al, 2009).

3.5.1 Beta-binomial and Negative-binomial Models

In the beta-binomial (BB) distribution, the probability of a 'success' for n trials are assumed to be random and follow the beta distribution, as opposed to the assumption that they

are fixed as is in the binomial distribution. This helps explain extra variability within the data that the binomial distribution does not. The BB distribution is used mainly in Bayesian statistics, but in the frequentist approach, it is used as an overdispersed binomial distribution.

Given $Y_i \sim$ independent $B(n_i, p_i)$ then $P(Y_i = y_i|n_i, p_i) = \binom{n_i}{y_i} p^{y_i}(1-p_i)^{n_i-y_i}$ where $0 < p_i < 1$ is a parameter that is randomly drawn from the beta distribution. The marginal PDF of BB is given by

$$\begin{aligned} f(y_i|n_i, \alpha, \beta) &= \int_0^1 P(Y_i = y_i|p_i, n_i)\pi(p_i|\alpha, \beta)dp \\ &= \binom{n_i}{y_i} \frac{1}{B(\alpha, \beta)} \int_0^1 p_i^{y_i+\alpha-1}(1-p_i)^{n_i-y_i+\beta-1}dp \\ &= \binom{n_i}{y_i} \frac{B(y_i+\alpha, n_i-y_i+\beta)}{B(\alpha, \beta)} \end{aligned}$$

for $\alpha, \beta > 0$ and where $\pi(p_i|\alpha, \beta) = Beta(\alpha, \beta) = \frac{p_i^{\alpha-1(1-p_i)^{\beta-1}}}{B(\alpha,\beta)}$ and

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

If we let the response probability be $\pi_i = \frac{\alpha}{\alpha+\beta}$, the mean and variance are then

$$E(Y_i) = n_i\pi_i$$

$$var(Y_i) = n_i\pi_i(1-\pi_i)[1+(n_i-1)\phi]$$

where $\phi = \frac{1}{\alpha+\beta+1}$ is the overdispersion parameter (Prentice, 1986).
The coefficients $\hat{\boldsymbol{\beta}}$ can be estimated in the same way as for the other GLMs, using maximum likelihood estimation.

The negative binomial (NB) model can be used as an alternative to the standard Poisson regression model, because it allows for unexplained variability between individuals, much in the same way as including an error term in normal linear regression (Coxe et al, 2009).

The NB assumes that the mean is a random variable that follows the gamma distribution with mean μ_i* and constant index ν_i (McCullagh and Nelder, 1989), therefore the NB model can be thought of as having an error function that is a mixture of the Poisson and gamma distributions.
If we have a response variable Y_i and if $Y_i \sim NB(\mu_i, \phi)$ then the probability mass function (PMF) is

$$P(Y_i = y_i; \mu_i, \phi) = \frac{\Gamma(y_i + \phi\mu_i)\phi^{\phi\mu_i}}{y_i!\Gamma(\phi\mu_i)(1+\phi)^{y_i+\phi\mu_i}}$$

for $y_i = 0, 1, 2, \ldots$. As $\phi \to 0$ then the PMF reduces to the Poisson model (Lawless, 1987). The mean and variance of the NB distribution are given by

$$E(Y_i) = \mu_i$$
$$var(Y_i) = \mu_i + \frac{\mu_i^2}{\nu_i}.$$

This shows that, given ν_i is not large, the variance increases more rapidly with the mean than in the Poisson model. For known ν_i, a GLM based on the NB distribution can be iteratively fit by weighted least squares. The Poisson log-likelihood can be used to obtain the parameter estimates, using the general Fisher scoring method.

3.5.2 Zero-inflated Models

There are situations when fewer or more values of zeros are observed in the data than would be expected in a given distribution with a specified mean and variance (Coxe et al, 2009). The problem of fewer zeros is referred to as 'truncated zeros' and many zeros is known as 'excess zeros'. In such a case, one may use zero-inflated models or hurdle models (see Ridout, Demetrio and Hinde (1998)). Lambert (1992) introduces the zero-inflated Poisson (ZIP) regression model to include covariates, something not considered previously by Johnson and Kotz (1969) who first described zero-inflated models. Zero-inflated models have been used quite a lot in the past, especially with regards to modelling count data with excess zeros, and so zero-inflated Possion (ZIP) models and zero-inflated negative binomial (ZINB) models are most commonly used. Hall (2000) looks at the ZIP and adapts it to obtain a zero-inflated binomial (ZIB) model. Zero-inflated models are thought to have two

processes (Coxe et al, 2009; Hall, 2000):

i. A 'zero state' for which the response variable is necessarily zero and

ii. a 'Poisson or binomial state' for which the response may be either a zero or a positive count.

We shall consider the ZIP and ZINB model first and then look at the ZIB model.

3.5.2.1 Zero-inflated Poisson and Negative Binomial Models

Given a response vector $\boldsymbol{Y}$ we have

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ Poisson(\mu_i) & \text{with probability } 1 - p_i \end{cases}$$

Suppose that p_i is the probability that the response Y_i for the ith individual is generated from the zero state, then a binary logsitic regression model can be used so that

$$\log(\frac{p_i}{1-p_i}) = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip}$$

where z_{ij} are the regressors for predicting those from the zero state and

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{ij}$$

where x_{ij} are the regressors from the Poisson state. $logit(p_i)$ and $\log(\mu_i)$ are the GLM canonical link functions. Given

$$p(y_i|x1, \ldots, x_n) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

for $y_i = 0, 1, 2, \ldots$ then

$$p(0) = P(Y_i = 0) = p_i + (1 - p_i)e^{-\mu_i}$$
$$p(y_i) = (1 - p_i)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

are the probabilities of observing a count of 0 and a nonzero count respectively.

The mean and variance are

$$E(Y_i) = (1 - p_i)\mu_i$$
$$var(Y_i) = (1 - p_i)\mu_i(1 + p_i\mu_i)$$

so that $var(Y_i) > E(Y_i)$ (unlike the equal mean and variance of the Poisson distribution).

The log-likelihood of the ZIP model is

$$\log L(\boldsymbol{\beta}, \gamma) = \sum_{y_i=0} \log[\exp(\boldsymbol{z_i}'\gamma) + \exp(-\exp(\boldsymbol{x_i}'\boldsymbol{\beta}))] + \sum_{y_i>0} [y_i \boldsymbol{x_i}'\boldsymbol{\beta} - \exp(\boldsymbol{x_i}'\boldsymbol{\beta})] - \sum_{i=1} \log[1 + \exp(\boldsymbol{z_i}'\boldsymbol{\gamma})] - \sum_{y_i>0} \log(y_i!)$$

where $\boldsymbol{z_i}' = (1, z_{i1}, \ldots, z_{ip})$, $\boldsymbol{x_i}' = (1, x_{i1}, \ldots, x_{ij})$, $\boldsymbol{\gamma_i} = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)^T$. The zero-inflated negative binomial (ZINB) model is written similarly to the ZIP model except now a dispersion parameter ϕ is included, so Y_i can be written as

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ NB(\mu_i, \phi) & \text{with probability } 1 - p_i \end{cases}$$

and where p_i is the probability that the response Y_i for the ith individual is generated from the zero state, and a binary logsitic regression model and log canonical link functions can be used so that

$$\log(\frac{p_i}{1 - p_i}) = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip}$$

where z_{ij} are the regressors for predicting those from the zero state and

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{ij}$$

where x_{ij} are the regressors from the Poisson state. The probabilities of observing a count of 0 and a nonzero count are

$$p(0) = P(Y_i = 0) = p_i + (1 - p_i)(1 + \phi\mu_i)^{-\frac{1}{\phi}}$$
$$p(y_i) = (1 - p_i)\frac{\Gamma(y_i + \frac{1}{\phi})}{\Gamma(y_i + 1)\Gamma(\frac{1}{\phi})}(1 + \phi\mu_i)^{y_i + \frac{1}{\phi}}$$

and the mean and variance are

$$E(Y_i) = (1 - p_i)\mu_i$$
$$var(Y_i) = (1 - p_i)\mu_i(1 + \mu_i(p_i + \phi))$$

3.5.2.2 Zero-inflated Binomial (ZIB)

If we have response variable that is a count which has an upper bound then the binomial distribution would be more appropriate than the Poisson distribution (Hall, 2000). Given Y_i we have

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ B(n_i, \pi_i) & \text{with probability } 1 - p_i \end{cases}$$

The mean and variance are given by,

$$E(Y_i) = (1 - p_i)n_i\pi_i$$

$$var(Y_i) = (1 - p_i)n_i\pi_i[1 - \pi_i(1 - p_i n_i)]$$

The canonical link functions for ZIB are both modelled via logistic functions as $logit(p_i)$ and $logit(\pi_i)$.

Lambert (1992) fits the ZIP model using maximum likelihood with the Expectation-maximization (EM) algorithm and Hall (2000) uses this method for the ZIB model as well, while the ZINB model is fit using the Newton-Raphson algorithm (Fang, 2008).

Chapter 4

Statistical Analysis of KZN AAT data

4.1 Description of Data

A study was carried out in the KwaZulu-Natal (KZN) province in South Africa to determine bovine trypanosomiasis prevalence in farms around the Hluhluwe-iMfolozi Game Park and along St. Lucia Wetlands Park, two of the major conservation and protected areas in the Umkhanyakude District (Ntantiso, 2012). The study area is situated within the 18,000 km^2 tsetse belt, where the two species of tsetse flies, *G.brevipalpis* and *G.austeni* are present (Eesterhuizen et al, 2005). The vegetation in the area is of natural bush and sand forest plantations, with average annual temperatures of 22 °C to 28 °C and rainfall of approximately 950 mm in summer and 260 mm in winter (Ntantiso et al, 2014).

4.1.1 Trypanosomiasis and Tsetse Abundance Survey

Three communal diptanks, where cattle move freely, located at various distances from the edge of the Hluhluwe-iMfolozi Game Park were surveyed. The three areas selected for surveillance were Ekhuphindisweni, Mvutshini and Ocilwane, and the cattle were surveyed regularly on a monthly basis for 15 months. The total number sampled were 1462. *T.congolense* was determined to be the dominant trypanosome species infecting the cattle and the buffaloes in the study area. Treatment against AAT is not a regular practice and mostly non-existent. AAT prevalence and the packed cell volume (PCV) were determined using buffy coat smears (Ntantiso et al, 2014). A PCV of greater than 24% is considered normal, whereas a PCV of 24% or less means the animal is anaemic (Marcotty et al, 2008).

Odour-baited H-traps were positioned at the three regions that were selected for the trypanosomiasis surveillance, and also at the Hluhluwe-iMfolozi Game Park, to determine the species (*G.brevipalpis* or *G.austeni*), the apparent density and the infection within the tsetse fly, which was done by dissecting the fly. The fly catches, with males and females, were collected every 2 weeks.

Buffaloes in the Hluhluwe-iMfolozi Game Park are considered to be reservoir hosts of the disease, that is they harbour the parasites but do not display any of the clinical symptoms exhibited by cattle. A total of 132 buffaloes were tested for trypanosomiasis as part of routine testing done for tuberculosis, and the infection rate among them was determined.

4.2 Exploratory Data Analysis

The three regions in KZN that were examined for tsetse fly populations and cattle, namely Ekuphindisweni, Mvutshini and Ocilwane, are renamed as Region 1, 2 and 3 respectively for the sake of ease. These areas surround the Hluhluwe-iMfolozi Game Park (Region 0), Regions 1 and 2 are situated just outside of the game park (3.143 km^2 and 3.172 km^2 respectively) and Region 3 a bit further away (4.744 km^2). Given their location and the tsetse fly population abundance, Regions 1 and 2 are considered high challenge areas, where there is a higher transmission rate and greater infectivity amongst the cattle, resulting in more cases of AAT. Region 3 is considered a low challenge area, with fewer cases of disease prevalence and a low tsetse fly population abundance. The winter months are considered to be from May to September (labelled 5, 6, 7, 8 and 9) and the summer months are from October to April (labelled 10, 11, 12, 1, 2, 3 and 4).

The herd average prevalence (HAP) is the average disease prevalence in a given region, and herd average PCV (HA-PCV) is the average PCV of each region. A higher HA-PCV means lower 'average' anaemia and therefore better general health of cattle in that region, and a lower HA-PCV means higher average anaemia in the region, representing a more sickly herd. Conversely, a higher HAP means a higher prevalence of disease in the region and a lower HAP means a lower prevalence of disease in a region. Regions 1 and 2 have lower HA-PCV's (25.9 and 26.97 respectively) than Region 3 (29.35), showing less healthy cattle. The HAP for Regions 1 and 2 are higher than for Region 3, with HAPs of 8.17%, 10.34%

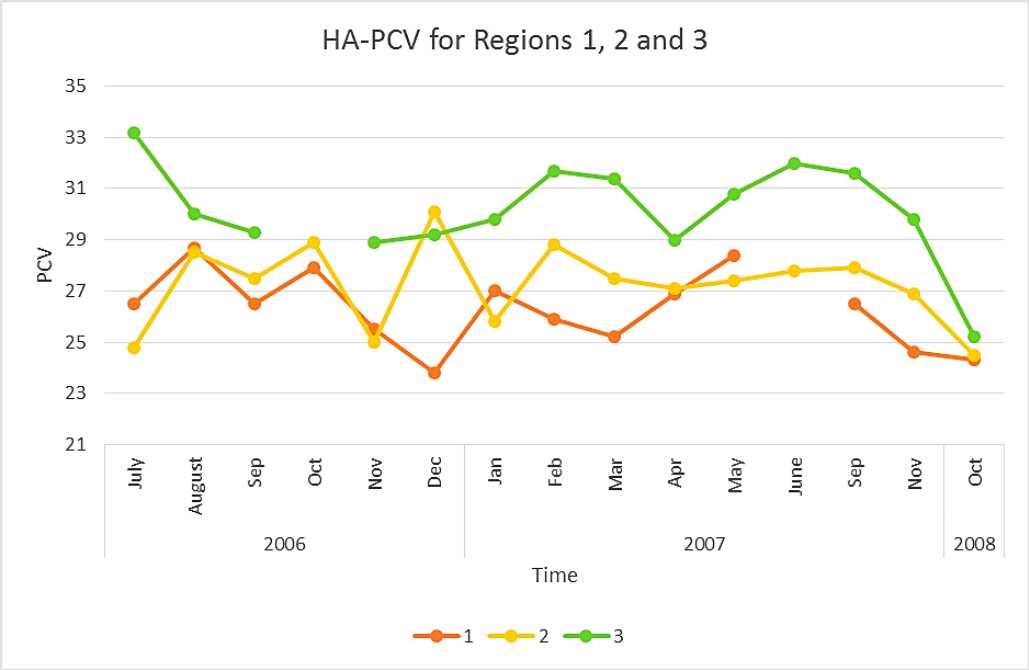


Figure 4.1: The herd average PCV for the three regions

and 1.84% respectively. This can be seen graphically in Figures 4.1 and 4.2, with Regions 1 and 2 having lower HA-PCVs and higher HAPs, whilst Region 3 has a higher HA-PCV and lower HAP. The average number of tsetse flies (*G.brevipalpis* and *G.austeni*) for those regions and Region 0 (the game park), show higher densities of flies in the game park (Region 0) and Regions 1 and 2, whilst the number of flies in Region 3 were significantly lower (Figure 4.3). The apparent density population (number of flies caught per trap)for *G.brevipalpis* was substantially higher, compared to the apparent density population of *G.austeni* (Tables 4.2 and 4.3), as shown for Regions 1 and 3, where no *G.austeni* were caught. This can be because the population of *G.austeni* is quite small or because the traps do not attract and catch the species (Ntantiso et al, 2014).

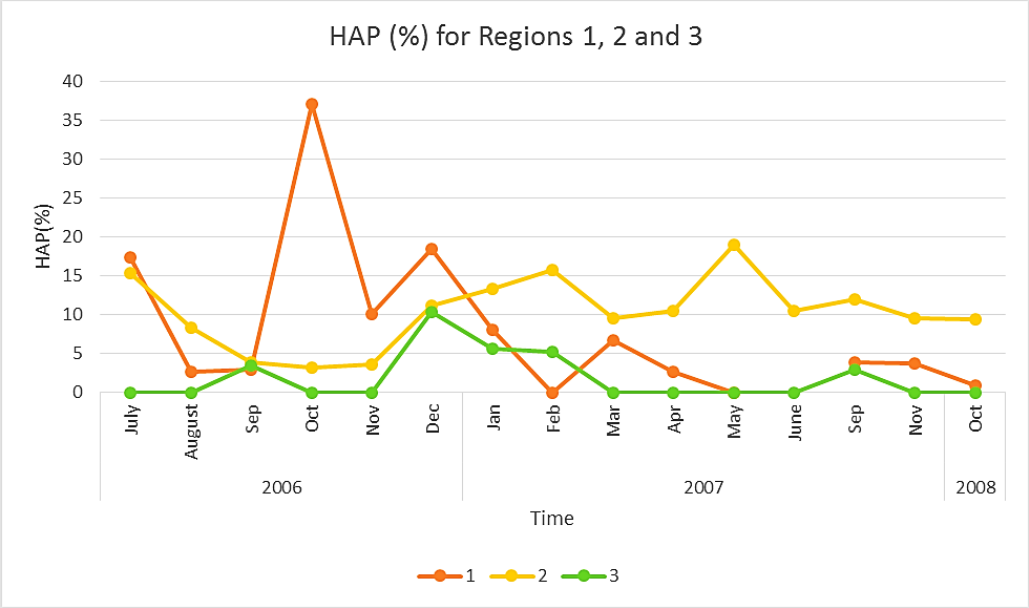


Figure 4.2: The herd average prevalence (HAP %) for the three regions

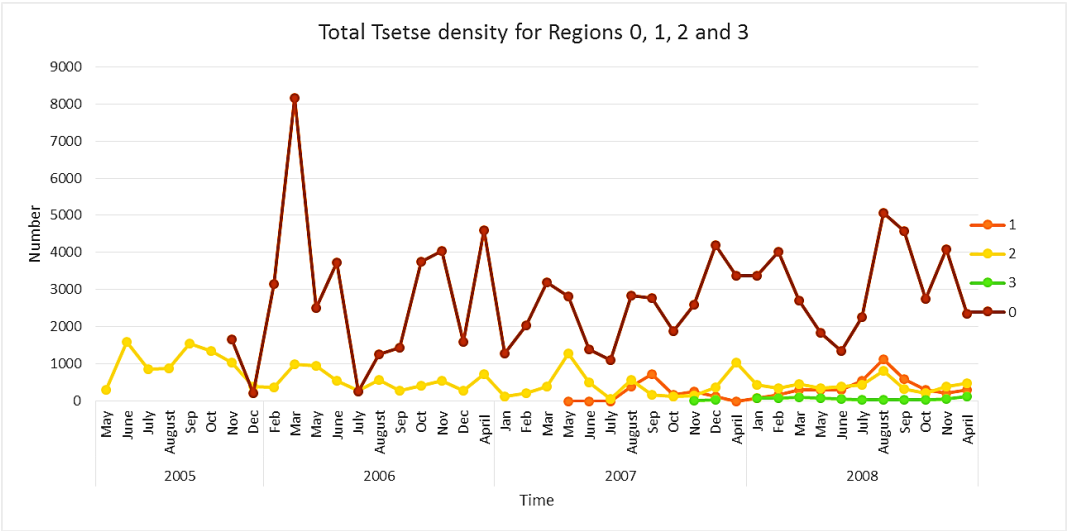


Figure 4.3: Total tseste density for the three regions and the game park (Region 0)

Table 4.1: Basic statistical analysis of PCV for the three regions

Summary Statistics of PCV					
Region	**N**	**Mean**	**Standard Deviation**	**Min**	**Max**
1	568	25.9	4.198	10	47
2	428	26.97	4.578	11	44
3	414	29.35	5.582	10	44

Table 4.2: Basic statistical analysis of *G.austeni* for the three regions and the game park

Summary statistics for *G.austeni*					
Region	**N**	**Mean**	**Standard Deviation**	**Min**	**Max**
0	36	8.53	10.377	0	43
1	20	-	-	-	-
2	40	2.23	3.15	0	13
3	13	-	-	-	-

4.3 Binomial Generalized Linear Model for AAT Prevalence Data

Since the prevalence of AAT is either 0 (negative) if *T.congolense* was absent or 1 (positive) if *T.congolense* was present, we use the binomial GLM with a logit link, and the PCV values, region and month used as the explanatory variables. The PCV values were tranformed to 0 for values > 24 and 1 for values ≤ 24, meaning a cow is non-anaemic and anaemic respectively.

The GENMOD procedure in SAS version 9.3 was used to fit the model. The parameter vector $\boldsymbol{\beta}$ is fitted using maximum likelihood estimation, and the scale/dispersion parameter is fit using either maximum likelihood, the residual deviance or Pearson's χ^2 divided by the degrees of freedom (Gordon, 2014).

The goodness of fit statistics for the binomial GLM are shown in Table 4.5. We can see that the deviance and Pearson Chi-square, divided by their degrees of freedom are 1.22 and 1.43 which are close to 1 and so we can say that this model fits the data reasonably well. Large values indicate model misspecification or overdispersion, whereas small values (< 1) also represent model misspecification and underdispersion. The model for our data is given as

Table 4.3: Basic statistical analysis of *G.brevipalpis* for the three regions and the game park

Summary statistics for *G.brevipalpis*					
Region	**N**	**Mean**	**Standard Deviation**	**Min**	**Max**
0	36	2775.5	1527.47	204	8129
1	16	367.69	265.655	71	1121
2	42	558.07	387.955	54	1603
3	13	58.54	30.953	14	120

Table 4.4: Table showing the class level information used in the binomial GLM for the AAT prevalence data

Class Level Information		
Class	**Levels**	**Values**
Month	12	1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12
Region	3	1; 2; 3

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \mu + pcv + month_i + region_j \tag{4.1}$$

for $i = 1, 2, \ldots, 12$ and $j = 1, 2, 3$ and where μ is the intercept term. So we are modelling the response, which is the prevalence of AAT infection in a cow, as a function of whether the cow is anaemic or non-anaemic (shown by PCV), where the cow is located (Region 1, 2 or 3) and the month it was sampled. Table 4.6 shows the parameter estimates and the Wald χ^2 values with the corresponding confidence intervals and p-values. The reference category for month is 9 and the reference category for region is 3. We can see that PCV is highly significant (p-value < 0.0001), and similarly, so is December (p-value of 0.0189) and Regions 1 and 2 (with p-values of 0.0121 and 0.0001 respectively) at a 5% level of significance. The odds ratio of PCV is $\exp(-1.1376) = 0.3206$ so the ratio of the odds of an animal not having the disease between an animal with a PCV of 1 and a PCV of 0 (i.e. of being anaemic and not being anaemic) is 0.3206, or the odds of an animal that is not anaemic is 0.3206 times the odds of an animal that is anaemic. Similarly, the odds of having the disease for an animal that is anaemic is $\frac{1}{0.3206} = 3.119$ times the odds of having the disease for animal that is not anaemic. A cow in Regions 1 and 2 has the odds of disease prevalence of 0.3035 and 0.1636 (respectively) times that of one in Region 3. Month 12 has an odds of disease prevalence of 0.2795 times that of one for Month 9. The scale parameter

Table 4.5: Table of goodness of fit measures for the binomial GLM

Criteria For Assessing Goodness of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	52	63.5018	1.2212
Scaled Deviance	52	52	1
Pearson χ^2	52	74.3609	1.43
Scaled Pearson χ^2	52	60.8922	1.171
Log Likelihood		-64.9272	
AIC		159.8543	

Table 4.6: SAS Proc GENMOD results for the binomial model for AAT Prevalence in KZN (significant parameters marked with *)

Analysis of Maximum Likelihood Parameter Estimates					
Parameter	**DF**	**Estimate**	**Std Error**	**Wald 95% CI**	**Pr >ChiSq**
Intercept	1	4.5198	0.5639	3.4147, 5.6250	$< .0001$
PCV*	1	-1.1376	0.2592	-1.6456, -0.6296	$< .0001$
Month 4	1	0.3266	0.7539	-1.1510, 1.8043	0.6648
Month 8	1	0.3807	0.7498	-1.0890, 1.8503	0.6117
Month 12*	1	-1.2748	0.5430	-2.3390, -0.2107	0.0189
Month 2	1	-0.0087	0.7622	-1.5025, 1.4852	0.9909
Month 1	1	-0.7746	0.6136	-1.9771, 0.4280	0.2068
Month 7	1	-0.7963	0.4792	-1.7354, 0.1429	0.0966
Month 6	1	-0.0734	0.8991	-1.8357, 1.6888	0.9349
Month 3	1	-0.1404	0.6358	-1.3865, 1.1057	0.8252
Month 5	1	-0.4367	0.5522	-1.5190, 0.6456	0.4290
Month 11	1	0.5380	0.58930	-0.6170, 1.6930	0.3613
Month 10	1	0.1520	0.4604	-0.7503, 1.0542	0.7413
Month 9 (ref)	-	-	-	-	-
Region 1*	1	-1.1924	0.4750	-2.1233, -0.2614	0.0121
Region 2*	1	-1.8101	0.4679	-2.7270, -0.8931	0.0001
Region 3 (ref)	-	-	-	-	-
Scale		1.1051		[1.1051, 1.1051]	-

(ϕ) was estimated as 1.1051, which shows slight overdispersion, which could be due to the differences in the regions and months.

Table 4.7 shows the overall significance of the variables used in the model. PCV is highly significant (p-value < 0.0001), and so is region (p-value $= 0.0003$), while month is not (p-

Table 4.7: Type 3 analysis of main effects of binomial GLM

Wald Statistics for Type 3 Analysis			
Source	**DF**	**Chi-square**	**Pr >ChiSq**
PCV	1	19.27	$< .0001$
Month	11	18.37	0.0735
Region	2	16.43	0.0003

value= 0.0735). This shows that prevalence of AAT is not necessarily seasonal, which could be expected as the parasites can be present in the blood, but the animal may not get sick until times of stress such as when it has poor nutrition, after which the infection can appear. For seasonality of AAT to be known cattle need to sampled and treated monthly in order to determine the incidence of AAT i.e. to look at the new infections appearing. Since region is significant, we can say that there is a relationship between where the cattle are located and whether they have the disease or not. Regions 1 and 2 were significant, so proximity to the game park could play a role, which is in accordance with literature that reservoir hosts play a big role in the epidemiology of AAT. Since PCV is considered an indicator of whether a cow is sick or not, it being a highly significant variable is not surprising.

4.4 Poisson Generalized Linear Model for Tsetse Fly Density Data

The tsetse fly abundance can be modelled using the Poisson regression model with a log link for both species. This is also done using Proc GENMOD in SAS. The explanatory variables used were month and year and the deviance scale parameter (ϕ) is estimated to assess whether overdispersion is present or not. Table 4.8 shows that the data for *G.austeni* is overdispersed when we use the Poisson GLM. The estimated scale is given as $\phi = 2.7049$ which shows significant overdispersion, and since we know that the Poisson assumption of an equal variance and mean is violated (we have a variance of 14.316 compared to a mean of 3.63), we look at using the negative binomial (NB) model to account for the overdispersion. For *G.austeni*, the zero-inflated Poisson and negative binomial (ZIP and ZINB) models can also be used because we have excess zeroes in the data. The 4 models are fit (Poisson, NB, ZIP and ZINB) to determine which model fits the best. For *G.brevipalpis*, only Poisson

and the NB model are fit, since we have significant overdispersion of $\phi = 29.53$ (Table 4.9).

Table 4.8: Poisson model for *G.austeni* with the estimated scale parameter

Analysis of Maximum Likelihood Parameter Estimates				
Parameter	**DF**	**Estimate**	**Wald 95% CI**	**Pr >ChiSq**
Scale	0	2.7049	[2.7049, 2.7049]	-

Table 4.9: Poisson model for *G.brevipalpis* with the estimated scale parameter

Analysis of Maximum Likelihood Parameter Estimates				
Parameter	**DF**	**Estimate**	**Wald 95% CI**	**Pr >ChiSq**
Scale	0	29.5283	[29.5283, 29.5283]	-

The goodness of fit criteria for the 4 models are compared to see which has the best fit (Table 4.10). For the Poisson GLM, we see that the deviance and the Pearson Chi-square (divided by their degrees of freedom) are 7.32 and 6.98 respectively, which are bigger than 1, compared to the NB model which has values of 1.64 and 1.20 which are closer to 1, so it has a better fit. For the ZIP and ZINB, we have Pearson Chi-squares of 3.57 and 1.31 respectively, so the ZIP is a better fit than the Poisson but is still greater than 1 so the model is misspecified whilst the ZINB model is closer to the value of 1 but a bit bigger than the value for the NB model. When we look at the AIC's, we see that the values are 375, 274, 337 and 272 for the Poisson, NB, ZIP and ZINB models, again showing that the NB and ZINB are the better fits. Looking at these criterion, the ZINB regression model is the best fit and we therefore use it to model our data (Table 4.11).

For *G.brevipalpis* the Poisson and NB model were compared in Table 4.13 and the NB model was fit (Table 4.14). The deviance/DF and the Pearson Chi-square/DF were 872 and 892 with a very high AIC of 25717.86 for the Poisson GLM, which were greatly reduced when the NB model was fit (1.5, 1.38 and an AIC of 744), so the model is said to have a good fit and not overdispersed.

From Table 4.11 we can see that the months that are significantly different from reference month 12, are months 7 and 8 for *G.austeni* and year was also significant (p-value= 0.0009). The regression coefficients for the months 7 and 8 are -1.5675 and -1.1399 respectively, so the exponentiation of those values are $e^(-1.5675) = 0.2086$ and $e^(-1.1399) = 0.3199$. This means that the multiplicative differences in the average number of flies for those months

Table 4.10: Goodness of fit Criteria of the Poisson, negative binomial, ZIP and ZINB for *G.austeni*

Criteria for Assessing Goodness of Fit				
Model	**Deviance/DF**	**Pearson χ^2/DF**	**Log Likelihood**	**AIC**
Poisson	7.3167	6.9761	88.1134	374.9772
Negative Binomial	1.6442	1.2046	423.4176	274.0185
ZIP	-	3.5748	664.4937	337.3808
ZINB	-	1.3092	-121.1595	272.3189

Table 4.11: SAS Proc GENMOD results of the Zero-inflated negative binomial model for *G.austeni* (significant parameters marked with *)

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimates**	**Std Error**	**Wald 95% CI**	**Pr >ChiSq**
Intercept	1	1215.864	287.8049	651.7773, 1779.952	< .0001
Month 1	1	0.4260	0.6325	-0.8137, 1.6658	0.5006
Month 2	1	0.4252	0.5590	-0.6704, 1.5208	0.4468
Month 3	1	0.9157	0.5329	-0.1287, 1.9601	0.0857
Month 4	1	0.7685	0.5315	-0.2731, 1.8102	0.1482
Month 5	1	-0.2377	0.6052	-1.4238, 0.9484	0.6945
Month 6	1	-0.4855	0.5895	-1.6409, 0.6698	0.4102
Month 7*	1	-1.5675	0.6025	-2.7484, -0.3866	0.0093
Month 8*	1	-1.1399	0.5806	-2.2778, -0.0020	0.0496
Month 9	1	-0.1461	0.5845	-1.2916, 0.9995	0.8027
Month 10	1	-0.6840	0.5547	-1.7712, 0.4032	0.2175
Month 11	1	0.4812	0.5020	-0.5028, 1.4651	0.3378
Month 12 (ref)	-	-	-	-	-
Year*	1	-0.6049	0.1435	-0.8860, -0.3237	< .0001

Table 4.12: Type 3 analysis of main effects of ZINB model for *G.austeni*

Wald Statistics for Type 3 Analysis			
Source	**DF**	**Chi-square**	**Pr >ChiSq**
Month	11	25.74	0.0071
Year	1	13.33	0.0003

Table 4.13: Goodness of fit Criteria of the Poisson and negative binomial for *G.brevipalpis*

Criteria for Assessing Goodness of Fit			
Model	**Deviance/DF**	**Pearson χ^2/DF**	**AIC**
Poisson	871.9206	891.6470	25717.8684
Negative Binomial	1.5010	1.3764	744.2998

Table 4.14: SAS Proc GENMOD results of the negative binomial model for *G.brevipalpis* (significant parameters marked with *)

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimates**	**Std Error**	**Wald 95% CI**	**Pr >ChiSq**
Intercept	1	-611.362	186.2405	-976.387, -246.338	0.0010
Month 1	1	-0.2591	0.5351	-1.3079, 0.7897	0.6282
Month 2	1	0.2157	0.4691	-0.7037, 1.1350	0.6456
Month 3	1	0.7971	0.4673	-0.1187, 1.7130	0.0880
Month 4	1	0.5119	0.4678	-0.4049, 1.4287	0.2738
Month 5	1	0.0045	0.4377	-0.8533, 0.8623	0.9918
Month 6	1	0.0915	0.4379	-0.7668, 0.9497	0.8345
Month 7	1	-0.6383	0.4383	-1.4974, 0.2207	0.1453
Month 8	1	0.1562	0.4399	-0.7059, 1.0184	0.7225
Month 9	1	0.1447	0.4378	-0.7134, 1.0028	0.7411
Month 10	1	0.1333	0.4372	-0.7237, 0.9903	0.7605
Month 11	1	0.4545	0.4374	-0.4027, 1.3117	0.2987
Month 12 (ref)	-	-	-	-	-
Year*	1	0.3086	0.0928	0.1266, 0.4905	0.0009

are 0.2086 and 0.3199 times less than the reference month 12. This can be attributed to the fact that month 7 has the lowest average temperature range ($10.64\,^{\circ}\text{C} - 24.01\,^{\circ}\text{C}$) (Figure 4.4), and *Glossina morsitans* are known to be active at temperatures of $18\,^{\circ}\text{C} - 32\,^{\circ}\text{C}$ (Fraumann, 2003) and at low temperatures the pupal period (or stage) can be more than 50 days and the rate of reproduction decreases (Pollock, 1982). To see whether months and years are significant to the models, we use the Wald Type 3 analysis in Tables 4.12 and 4.15. We can see that both month and year is significant for *G.austeni*, whereas only year is significant for *G.brevipalpis*.

Table 4.15: Type 3 analysis of main effects of NB model for *G.brevipalpis*

Wald Statistics for Type 3 Analysis			
Source	**DF**	**Chi-square**	**Pr >ChiSq**
Month	11/29	11.81	0.3781
Year	1/29	9.36	0.0022

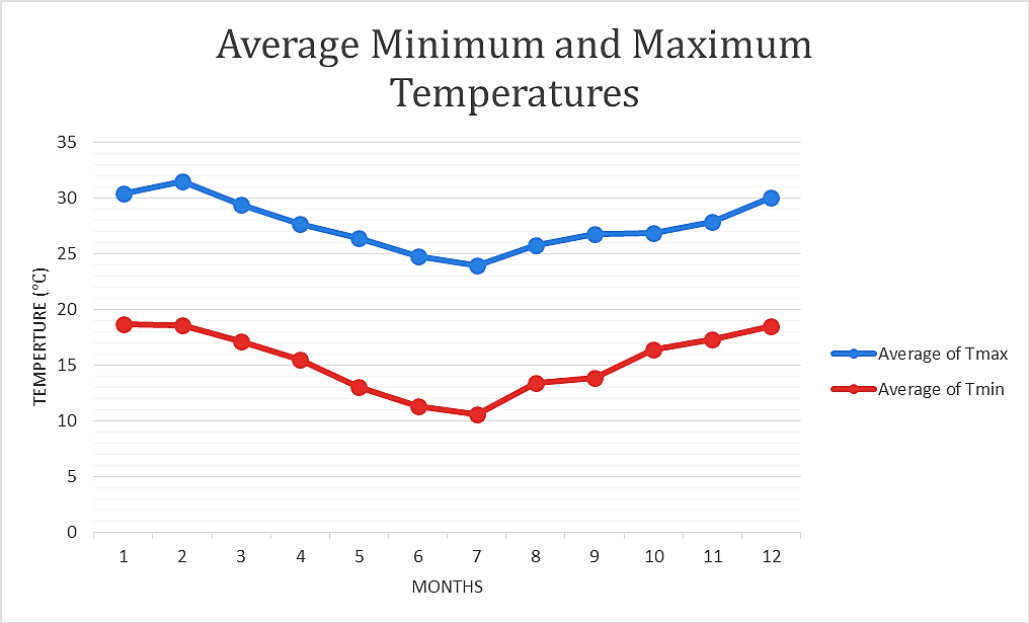


Figure 4.4: Average minimum and maximum temperatures for Bushlands station (longitude and latitude: -28.13938, 32.2949)

4.5 Summary

The binomial GLM was used to model AAT prevalence data for three regions in the KwaZulu-Natal Province. The explanatory variables that were used were PCV, which is considered to be an indicator of AAT in trypanosome endemic areas, month and region (where the cattle are located to the Hluhluwe-iMfolozi Game Park). PCV was found to be highly statistically significant at a 5% level of significance, with a p-value of $< .0001$. Regions 1 and 2, which are the closest to the game park, were also found to be statistically significant, so where the cattle are located seems to be important. Month was not found to be statistically significant, and since this was a study of prevalence and not incidence of AAT, the parasites could be present and resurface at times when the cow is under stress. The scaling parameter ϕ was estimated to determine whether there was overdispersion in the model and ϕ was found to be 1.1051, which shows slight overdispersion and this could be as a result of the heterogeneity in the regions.

To model the tsetse fly abundance of both species, the Poisson GLM was used with month

and year as the explanatory variables, and the scaling parameter was estimated to determine whether they were overdispersed. Both the Poisson models for both species had high values of ϕ and so alternative count models were fit. *G.austeni* has a high frequency of zeros and lower counts, due to either a low population density or because the traps are not efficient in catching them, while *G.brevipalpis* has a substantially higher abundance. For these reasons the negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial models (ZINB) were fit for *G.austeni* while only the NB model was fit for *G.brevipalpis*. The ZINB regression model had the best fit for *G.austeni*, with both months 7 and 8 and year found to be statistically significant. This could be because month 7 has the lowest minimum and maximum temperatures during the year and at lower temperatures, tsetse flies become less active and the pupal stage lengthens to around 50 days and the reproductive rate decreases. For *G.brevipalpis* the NB regression model was a much better fit than the Poisson regression model. Only year was found to be statistically significant in this model.

Chapter 5

Model fitting and Climate Change

5.1 Data Fit to the Mathematical Model

Many assumptions go into constructing a disease model, and initial parameter values are usually derived from various literature sources, but there is a need to obtain data that will help construct, verify and validate the disease model (Carpenter, 1994). These models need to be developed with both the biological systems in mind and the data that is collected through the epidemiological studies. Statistical analyses are used to identify relevant risk factors and help to determine the time to a particular event example incubation, latent and infectious periods (Carpenter, 1994). In estimating the model parameters, the model needs to represent the system we are studying so that it can be used to both inform and predict so that they can be used to develop control strategies (Dransfield and Brightwell, 1989).

As a starting point to judge how the model fits our actual data, we fit it to the KZN AAT prevalence data using least squares and the parameters that were used in the model were estimated from that fit. Figure 5.1 shows the real data for 16 months and the model fit. The model fit seems to increase sharply from month 16, which might not be a true representation of what happened in those months. More data points need to be added to see whether it follows the model fit. The estimated parameters from the fit are in Table 5.1, which also shows the values from the literature. The values that do not match the literature values are μ_C, α and which are too high to make sense. The parameters that apply to the tsetse fly i.e. μ_F and β are within the range, the mortality rates of tsetse flies vary depending on the mean temperature (Rogers, Randolph and Kuzoe, 1984) and the biting rate is 2 or 3 days and up to 7 days in cooler seasons (Rogers, Hendrickx and

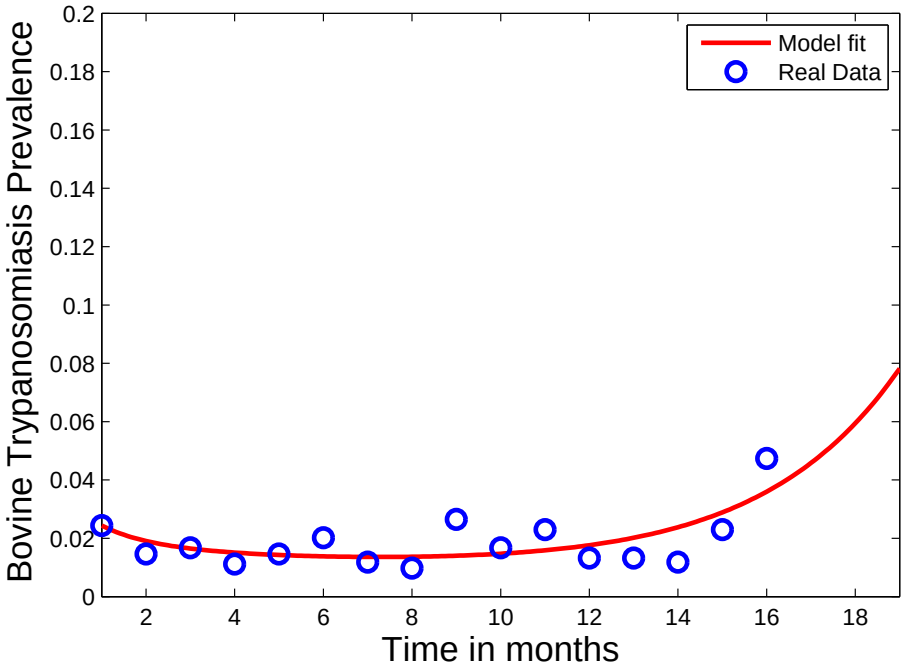


Figure 5.1: The data was fit to the mathematical model using least squares using the initial parameter values from the literature sources

Table 5.1: The parameters used in the mathematical model with the initial values (sources given in Chapter 2) and the values estimated from the model fit to the data.

Parameter	**Description**	**Value from Literature**	**Estimated Value**
μ_C	Mortality rate of cattle (excluding disease-induced) $days^{-1}$	0.0005	0.2875
μ_B	Mortality rate of buffalo $days^{-1}$	0.0001053	0.0001589
μ_F	Mortality rate of tsetse fly $days^{-1}$	0.030	0.0240
α	Mortality rate of cattle due to disease $days^{-1}$	0.002	0.9785
β	Biting rate $days^{-1}$	0.25	0.1378
η	Weight (distance from buffalo)	1.3	2.9746
$\hat{f}$	Probability of infected fly producing infection in cow	0.2	0.0544
$\hat{f}_1$	Probability of infected fly producing infection in buffalo	0.46	0.5597
$\hat{f}_2$	Probability of infected blood meal from host producing infection in fly	0.025	0.0814

Slingenbergh, 1994). The values for $\hat{f}$ and $\hat{f}_1$ seem to be reasonable but need to be verified. The interesting parameter value is that of $\hat{f}_2$ of 0.0814, since the infection with mature parasites in *G. austeni* in KZN was found to be 8% (Ntantiso, 2012). The estimated values can be used to calculate R_0 from the NGM $\boldsymbol{K}$, so we have

$$\boldsymbol{K} = \begin{bmatrix} 0 & 0 & R_{CF} \\ 0 & 0 & R_{BF} \\ R_{FC} & R_{FB} & 0 \end{bmatrix}$$

$$\boldsymbol{K} = \begin{bmatrix} 0 & 0 & 0.9316 \\ 0 & 0 & 408.1286 \\ 0.0089 & 0.5558 & 0 \end{bmatrix}$$

from which we get $R_0 = 15.062 > 1$. For $R_{FC} = 0.0089$ we expect to get a low number because of the high values of μ_C and α, but for R_{CF} we have a value of 0.9316 which is

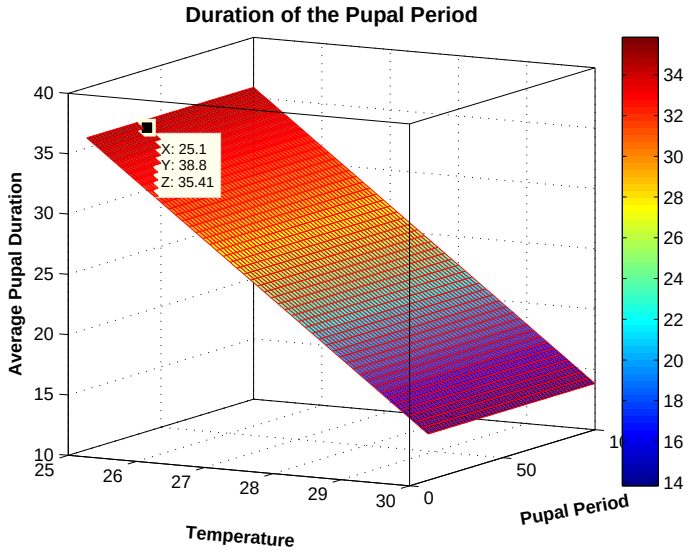
less than one, so if the transmission is from cattle to the fly then AAT will die out. We also expect R_{FB} (transmission from fly to buffalo) to be low since the buffalo is a reservoir host and is not affected by trypanosomiasis. Our important element is $R_{BF} = 408.1286$, which contributes the most to R_0 and so the transmission from the buffalo to the fly is significant in maintaining transmission of AAT in KZN.

5.2 Climate Change and Tsetse Population Dynamics

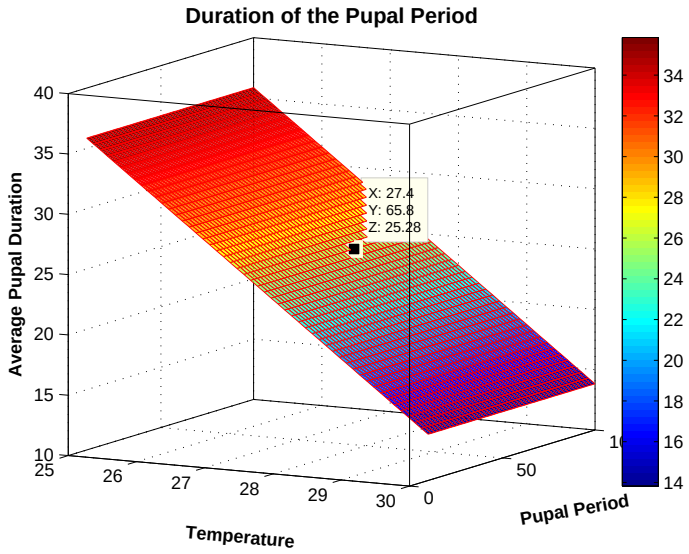
As a result of climate change, global mean surface temperature is likely to increase $0.3\,^{\circ}\text{C} - 1.7\,^{\circ}\text{C}$ by the end of the 21st century (IPCC, 2014). Increases in temperature affect tsetse fly population dynamics such as rates of larval production, pupal development, mortality rates etc. (Hargrove, 2004). These changes in the dynamics may mean that the geographical distribution of the tsetse flies may be altered in a way that will expose naive animals and humans to infection (Moore et al, 2012), and result in an increased African trypanosomiasis burden on these vulnerable populations (TDR, 2012).

High temperatures affect the rates of larval production, shorten the pupal period, increase mortality rates in pupa and adults (Pollock, 1992; Hargrove, 2004). Those factors affecting the birth rates are easier to predict than the ones affecting mortality rates (Rogers et al, 1994; Hargrove, 2004), so at this point we will look at one the factors affecting birth rates i.e. the pupal period. Artzrouni and Gouteux (2006) estimate the average duration of the pupal period as a function of temperature. Surface plots were done to show how for different temperatures (X), the pupal period (Y) changes, resulting in an average pupal duration (Z). From the surface plots (Figures 5.2a-5.2c), we see that there is a linear relationship between temperature and pupal period and the average pupal duration. Different points on the surface plot were chosen to illustrate our point. For the temperature range $25\,^{\circ}\text{C} - 30\,^{\circ}\text{C}$, we see that the average duration of the pupal range is 35 days for $25\,^{\circ}\text{C}$ (Figure 5.2a), then goes down to 27 days (Figure 5.2b) and for $30\,^{\circ}\text{C}$ it is about 14 days (Figure 5.2c). So for higher temperatures, the average pupal duration is reduced and if the temperature becomes too hot, the pupa will die.
To predict how the tsetse population will change in relation to climate change, the different factors need to be estimated and considered separately (Hargrove, 2004), so that appropriate control measures can be put in place, for as Rogers (1994) stated: "the study of

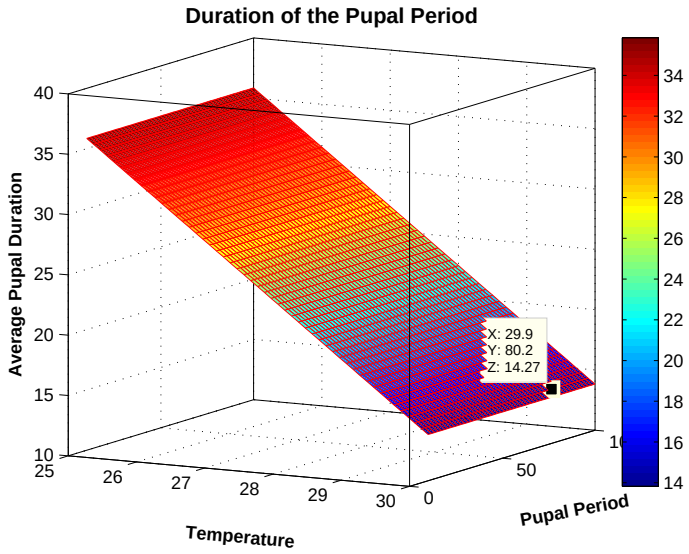
vector populations dynamics therefore forms a vital part of vector-borne disease epidemiology".



(a) Temperature (X) 25 °C, average pupal duration (Z) 35.41 days



(b) Temperature (X) 27 °C, average pupal duration (Z) 25.28 days



(c) Temperature (X) 29.9 °C, average pupal duration (Z) 14.27 days

Figure 5.2: Surface plots showing how different temperatures affect the average pupal duration.

Chapter 6

Discussions and Conclusions

6.1 Discussion and Conclusions

African animal trypanosomiasis, restricted to parts of the KwaZulu-Natal Province, is a disease which contributes significantly to the disease burden of cattle, and affects resource-poor farmers especially whom do not have access to treatment. Drug resistance is also a problem and dipping of cattle using insecticides has proved to be unsustainable. Even though the incidence of AAT has increased, little is known about the epidemiology of the disease in the region. To better understand the dynamics of AAT, mathematical modelling was done to investigate the interactions between the cattle, tsetse fly and buffalo which is considered to be the reservoir host, and a statistical analysis of the data collected from three sites around the Hluhluwe-iMfolozi Game Park was done to assess the relationships between the variables.

In Chapter 2, an SI model was constructed for the different classes of the population i.e. susceptible and infected cattle and tsetse flies and infected buffaloes. The basic reproduction number R_0 was derived using the next-generation matrix to assess how the different elements contribute to R_0 and although it was found that cattle and flies are able to maintain the transmission of AAT, the important element was the transmission from buffaloes to flies which contributed the most to R_0. This was confirmed when we used the model to see how the different classes vary over time, and when we considered transmission from just the buffalo population, infected cattle and flies increased a lot more steeply than when the transmission was just between the cattle and flies.
A sensitivity analysis was done, and contour plots were plotted, to identify the parameters which affect the prediction imprecision of the outcome variables using partial rank corre-

lation coefficients (PRCC). It was found that the mortality rate of tsetse flies and their biting rate were the most important parameters, so by getting more accurate measurements we can reduce the prediction imprecision of our outcome variables such as infected cattle and R_0. PRCCs can show which parameters to target if we want to look at intervention measures and determine how to efficiently reduce AAT.

In Chapter 4, generalized linear models (GLMs) were used to analyse the prevalence and tsetse abundance data collected from the three regions, since we had binary and count data. The prevalence data was modelled using a binomial GLM and the packed cell volume (PCV) and region were significant, so PCV is a useful indicator of prevalence and where the cattle are located also determines prevalence. The data was found to be slightly overdispersed however, which could be because of the cluster effect of 'region'. The tsetse abundance data was modelled using Poisson GLMs but since there was significant overdispersion present, alternative models were considered. Since there were excess zeroes for *G. austeni*, zero-inflated models were done and the best fit was found to be the zero-inflated negative binomial, while the negative binomial model was used for *G. brevipalpis* to account for the overdispersion. Accounting for overdispersion is important, because if we don't we could get misleading inferences about the data. Month and year were found to be significant for *G. austeni*, but only year was significant for *G. brevipalpis*. Since *G. austeni* is considered the main vector of AAT, seasonal changes in the population need to be looked at in relation to the incidence of AAT in the cattle to determine the pattern of the disease.

The AAT prevalence data was fit to the mathematical model using least squares in Chapter 5, and the input parameters were estimated and used to calculate R_0 again so that it is more site-specific. From the next-generation matrix we could see that the transmission from the buffaloes to the flies was again the important element and this time transmission between cattle and flies did not contribute to R_0, since their values were less than one.
Since climate change is predicted to affect the geographical distribution of tsetse flies, the relationship between temperature and tsetse fly abundance was looked at, specifically the duration of the pupal period, since it is one of the factors that affects birth rates. For higher temperatures the duration was shorter than for cooler temperatures. This could have a big impact on the AAT situation because tsetse flies might modify their behaviour and shift their geographical range to regions that are cooler, which might put cattle populations in other regions at risk of AAT outbreaks.

In conclusion, it was found that the buffalo is an important factor in maintaining AAT transmission in the sites around the game park, especially when using the parameters that were estimated from the model fit as it was more site-specific. It was also found that if interventions are to be put in place, looking at the parameters that affect the tsetse fly are critical, and seasonal patterns in relation to the incidence of AAT need to be assessed. It is important to note that for model predictions to be strengthened and to reduce uncertainty in our outputs, better estimates of the parameters are needed by collecting additional data so that the model performs more realistically. This initial model and statistical analysis shows however that if control measures are put in place, they should consider both the reservoir hosts and the tsetse flies, especially if the implications of climate change are taken into account, which would lead to naive cattle populations being exposed to AAT.

6.1.1 Future Work

This study can be extended to look at how interventions would affect R_0 and include climate-dependent parameters to the tsetse fly dynamics. To get a more clear picture of the prevalence of AAT in the different regions, those near and further away from the game parks, statistical analysis needs to be done on additional data from other regions. Spatial and spatio-temporal statistical models can also be used in an extension of the analysis.

Bibliography

Artzrouni, M. and Gouteux, J. P. (2006). A parity-structured matrix model for tsetse populations. *Mathematical Biosciences* **204**, 215-231. (doi:10.1016/j.mbs.2006.08.022).

Bewick, V., Cheek, L. and Ball, J. (2005). Statistics review 14: logistic regression. *Critical care* **9**, 112-118. (doi:10.1186/cc3045).

Blower, S. M. and Dowlatabadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *International Statistical Review* **62**, 229-243.

Brightwell, R., Dransfield, R. D. and Williams, B. G. (1992). Factors affecting seasonal dispersal of the tsetse flies Glossina pallidipes and G. longipennis (Diptera: Glossinidae) at Nguruman, south-west Kenya. *Bulletin of Entomological Research* **82**, 167-182.

Byers, A. L., Allore, H., Gill, T. M. and Peduzzi, P. N. (2003). Application of negative binomial modelling for discrete outcomes, a case study in aging research. *Journal of Clinical Epidemiology* **56**, 559-564. (doi:10.1016/S0895-4356(03)00028-3).

Carpenter, T. E. (1994). Collection, collation, analysis and dissemination of data on vector-borne and other parasitic diseases. In *Modelling Vector-borne and other Parasitic Diseases* (eds B. D. Perry and J. W. Hansen), pp. 249-260. Nairobi: ILRAD.

Clair, M. (1988). The epidemiology of African animal trypanosomiasis. In *ILCA/ILRAD*, 77-86.

Connor, R.J. and Van Den Bossche, P. (2004). African animal trypanosomoses. In *Infectious diseases of livestock* (eds J.A.W. Coetzer, and R.C. Tustin), 251-296. Oxford University Press, Cape Town, Southern Africa.

Coxe, S., West, S. G. and Aiken, L. S. (2009). The analysis of count data: a gentle introduction to Poisson regression and its alternatives. *Journal of personality Assessment* **91**, 121-136. (doi:10.1080/00223890802634175).

Davis, S., Aksoy, S. and Galvani, A. (2011). A global sensitivity analysis for African sleeping sickness. *Parasitology* **138**, 516-526. (doi:10.1017/S0031182010001496).

Deveze, J. C. (2011). *Challenges in African agriculture. Washington: The International Bank for Reconstruction and Development*. eISBN 978-0-8213-8515-9.

Diekmann, O., Heesterbeek, J. A. P. and Metz, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365-382.

Diekmann, O., Heesterbeek, J. A. P. and Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *J. R. Soc. Interface* **7**, 873-885. (doi:10.1098/rsif.2009.0386).

Dransfield, R. D. and Brightwell, R. (1989). Problems of field testing theoretical models: a case study. *Ann. Soc. Belge Med. Trop.* **69**, 147-154.

Du Toit, R. (1954). Trypanosomiasis in Zululand and control of tsetse flies by chemical means. *Onderstepoort Journal of Veterinary Research* **26**, 317–378.

Eesterhuizen, J., Kappmeier-Green, K., Marcotty, T. and Van den Bossche, P. (2005). Abundance and distribution of the tsetse flies, Glossina austeni and G. brevipalpis, in different habitats in South Africa. *Medical and Veterinary Entomology* **19**, 367-371.

Fang, R. (2008). Zero-inflated Negative Binomial (ZINB) regression model for over-dispersed count data with excess zeros and repeated measures, an application to human mircobiota sequence data. MSc Dissertation, University of Colorado.

FAO. (2004). The state of food insecurity in the world. Rome: Food and Agriculture Organization of the United Nations.

Fraumann, R. (2003). Glossina morsitans (On-line). See http://animaldiversity.org/accounts/Glossina_morsitans/.

Gardner, W., Mulvey, E. P. and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin* **118**, 392-404.

Gettinby, G. (1989). Understanding infectious diseases: modelling approaches for the trypanosomiases. *Ann. Soc. Belge Med. Trop.* **69**, 21-30.

Gettinby, G., Revie, C. W. and Forsyth, A. J. (1994). Modelling: a review of systems and approaches for vector-transmitted and other parasitic diseases in developing countries. In *Modelling Vector-borne and other Parasitic Diseases* (eds B. D. Perry and J. W. Hansen), pp. 249-260. Nairobi: ILRAD.

Gordon, J. (2014). *SAS software to fit the generalized linear model*. Cary, NC: SAS Institute Inc.

Gouteux, J. P., Artzrouni, M. and Jarry, M. (2001). A density-dependent model with reinvasion for estimating tsetse fly populations (diptera: Glossinidae) through trapping. *Bulletin of Entomological Research* **91**, 177-183. (doi:10.1079/BER200185).

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

Hargrove, J. (2004). Tsetse population dynamics. In *The trypanosomiases* (eds I. Maudlin, P. H. Holmes and M. A. Miles), 113-138. CABI Publishing, Wallingford, UK.

Hassan, A. S., Garba, S. M., Gumel, A. B. and Lubuma, J. M. S. (2014). Dynamics of Mycobacterium and bovine tuberculosis in a human-buffalo population. *Compuational and Mathematical Methods in Medicine* **2014**.

Hendrickx, G., Nevill, E., Biesemans, I., Kappmeier-green, K., Van Camp, N. and Williams, R. (2003). The use of geostatistics and remote sensing to optimise tsetse field survey results: The example of KwaZulu-Natal. *Newsletter on Integrated Control of Pathogenic Trypanosomes and their Vectors* **7**, 26–29.

IPCC. (2014). The synthesis report of the fifth assessment report of the intergovernmental Panel on Climate Change. See http://www.ipcc.ch/publications_and_data/publications_and_data_reports.shtml.

Jordan, A. M. (1988). The role of tsetse in African animal trypanosomiasis. In *ILCA/ILRAD*, 37-42. Bristol, ODA/University of Bristol.

Kappmeier, K., Nevill, E. M. and Bagnall, R. J. (1998). Review of tsetse flies and trypanosomosis in South Africa. *Onderstepoort Journal of Veterinary Research* **65**, 195-203.

Kappmier-Green, K., Potgieter, F.T. and Vreysen, M.J.B. (2007). A strategy for an area-wide control campaign with an SIT-(*Glossina austeni* and *Glossina brevipalpis*) free South Africa. In *Area-wide control of insect pests* (eds M.J.B. Vreysen, A.S. Robinson and J. Hendrichs), 308–323. Springer, Vienna.

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

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15**, 209-225.

Leak, S.G.A. (1999). *Tsetse Biology and Ecology: Their Role in the Epidemiology and Control of Trypanosomosis.* London, UK: CABI Publishing.

Mamabolo, M. V., Ntantiso, L., Latif, A. A. and Majiwa, P. A. O. (2009). Natural infection of cattle and tsetse flies in South Africa with two genotypic groups of Trypanosoma congolense. *Parasitology* **136**, 425-431.

Marcotty, T., Simukoko, H., Berkvens, D., Vercruysse, J., Praet, N. and Van den Bossche, P. (2008). Evaluating the use of packed cell volume as an indicator of trypanosomal infections in cattle in eastern Zambia. *Preventive Veterinary Medicine* **87**, 288-300. (doi:10.1016/j.prevetmed.2008.05.002).

Marino, S., Hogue, I. B., Ray, C. J. and Kirschner, D. E. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology* **254**, 178-196. (doi:10.1016/j.jtbi.2008.04.011).

Masumu, J., Marcotty, T., Geysen, D., Geerts, S., Vercruysse, J., Dorny, P. and Van den Bossche, P. (2006). Comparison of the virulence of Trypanosoma congolense strains isolated from cattle in a trypanosomiasis endemic area of eastern Zambia. *International Journal for Parasitology* **36**, 497-501.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* London: Chapman and Hall, 2nd edition. ISBN 0-412-31760-5.

Mckay, M. D., Beckman, R. J. and Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **42**, 55-61.

Milligan, P. J. M. and Baker, R. D. (1988). A model of tsetse-transmitted animal trypanosomiasis. *Parasitology* **96**, 211-239.

Moore, S., Shrestha, S., Tomlinson, K. W. and Vuong, H. (2012). Predicting the effect of climate change on African trypanosomiasis: integrating epidemiology with parasite and vector biology. *J. R. Soc. Interface* **9**, 817-830. (doi:10.1098/rsif.2011.0654).

Morris, R. S. and Marsh, W. E. (1994). The impact of modelling on animal disease control. In *Modelling Vector-borne and other Parasitic Diseases* (eds B. D. Perry and J. W. Hansen), pp. 249-260. Nairobi: ILRAD.

Motloang, M. (2012). Vector competence of Glossina austeni and G. brevipalpis and characterization of Trypanosoma congolense strains from northern KwaZulu-Natal, South Africa. MSc Dissertation, University of Pretoria.

Motloang, M. Y., Masumu, J., Mans, B. J. and Latif, A. A. (2014). Virulence of Trypanosoma congolense strains isolated from cattle and African buffaloes (Syncerus caffer) in Kwazulu-Natal. *Onderstepoort Journal of Veterinary Research* **81**. (doi:10.4102/ojvr.v81i1.679).

Murray, M., Morrison, W.I. and Whitehead, D.D. (1982). Host susceptibility to African trypanosomiasis: trypanotolerance. *Advances in Parasitology* **21**, 1-68.

Murray, M., Morrison, W.I. and Whitelaw, D.D. (1982). Host susceptibility to African trypanosomiasis: trypanotolerance. *Advances in Parasitology* **21,** 1-68.

Murray, M., Trail, J. C. M. and d'Ieteren, G. D. M. (1990). Trypanotolerance in cattle and prospects for the control of trypanosomiasis by selective breeding. *Rev. sci. tech. Off. Int. Epiz.* **9**, 369-386.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A (General)* **135**, 370-384.

Ntantiso, L., De Beer, C., Marcotty, T. and Latif, A. A. (2014). Bovine trypanosomosis prevalence at the edge of Hluhluwe-iMfolozi Park, Kwazulu-Natal, South Africa. *Onderstepoort Journal of Veterinary Research* **81**. (doi:10.4102/ojvr.v81i1.762).

Perry, B., Randollph, T., Omore, A., Perera, O. and Vatta, A. (2005). Improving the health of livestock kept by the resource-poor in developing countries. In *Livestock and Wealth Creation* (eds E. Owen, A. Kitalyi, N. Jayasuriya, and T. Smith), 233-261. Department of National Development Press, Nottingham.

Pollock, J. N. (1992). Training manual for tsetse control personnel. Rome: Food and Agriculture Organization of the United Nations.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* **81**, 321-327.

Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. International Biometric Conference. Cape Town, South Africa.

Robinson, T. (1998). Practical applications of geopgraphic information systems in tsetse and trypanosomiasis control. In *Towards livestock disease diagnosis and control in the 21st century* (Proceedings of an International Symposium on Diagnosis and Control of Livestock Diseases using Nuclear and related techniques), 421-437. Vienna, Austria.

Rogers, D. J. (1979). Tsetse population dynamics and distribution: a new analytical approach. *Journal of Animal Ecology* **48**, 825-849.

Rogers, D. J. (1988). A general model for the African trypanosomiases. *Parasitology* **97**, 193-212.

Rogers, D. J. (1991). Satellite imagery, tsetse and trypanosomiasis in Africa. *Preventive Veterinary Medicine* **11**, 201-220.

Rogers, D. J. (1994). The modelling of vector dynamics in disease research. In *Modelling Vector-borne and other Parasitic Diseases* (eds B. D. Perry and J. W. Hansen), pp. 249-260. Nairobi: ILRAD.

Rogers, D. J. (2000). Satellites, space, time and the African trypanosomiases. *Advanced Parasitology* **47**, 129-171.

Rogers, D. J. (2006). Models for vectors and vector-borne diseases. *Advances in Parasitology* **62**. (doi:10.1016/S0065-308X(05)62001-5).

Rogers, D. J., Hendrickx, G. and Slingenbergh, J. H. W. (1994). Tsetse flies and their control. *Rev. sci. tech. Off. Int. Epiz.* **13**, 1075-1124.

Rogers, D. J., Randolph, S. E. and Kuzoe, F. A. (1984). Local variation in the population dynamics of Glossina palpalis palpalis (Robineau-Desvoidy) (Diptera: Glossinidae). I. Natural population regulation. *Bulletin of Entomological Research* **74**, 403-423.

Sanchez, M. A. and Blower, S. M. (1997). Uncertainty and sensitivity analysis of the basic reproductive rate. *American Journal of Epidemiology* **145**, 1127-1137.

Simarro, P. P., Cecchi, G., Franco, J. R., Paone, M., Diarra, J. A. R. P., Fevre, E. M., Mattioli, R. C. and Jannin, J. G. (2012). Estimating and mapping the population at risk of sleeping sickness. *PLos Negl Trop Dis* **6**, e1859. (doi:10.1371/journal.pntd.0001859).

Steverding, D. (2008). The history of African trypanosomiasis. *Parasites and Vectors* **1**.
(doi:10.1186/1756-3305-1-3).

Sutherst, R. (2004). Global change and human vulnerability to vector-borne diseases. *Clinical Microbiology Reviews* **17**, 136-173. (doi:10.1128/CMR.17.1.136-173.2004).

TDR. (2012). Assessment of research needs for public health adaptation to social, environmental and climate change impacts on vector-borne diseases in Africa. Addis Ababa: World Health Organization on behalf of the Special Programme for Research and Training in Tropical Diseases.

Van den Bossche, P. (2001). Some general aspects of the distribution and epidemiology of bovine trypanosomosis in sourthern Africa. *International Journal for Parasitology* **31**, 592-598.

Van den Bossche, P. and Rowlands, G. J. (2001). The relationship between the parasitological prevalence of trypanosomal infections in cattle and herd average packed cell volume. *Acta Tropica* **78**, 163-170.

Van den Bossche, P., Eesterhuizen, J., Nkuna, R., Matjila, T., Penzhorn, B., Geerts, S. and Marcotty, T. (2006). An update of bovine trypanosomosis situation at the edge of Hluhluwe-Mfolozi Park, Kwazulu-Natal Province, South Africa. *Onderstepoort Journal of Veterinary Research* **73**, 77-79.

van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences* **180**, 29-48.

Vreysen, M. J. B., Balenghien, T., Saleh, K. M., Maiga, S., Koudouou, Z., Cecchi, G. and Bouyer, J. (2013). Release-recapture studies confirm dispersal of Glossina palpalis gambiensis between river basins in Mali. *PLos Negl Trop Dis* **7**, e2022. (doi:10.1371/journal.pntd.0002022).

Williams, B. G, Dransfield, R. D. and Brightwell, R. (1990). Tsetse fly (Diptera: Glossinidae) population dyanimcs and the estimation of mortality rates from life-table data. *Bulletin of Entomological Research* **80**, 479-485.

Appendix A

Scatter Plots

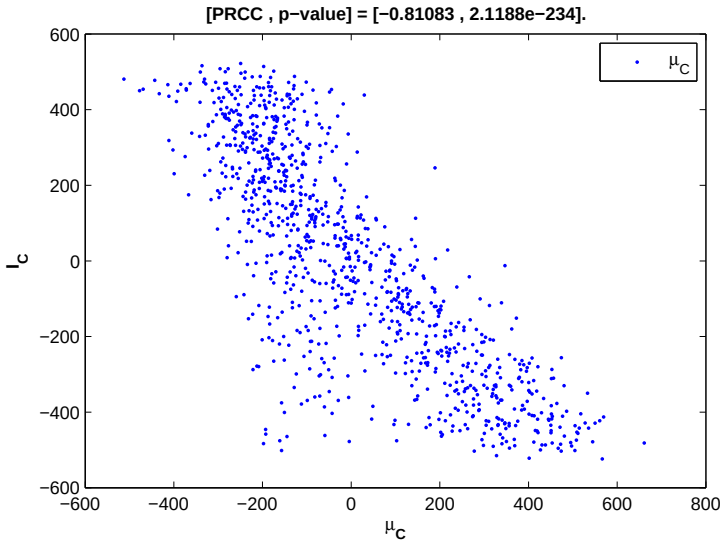


Figure A.1: Scatter Plot for infected cattle and the natural mortality rate of cattle

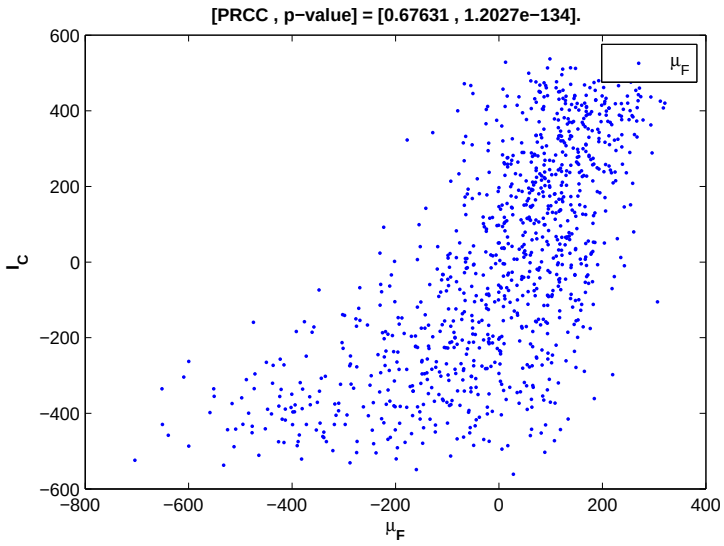


Figure A.2: Scatter Plot for infected cattle and the natural mortality rate of tsetse flies

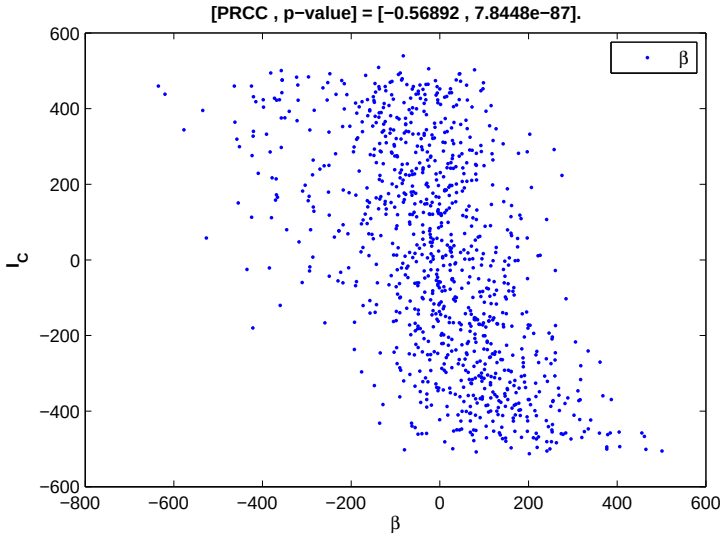


Figure A.3: Scatter Plot for infected cattle and the biting rate

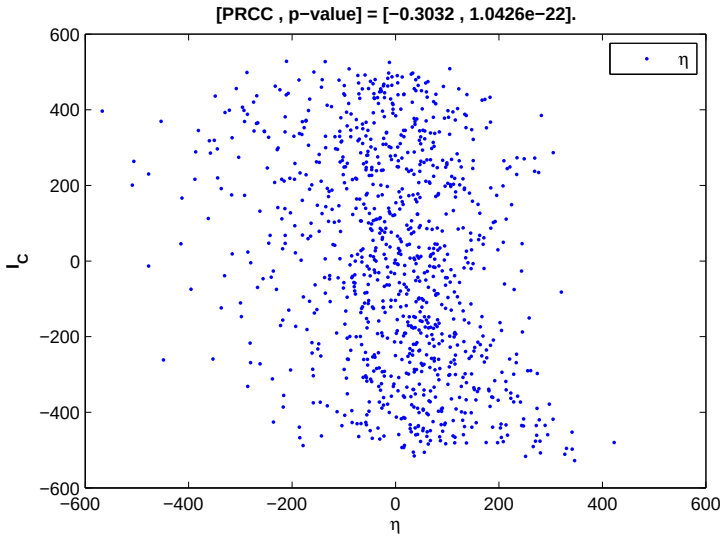


Figure A.4: Scatter Plot for infected cattle and the weight of infectivity

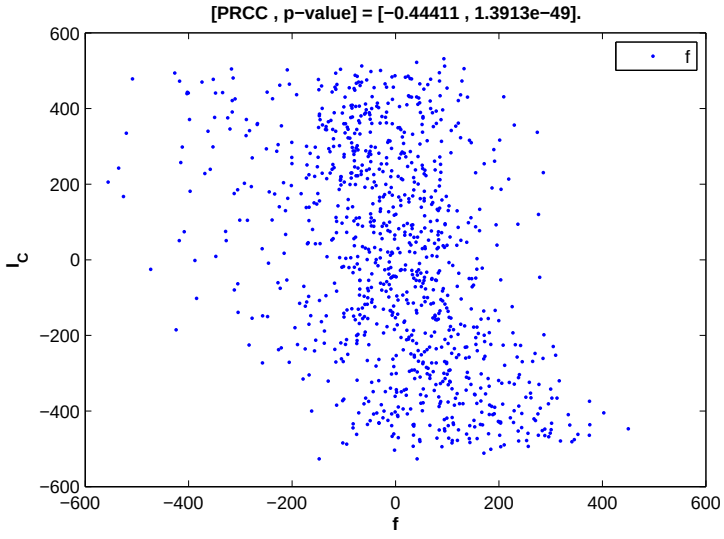


Figure A.5: Scatter Plot for infected cattle and the probability of infected fly producing infection in cow

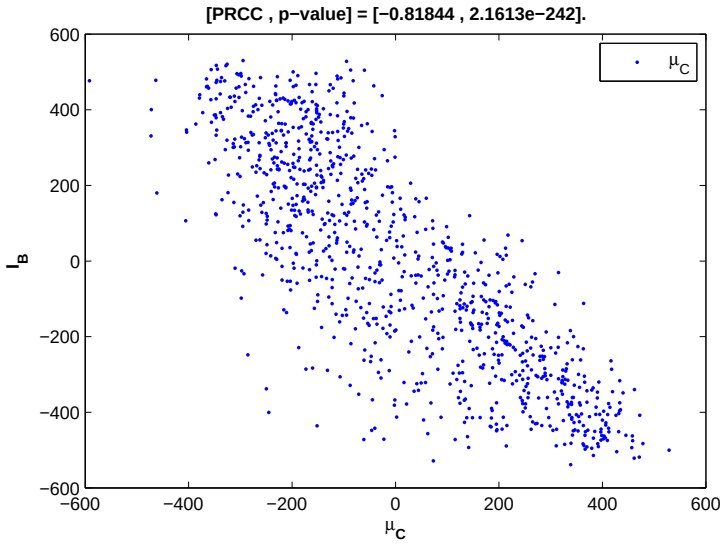


Figure A.6: Scatter Plot for infected buffaloes and the natural mortality rate of cattle

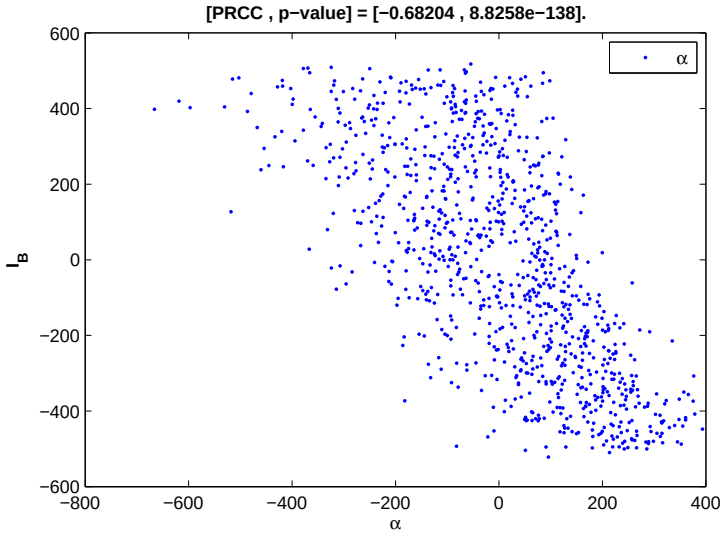


Figure A.7: Scatter Plot for infected buffaloes and the disease induced mortality rate of cattle

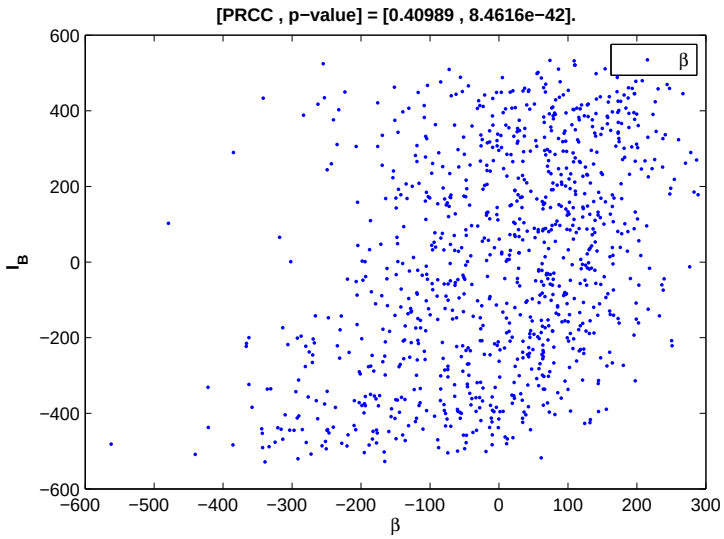


Figure A.8: Scatter Plot for infected buffaloes and the biting rate

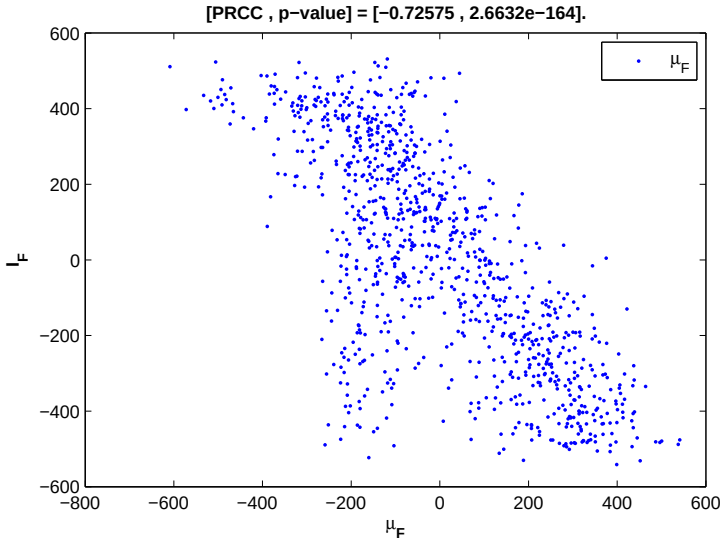


Figure A.9: Scatter Plot for infected tsetse flies and the natural mortality rate of flies

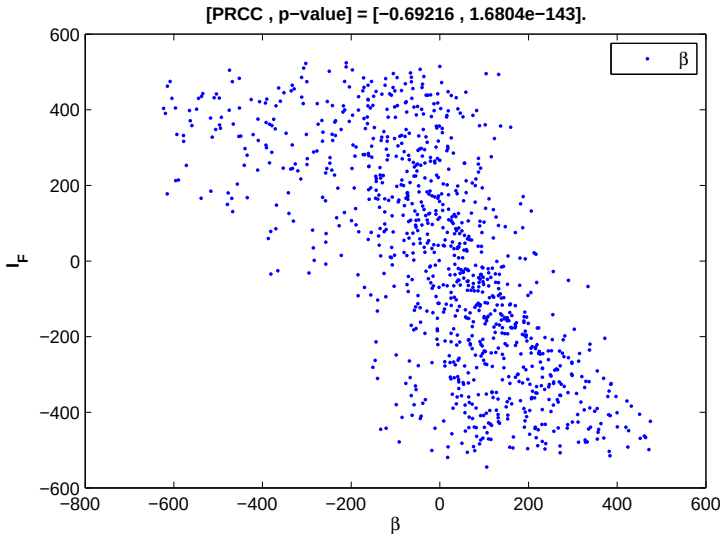


Figure A.10: Scatter Plot for infected tsetse flies and the biting rate

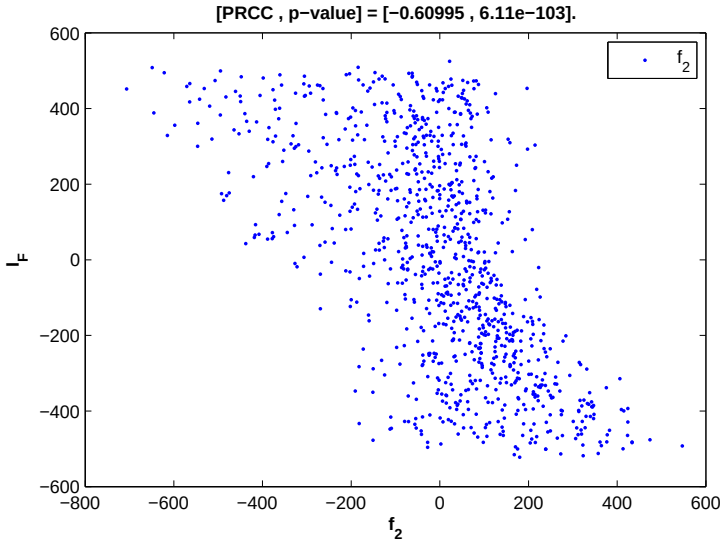


Figure A.11: Scatter Plot for infected tsetse fly and the probability of infected host infecting the fly

Appendix B

SAS Code

```
*****AAT code***** ;
data tryps ;
set tryps_by_yearandmonth ;

proc genmod data=tryps descending ;
class month region ;
model prevalence = pcv1 month region / aggregate=(pcv1 month region) dscale dist=bin
link=logit type3 type1 wald ;
title1 'Binomial Model: Tryps' ;
run ;

*****Tsetse abundance code*****;
data tst2 ;
set calc2 ;

***G.austeni*** ;
proc genmod data=tst2 ;
class month ;
model gatotal = month year / dscale dist=poi link=log type3 wald ;
title1 'Poisson:G.austeni' ;
run ;

proc genmod data=tst2 ;
```

```
class month ;
model gatotal = month year / dscale dist=negbin link=log type3 wald;
title1 'Negative Binomial:G.austeni' ;
run ;

proc genmod data=tst2 ;
class month ;
model gatotal = month year / dscale dist=zip type3 wald;
title1 'Zero-inflated Poisson:G.austeni' ;
zeromodel ;
run ;

proc genmod data=tst2 ;
class month ;
model gatotal = month year / dscale dist=zinb type3 wald;
title1 'Zero-inflated Negative Binomial:G.austeni' ;
zeromodel ;
run ;

***G.brevipalpis*** ;
proc genmod data=tst2 ;
class month ;
model gbtotal = month year / dscale dist=poi link=log type3 wald ;
title1 'Poisson:G.brevipalpis' ;
run ;

proc genmod data=tst2 ;
class month ;
model gbtotal = month year / dscale dist=negbin link=log type3 wald ;
title1 'Negative Binomial:G.brevipalpis' ;
run ;
```