The spatial epidemiology of Rhodesian sleeping sickness in recently affected areas of central and eastern Uganda

Nicola Ann Batchelor

Submitted in fulfilment of the Requirements for the Degree of Doctor of Philosophy

The University of Edinburgh 2009

Declaration:

I declare that all research within this thesis is my own work and that the thesis was entirely composed by me. The work has not been submitted for any other degree or professional qualification.

The work discussed in Chapter 3 has been published in PLoS Neglected Tropical Diseases. Further details of the paper, including a URL, are included in Appendix C.

The analysis discussed in Chapter 4 is in preparation for submission to PLoS Neglected Tropical Diseases.

Part of the analysis in Chapter 5 (exploratory analysis and multivariate analysis sections) is in preparation for submission to Nature.

Part of the analysis in Chapter 5 (identification and characterisation of potential high transmission zones) is in preparation for submission to Geospatial Health.

Nicola A Batchelor
Edinburgh

“**Tsetse flies are more numerous and troublesome than we have ever before found them. They accompany us on the march, often buzzing round our heads like a swarm of bees. They are very cunning, and when intending to bite, alight so gently that their presence is not perceived till they thrust in their lance-like proboscis. The bite is acute, but the pain is over in a moment; it is followed by a little of the disagreeable itching of the mosquito’s bite. This fly invariably kills all domestic animals except goats and donkeys; man and the wild animals escape. We ourselves were severely bitten on this pass, and so were our donkeys, but neither suffered from any after effects.”**

David Livingstone (1813 – 1873)

From *A Popular Account of Dr. Livingstone's Expedition to the Zambesi and Its Tributaries.*

Table of contents

Acknowledgements

I owe a lot of thanks to the man in my life, William Wardrop. He has provided a lot of encouragement and support throughout my PhD. I really appreciate everything he has done for me. I'd like to thank my parents for giving me the opportunities that led me to where I am now and also, thanks to my sister Heather for always being encouraging and interested in my work.

Thanks to Sue Welburn, who has allowed me the flexibility to advance my knowledge and skills immensely during the PhD. Her frequent astute insights have been of great help in planning, analysis and interpretation. Peter Atkinson has pushed me when I needed it most to do the best job possible and along with Peter Gething has encouraged me to learn Bayesian geostatistics and the use of R for the data analysis. Both also gave assistance regarding remote sensing data and analysis; a topic to which I was a complete novice at the start of my PhD. Thanks to Eric Fèvre, who has provided a lot of encouragement for the project and has dedicated a lot of time helping me with all manner of things, from statistics to land cover classifications and thesis formatting. I should also say that it was a lecture Eric gave on sleeping sickness epidemiology during my undergraduate studies that first prompted my fascination with the disease. Kim Picozzi has also been of great assistance during the PhD, both in and out of the laboratory.

Thanks to Jenna Fyfe and Beatrix Wissemann for assistance with the planning of my research, data collection and fieldwork which was much needed towards the beginning of my PhD. Many thanks also to Richard Selby for a great deal of help with data collection, geo-referencing of villages and good company while on fieldwork. Joseph Ssempijja also deserves a special thank you. He was an expert off-road driver during geo-referencing fieldwork during both the dry and wet seasons and managed to get us out of a few muddy situations! Thanks to Dr Charles Waiswa of the University of Makerere for taking the time out to discuss my research with me. He also kindly made himself available for any type of assistance which may have been needed during fieldwork. Additionally, Dr Abbas Kakembo and Dr Wambogo

of the Uganda Ministry of Health, Coordinating Office for the Control of Trypanosomiasis in Uganda (COCTU), have spent much time talking to me about the current sleeping sickness situation in Uganda.

I would like to thank a number of people for their assistance in data collection by conducting hospital visits when I was in the UK: Richard Selby, Beatrix Wissemann, Jenna Fyfe, Sally Wastling and Christine Amongi Acup have all helped to keep the datasets as up to date as possible. I would also like to thank Dr Eric Fèvre for the provision of a geo-referenced dataset which he collected as part of his PhD thesis and post doctoral research. Many thanks also to the local Government staff and local guides in Uganda (of whom there are too many to mention individually) whose knowledge of the study areas was instrumental for the geo-referencing of villages. Also thanks to the many hospital employees who assisted us with data collection and additional information, particularly those at Lwala and Serere Hospitals.

Thanks to Dr Ellie Biggs for many helpful chats about GIS, spatial analysis and other general issues and Gary Watmough for numerous discussions and a lot of assistance regarding GIS, spatial analysis, remote sensing and statistical methods. Gary also showed me how to use ATCOR atmospheric correction software, and provided a lot of help and input into the use of eCognition for land cover classification purposes. In addition, thanks to my office mates Richard Selby, Ellie Biggs, Gary Watmough, Sam Bateman and Niamh Burke for being a good support network.

For assistance and guidance regarding statistical methods and particularly model-based geostatistics I would like to thank Professor Peter Diggle and Dr Sujit Sahu. They both provided a great deal of assistance when it was most needed. Professor Murray Lark also deserves thanks for his input regarding autologistic regression in response to an out of the blue email. The MCMC iterations and final predictions were run on the University of Southampton IRIDIS Beowulf computing cluster. Many thanks to David Baker for his patience when it came to showing me how to use the IRIDIS cluster and providing me with the UNIX commands necessary to complete the analysis.

Abstract

The tsetse transmitted fatal disease of humans, sleeping sickness, is caused by two morphologically identical subspecies of the parasite *T. brucei; T. b. rhodesiense* and *T. b. gambiense*. Current distributions of the two forms of disease are not known to overlap in any area, and Uganda is the only country with transmission of both. The distribution of Rhodesian sleeping sickness in Uganda has expanded in recent years, with five districts newly affected since 1998. This movement has narrowed the gap between Rhodesian and Gambian sleeping sickness endemic areas, heightening concerns over a potential future overlap which would greatly complicate the diagnosis and treatment of the two diseases. An improved understanding of the social, environmental and climatic determinants of the distribution of Rhodesian sleeping sickness is required to allow more effective targeting of control measures and to prevent further spread and possible concurrence with Gambian sleeping sickness. The work presented in this thesis investigates the drivers of the distribution and spread of Rhodesian sleeping sickness in districts of central and eastern Uganda which form part of the recent disease focus extension.

The spatial distribution of Rhodesian sleeping sickness was examined in Kaberamaido and Dokolo districts where the disease was first reported in 2004, using three different methodologies. A traditional one-step logistic regression analysis of disease prevalence was compared with a two-step hierarchical logistic regression analysis. The two-step method included the analysis of disease occurrence followed by the analysis of disease prevalence in areas with a high predicted probability of occurrence. These two methods were compared in terms of their predictive accuracy. The incorporation of a stochastic spatial effect to model the residual spatial autocorrelation was carried out using a Bayesian geostatistical approach. The geostatistical analysis was compared with the non-spatial models to assess the importance of spatial autocorrelation, to establish which method had the highest predictive accuracy and to establish which factors were the most significant in terms of the disease's distribution. Links between Rhodesian sleeping sickness and

landcover in Soroti district were also assessed using a matched case-control study design. Temporal trends in these relationships were observed using an annually stratified analysis to allow an exploration of the disease's dispersion following its introduction to a previously unaffected area. This work expands on previous research that demonstrated the source of infection in this area to be the movement of untreated livestock from endemic areas through a local livestock market.

With regards to the comparison of regression frameworks, the two-step regression compared favourably with the traditional one-step regression, but the Bayesian geostatistical analysis outperformed both in terms of predictive accuracy. Each of these regression methods highlighted the importance of distance to the closest livestock market on the distribution of Rhodesian sleeping sickness, indicating that the disease may have been introduced to this area via the movement of untreated cattle from endemic areas, despite the introduction of regulations requiring the treatment of livestock prior to sale. In addition, several other environmental and climatic variables were significantly associated with sleeping sickness occurrence and prevalence within the study area. The temporal stratification of the matched case-control analysis highlights the dispersion of sleeping sickness away from the point of introduction (livestock market) into more suitable areas; areas with higher proportions of seasonally flooding grassland, lower proportions of woodland and dense savannah and lower elevations. These findings relate to the habitat preferences of the predominant vector species in the study area; *Glossina fuscipes fuscipes*, which prefers riverine vegetation.

The findings presented highlight the importance of the livestock reservoir as well as the climatic and environmental preferences of the tsetse fly vector for the introduction of Rhodesian sleeping sickness into previously unaffected areas, the subsequent spread of infection following an introduction and the equilibrium spatial distribution of the disease. By enhancing the knowledge base regarding the spatial determinants of the distribution of Rhodesian sleeping sickness within newly affected areas, future control efforts within Uganda may be better targeted to decrease prevalence and to prevent further spread of the disease.

List of Figures

List of Tables

Abbreviations

AIDS	Acquired immune deficiency syndrome
ATCOR-2	Atmospheric & Topographic Correction for Small FOV Satellite Images
AUC	Area under the ROC curve
AVHRR	Advanced Very High Resolution Radiometer
CI	Confidence interval
COCTU	Coordinating Office for the Control of Trypanosomiasis in Uganda
CrI	Credible interval (Bayesian statistics)
DALY	Disability adjusted life years
FAO	Food and Agricultural Organization of the United Nations
GIS	Geographical information system
GLM	Generalised linear model
GPS	Global positioning system
HIV	Human immunodeficiency virus
KDE	Kernel density estimation
Landsat ETM+	Landsat Enhanced Thematic Mapper Plus
Landsat TM	Landsat Thematic Mapper
LC I, II, III, IV and V	Local council level 1 (village), 2 (parish), 3 (sub county), 4 (county) and 5 (district)
LST	Land surface temperature
MCMC	Markov chain Monte Carlo
MIR	Middle infra-red (AVHRR channel 3)
NaLIRI	National Livestock Research Institute (previously LIRI – Livestock Research Institute)
NDVI	Normalised difference vegetation index
NOAA	National Oceanographic and Atmospheric Administration
OR	Odds ratio
ROC curve	Reciever operator characteristic curve
SRTM	Shuttle Radar Topography Mission
USGS	United States Geological Survey
UTM	Universal Transverse Mercator

VSG	Variable surface antigen
WGS	World Geodetic System

Chapter 1: General introduction

“The beginning of knowledge is the discovery of something we do not understand”

Frank Herbert (1920 – 1986)

1.1. Sleeping sickness

Of all the tropical infectious diseases, sleeping sickness (also known as human African trypanosomiasis) is perhaps one of the most important, yet most neglected, with an estimated 50,000 to 70,000 new human infections annually (including unreported cases) (World Health Organization 2006a; Fèvre *et al.* 2008), and resulting in an estimated 1.67 million disability adjusted life years (DALYs) lost in the year 2004 (World Health Organization 2008). The disease results in severe morbidity and mortality in humans if left untreated and is ranked sixth worldwide in terms of DALYs lost due to parasitic diseases, based on DALY estimates for 2004 (World Health Organization 2008). In addition to this, certain parasite species within the same genus as the parasite responsible for human sleeping sickness (trypanosomes) can also infect animals, causing a disease called nagana (Bruce 1895; Bruce *et al.* 1910). The resulting losses in terms of agricultural production as well as human lives cause considerable impacts on affected communities and act as a restraint to agricultural and economic development in many areas, propagating underdevelopment in the poorest rural areas of some of the world's least developed countries.

The human form of the disease, sleeping sickness, is caused by two distinct, yet morphologically identical, subspecies of the haemoflagellate *Trypanosoma brucei*: *Trypanosoma brucei rhodesiense* (which was first seen by Forde (1902)) and *Trypanosoma brucei gambiense* (which was identified initially by Stephens and Fantham (1910)). Both of these parasites are transmitted by the tsetse fly (*Glossina* spp.), of which there are many different species, each with varying habitat preferences and environmental constraints, although not all of them can act as vectors for *T. brucei* spp (Leak 1999). The two human infective *T. brucei* subspecies result in clinically and epidemiologically different diseases, and are geographically separated by the Great African Rift Valley. The trypanosome species which can infect animals but are not infective to humans include a third subspecies of *T. brucei: Trypanosoma brucei brucei.*

Rhodesian sleeping sickness, caused by *T. b. rhodesiense,* is found in eastern countries of sub-Saharan Africa and is characterised by an acute progression of the disease, with over 80% of fatalities occurring within six months of onset in the absence of treatment (Odiit *et al.* 1997). A reservoir exists in wild and domestic animals for this form of the disease, and human cases are usually the consequence of a sporadic infection (the vectors feed preferentially on animals) (Onyango *et al.* 1966; Welburn *et al.* 2001a). In contrast, Gambian sleeping sickness, caused by *T. b. gambiense,* is found in western and central Africa and results in a far slower progression to death (can be up to several years after the initial symptoms develop) (Apted 1970). At the moment it is thought that humans are the main reservoir of *T. b. gambiense*, with animals playing a negligible role in the epidemiology.

1.1.1. Parasite lifecycle and clinical aspects of infection

The lifecycle of the parasite is shown in Figure 1.1. The parasite is picked up from the blood of an infected human or animal by a tsetse fly and undergoes an essential maturation process within the vector, resulting in the infectious stage being present in the salivary glands. The parasites are then injected into a human or animal host when the vector takes its next blood meal. The local multiplication of the trypanosomes following inoculation by the tsetse fly may cause a small painful lesion (chancre) around the site of the tsetse bite, although this is not always present and may be very difficult to see on dark skin (Godfrey and Fairbairn 1958; Greenwood and Whittle 1980). This skin lesion may be accompanied by regional lymphadenopathy.

Following invasion of the blood stream, the parasites begin to multiply in the lymphatic system. This haemolymphatic stage of disease is known as early stage infection, and leads to non-specific symptoms such as fever, headache and malaise (Greenwood and Whittle 1980). As a consequence of infection and the initial multiplication of the parasites in the bloodstream, the immune system begins to bring the infection under control. However, alterations in the surface antigens of the parasites (variable surface antigens – VSG's) allow them to escape the immune reaction which has been initiated. This allows rapid multiplication of the parasites once again, giving waves of high parasitaemia and the resulting irregular fever. For

further information regarding VSG's and antigenic variation in trypanosomes see the review by Borst *et al* (1996). These symptoms gradually worsen until the parasites cross the blood brain barrier, resulting in involvement of the central nervous system and a chronic encephalopathy. This is known as the meningoencephalitic (late) stage of disease and gives rise to more severe symptoms including headache, mental changes, difficulty concentrating, and disruption of the sleep cycle, eventually leading to coma and death. The progression through these stages differs markedly between the two forms of sleeping sickness, with *T. b. gambiense* producing a chronic infection with a long asymptomatic period, which results in death after several years, whereas *T. b. rhodesiense* will bring about death within months (Apted 1970; Odiit *et al.* 1997).

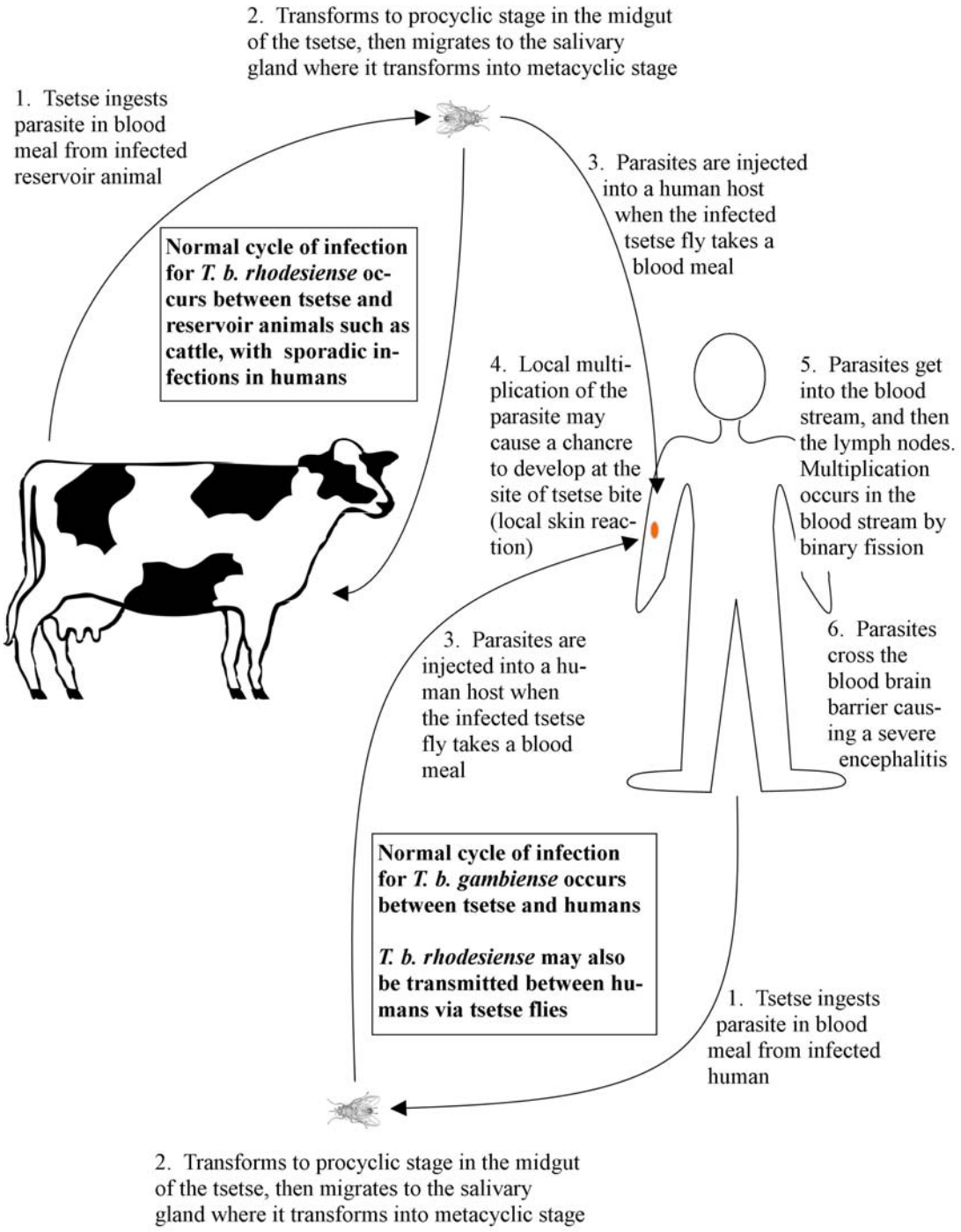


Figure 1.1: The lifecycle of *Trypanosoma brucei rhodesiense* and *Trypanosoma brucei gambiense* in human and animal hosts and tsetse fly vector.

1.1.2. Sleeping sickness control

The mainstays of sleeping sickness control are: case detection, diagnosis and treatment; vector control and control in the animal reservoir. The two distinct forms of sleeping sickness require differing control measures due to their epidemiology and clinical progression. For *T. b. gambiense* sleeping sickness, the detection and treatment of the asymptomatic human reservoir alone will have a considerable impact on the incidence of disease (Cattand *et al.* 2001). Due to the long asymptomatic period for Gambian sleeping sickness, epidemics may go undetected for some time and, thus, active detection and treatment is the preferred control option. However, due to the rapid onset of *T. b. rhodesiense* sleeping sickness and the involvement of an animal reservoir, the use of case detection and treatment alone is not sufficient to control this form of the disease, as has been demonstrated recently by Welburn *et al* (2001a) using mathematical modelling. Control of *T. b. rhodesiense* in the animal reservoir, along with vector control measures must also be used to support the use of passive detection and treatment for Rhodesian sleeping sickness (Welburn *et al.* 2001a).

1.2. The general spatial distribution of sleeping sickness

Sleeping sickness affects principally rural areas, and is found only in sub-Saharan Africa between 15° North and 29° South (Barrett *et al.* 2003). The occurrence of sleeping sickness within these boundaries occurs in discrete foci due to the climatic and environmental requirements of the vector as well as the distribution of animal reservoirs and human hosts. Infection prevalence is heterogeneous between different foci of disease, as well as between different villages within the same foci. The ecological and environmental determinants of such differences in transmission intensity are as yet poorly understood (Department of Communicable Disease Surveillance and Response 2000). Further discussion regarding the distribution of tsetse, sleeping sickness and animal trypanosomiasis with respect to external factors is provided in Section 1.5.

Throughout the disease's geographical range, 36 countries are known to be affected by sleeping sickness with over 250 endemic foci of the disease (Cattand *et al.* 2001). From the most recent published estimates, 60 million people are thought to be at risk of sleeping sickness within these areas (this figure is currently under revision (World Health Organization 2006b)), with a very small proportion of these benefiting from regular case finding, surveillance or adequate vector control measures (Pepin and Meda 2001; Fèvre *et al.* 2008). The lack of infrastructure in the remote and rural areas most affected by this disease complicates its epidemiological surveillance, resulting in underreporting and, thus, inaccurate information on the incidence of disease and its spatial distribution within foci of active transmission. For a more complete discussion of the underreporting of sleeping sickness, see Chapter 2, Section 2.1.2.1.

The general geographical separation of the two human infective subspecies of *T. brucei* means that presently, the two subspecies do not coincide in any area. The only country with active foci of both types of sleeping sickness is Uganda, and the recent introduction of Rhodesian sleeping sickness into areas which were previously free of the disease has raised concerns about the potential future overlap of the two subspecies (Picozzi *et al.* 2005). The two subspecies require different diagnostic methods, treatment regimes and control methods. Currently, the selection of diagnostics and treatment is carried out based on knowledge of the sleeping sickness subspecies endemic in the area in which the patient is thought to have been infected. The overlap of the two forms of sleeping sickness would greatly complicate an already difficult situation as definitive subspecies differentiation would have to be carried out prior to treatment. However, the two subspecies are morphologically identical and, thus, expensive and complex molecular diagnostics which are inappropriate for the rural settings in which the disease occurs would be required (Hutchinson *et al.* 2003). As a result, inappropriate drugs may be administered to patients where the true subspecies of trypanosome is not known, consequently increasing the rate of treatment failures.

Recent control activities have been undertaken in Uganda by a public-private partnership involving CEVA Santé Animale, IK Investment Partners, the Coordinating Office for the Control of Trypanosomiasis in Uganda (COCTU), IKARE, Makerere University and the University of Edinburgh (Ceva Sante Animal *et al.* 2006). This involved the mass treatment of cattle with trypanocidal drugs along with the restricted application of insecticide to cattle within areas at highest risk of convergence. The programme aims were to reduce the prevalence of *T. b. rhodesiense* in cattle populations, thus, reducing the number of human infections and preventing further spread of the disease (Kabasa 2007). The control programme was instigated in 2006 and since then there have been no further recorded long distance movements of *T. b. rhodesiense* in Uganda and concurrence of the two parasite subtypes has not occurred.

1.3. Sleeping sickness in Uganda

Traditional foci of disease exist in south east Uganda (the Busoga focus) where *T. b. rhodesiense* is present, and the north west of the country (the West Nile focus), where *T. b. gambiense* occurs (see Figure 1.2). Further information regarding Uganda and the specific study areas for each chapter is provided in Chapter 2, Section 2.3.

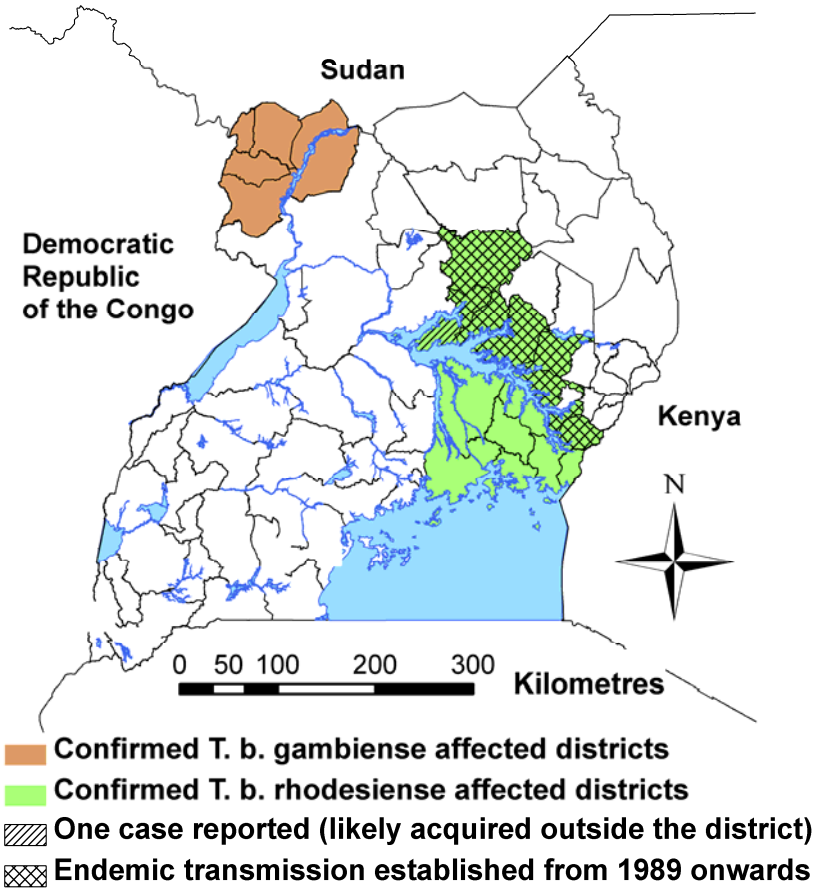


Figure 1.2: Map of Uganda highlighting areas of Rhodesian and Gambian sleeping sickness transmission.

1.3.1. The Busoga focus, south east Uganda

The first recognition of sleeping sickness in Uganda occurred in 1898, in the Busoga region, south east Uganda (Cook 1901). At the time, the disease in this area was thought to be caused by *T. gambiense* (later named *T. b. gambiense*), with the importation of *T. rhodesiense* (later called *T. b. rhodesiense*) occurring in 1940 via migrant labourers (MacKichan 1944). However, this theory has recently come under dispute with the examination of historical records suggesting that Rhodesian sleeping sickness has been present in the Busoga focus all along with only occasional cases of imported Gambian sleeping sickness seen in the area (Koerner *et al.* 1995; Fèvre *et al.* 2004).

The Busoga region has seen several epidemics of sleeping sickness in the years since the disease was first documented in this area; one beginning in the early 1900s (Langlands 1967); a second which began in 1940 (MacKichan 1944) and a third beginning in 1971, this time outside of the traditional tsetse fly belt. The political and economic situation in the country during the third epidemic hampered control efforts, with a lack of resources and trained personnel, as well as large volumes of uncontrolled population movements through tsetse infested bush (Matovu 1982; Okiria 1985). As a result of insufficient control measures, the disease spread northwards to Tororo district (bordering Kenya) in 1984 (Mbulamberi 1989).

Since the introduction of Rhodesian sleeping sickness to Tororo district in 1984 the disease has spread further, reaching Soroti district on the north east shores of Lake Kyoga in December 1998. This was found to be due to the movement of untreated livestock from endemic areas through the local livestock market (Fèvre *et al.* 2001a). In 2004 the disease spread further, to Kaberamaido and Dokolo districts on the northerly shore of Lake Kyoga (Fèvre *et al.* 2005; Picozzi *et al.* 2005). Following this spread, there is a distance of only 150 km separating the active foci of Rhodesian sleeping sickness north of Lake Kyoga and Gambian sleeping sickness in the north west of the country (Picozzi *et al.* 2005). Figure 1.2 illustrates the spatial distribution of Rhodesian sleeping sickness in Uganda, including the recent spread of the disease.

1.3.2. The West Nile focus, north west Uganda

The first identification of *T. b. gambiense* sleeping sickness in the north west of Uganda occurred a few years after the recognition of the epidemic in the Busoga region, in 1904. However, it is likely that the disease was present in this area prior to its identification. Following the introduction of the parasite to the Nile river area, it spread rapidly, affecting areas to the north east of Lake Albert and bordering Sudan in 1914 (Adams 1907).

The spread of sleeping sickness westwards into West Nile district was first recorded in 1925, resulting in a severe epidemic following large scale population movements into this area. Despite a considerable reduction in sleeping sickness incidence in this

area from the late 1940s (Morris 1962), in more recent years (from 1980 onwards), civil unrest and population movements due to violence in the region has resulted in increasing case numbers and the Uganda West Nile focus is now contiguous with the southern Sudan focus of disease (Moore and Richer 2001; Picozzi *et al.* 2005). Figure 1.2 demonstrates the districts of north west Uganda known to be affected by Gambian sleeping sickness.

1.4. Spatial epidemiology

Maps and spatial analysis have been used to support infectious disease epidemiology for many years, from the rudimentary use of maps of Cholera cases in London by John Snow in 1845 (Newsom 2006) to the use of complex mapping software and geostatistical methods in modern times (Elliott and Wartenberg 2004). The discipline of 'spatial epidemiology' has developed greatly since the 19th century with the development of spatial statistics and advances in specialist geographical information system (GIS) software and handheld global positioning system (GPS) receivers in the past two decades. The utilisation of these methods and software packages can allow the spatial linkage of various datasets of relevance to disease distributions (e.g. land cover and climatic data) and the description and understanding of complex spatial patterns in the risk of disease (Bergquist 2001; Rushton 2003).

1.4.1. Descriptive spatial epidemiology

Basic mapping of disease occurrence as either point patterns (the precise location of actual cases of disease) or aggregate data (case counts or risk, rates or prevalence of disease for different areas) can be used to summarise the observed variation in disease both spatially and temporally (Gatrell *et al.* 1996; Elliott *et al.* 2000). Descriptive spatial methods may be used to provide information to support the formation of policy and allocation of resources or to highlight potential or actual high risk areas in need of intervention. By comparing disease distributions with maps of potential risk factors (e.g. the location of water bodies), it is also possible to develop hypotheses for further study (Ostfeld *et al.* 2005).

The adjustment of disease measures for differences in population composition may be carried out in descriptive spatial epidemiology in the same manner as in classical epidemiology studies; for example the calculation of rates or standardisation based on a confounding factor such as age. This allows a more direct comparison of different areas or locations which may have heterogeneous population structures. In addition, kernel density estimation (KDE) can be used to smooth observed disease patterns across an area. This is particularly useful in circumstances where a high point density hinders the direct interpretation of the point pattern; KDE can allow a straightforward visualisation of the density of cases across the area of interest (Gatrell *et al.* 1996; Rushton 2003). The introduction of temporal data via a time-series of maps can also be used to allow the examination of spatio-temporal patterns in disease occurrence, for example, the spatial movement of an epidemic across a defined area through time (Ostfeld *et al.* 2005).

1.4.2. Geographical correlation studies

With an *a priori* hypothesis, spatial epidemiology may be used to examine spatial relationships between measures of disease (e.g. occurrence or prevalence) and potential risk factors using statistical modelling. There is a wealth of spatially referenced data which may be used in these types of studies and the use of remotely sensed data, which can provide continuous coverage for environmental factors including land cover and climatic variables, has increased a great deal in recent years (Brooker *et al.* 2002c; De La Rocque *et al.* 2004). The resulting models may also be used to predict the distribution of disease over a continuous area. This can be particularly useful where data on the distribution of a disease is patchy by allowing the prediction of disease occurrence or prevalence across ecologically similar areas for which no or little data are available (Brooker 2007). Several methods are available for geographical correlation studies and predictive mapping; a more detailed discussion of the use of regression modelling is provided in Chapter 3, Section 3.1.2 and spatial regression modelling in Chapter 4, Section 4.1.1.

1.4.3. Point or line source exposures

In some situations, exposures may occur from a point or linear source, for example, a water body or river. In these cases, it is possible to examine the risk of disease in relation to the source by carrying out highly localised studies in the area of the potential exposure (Elliott *et al.* 2000). These types of studies may provide evidence of a higher risk of disease acquisition in proximity to the potential exposure site than in areas further away.

1.4.4. Cluster detection

Statistical methods can be used to look for clustering within disease data; the non-random spatial distribution of disease cases in comparison with the distribution of the general population (non-cases). Although interpretation of cluster analyses may be difficult, these types of studies can provide useful insights into the distribution and epidemiology of disease and may be used for the generation of hypotheses. Cluster analysis may also be carried out using spatio-temporal data, thus allowing the detection of clustering in space and time (Kulldorff and Nagarwalla 1995; Gatrell *et al.* 1996).

1.4.5. The use of spatial epidemiology in disease control

In recent years, the utility of spatial epidemiology and GIS in disease control have been realised, with the application of spatial methods to the investigation of infectious disease distributions and an increase in the number of disease control programmes utilising spatial analyses (Thomson and Connor 2000). The mapping of disease patterns allows the examination of spatial heterogeneity of infection risk, and can highlight areas of high risk which have the greatest need for intervention.

Problems in the planning of disease control programmes may be encountered where disease distribution data are incomplete or patchy. In such cases, predictive modelling can be used to provide maps of the predicted distribution of disease over areas for which no information is available, based on external factors including

ecological and environmental variables (Brooker *et al.* 2001; Herbreteau *et al.* 2007). This is of particular importance in lower income countries where it may be difficult to obtain complete data regarding the distribution of a disease due to poor surveillance systems combined with logistical difficulties in the implementation of extensive surveys. There are many examples in the literature of the integration of spatial epidemiology into disease control programmes, for example, helminth (Brooker *et al.* 2002a; Brooker *et al.* 2002c), schistosomiasis (Kabatereine *et al.* 2004; Clements *et al.* 2006a) and onchocerciasis control (Gemade *et al.* 1998; Noma *et al.* 2002). The distributions of these examples are spatially heterogeneous; areas where more intense transmission occurs are clearly priority areas for the implementation of control efforts. The use of spatial epidemiology and GIS can assist in the identification of these priority areas and allow the targeting of control efforts, thus, allowing a more cost effective use of resources (Bergquist 2001).

1.5. The spatial epidemiology of sleeping sickness

1.5.1. Vector distribution

It is well documented that the focal distribution of human sleeping sickness (and animal trypanosomiasis) is determined largely by the ecological and environmental requirements of the tsetse fly vectors. There is a wide range of *Glossina* species present across the sub-Saharan fly belt, each with different environmental niches. The three main groups of tsetse fly species are the Morsitans group (found in savannah areas), the Fusca group (found in dense forests) and the Palpalis group (found in riverine forest areas) (Leak 1999).

Climate is a very important determinant of tsetse distributions and as climate is dependent on altitude, this also influences their distributions, with very low vector densities (if any at all) in highland areas. Several studies have used a combination of ground measured variables and remotely sensed imagery to investigate and quantify relationships between the occurrence of tsetse fly vectors and external factors (i.e. environmental or climatic). Environmental factors demonstrated to have a significant influence on vector occurrence or density include vegetation cover (Rogers *et al.*

1996; Kitron *et al.* 1996), land use patterns, land cover types (Mahama *et al.* 2005), humidity, normalised difference vegetation index (NDVI; a surrogate measure for the greenness of vegetation) (Rogers and Randolph 1991; Hendrickx *et al.* 1999a), temperature, rainfall (Rogers *et al.* 1996; Robinson *et al.* 1997a; Hendrickx *et al.* 2001b) and elevation (Robinson *et al.* 1997a). Findings suggest that tsetse occurrence is often dependent on just one variable (e.g. temperature) at the limits of its distribution, but within the limits (e.g. in areas with a suitable temperature) the distribution is also influenced by additional environmental and climatic variables such as rainfall and NDVI (Rogers and Randolph 1993; Rogers *et al.* 1996). However, the relationships found vary between vector species, studies, and geographical areas, highlighting the importance of local factors in vector occurrence.

Following the description of relationships between tsetse occurrence or abundance and external variables, predictive maps illustrating the distribution of *Glossina* sp at the continental, national, regional and local scales have been produced (Rogers *et al.* 1996; Robinson *et al.* 1997a; Wint and Rogers 2000; Hendrickx *et al.* 2001b; Mahama *et al.* 2005). This has allowed a greater understanding of the spatial heterogeneity in vector distributions and the circumvention of problems in disease control programmes which may arise through a lack of comprehensive and up-to-date information. Figures 1.3 to 1.5 illustrate the predicted suitability for *G. fuscipes*, *G. pallidipes* and *G. morsitans* in Uganda from one such study, which has been made available by the Food and Agriculture Organization of the United Nations (FAO) (Wint and Rogers 2000).

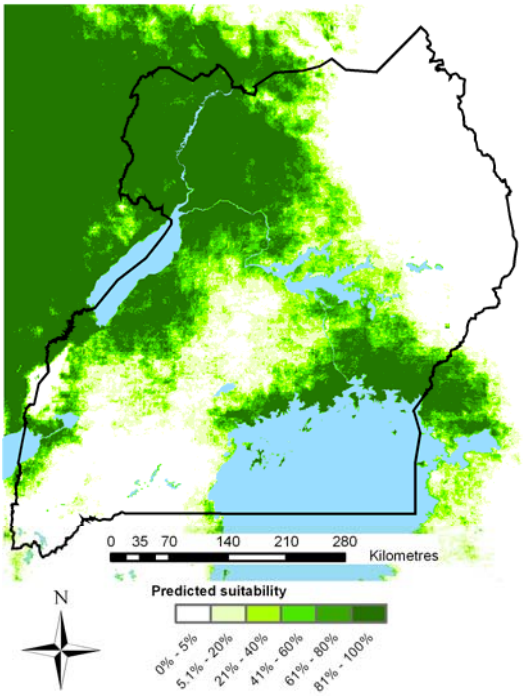


Figure 1.3: Predicted suitability for *Glossina fuscipes fuscipes* in Uganda. Data source: (Wint and Rogers 2000)

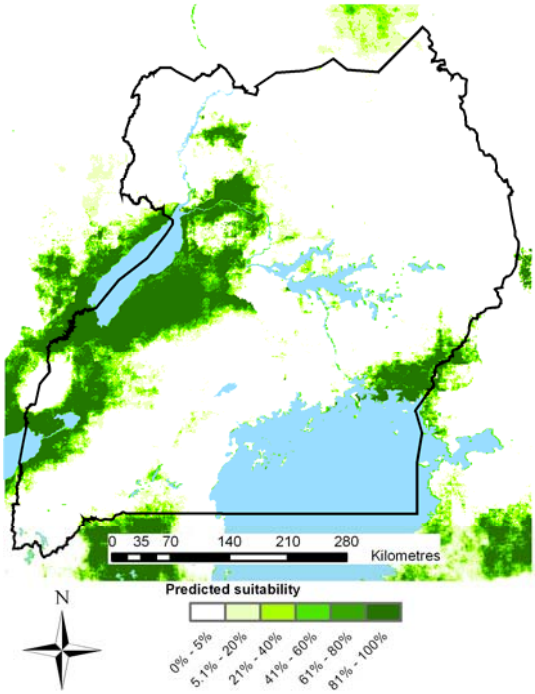


Figure 1.4: Predicted suitability for *Glossina pallidipes* in Uganda. Data source: (Wint and Rogers 2000)

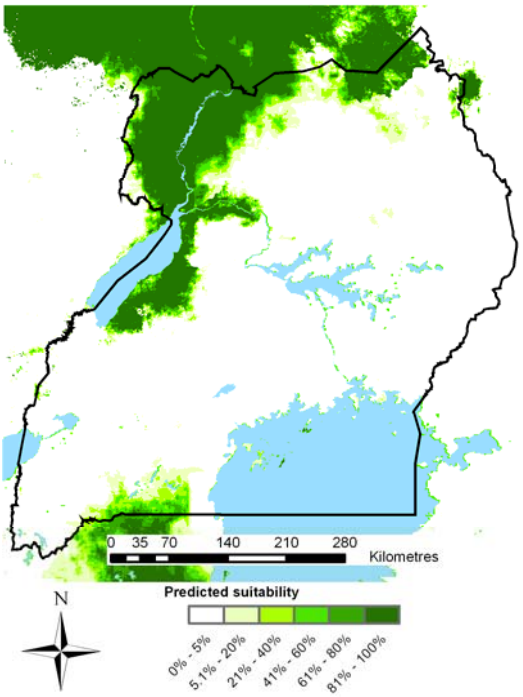


Figure 1.5: Predicted suitability for *Glossina morsitans* in Uganda. Data source: (Wint and Rogers 2000)

1.5.2. Trypanosomiasis distributions

The spatial distribution of trypanosomiasis infections in both humans and animals, although not controlled directly by environmental factors (as is the distribution of vector species), does have an indirect relationship with these factors. Environmental factors influencing the presence of the disease vector will similarly impact on the presence of trypanosomiasis, and spatial analysis techniques have been used to quantify the relationships between the occurrence or prevalence of trypanosomiasis and various environmental and climatic variables. Several significant factors have been established including altitude, land pressure from human populations, the availability of vector blood meals and animal herd information (Hendrickx *et al.* 1999b; Hendrickx *et al.* 2000).

Predictive mapping has not been utilised as widely for trypanosomiasis distributions as for tsetse distributions or other diseases such as malaria or schistosomiasis. However, one such study which predicted spatially the prevalence of animal trypanosomiasis, based on quantified relationships with external factors, has been published by Hendrickx *et al* (2000). It was found that the predictive accuracy for trypanosomiasis prevalence was lower than the predictive accuracy for tsetse abundance (Hendrickx *et al.* 2000). The use of similar methods for the prediction of human sleeping sickness risk has not yet been realised, with the majority of such studies focusing on the distribution of the tsetse fly vector, as described in section 1.5.1.

1.5.3. Spatial risk factors

Within smaller areas (e.g. within endemic disease foci), areas with high sleeping sickness prevalence tend to occur where there is a lot of contact between humans, tsetse flies and animal reservoirs (for Rhodesian sleeping sickness), for example, animal watering points (Rogers and Williams 1993). Although it is widely accepted that the focal distribution of sleeping sickness is dependent on environmental and ecological factors, there have been relatively few studies looking at local spatial risk factors for sleeping sickness. In studies which have been undertaken at the village

and/or household level, several different risk factors have been identified, although knowledge regarding the determinants of the spatial distribution of sleeping sickness within small areas is scarce.

Gambian sleeping sickness was found to be associated with the cultivation of cash crops in the Cote D'Ivoire, demonstrating the correlation with certain types of land cover, and the location of homesteads in relation to plantations and agricultural land was shown to have an influence on the occurrence of sleeping sickness (Fournet *et al.* 2000). Population mobility patterns have also been shown to influence the epidemiology of sleeping sickness, with infection of agricultural workers who reside in a large town occurring in the peri-urban areas outside of the town where they travel to work each day (Fournet *et al.* 2000). In addition to this, the tendency for cases to cluster within families highlights the importance of shared risk factors or, potentially, hereditary factors (Pepin and Meda 2001).

Familial clustering has also been demonstrated for Rhodesian sleeping sickness and further studies have indicated the propensity for infection in individuals with certain occupations or patterns of mobility. Individuals who travel regularly outside of the village, particularly into an endemic area, are at a higher risk of acquiring sleeping sickness (i.e. hunters or those collecting firewood) due to their increased contact with the tsetse fly vector (Wyatt *et al.* 1985; Okia *et al.* 1994). Increased risk of disease and clustering of cases has been observed in areas close to wetlands and swamps, highlighting the importance of tsetse habitat requirements within the study areas (Odiit *et al.* 2006; Zoller *et al.* 2008).

The importation of sleeping sickness to an area previously free of infection in Uganda via the movement of untreated, infected cattle from endemic areas has been demonstrated by Fèvre *et al* (Fèvre *et al.* 2001a). This finding, with the subsequent clustering of cases around the livestock market where the cattle were traded, demonstrates the vital role of the animal reservoir of infection for the spread of human disease (Fèvre *et al.* 2001a). However, recent research by Zoller *et al* (2008) found a contradictory protective effect due to regular visits to a cattle market. The

same study also demonstrated a protective effect due to regular travel outside of the village of residence (Zoller *et al.* 2008), in contradiction to findings from Okia *et al* (1994). Other protective factors detected include an increasing population density (which leads to the destruction of tsetse habitats) (Reid *et al.* 2000; Odiit *et al.* 2006) and the presence of grazing animals near places of high human-fly contact (Okia *et al.* 1994).

Although several risk factors have been identified, there is a lack of consistency between associations found in various studies. This is further complicated by the number of species of *Glossina* vector, as well as the two dissimilar subspecies of human infective trypanosome. As all three vector groups (Morsitans, Fusca and Palpalis) have differing environmental niches, the spatial determinants influencing their distributions are likely to vary. Thus, any risk factors for sleeping sickness (or animal trypanosomiasis) found in one area may not apply to other areas due to the habitats and behaviours of the local vector species.

1.5.4. The use of spatial epidemiology for the control of sleeping sickness

The prioritisation of areas for the control of tsetse and trypanosomiasis has been aided by the use of GIS and mapping. The integration of data layers representing agricultural intensity, cattle stocking levels, soil erosion rates and biodiversity with vector and animal trypanosomiasis distributions (or predicted distributions) in a GIS, has allowed the selection of areas where agriculture is constrained due to the presence of trypanosomes, and where successful control would result in the greatest benefits, with the least chance of adverse impacts on the environment such as soil erosion, or a significant reduction in biodiversity (Robinson 1998; Robinson *et al.* 2002; Symeonakis *et al.* 2007).

The Programme Against African Trypanosomiasis information system (PAAT-IS) was developed specifically to assist in the targeting of control efforts, and encompasses a large number of data sources including cattle density, tsetse distribution and NDVI. The information system integrates GIS with country level tsetse and trypanosome information to allow the selection of areas where the most

rapid impact of tsetse removal will be obtained and also predicts the impacts of tsetse control (Gilbert *et al.* 1999). GIS-based methods similar to those discussed above have also been used to identify the potential impacts of increasing human population densities on tsetse populations, indicating a reduced risk of human trypanosomiasis due to the destruction of tsetse habitats (Reid *et al.* 2000).

The likely effectiveness of different control measures has also been investigated using spatial and epidemiological modelling by Yu *et al* (1995). The integration of models representing tsetse dispersal with the modelled effects of various barrier methods (to prevent the reinvasion of tsetse into areas free of the vector) were used to demonstrate the most effective methods for the prevention of re-invasion. A review by Hendrickx *et al* (2001a) details further applications of GIS to the control of tsetse and animal trypanosomiasis at local, national and continental levels.

1.5.5. Potential uses of spatial epidemiology for sleeping sickness

As discussed, sleeping sickness occurs in distinct foci with a marked heterogeneity in its spatial distribution and prevalence. The determinants of this heterogeneity are not yet fully understood and involve a combination of vector, parasite and host related factors. Although the more general determinants of the distribution of sleeping sickness are widely known – for example the distribution of suitable habitats for the tsetse fly vector – there are several questions remaining.

The use of statistical models describing the relationship between vector occurrence or disease measures and environmental variables has been illustrated for other parasitic and vector-borne diseases, including malaria and schistosomiasis (Hay *et al.* 1996; Brooker 2002; Noor *et al.* 2008a). However, there have been few applications of these methods for sleeping sickness research; the majority of those studies which have been carried out look principally at the presence or abundance of the *Glossina* sp. vectors. It is documented that the presence of tsetse fly vectors does not always equate to presence of human sleeping sickness (or of animal trypanosomiasis) (Lancien *et al.* 1990). The use of statistical models for the investigation of spatial risk factors would be of great benefit to the understanding of sleeping sickness

distributions (and animal trypanosomiasis). Predictive models which could be used to highlight areas with a high risk of human sleeping sickness would also be of assistance for the planning of sleeping sickness control activities.

As discussed in Section 1.2, concerns have been heightened over the potential future overlap of Rhodesian and Gambian sleeping sickness in Uganda following the recent spread of the Rhodesian form of the disease (Fèvre *et al.* 2001a; Fèvre *et al.* 2005; Picozzi *et al.* 2005). Thus far, surveillance efforts have not detected concurrent transmission of the two forms of the disease. However, there have not yet been any studies aiming to quantify the risk of concurrence of Rhodesian and Gambian sleeping sickness, although this is evidently an important consideration for the provision of health care services, sleeping sickness diagnostic services, treatment facilities, disease surveillance and vector control measures (Picozzi *et al.* 2005). The application of GIS and spatial epidemiology would be of great value in this setting, by increasing our knowledge of the dynamics of disease transmission and the determinants leading to the observed spatial heterogeneity of disease prevalence. In addition, by determining factors influencing the continuing spread of Rhodesian sleeping sickness within Uganda and quantifying relationships with external factors, it may be possible to provide information for the focusing of disease control efforts. An improved understanding would allow the most cost efficient use of resources to decrease the prevalence of sleeping sickness and prevent further spread of the disease and a potential future overlap of the two subspecies.

1.6. Research Objectives

This thesis focuses on the spatial epidemiology of Rhodesian sleeping sickness in areas of Uganda to which the disease has recently been introduced. The main objectives of the research were to investigate the relationships between the spatial distribution of Rhodesian sleeping sickness within these areas and several environmental, climatic and social factors and to identify factors involved in disease epidemics and the extension of existing disease foci into new areas. More specifically, the objectives of the research were to:

- Investigate the effects of environmental, climatic and social factors on the distribution of Rhodesian sleeping sickness in recently affected areas of Uganda.
- Produce predictive maps to highlight areas outside of the study areas which may support a high prevalence of sleeping sickness if the disease was to spread further.
- Compare different regression methodologies (including non-spatial and spatial regression) for the investigation of relationships and to assess the predictive accuracy of each.
- Identify factors which may be involved in the continuing spread of Rhodesian sleeping sickness in Uganda.
- Investigate links between the distribution of sleeping sickness and land cover types in a recently affected area of Uganda
- Investigate temporal changes in observed relationships between the distribution of sleeping sickness, landcover and other factors to give insight into the disease's dispersion following its introduction.
- Identify and characterise potential high transmission zones in a recently affected area of Uganda.

Chapter 2. Data acquisition and management: sleeping sickness and supplementary data

"To write it, it took three months; to conceive it -- three minutes; to collect the data in it -- all my life."

F. Scott Fitzgerald (1896 - 1940)

2.1. Introduction

2.1.1. Sleeping sickness data sources in Uganda

The collection of health related data in developing countries, including Uganda, is often unreliable. For the case of sleeping sickness in Uganda, individual patient records are collected at the treating health facility using written records which are normally maintained by the laboratory staff. These records will normally include the patient's name, age, sex, village of residence (as well as the parish and district), date of admission, stage of illness, method of detection (active or passive), outcome of infection and date of discharge (or death). In addition to these records, which are normally recorded in a specific sleeping sickness log book, further details regarding the patient's treatment, progress and outcome can be found in individual patient records. The national collation of Rhodesian sleeping sickness records is conducted periodically by the Ministry of Health's Coordinating Office for the Control of Trypanosomiasis (COCTU), resulting in a spreadsheet containing patient records with a similar level of detail as the laboratory log books, although usually the village of residence is omitted (A. Kakembo, pers. com.).

In addition to sleeping sickness hospital records, intensive sleeping sickness screening programmes are occasionally used for various purposes, such as the assessment of disease burden in areas which have been shown to be newly affected by the disease. In such circumstances, a large volume of data can be generated, with information regarding each of the individuals screened, including village of residence and occasionally also the location of the individual homesteads.

2.1.2. Data quality considerations

One of the most important aspects of data acquisition and collation is data quality. Prior to any data analysis, one needs to account for any aspects of the data acquisition which may result in inaccuracies, missing data or bias. This is of particular importance in a developing country setting where electronic hospital

systems and patient records are not available, and the documentation of health care related information is dependent on the routine manual recording by laboratory and other health care staff within individual health centres.

2.1.2.1. Under-reporting of sleeping sickness

The under-reporting of sleeping sickness is well documented: a recent estimation based on data from south eastern Uganda suggested that 39% of all Rhodesian sleeping sickness cases in the study area were not reported and for every death recorded, another 12 deaths due to sleeping sickness went unrecorded (Odiit *et al.* 2005). Sleeping sickness tends to occur in poor, rural communities (Okia *et al.* 1994; Welburn *et al.* 2006) where control measures may be difficult to implement and sustain. The poor health care networks often present in these remote areas can lead to problems in the detection and treatment of sleeping sickness cases, contributing to the under-reporting of the disease. Additionally, civil unrest and conflict are known to create conditions which promote the spread and transmission of sleeping sickness (i.e. population movements, poor nutrition and increased contact with vectors), which is compounded by the further deterioration of already substandard control measures and health care systems (Berrang-Ford 2007). The diagnosis and treatment of sleeping sickness is difficult (Pepin and Milord 1994; Buscher and Lejon 2004; Burri *et al.* 2004) and can only be performed in health centres which have been provided with specialist training and equipment. Therefore, even though a patient may reach a health centre, it may not necessarily be possible to diagnose the infection accurately or provide treatment, resulting in significant service-provider delays (Odiit *et al.* 2004b).

Aside from problematic access to health care, there are various other points in the care-seeking pathway where people infected with sleeping sickness may miss the opportunity for diagnosis (and thus treatment), contributing to under-reporting (see Figure 2.1 for an illustration of the diagnosis and treatment pathway for patients including the main reasons for missed opportunities for treatment) (Odiit *et al.* 2004b; Bukachi *et al.* 2009). The initial symptoms of sleeping sickness are non-

specific and are often mistaken for several other conditions: in many cases early stage sleeping sickness cases will be treated (either via self-treatment or by a local health centre) for malaria, due to the similarity in the cyclic fevers in both diseases (Odiit *et al.* 2004b; Bukachi *et al.* 2009). Late stage infection is also frequently misdiagnosed: affected individuals and their families may assume that the illness is caused by HIV or AIDS and presume that attendance at a health facility will be fruitless (Odiit *et al.* 2004b; Bukachi *et al.* 2009). In some cases the illness may be attributed to witchcraft, with some patients visiting witch doctors or traditional healers prior to, or instead of, seeking care from the health services (Bukachi *et al.* 2009). This highlights the importance of knowledge of sleeping sickness amongst health service staff and the general population, and in the absence of community awareness of the disease it is possible for individuals to reach late stage infection without being aware of the possibility that they have sleeping sickness.

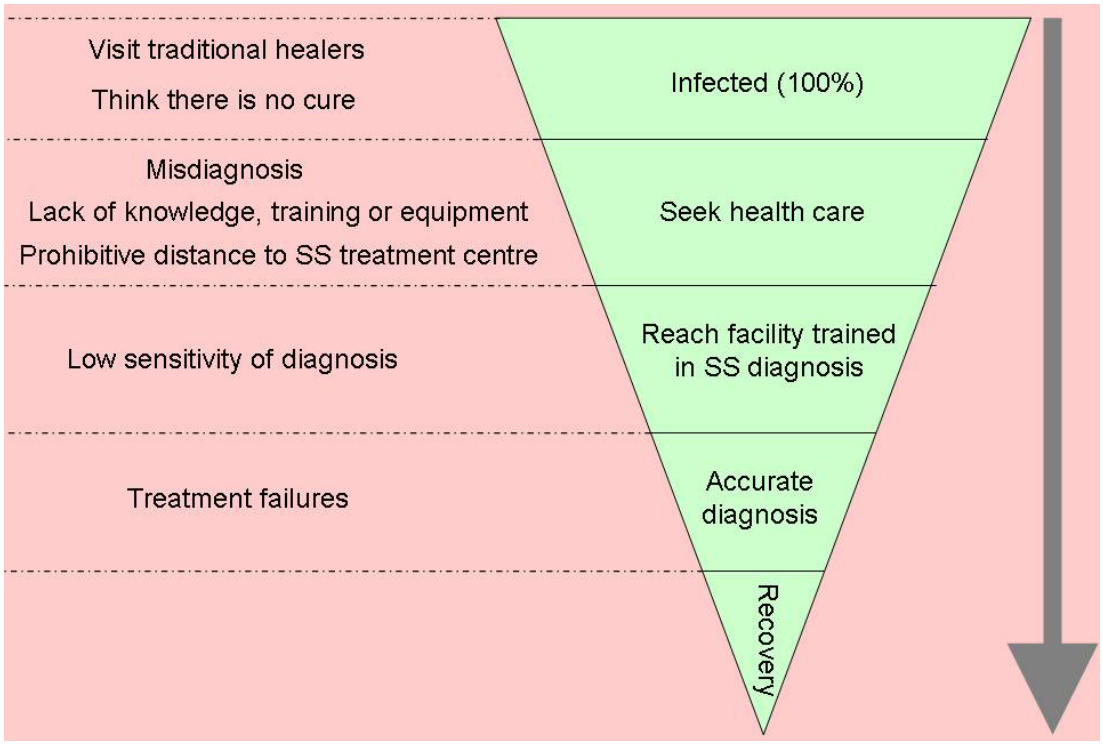


Figure 2.1: Diagnosis and treatment pathway for sleeping sickness patients, including main reasons for lost opportunity of diagnosis and treatment to the left (Cattand and De Raadt 1991; Odiit *et al.* 2004b; Bukachi *et al.* 2009).

Those affected by sleeping sickness generally live in poor, rural communities and are primarily subsistence farmers. In the event that they have been referred to the

sleeping sickness treatment centre (by a health professional, or by a member of their community) the cost of transport to the clinic which can be over 100 km away in some cases may prove prohibitive. In addition, the time required for successful treatment (approximately one month in most cases) can add a considerable burden on the individual's family in terms of time away from their home and gardens (Odiit *et al.* 2004b). These factors may add to the under-reporting of sleeping sickness, particularly in areas which are poorly served by sleeping sickness diagnostic and treatment facilities (i.e. areas remote from the closest centre trained and equipped to deal with sleeping sickness) (Odiit *et al.* 2004a).

The sensitivity of diagnostic methods may also contribute to the under-reporting of sleeping sickness. In Uganda, Rhodesian sleeping sickness is diagnosed via direct detection of the parasite in blood, lymphatic fluid or cerebrospinal fluid using microscopy. This method is known to have poor sensitivity, yet in the absence of a cheap, rapid and easy to use diagnostic test, microscopy remains the only feasible method for the diagnosis of Rhodesian sleeping sickness in poor, rural areas such as south east Uganda (Cattand and De Raadt 1991).

The under-reporting of the disease, caused by a combination of each of the factors discussed means that for every patient who is treated for sleeping sickness a number more remain in their communities, infective to the tsetse vector species and ultimately resulting in death (Odiit *et al.* 2005). Under-reporting also creates difficulties for the estimation of the population at risk of disease, the total population affected and the geographical distribution of the disease for the purposes of health service planning and provision. When considering the spatial distribution of sleeping sickness, differential rates of reporting depending on the accessibility of health services from different areas can create false impressions of the distribution, with distance to the treatment centre having been shown to influence attendance rates in previous studies (Odiit *et al.* 2004a; Odiit *et al.* 2006). This can create a confounding effect on any spatial analysis and must be considered in the methodology and interpretation of any such work.

2.1.2.2. Inaccuracies in sleeping sickness records

Record keeping in the majority of health centres and hospitals in Uganda relies on the use of hand-written patient records and log books, which are prone to inaccuracies. Problems encountered when using this type of data include illegible handwriting, spelling errors, wrongly recorded information (e.g. if the health worker mis-heard something the patient told them) and data omissions. Besides inaccuracies such as these, the collation of sleeping sickness records at the clinic level or the national level can also result in transcription errors or missed records. It is important that such data quality issues be considered, and every effort should be made to ensure the highest level of accuracy in data acquisition and collation.

2.1.2.3. Other considerations

The method of detecting sleeping sickness cases should also be considered prior to the analysis of case records to avoid bias. The majority of cases of Rhodesian sleeping sickness are detected passively (where a symptomatic person seeks care), although in the past active detection (where screening teams visit villages and actively test the population) has been used in some areas. The use of active surveillance teams results in higher rates of reporting for the areas in which they are working, as a number of infected individuals may be diagnosed who would not have otherwise sought treatment, or who may have missed the opportunity of treatment due to one of the reasons discussed in Section 2.1.2.1. Ideally, the method of detection of each case of sleeping sickness would be recorded in the sleeping sickness records (and in many areas this is the case). However, as mentioned previously, data quality issues may result in a lack of information regarding the detection method for sleeping sickness cases. For these instances, supplementary knowledge regarding the historical use of active screening in different areas can assist in the analysis and interpretation of data.

2.1.3. Village level maps of Uganda

For the purposes of spatial analysis, village level maps of the study area can be used to allow geo-referencing of disease data where the village of residence for each case is known. The most recent survey maps that identify individual village locations in Uganda at a spatial scale of 1:50,000 were carried out by the Department of Lands and Surveys from the late 1950s through to the 1970s. Uganda has seen a significant population expansion from approximately 9.5 million in 1969 to 24.4 million at the most recent census in 2002 (Uganda Bureau of Statistics 2008). In addition, since the publication of the most recent survey maps, Uganda has experienced frequent periods of civil and political unrest (Klugman *et al.* 1999). The population growth, coupled with civil strife (and the resulting temporary and permanent population movements) since the publication of the survey maps is likely to have contributed to a substantial amount of variation in the number and spatial distribution of villages during the intervening years. The names of villages are also likely to have changed since the publication of these survey maps. The differences between the village structure when the survey maps were produced and the present day discounts them as a potential source of village coordinates for the geo-referencing of sleeping sickness data.

2.2. Specific objectives

The work detailed in this chapter formed the basis of all further research within the thesis by providing the specific datasets used.

The specific aims addressed in this chapter were to develop a database suitable for the management of human sleeping sickness (and animal trypanosomiasis) data, linked to village coordinates where available and to populate this database via the acquisition and collation of sleeping sickness patient records from various sources within Uganda. The database was created to allow the straightforward entry of relevant data and enable the extraction of explicit datasets for further analysis. Previously collected sleeping sickness records and village coordinates from an area

recently affected by Rhodesian sleeping sickness were supplemented by additional data acquisition and geo-referencing. Additional datasets of relevance to tsetse and sleeping sickness distributions were also obtained for use in the subsequent analyses. An exploratory analysis was used to illustrate the spatial spread of Rhodesian sleeping sickness into the study areas since 1998.

2.3. Study areas

The research described in this thesis refers specifically to Uganda, a landlocked country of East Africa, bordered by Kenya to the east, Sudan to the north, the Democratic Republic of the Congo to the west and Lake Victoria, Tanzania and Rwanda to the south (see Figure 2.2). As discussed in Chapter 1, Section 1.2, Uganda is the only country that sustains transmission of both Gambian and Rhodesian sleeping sickness; Gambian sleeping sickness occurs in the north west of the country and Rhodesian sleeping sickness traditionally occurs in the Busoga focus in the south east of Uganda (Welburn *et al.* 2001a). However, since the mid 1980s *T. b. rhodesiense* has extended its range in Uganda and now also occurs in parts of the eastern and northern regions of Uganda (see Chapter 1, Figure 1.2) (Mbulamberi 1989; Fèvre *et al.* 2001a; Hutchinson *et al.* 2003; Picozzi *et al.* 2005; Berrang-Ford *et al.* 2006). Further information on historical sleeping sickness distributions in Uganda and the recent spread into previously unaffected areas can be found in Chapter 1, Section 1.3.

Figure 2.2: World map indicating the location of Uganda

A comprehensive database of sleeping sickness patient records from Soroti, Kaberamaido, Dokolo, Amolatar and Lira districts was created. Records were obtained from Serere hospital, Soroti district; Lwala hospital, Kaberamaido district and the National Livestock Health Research Institute (NaLIRI), Tororo district. The analyses discussed in Chapters 3 and 4 focus specifically on Kaberamaido and Dokolo districts (two of the most recently affected districts, with the first cases reported in 2004; see Figure 2.3). The analyses in these two chapters exclude Amolatar and Lira districts which have also been newly affected by Rhodesian sleeping sickness since 2004, due to the small number of cases occurring in each. The analysis discussed in Chapter 5 focuses on Soroti district (which neighbours Kaberamaido district and was first affected in 1998; see Figure 2.3).

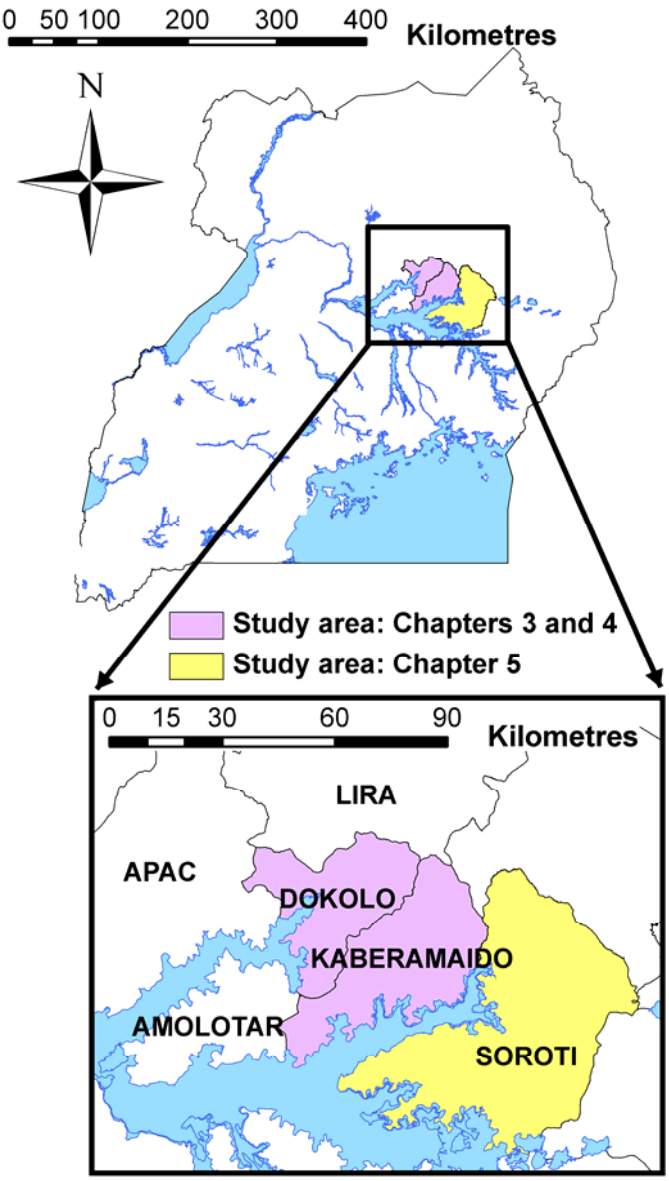


Figure 2.3: Map of Uganda highlighting the study districts; Dokolo, Kaberamaido and Soroti.

The study districts are in the Eastern (Soroti and Kaberamaido districts) and Northern (Dokolo, Amolatar and Lira districts) regions of Uganda, with Soroti, Kaberamaido, Dokolo and Amolatar all bordering Lake Kyoga. Lake Kyoga is a large, shallow lake (maximum depth of 5.7 m) with a network of rivers, streams and swamps covering much of the surrounding land (International Lake Environment Committee 2008). The populations within the study districts during the 2002 national census were approximately (rounded to the nearest 1,000); 370,000 in Soroti district; 132,000 in Kaberamaido district; 129,000 in Dokolo district; 96,000 in Amolatar

district and 515,000 in Lira district (Uganda Bureau of Statistics 2008). Primary economic activities within these five districts are subsistence farming and fishing (in areas in close proximity to the lake) (Fountain Publishers 2007). Transport networks in the study areas are of variable quality; Lira and Soroti districts have the most developed transport infrastructure, but rural areas within all five districts suffer from poor quality road networks.

2.4. Methods

2.4.1. Trypanosomiasis database

A relational database was created with data tables to hold spatial information, human sleeping sickness case data and also animal trypanosomiasis data in the Microsoft® Access software (Microsoft Corporation, Redmond, Washington). Lookup tables were included for several fields in which repeated information was to be stored (e.g. district of residence). The overall structure of the database was illustrated using an entity relationship diagram, created using the DeZign for Databases software (Datanamic Solutions BV, the Netherlands).

The database was designed around three major data storage tables: village; human sleeping sickness cases and animal trypanosomiasis data, with the village table taking the predominant role (storing all spatial information) and the two subsequent tables linking to the village table via a unique village identifier. The village table was constructed to allow storage of individual village information, including village name, the parish, district and region the village is located in, coordinates, altitude and population. A one-to-many relationship was used between the village table and the human sleeping sickness and animal trypanosomiasis tables. This was to allow the inclusion of multiple human sleeping sickness and animal trypanosomiasis records for each individual village, linked via the unique village identifier. A search facility was also added to the main data entry form to allow easy searching for specific villages when adding additional human or livestock records and to prevent village duplication.

The sleeping sickness data table was designed to allow each individual sleeping sickness patient's record to be stored, along with all data relating to that case, including gender, age, date of admission, site of admission and outcome. A free text field was also included to allow additional comments to be added. The animal trypanosomiasis table was designed to allow the entry of individual animal data, or aggregated animal sampling data (recording the numbers of animals tested for each species of trypanosome and the numbers positive) to allow flexibility. Data fields were added to allow additional information such as the testing methods used to be included.

2.4.2. Acquisition of sleeping sickness data

No patient names were recorded within the database or as part of the data acquisition process to maintain patient confidentiality and to adhere to the international ethical guidelines for biomedical research involving human subjects.

2.4.2.1. Kaberamaido, Dokolo, Amolatar and Lira districts sleeping sickness data

The sleeping sickness treatment centres serving Kaberamaido, Dokolo, Amolatar, Lira and Soroti districts; Lwala Hospital, Kaberamaido district and Serere Hospital (formerly Serere Health Centre), Soroti district, were visited between June 2007 and June 2008. During meetings with the medical records staff and laboratory personnel at these clinics, copies of the sleeping sickness patient log books were obtained with details including the patient's age, sex, village of residence, date of admission and stage of infection. Each sleeping sickness case record was entered into the trypanosomiasis database, linked via a unique identifier to the case's village of residence (the village of residence was entered at this point if not already present in the database – see also Section 2.4.3 for information regarding the collection of village information and coordinates). The patient records from the NaLIRI clinic (in Tororo district) were also examined and any cases resident in Kaberamaido, Dokolo,

Amolotar or Lira districts were obtained and entered into the trypanosomiasis database.

2.4.2.2. Soroti district sleeping sickness case-control data

Previously acquired Rhodesian sleeping sickness case records from Soroti district (collected from Serere Hospital), as published by Fèvre *et al* (Fèvre *et al.* 2001a) were obtained. The case records covered the time period December 1998 to November 2002 and included details of each patient's age, sex, date of admission and village of residence. The records were reformatted and imported into the trypanosomiasis database.

In addition to the sleeping sickness case records from Soroti district, control data were also obtained. Controls were selected from the hospital inpatient, outpatient, tuberculosis and maternity ward records (in this order of preference). Any patients with a primary diagnosis of a vector-borne disease were excluded to prevent a spatial bias in the results, which may arise due to the similarities between tsetse habitat and other vector habitats. One control was selected for every case, matched on age group (<1, 1-9, 10-14, 15-19, 20-49, 50-64 and ≥65 years), sex and month of admission. The control data were stored separately, in a Microsoft® Excel (Microsoft Corporation, Redmond, Washington) spreadsheet. Additional information regarding the collection of this dataset, the case-control study design and the matching of controls can be found in Fèvre *et al* (Fèvre *et al.* 2001a) and Chapter 5.

The sleeping sickness data for Soroti district were supplemented via the collection of patient records covering the period between 2002 and the time of collection (2008). These were added manually to the trypanosomiasis database. Additionally, patient records from the NaLIRI clinic were examined for any cases resident in Soroti district. The case records collected as mentioned in section 2.4.2.1, the previously collected sleeping sickness data and the supplementary case records were conflated into one data holding using the trypanosomiasis database.

2.4.3. Village geo-referencing

2.4.3.1. Kaberamaido and Dokolo districts

All villages within Kaberamaido and Dokolo districts were geo-referenced using a hand held GPS (Garmin, Olathe, KS) between February and June 2008. Comprehensive and up-to-date lists of village names were obtained for each parish within Dokolo and Kaberamaido districts from the Uganda Bureau of Statistics and by meeting with the LCIII (Sub County) chairpersons. Local guides were recruited from the Sub County local government offices to assist with geo-referencing: these were mainly LCI (village), LCII (parish) or LCIII chairpersons with a good knowledge of the villages and routes in the areas to be covered. For each village, the coordinates were recorded at a location in the centre of the village: this was described to each of the guides as being 'the middle of the village, so that the village surrounds you on all sides' for consistency.

Village coordinates were saved in decimal degrees, using the World Geodetic System 84 (WGS 84) datum with no projection. The coordinates were exported from the GPS unit into the Geocaching Swiss Army Knife software (available from http://www.gsak.net/index.php) for waypoint management. The details of all villages within these districts (coordinates, parish, sub-county, county and district) were then formatted and imported into the trypanosomiasis database village table. The village information and coordinates were linked to the sleeping sickness case records via a unique identifier as discussed in Section 2.4.1.

2.4.3.2. Soroti district geo-referencing

The previously collected patient data from Soroti district also contained village coordinates for all sleeping sickness cases and their matched controls. A hand-held GPS was used to record the location at each village's central meeting point, as determined by the village chairperson. These coordinates were provided in decimal degrees with the datum WGS 84 and no projection. For further information on the geo-referencing of case and controls in Soroti district see Fèvre *et al* (Fèvre *et al.*

2001a). The village details from this dataset along with each village's coordinates were reformatted and imported into the trypanosomiasis database.

2.4.4. Collection of supplementary data

Supplementary data sources for factors thought to contribute to the spatial distribution of tsetse and sleeping sickness within the study area were collected from various sources. These consisted of continuous surface (raster), point and line (polygon) data. See Tables 2.1 and 2.2 for a list of the continuous data surfaces which were obtained along with their sources.

2.4.4.1. Fourier processed Advanced Very High Resolution Radiometer data

Several temporal Fourier-processed indices were obtained from Advanced Very High Resolution Radiometer (AVHRR) imagery: land surface temperature (LST), normalised difference vegetation index (NDVI) and middle-infrared (MIR, AVHRR channel 3). NDVI is a measure of the amount of green vegetation (Tucker 1979) and reflectance in the MIR band has also been linked to vegetation cover (Boyd *et al.* 1999). Both vegetation cover (in terms of suitable tsetse habitat) and temperature have been shown to influence the distribution of sleeping sickness (Rogers *et al.* 1996). Temporal Fourier processing reduces the number of data to be processed by eliminating redundancy and characterising seasonality. The minimum, mean, maximum, phase (the timing of the cycle) and amplitude (the amount of variation around the mean) of the annual and biannual cycles were used for each of LST, NDVI and MIR. These indices were calculated using raw data from the AVHRR sensor onboard the National Oceanographic and Atmospheric Administration (NOAA) satellites at a 1 km spatial resolution. The raw data are available as decads (10 day units) from April to December 1992, January to September 1993, February to December 1995 and January to April 1996. Full details regarding the data used and the Fourier analysis can be found in Hay *et al* (2006).

2.4.4.2. Landsat images and NDVI

Three Landsat ETM+ images (path 171, row 59) from 27th January, 17th April and 27th November 2001 were downloaded from the U.S. Geological Survey global visualisation viewer for use in a land cover classification, as detailed in Chapter 5. The images were all level 1 terrain (corrected) products (level 1T), which have undergone radiometric and geometric correction. Landsat images have a spatial resolution of 30 m for all bands apart from band 6 (60 m) and the panchromatic band (15 m). Further information on the Landsat ETM+ sensor is provided in Chapter 5, Section 5.1.2. NDVI was also calculated using the red and near-infrared wavebands of the November 2001 Landsat ETM+ image using the following formula: NDVI = (near-infrared – red)/(near-infrared + red) (Tucker 1979).

2.4.4.3. Population density

Population density at a 30 arc second (approximately 1 km) spatial resolution was obtained from the Landscan™ 2006 global population dataset (Oak Ridge National Laboratory 2006). Increasing population density has previously been demonstrated to have a protective effect against sleeping sickness, as areas with higher population densities have increased levels of human disturbance of potential tsetse habitats and, therefore, reduced tsetse populations and lower levels of transmission (Okia *et al.* 1994; Reid *et al.* 2000; Odiit *et al.* 2006). The global population dataset consists of population counts on a 30 arc second by 30 arc second spatial grid, calculated using sub-national census data and a number of likelihood coefficients such as land cover, nighttime lights and slope (Dobson *et al.* 2000; Oak Ridge National Laboratory 2006). An additional grid file containing the area of each cell was also obtained with the data to allow the calculation of population density per squared kilometre as follows:

Population density = grid population count / grid area

2.4.4.4. Predicted tsetse suitability

A continuous surface of predicted suitability for different tsetse species within Uganda at a 1 km spatial resolution was obtained from the FAO (Wint and Rogers 2000). This coverage was created using a logistic regression analysis of fly presence data and several environmental covariates, including remotely sensed climatic and environmental indices. The predicted suitability for *G. f. fuscipes* was acquired; this is the predominant vector species responsible for transmission of Rhodesian sleeping sickness within the study areas in Uganda. Predicted suitability for *G. pallidipes* and *G. morsitans* were also obtained, although these are not thought to be present within the study areas. Full details of the methods used for these predictions can be found in Wint and Rogers (2000).

2.4.4.5. Elevation

Elevation data for Uganda was acquired from the Shuttle Radar Topography Mission (SRTM) at a 3 arc second spatial resolution (approximately 100 m) (US Geological Survey 2006). Previous studies have illustrated the constraining effect increasing elevation can have on tsetse and sleeping sickness distributions (Rogers and Randolph 1993; Robinson *et al.* 1997b). However, the scale of this effect within the study area is likely to be smaller and the biological interpretation of any relationship detected different, due to the relatively small changes in elevation across the study area.

2.4.4.6. Nighttime lights of the world

The Global Nighttime Lights of the World dataset was obtained from the Defence Meteorological Satellite Program as a proxy measure of poverty (Defence Meteorological Satellite Program 2004). This dataset uses an operational linescan system to observe sources of near infra-red emissions at the Earth's surface, which may be generated by cities, towns and villages. The system captures a night time image of lights at a 1 km spatial resolution. It has previously been shown that this

data relates to measures of poverty when linked with population data and hence may be useful as a proxy for poverty in areas for which reliable poverty data is not available, or where poverty data is not available at the required spatial resolution (Noor *et al.* 2008b).

2.4.4.7. Health centre and livestock market locations

A list of all hospitals and health centres in Uganda was obtained from the Ugandan Ministry of Health. Coordinates for health centres and hospitals in Lira, Kaberamaido, Dokolo, Amolatar, Apac and Soroti districts were obtained from the Ugandan Ministry of Health and by using a handheld GPS. Meetings with COCTU staff were used to obtain a comprehensive list of the facilities which were trained and equipped for the diagnosis and treatment of sleeping sickness (A. Kakembo, pers. com.).

The coordinates of livestock markets within Lira, Dokolo, Kaberamaido, Amolatar and Soroti districts were also recorded using a handheld GPS during fieldwork in February 2008. All coordinates were collected and stored in decimal degrees using the WGS 84 datum and no projection.

2.4.4.8. Village population data

Village population data, collected during the most recent national census (2002), were obtained for Lira, Dokolo, Amolatar and Kaberamaido districts from the Uganda Bureau of Statistics (Uganda Bureau of Statistics 2008). These supplementary data (number of households, male population, female population and total population) were added to the village data table in the trypanosomiasis database.

2.4.4.9. Landcover data

Landcover data for south east Uganda were obtained from the National Biomass Survey, which was conducted by the Uganda Forest Department between 1995 and

2002 (Forest Department 2002). Remote sensing images were used to carry out landcover classifications, with resulting coverage of 13 different landcover classes: deciduous plantation or woodlot; coniferous plantation or woodlot; tropical high forest (fully stocked); tropical high forest (depleted); woodland; bushland; grasslands and savannah; wetland; subsistence farmland; uniform farmland; urban or rural built-up areas; open water and impediments (e.g. barren soil or bare rock). These classifications were the result of a quantitative interpretation of remotely sensed images along with ground data and supplementary data layers and, thus, their accuracy may be variable (Forest Department 2002). An additional shapefile was obtained that delineated gazetted land: land held in trust by the Ugandan Government for the people of Uganda (generally forest reserves or national parks).

	Source	Scaling required	Units
NDVI phase of biannual cycle	AVHRR	(x/100)	Months
NDVI phase of annual cycle	AVHRR	(x/100)	Months
Maximum NDVI	AVHRR	(x/1000) – 1	No units
Minimum NDVI	AVHRR	(x/1000) – 1	No units
NDVI biannual amplitude	AVHRR	(x/1000)	No units
NDVI annual amplitude	AVHRR	(x/1000)	No units
Mean NDVI	AVHRR	(x/1000) – 1	No units
LST phase of biannual cycle	AVHRR	(x/100)	Months
LST phase of annual cycle	AVHRR	(x/100)	Months
Maximum LST	AVHRR	(x/10) - 273	°C
Minimum LST	AVHRR	(x/10) - 273	°C
LST biannual amplitude	AVHRR	(x/10)	°C
LST annual amplitude	AVHRR	(x/10)	°C
Mean LST	AVHRR	(x/10) - 273	°C
MIR phase of biannual cycle	AVHRR	(x/100)	Months
MIR phase of annual cycle	AVHRR	(x/100)	Months
Maximum MIR	AVHRR	(x/10) - 273	°C
Minimum MIR	AVHRR	(x/10) - 273	°C
MIR biannual amplitude	AVHRR	(x/10)	°C
MIR annual amplitude	AVHRR	(x/10)	°C
Mean MIR	AVHRR	(x/10) - 273	°C

Table 2.1: AVHRR indices obtained (Hay *et al.* 2006)

	Source	Scaling/Calculations required	Units
Landsat ETM+ Bands 1-7	Landsat ETM+		Radiance
NDVI	Landsat ETM+	(Band 4 – Band 3) /(Band 4 + Band 3)	No units
Population density	Landscan (Oak Ridge National Laboratory 2006)	Population/area	Population per km^2
Suitability for G. fuscipes	Predicted tsetse suitability (Wint and Rogers 2000)		%
Elevation	SRTM (US Geological Survey 2006)		Metres
Night time lights	Nighttime lights of the world (Defence Meteorological Satellite Program 2004)		%
Landcover	National Biomass survey (Forest Department 2002)		No units

Table 2.2: Additional supplementary datasets obtained.

2.5. Results

2.5.1. Trypanosomiasis database

As described in Section 2.4.1, the structure of the database was based on three main data tables storing village information, human sleeping sickness case records and

animal trypanosomiasis data. Each record in the human sleeping sickness and animal trypanosomiasis tables was linked explicitly to the village information table using unique village ID numbers, thus, allowing the spatial visualisation and analysis of the data. Additional lookup tables were used to store data which are frequently repeated in the records (for example gender). The animal trypanosomiasis data storage tables were included within the database to allow multi-functionality of the database and to provide a single data holding for all available data. However, the spatial and temporal coverage of the available animal trypanosomiasis data was not adequate for inclusion in the analyses, and so is not discussed.

An entity relationship diagram created using DeZign for Databases illustrates the main structure of the database (Figure 2.4). The majority of table attributes (fields) and lookup tables have been omitted from this diagram for simplicity. Forms were also incorporated into the design of the database to allow a user-friendly interface and to simplify the process of data entry. A screen shot of the main data entry form can be seen in Figure 2.5. For full details of the structure of the trypanosomiasis database, along with field names, definitions, and relationships, see Appendix A.

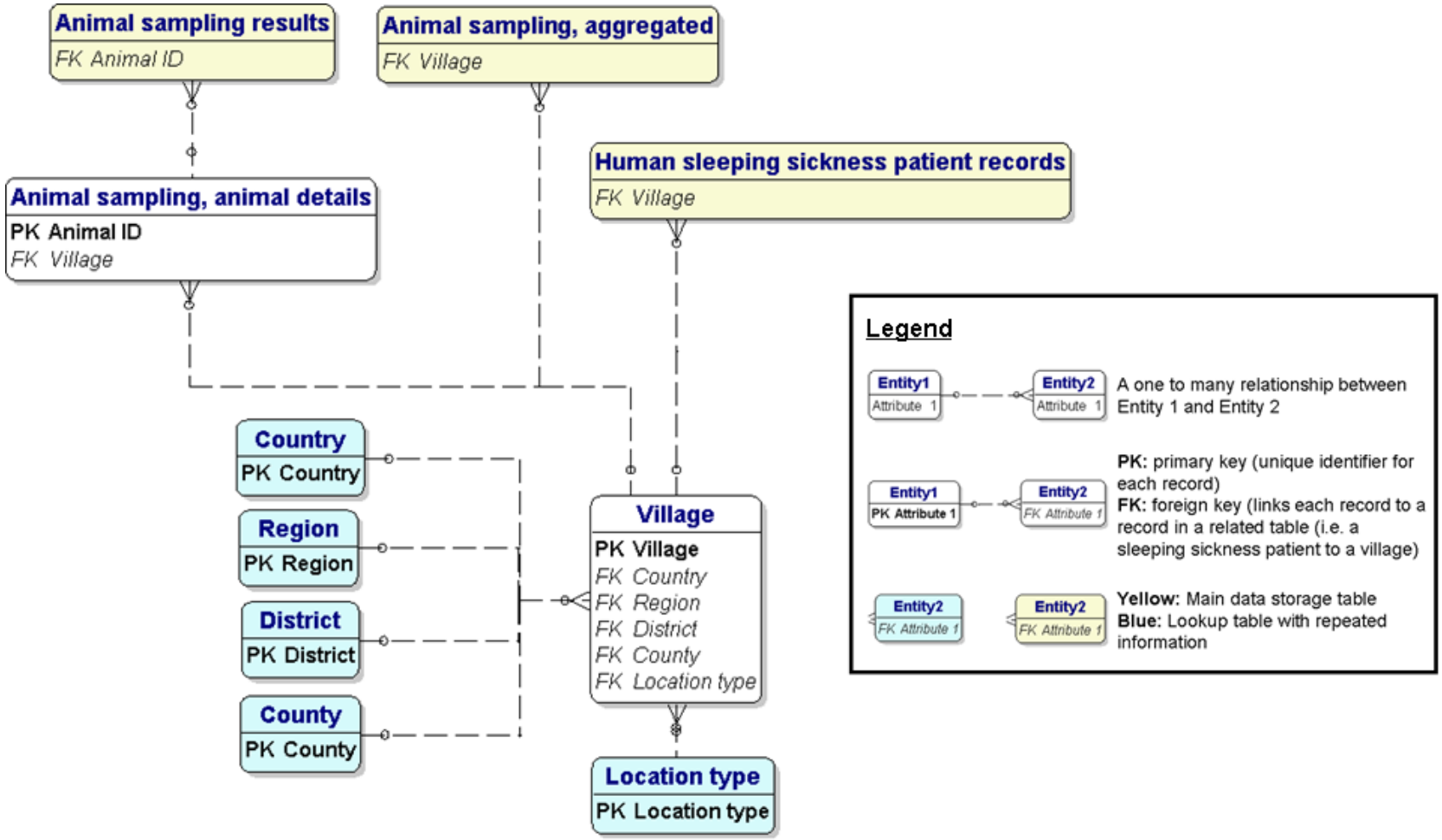


Figure 2.4. Entity relationship diagram representing the main data tables and relationships in the trypanosomiasis database.

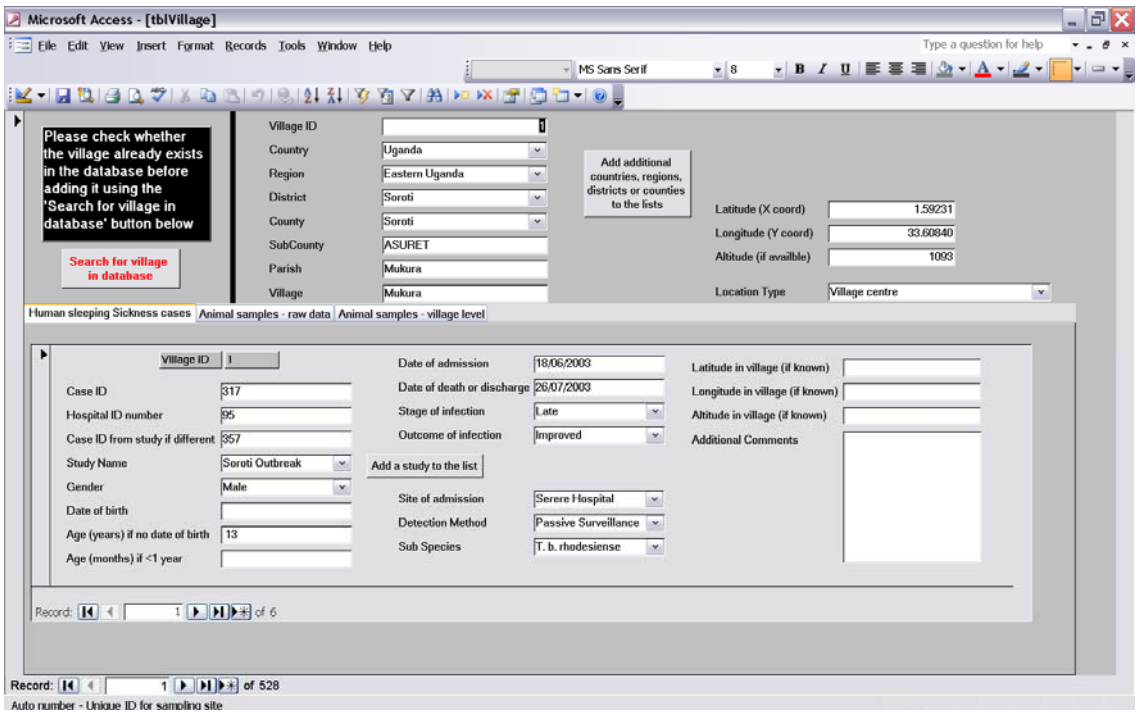


Figure 2.5. Main data entry form in the trypanosomiasis database. Village information is in the top half of the screen shot, and related human sleeping sickness case data in the lower half of the screen.

2.5.2. Sleeping sickness data

The review of Rhodesian sleeping sickness case records allowed the creation of a comprehensive database of sleeping sickness cases occurring within the study areas. Records are available from the first reported cases in Soroti district in 1998, through the spread of the disease into Kaberamaido and Dokolo districts in 2003 and 2004, up to the end of 2008 when the most recent fieldwork was undertaken. Table 2.3 shows the total annual case counts by district of residence from 1998 to 2008, and also highlights the number of cases which could not be matched to their parish of residence. The problems with matching to parish may have resulted from missing data in the patient records or errors in the recording of patient details in the hospital or during transcription to the database. The majority of cases occurred in Soroti, Kaberamaido and Dokolo districts, with only one case reported from Amolatar district (in 2007) and 17 cases from Lira district between 2004 and 2008.

A seasonal trend in the monthly sleeping sickness case counts can be observed in Soroti district (see Figure 2.6) with pronounced peaks during January and February of each year except 2004. The trend is less clear in Kaberamaido, Dokolo, Lira and Amolatar districts (see Figure 2.7) although peaks can be seen during January in 2005, 2006 and 2008. In relation to precipitation trends within the study area, the long rains normally run between March and July, and the short rains from September to November, with a dry season from December to April. The January and February peak in sleeping sickness cases may be linked to the periods of increased rainfall the preceding year. Following periods of rainfall swamps, streams and seasonally flooding areas are wet, thus, allowing the extension of the spatial range of tsetse. This in turn results in increased contact between tsetse, livestock and humans which can promote the transmission of sleeping sickness (Food and Agricultural Organization of the United Nations 1982).

District	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Unknown year
Soroti	2	72 (1)	54 (2)	47	88	121	54 (1)	86 (3)	31 (1)	29 (2)	24 (2)	0
Kaberamaido						1	114 (1)	124 (1)	41	36 (1)	35 (1)	0
Dokolo							17 (3)	41 (3)	18 (4)	18 (1)	22 (3)	1
Amolatar										1		0
Lira							2	3 (1)	11		1 (1)	0

Table 2.3: Annual case counts, aggregated to district of residence, for Soroti, Kaberamaido, Dokolo, Lira and Amolatar districts, 1998 to 2008. Values in brackets indicate the number of cases which could not be matched to the parish of residence due to recording errors. Shaded areas represent time periods were no diagnostic and treatment facilities were available within the direct vicinity (i.e. within that or a neighbouring district)

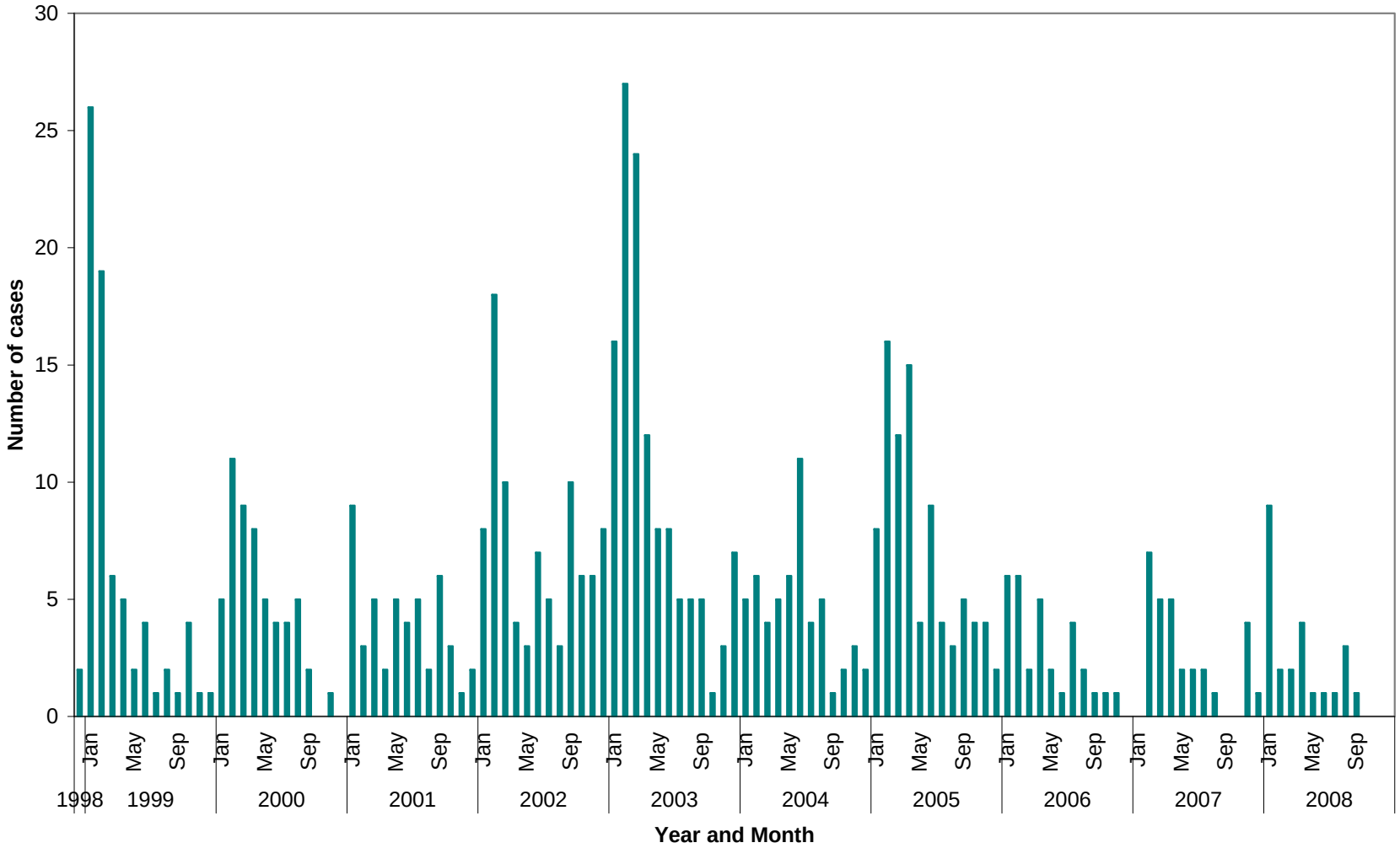


Figure 2.6: Monthly sleeping sickness case counts from Soroti district, December 1998 to September 2008.

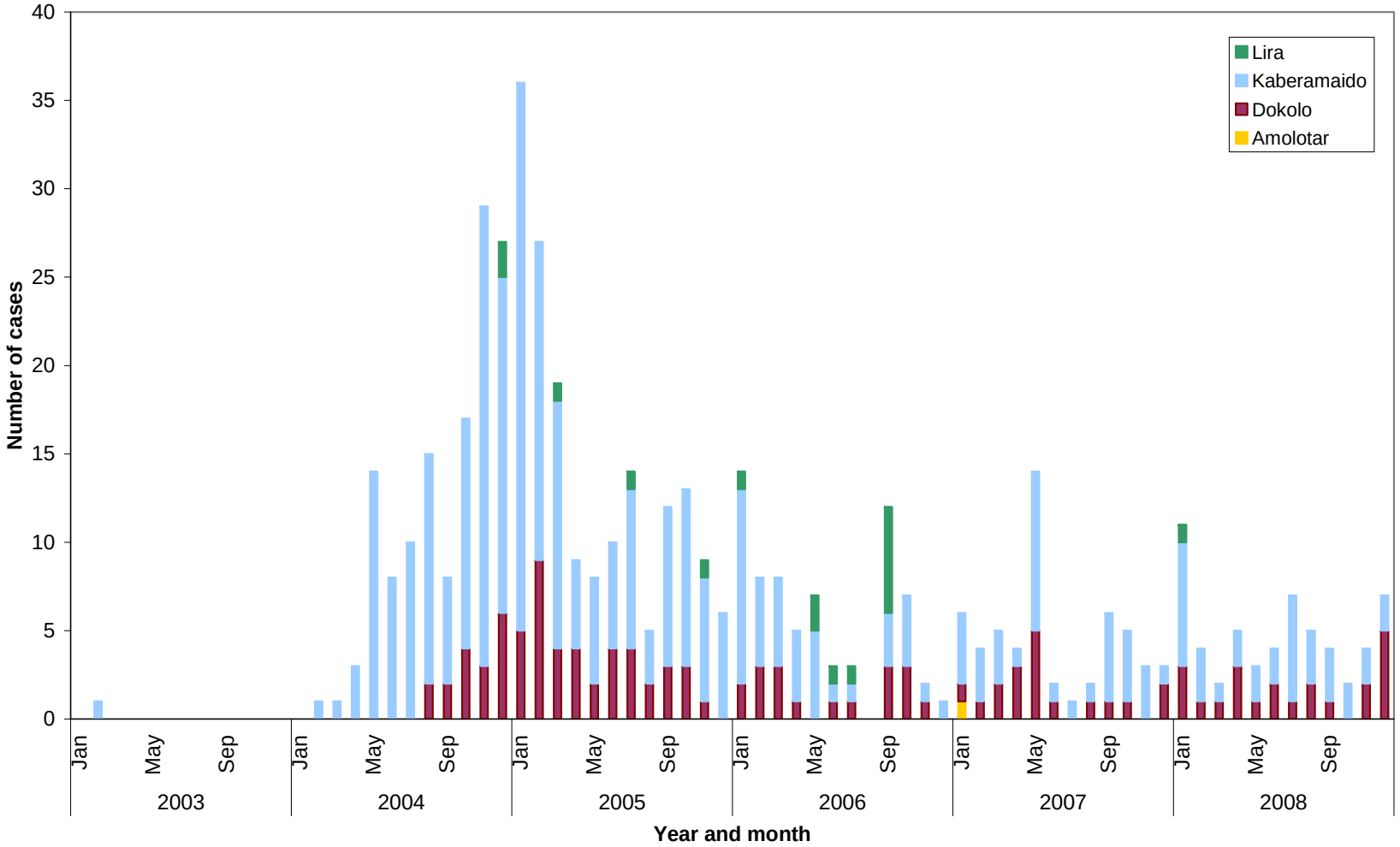


Figure 2.7: Monthly sleeping sickness case counts from Kaberamaido, Dokolo, Amolatar and Lira districts, January 2003 to December 2008.

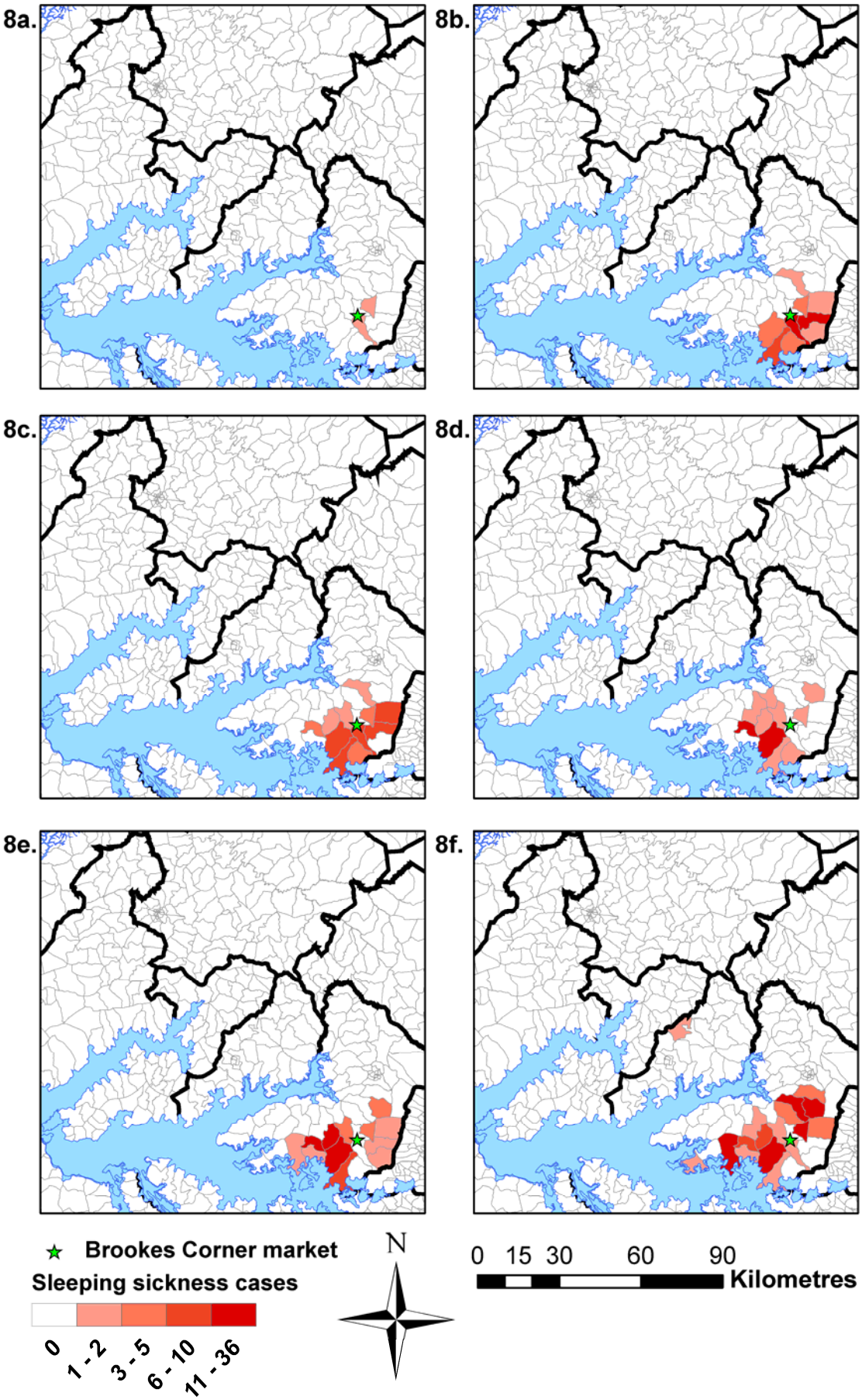
The spatial visualisation of Rhodesian sleeping sickness cases within the study districts over time demonstrates the spatio-temporal evolution of the distribution of disease (see Figure 2.8). It has been demonstrated previously that the parasite was introduced initially to Soroti district via the movement of infected cattle through Brookes Corner livestock market (see Figure 2.8 for market location) (Fèvre *et al.* 2001a). The significant spatial clustering of cases around the livestock market has been described by Fèvre *et al* (Fèvre *et al.* 2001a); this clustering can be observed from 1998 to 2001 (Figures 2.8a to 2.8d), with subsequent dispersion away from the livestock market. The distribution, clustering and dispersion can be seen more clearly using village point data (see Chapter 5). Over time, although the disease appears to spread and disperse within the district, a residual focus remains in the parishes closest to the livestock market with increased case numbers in this area in 2003 and 2005.

In February 2003, a single case of sleeping sickness from Kaberamaido district was reported at the NaLIRI sleeping sickness treatment centre; more than 150 km (Euclidean distance) from the district (see Figure 2.8f). From February 2004 onwards, increasing numbers of cases were recorded from Kaberamaido district (diagnosed at Serere hospital in Soroti district) and from August 2004 the disease had also spread to Dokolo district (see figures 2.8g to 2.8k). The affected areas in Kaberamaido and Dokolo districts were not contiguous with areas affected within Soroti district; the minimum distance between affected districts was approximately 12 km, across a main finger of Lake Kyoga.

The subsequent distribution of sleeping sickness within Kaberamaido and Dokolo districts extended outwards, with a clear focus of infection along the border between the two districts throughout the time period (see Figures 2.8f to 2.8k). The single case reported from Amolatar district can be seen in Figure 2.8j, and the cases residing in Lira district in Figures 2.8g to 2.8i (note there was a case reported from Lira district in 2008 but this could not be matched to parish of residence and so could not be mapped). Similarly to the spread of disease into Kaberamaido and Dokolo districts, the parishes affected in Amolatar and Lira were not adjacent to any other

affected parishes (minimum distance from another affected parish was approximately 26 km for the Amolatar case and 22 km for the cases in Lira).

As discussed in Chapter 1, Section 1.2 a large scale control programme involving the mass treatment of livestock in Kaberamaido, Dokolo, Amolatar, Lira and Apac (to the west of Lira and Dokolo, see Figure 2.3) with trypanocidal drugs was instigated in November 2006 (Ceva Sante Animal *et al.* 2006). By reducing the prevalence of *T. b. rhodesiense* in the livestock reservoir, this programme aimed to reduce the burden of sleeping sickness within treated areas, prevent any further spread of the disease and ultimately, prevent the overlap of Rhodesian and Gambian sleeping sickness (Ceva Sante Animal *et al.* 2006). Figures 2.8i, 2.8j and 2.8k illustrate a reduced number of cases annually in 2006, 2007 and 2008. Some of the observed decrease in cases after 2006 may be attributed to the control programme (see also Table 2.3 and Figure 2.7) and therefore data from 2007 and 2008 were excluded from all analyses.


8a.
8b.
8c.
8d.
8e.
8f.
Brookes Corner market
Sleeping sickness cases
0
1 - 2
3 - 5
6 - 10
11 - 36
N
0 15 30 60 90
Kilometres

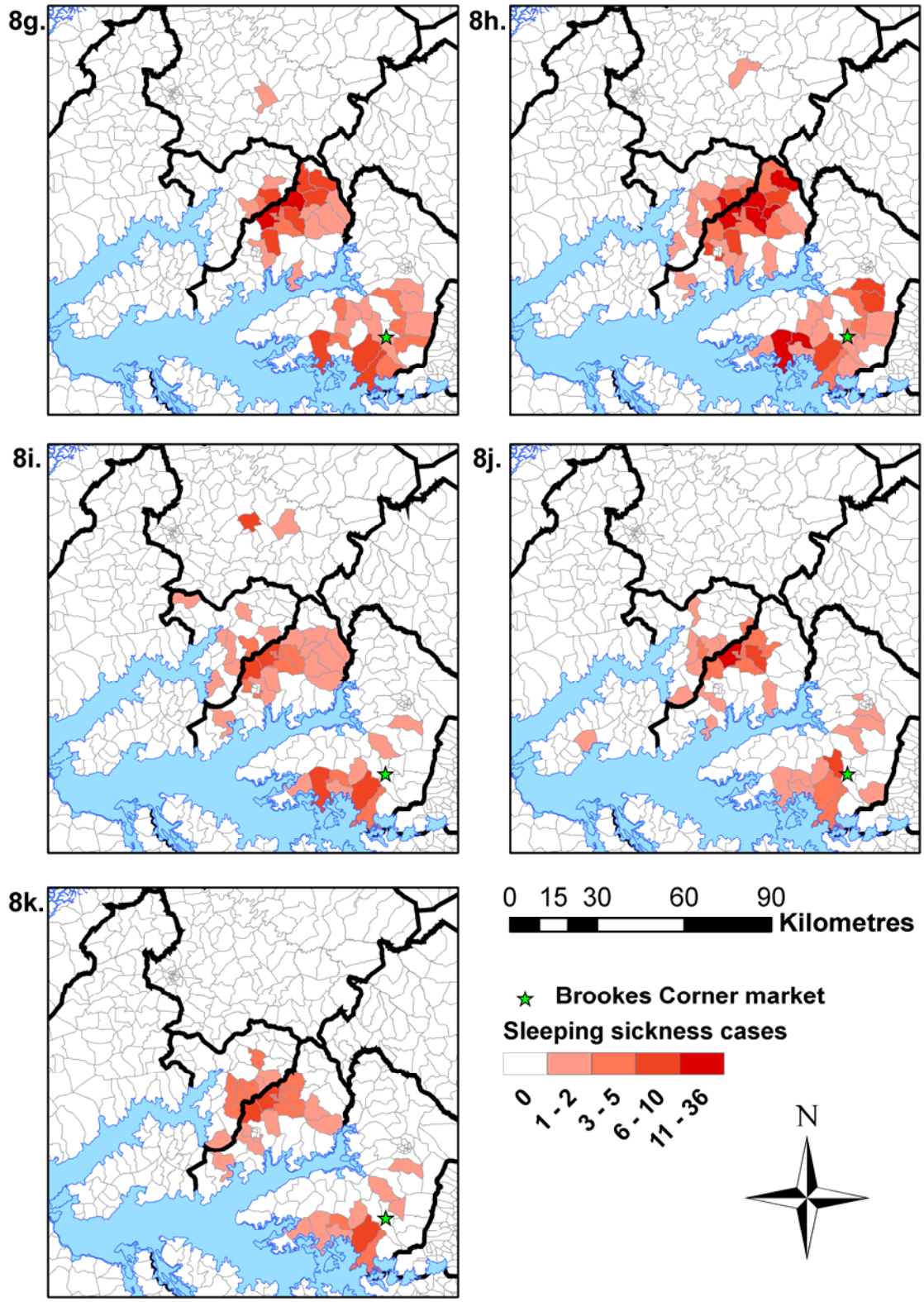


Figure 2.8: Annual parish level Rhodesian sleeping sickness case counts for Soroti, Kaberamaido, Dokolo, Amolatar and Lira districts. Years 1998 (2.8a); 1999 (2.8b); 2000 (2.8c); 2001 (2.8d); 2002 (2.8e); 2003 (2.8f); 2004 (2.8g); 2005 (2.8h); 2006 (2.8i); 2007 (2.8j) and 2008 (2.8k).

2.6. Discussion

The Rhodesian sleeping sickness case records obtained illustrate the spatial and temporal distribution of Rhodesian sleeping sickness in Soroti, Kaberamaido, Dokolo, Lira and Amolatar districts since the initial introduction to Soroti district in 1998. Prior to 1998, these districts were free from sleeping sickness (discounting a single case in Soroti district in the 1960s which was thought to have been acquired outside of the district (Hutchinson *et al.* 2003)), but transmission now appears to be established within Soroti, Kaberamaido and Dokolo districts.

The single case reported from Amolatar district is most likely to have been acquired outside of the district. The reported case was resident in a parish which was not adjacent to any other affected areas, suggesting either the importation of the parasite via human or livestock movements as opposed to spread via vector movements. In the absence of any further cases from Amolatar it is more probable that the infection was acquired outside of the district. The cases occurring in Lira district were again in areas not contiguous with areas already affected by the parasite. The number of cases throughout the time period (17 in total, occurring in 4 adjoining parishes) indicates that active transmission may have been occurring in this area as opposed to the acquisition of disease outside the district. The distance from the closest areas which had previously been affected by sleeping sickness would suggest that the spread of disease into Lira district was brought about by livestock or population movements rather than the movement of the tsetse vector. The low number of cases reported from Lira district leads to the question; if active transmission of sleeping sickness was occurring in Lira district, why was incidence so low when compared with that observed in Kaberamaido and Dokolo districts? It is possible that tsetse abundance in this district was lower than in Kaberamaido and Dokolo districts; Figure 1.3 illustrates that Lira district lies on the boundary of the predicted spatial range of the predominant vector species, *G. f. fuscipes*. The lack of sleeping sickness diagnostic and treatment facilities within the district may also contribute to the low observed incidence of sleeping sickness in this area. It is likely that the prevalence in livestock was higher than the human prevalence of infection, with a low level of transmission

being maintained in the cattle population and an occasional sporadic infection occurring in humans. The regular assessment of the spatial distribution of sleeping sickness within the area, using hospital records, would be beneficial to ensure that the area does not experience a resurgence in the number of cases in the future.

The spread of Rhodesian sleeping sickness, as illustrated, has extended the focus of Rhodesian sleeping sickness by approximately 130 km, and has narrowed the gap between areas of Rhodesian and Gambian sleeping sickness transmission. The narrowing of the buffer area poses an immediate risk to communities in this area and endangers the utility of sleeping sickness diagnosis and treatment facilities in the potential overlap area. Further spread of Rhodesian sleeping sickness (or Gambian sleeping sickness) may narrow the buffer zone further, or result in a concurrence of the two forms of sleeping sickness. Within any area of concurrent infection, diagnostic and treatment options will be severely compromised.

The further spread of Rhodesian sleeping sickness should be prevented in order to prevent a future overlap with the Gambian form of the disease. To enable the most appropriate control measures to be established, improved knowledge regarding the spatio-temporal dynamics of the disease, reasons for disease spread and the environmental, climatic and social aspects of disease distribution is required. Such knowledge may allow more effective targeting of resources to prevent a future concurrence of the two forms of sleeping sickness and the devastating consequences that this would entail.

Chapter 3. Investigating the spatial distribution of *T. b. rhodesiense* in a newly affected area of Uganda using logistic regression

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

William Lawrence Bragg (1890 – 1971)

3.1. Introduction

3.1.1. The spread of Rhodesian sleeping sickness in Uganda

As discussed in Chapter 1, Section 1.3, Uganda has experienced a resurgence of sleeping sickness in the past two decades. Since Rhodesian sleeping sickness was introduced into Tororo District in 1987, the disease has spread persistently north west into previously unaffected areas of Uganda (Picozzi *et al.* 2005; Fèvre *et al.* 2005). The further spread into Kaberamaido, Dokolo, Lira and Amolatar districts from 2003 onwards has heightened concerns over the potential future overlap with the Gambian form of the disease and raised questions regarding the factors governing the continuing spread of the disease. Following the most recent expansion of the area affected by Rhodesian sleeping sickness, the buffer between active foci of Rhodesian and Gambian sleeping sickness has narrowed, with an estimated 150 km now separating the two forms of the disease (Picozzi *et al.* 2005).

Evidence suggests that the introduction of Rhodesian sleeping sickness into Soroti district in 1998 can be attributed to the movement of untreated cattle from endemic areas through the local livestock market (Fèvre *et al.* 2001a; Welburn *et al.* 2001b). Since this finding, the Government of Uganda has introduced new regulations requiring the treatment (with trypanocidal drugs) of all cattle from *T. b. rhodesiense* endemic areas being sold at markets (Wendo 2002), although adherence to these regulations is believed to be less than 100% (R. Selby, pers. com.). The further spread since the introduction of these regulations (from 2003 onwards) has raised concerns over their implementation and stimulated the creation of a public-private partnership to reduce the prevalence of Rhodesian sleeping sickness in the livestock reservoir, as discussed in Chapter 1, Section 1.2 (Kabasa 2007).

A robust evidence base is required to allow the most effective allocation of resources and the targeting of control efforts. Priorities with regards to the control of sleeping sickness and the prevention of a future overlap between Rhodesian and Gambian sleeping sickness include: establishing the mechanism of continuing spread of

Rhodesian sleeping sickness; assessing the likelihood of further spread; quantifying the risk of future concurrence and highlighting priority areas for control activities to prevent further spread and a future overlap. To meet each of these needs an enhanced understanding of the factors involved in the disease's spatial distribution and spread is required.

The spatial distribution of sleeping sickness is driven by complex interactions of many factors. The occurrence of disease in an area is dependent on the establishment of disease transmission, which in turn is reliant on the suitability of an area for the disease vectors. Within affected areas, a spatially varying intensity of transmission can result in the heterogeneous village level prevalence of disease. These two processes giving rise to i) the establishment of sleeping sickness transmission and ii) the heterogeneous prevalence of sleeping sickness in an area are likely to be driven by different environmental, climatic and social factors associated with the presence and density of tsetse flies (Rogers and Randolph 1986; Rogers 1988; Rogers and Williams 1993; Berrang-Ford *et al.* 2006), the introduction of the parasite (Fèvre *et al.* 2001a; Fèvre *et al.* 2005), the presence of reservoir host species and the frequency of human-fly contact (Courtin *et al.* 2005). However, the factors that control the heterogeneous distribution of sleeping sickness within small areas (i.e. between village variation at a district level) are poorly understood, though this knowledge would be of practical use for the targeting of control efforts and the prevention of further spread.

3.1.2. Regression analysis in spatial epidemiology

The purposes of epidemiological studies can be described as clarifying the etiology of disease, evaluating etiological hypotheses, providing an evidence base for the implementation of preventative or control measures and assessing the impact of interventions (Lilienfeld and Stolley 1994). There is a wide array of qualitative and quantitative methods available for the analysis of epidemiological data: a commonly used quantitative method is regression analysis. Regression methods statistically model the distribution of a dependent outcome variable (e.g. weight) as a function of one or more independent variables (covariates; e.g. age) (Fox 1997). The use of such

methods can allow an assessment of the influence the covariates have on the outcome variable (e.g. how does age affect weight?), following which, known values of the covariates can be used to predict the outcome variable (Fox 1997). Regression analysis can also be applied to spatial data; where the outcome variables have a known spatial distribution and the covariate data are obtained from the same locations (implicitly spatial), or with the inclusion of the spatial locations of the data observations in the analysis (explicitly spatial), thus taking into consideration spatial autocorrelation (where observations in close proximity to one another are more similar than observations further from one another) (Pfeiffer *et al.* 2008b). Chapter 4, Section 4.1.1 provides further information regarding the explicit inclusion of spatial information in regression analyses. The primary purposes of regression analysis in spatial epidemiology are normally to quantify the effects of spatial covariates on the outcome variable and to produce a spatially predictive model. Hypothesis testing may also be carried out, for example, to test the hypothesis of a specific source of infection.

3.1.2.1. Assessing geographical correlation and spatial risk factors

As discussed in Chapter 1, Section 1.4, the spatial distributions of many infectious diseases are linked to specific factors which can also be measured on a geographical scale. Examples include precipitation, temperature, land cover type, population density and proximity to landscape features (e.g. a river) (Bergquist 2001; Elliott and Wartenberg 2004; Pfeiffer *et al.* 2008b), each of which may affect the spatial distribution of disease agents, vectors or hosts, or influence the probability of interaction between them (Graham *et al.* 2004; Ostfeld *et al.* 2005). The use of regression analysis in geographical correlation studies and spatial epidemiology can aid the understanding of these relationships by quantifying correlations between disease risk and external factors on a spatial scale. In addition, the assessment of spatial risk factors in such an analysis may assist with the generation or testing of hypotheses (e.g. regarding a potential point source of infection).

3.1.2.2. Predictive models

The regression process models the variability in the outcome or response variable as a function of one or more explanatory variables. The outcome of this process is a regression equation, which quantifies the relationships between the outcome variable and the explanatory variables. For example, a simple (non spatial) regression analysis may result in an equation describing the relationship between weight and age. From this equation, it is possible to extrapolate and predict the value of the response variable (weight) if the values of the explanatory variables are known (age). When considering the application of regression methods in spatial epidemiology, the possibilities for spatial prediction become clear: a regression equation which is either explicitly or implicitly spatial can be used to predict the outcome variable at spatial locations for which the outcome has not been measured (Curran *et al.* 2000; Brooker 2007).

The resulting predicted disease distribution maps are frequently called 'risk' maps and they can be used to highlight areas with the potential for high prevalence based on the explanatory variables, to provide estimates of disease burden or populations at risk in areas where large scale surveys are not possible and to assist in the planning of disease control activities (Kitron 2000). An increased knowledge of disease distributions can greatly assist in the distribution of scarce resources; interventions can be better planned and targeted to the areas which will benefit the most (Bergquist 2001). This is of particular importance in settings where health care data are incomplete and the use of large scale surveys to assess the disease prevalence within an area would be technically challenging, expensive and time consuming (Rogers 2000; Brooker 2002).

The use of regression methods for the quantification of correlations and to provide predictive maps for epidemiological applications is linked closely with the use of similar methods in ecology; regression methods are widely used in ecological research to study species distributions in relation to one or more explanatory factors (Guisan and Zimmermann 2000; Guisan *et al.* 2002; Lehmann *et al.* 2002; Moisen *et*

al. 2006). The resulting models can then be used to predict the spatial distribution of the species over larger areas based on external variables such as temperature, precipitation or land cover types, thus, negating the need for costly, time intensive surveys (Guisan and Zimmermann 2000; Guisan *et al.* 2002; Moisen *et al.* 2006). The probability distributions of variables used in many epidemiological and ecological applications are very similar, thus, necessitating the use of specific statistical models. In some cases, presence-absence data may be used: for an epidemiological application an individual or a location may be recorded as either having or not having a disease (Odiit *et al.* 2006), and for an ecological application, species will be recorded as either present or absent at specific sites (Mladenoff *et al.* 1999). Counts of disease cases at defined locations (i.e. settlements) may also be used in epidemiological analyses, similarly to species abundance data in ecological analyses (Brown *et al.* 1995), although in epidemiological analyses it is more common to use prevalence or incidence data than counts (Noor *et al.* 2008a).

3.1.2.3. Case study: urinary schistosomiasis in Tanzania

Several recent studies have examined the spatial epidemiology of schistosomiasis to further the understanding of the spatial distribution of the disease in relation to environmental and climatic factors, produce estimates of the disease burden in areas where large scale surveying is not feasible and to assist in the spatial targeting of schistosomiasis control efforts (Brooker *et al.* 2001; Clements *et al.* 2006b; Brooker 2007; Clements *et al.* 2008). One such study by Brooker *et al* (2001) exemplifies the use of logistic regression analysis for the assessment of covariate effects and spatial prediction based on environmental data.

Using *Schistosoma haematobium* prevalence data from school based surveys, geo-referenced schools were defined as being either above or below the World Health Organization's mass treatment threshold of 50% prevalence. Logistic regression was used to assess the significance of several environmental covariates (LST, NDVI, rainfall and elevation) in relation to the spatial distribution of high prevalence schools. The final regression model, which was validated using independent data,

highlighted the significance of elevation, minimum LST and mean NDVI to the probability of a school having infection prevalence of more than 50% (Brooker *et al.* 2001). The resulting model was used to predict the probability of areas having a high prevalence of urinary schistosomiasis (over 50%) and the number of school-aged children living in areas predicted to have high prevalence was calculated.

Aside from providing information on geographical correlations and spatial risk factors, the predictive mapping in this example was used to overcome a lack of spatially comprehensive prevalence data and to identify areas which are likely to sustain a high prevalence of urinary schistosomiasis. The identification of high prevalence areas is of assistance for the prioritisation of additional surveys and the targeting of control measures. Estimates of the (maximum) costs of control within Tanzania were calculated to be between $1 million and $3.2 million, based on the estimated populations at risk and prospective cost analyses of the school-based delivery of drugs to treat schistosomiasis (praziquantel) (Azene *et al.* 1999; Brooker *et al.* 2001).

3.1.2.4. Geographical correlation assessment and predictive mapping of sleeping sickness: previous studies

There are several examples in the literature of the application of regression analysis to the study of sleeping sickness epidemiology (Okia *et al.* 1994; Moore *et al.* 1999; Fèvre *et al.* 2001a; Robays *et al.* 2004; Odiit *et al.* 2004a; Odiit *et al.* 2006; Zoller *et al.* 2008), although several are wholly or partially non-spatial analyses focusing on individual risk factors in a non-geographical context (Okia *et al.* 1994; Moore *et al.* 1999; Robays *et al.* 2004; Zoller *et al.* 2008). Fèvre *et al* (Fèvre *et al.* 2001a) used logistic regression to assess proximity to the main livestock market as a spatial risk factor following the spread of Rhodesian sleeping sickness into a previously unaffected area. Results illustrated an increased risk of sleeping sickness in villages closer to the livestock market with a lessening of the association over time. The use of regression analysis to study geographical correlations in this study indicated that the introduction of Rhodesian sleeping sickness to the study area occurred due to the

movement of infected, untreated livestock from endemic areas being sold in the livestock market.

A spatial analysis of the stage of infection at the time of sleeping sickness diagnosis (early or late stage) has also been carried out using regression methods. The study by Odiit *et al* (2004a) demonstrated a correlation between proximity to the sleeping sickness treatment centre and the proportion of cases which were diagnosed in the early stage of infection; a higher proportion of cases were diagnosed in the early stage from villages within 10 km of the treatment centre than in villages further away, highlighting the importance of health care accessibility for the diagnosis and surveillance of sleeping sickness. However, neither of these studies demonstrates the utility of regression analyses for the investigation of environmental or climatic risk factors and, thus, do not enable the prediction of risk surfaces as described in Section 3.1.2.2.

Further research by Odiit *et al* (2006) used regression to assess the significance of several spatial risk factors in relation to sleeping sickness occurrence in an area of south east Uganda, including environmental variables. Increasing population density demonstrated a protective effect as did increasing distance to the sleeping sickness treatment centre (illustrating a confounding effect of accessibility to diagnostic and treatment facilities) and increasing distance to long vegetation swamp. Additional analysis within a smaller sub-section of the same area (at the village level) investigated the relationships between the occurrence of sleeping sickness and the proportion of buffer zones of different radii (measured from homestead locations) that intersected with areas of wetland (Zoller *et al.* 2008). The observed relationship (sleeping sickness affected homesteads were closer to the wetland than other homesteads within the same village) was most significant using a buffer of 800 to 900 m, and retained significance up to a buffer size of 3 km. These two studies illustrate the links between the spatial distribution of sleeping sickness and environmental variables, which have been assessed by way of regression analysis. Thus far, regression analysis has not been used to produce predictive maps for sleeping sickness based on environmental and climatic covariates.

Several other publications regarding the spatial epidemiology and risk mapping of tsetse and animal trypanosomiasis utilise discriminant analysis techniques (Rogers *et al.* 1996; Robinson *et al.* 1997a; Hendrickx *et al.* 2000; Hendrickx *et al.* 2001b) for predictive mapping; discriminant analysis allows the classification of areas for which no observations are available into one of two (or more) groups (e.g. disease present or absent) based on several predictor variables. This method seeks to identify the set of independent variables which best differentiates between two (or more) classes. However, this process does not necessarily allow the researcher to make inferences about the relationships between the response variable and independent predictor variables, as can be done following a regression analysis. Additional information on discriminant analysis is provided in Chapter 4, Section 4.1.1.3.

3.2. Specific objectives

The analysis discussed in this chapter examines the spatial distribution of *T. b. rhodesiense* sleeping sickness in two newly affected districts of Uganda (Kaberamaido and Dokolo) in relation to several environmental, climatic and social variables. Overall, these analyses aimed to investigate the effects of various spatial covariates on the occurrence and prevalence of Rhodesian sleeping sickness within Kaberamaido and Dokolo districts. More specifically, these analyses were also used to generate hypotheses regarding the initial source of the parasite within the study area. Prevalence of sleeping sickness was predicted spatially to highlight areas with the potential for high prevalence of sleeping sickness and to enable the targeting of future control efforts. The utilities of two different methodologies were compared: a two-step regression method and a traditional one-step regression method. The two-step regression was used to allow the separate analysis of factors governing the occurrence and prevalence of sleeping sickness. The prevalence analysis in the two-step regression model was conducted solely on areas that had a high predicted probability of occurrence. This was anticipated to provide an increase in predictive accuracy (for predicted prevalence) due to the exclusion of large areas with little or no sleeping sickness transmission.

3.3. Methods

The study area included Kaberamaido and Dokolo districts in Uganda (see Chapter 2, Section 2.3); two of the districts most recently affected by *T. b. rhodesiense* sleeping sickness.

3.3.1. Data

A spatially referenced dataset containing all villages within Kaberamaido and Dokolo districts, linked to sleeping sickness case records (collected from Lwala and Serere hospitals), was used for the analysis in this chapter. All cases which could be matched to their village of residence and were diagnosed between January 2004 and December 2006 were used. Those recorded from January 2007 onwards were excluded to avoid the introduction of spatial bias following the implementation of a control programme in 2006. The control programme involved the mass treatment of livestock in Kaberamaido, Dokolo and neighbouring districts and aimed to decrease the prevalence of *T. b. rhodesiense* in the animal reservoir.

The resulting dataset contained all villages within Kaberamaido and Dokolo districts except for those which could not be geo-referenced, with their total populations (village population data from the most recent national census (Uganda Bureau of Statistics 2008)) and the number of diagnosed sleeping sickness cases recorded as being resident within each village during the study period (from hospital records). Further information on the dataset used in this chapter can be found in Chapter 2, Sections 2.4.2.1 (sleeping sickness case records) and 2.4.3.1 (village geo-referencing).

The geo-referenced sleeping sickness case data were visualised using ArcMap 9.1 (ESRI, Redlands, CA). External covariate data (as detailed in Table 1) were extracted for each village and linked with the sleeping sickness case data for use in the regression analyses. Fourier-processed AVHRR indices; LST, NDVI and MIR were

used along with NDVI calculated using a Landsat ETM+ image as described in Chapter 2, Sections 2.4.4.1 and 2.4.4.2 (NASA Landsat Programme 2004; Hay *et al.* 2006). In addition, predicted tsetse suitability (Wint and Rogers 2000), elevation (US Geological Survey 2006), population density (Oak Ridge National Laboratory 2006), nighttime lights of the world (Defence Meteorological Satellite Program 2004; Noor *et al.* 2008b), Euclidean distances to physical features (gazetted land, rivers, bush, woodland, swamps, permanently wet land and seasonally wet land) and Euclidean distances to the closest livestock market and health centre (of any type) were extracted. Further information can be found on each of these datasets in Chapter 2, Section 2.4.4.

The landcover covariates were selected as potential tsetse habitats allowing the investigation of the effect of proximity of villages to these types of landcover on sleeping sickness occurrence and prevalence. The distance to the closest health centre (of any type, i.e. not necessarily trained or equipped to diagnose or treat sleeping sickness) was used to deal with the confounding effect of access to health care and the distance to the closest livestock market to investigate the possibility that cattle movements in this area may have caused or contributed to the introduction and establishment of sleeping sickness transmission, as was found in a neighbouring district (Fèvre *et al.* 2001a). The distance to the sleeping sickness treatment centre was not used as there was only one treatment centre within the study area and an additional treatment centre in the neighbouring district serving the study population, which would affect the final predictions and prevent extrapolation over a larger area. The covariates used are listed in Table 3.1: all were continuous variables.

Source	Variable	Spatial resolution	Units	Used in regression
Fourier processed AVHRR data sets	NDVI phase of annual and biannual cycle	1 km	Months	X
	Minimum and maximum NDVI	1 km	No units	X
	Annual and biannual amplitude of NDVI	1 km	No units	X
	Mean NDVI	1 km	No units	X
	MIR phase of annual and biannual cycle	1 km	Months	
	Minimum and maximum MIR	1 km	°C	
	Annual and biannual amplitude of MIR	1 km	°C	
	Mean MIR	1 km	°C	
	LST phase of annual and biannual cycle	1 km	Months	X
	Minimum and maximum LST	1 km	°C	X
	Annual and biannual amplitude of LST	1 km	°C	X
	Mean LST	1 km	°C	X
Predicted tsetse suitability coverages	Predicted suitability for *G. fuscipes*	1.1 km	Predicted % suitability	
	Predicted suitability for *G. morsitans*	1.1 km	Predicted % suitability	
	Predicted suitability for *G. pallidipes*	1.1 km	Predicted % suitability	
Shuttle Radar Topography Mission	Elevation	3 arc seconds	Metres	X
Landsat	NDVI	30 m	No units	
Landscan	Population density	30 arc seconds	People per Km^2	X
Nighttime lights of the world	Nightlights	30 arc seconds	Percentage	
National biomass study	Distance to gazetted land	Continuous	Kilometres	X
	Distance to river	Continuous	Kilometres	
	Distance to bush areas	Continuous	Kilometres	X
	Distance to wooded areas	Continuous	Kilometres	X
	Distance to swamp land	Continuous	Kilometres	
	Distance to permanently wet land	Continuous	Kilometres	X
	Distance to seasonally wet land	Continuous	Kilometres	X
Other geo-referenced locations	Distance to health centre (any type)	Continuous	Kilometres	X
	Distance to livestock market	Continuous	Kilometres	X

Table 3.1: Covariates used, indicating those included in model development.

3.3.2. Statistical analysis

An exploratory analysis was conducted for each of the covariates listed in Table 3.1: i) scatterplots to examine relationships with sleeping sickness prevalence; ii) box and whisker plots to examine the distributions of covariate data in villages which have had cases of sleeping sickness compared to villages which have not and iii) visualisation of the geographical distributions of the outcome variables in relation to the external covariates (plots were created in the R statistical software (R Development Core Team 2006) and visualisations in ArcMap 9.1). Seventeen covariates were selected for use in the regression analyses, as indicated in Table 3.1, based on observed relationships with sleeping sickness occurrence and prevalence and previous knowledge of significant variables from published research.

The statistical modelling was carried out using logistic regression: a generalised linear model (GLM) which can be used where the outcome variable is the result of a series of Bernoulli trials (a single occurrence which can have one of two outcomes; success or failure). In relation to the analysis of sleeping sickness data at a village level, logistic regression can be used to model disease occurrence (where each village either has or has not had a case of the disease reported) or disease prevalence (where a number of people within each village of known population either have or have not had the disease) (Hosmer and Lemeshow 1989a). The modelling process describes the variability in the response variable as a function of the explanatory variables. Odds ratios (ORs) are calculated by exponentiation of the regression parameters associated with each covariate; these illustrate the strength and direction of associations between the explanatory and outcome variables. Statistical significance was judged at the 95% level in all analyses. Each step of the logistic regression analyses was carried out using the R statistical software (R Development Core Team 2006) with the *base* package. The main steps of the analysis are summarised in Figure 3.1.

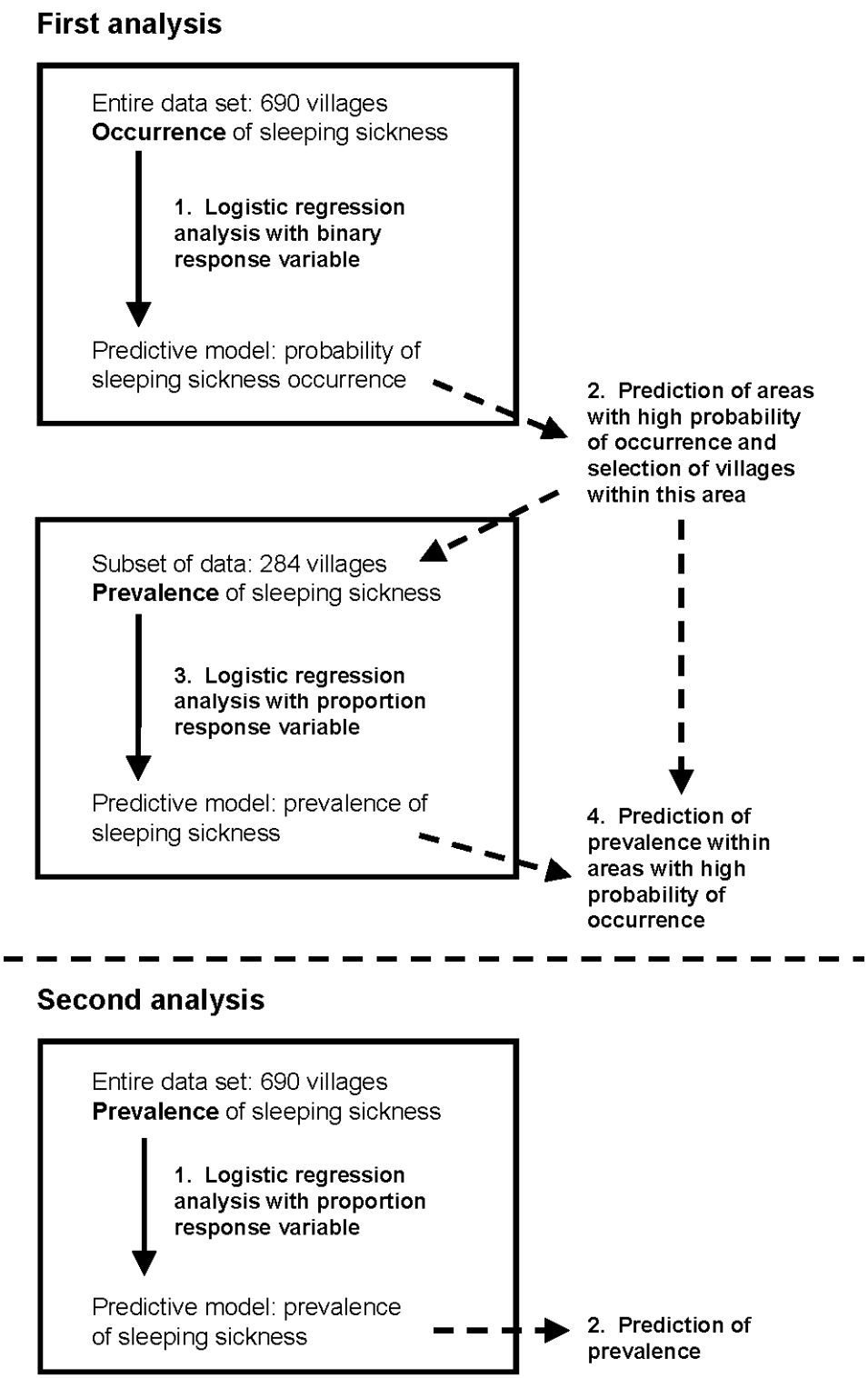


Figure 3.1: Diagram illustrating the two regression methodologies, including the main steps involved in each.

3.3.2.1. Two-step analysis of sleeping sickness occurrence and prevalence

The first method comprised two logistic regression models applied sequentially (first analysis, Figure 3.1). An initial model was fitted that predicted probability of sleeping sickness occurrence using the sleeping sickness status of all villages in the study area as the outcome of interest. Villages for which at least one case of sleeping sickness was reported during the study period were classified as case villages, while villages for which no cases were reported were treated as controls (giving a binary outcome). The two-step model was developed to test its predictive capability against a traditional regression analysis and to investigate aspects of the underlying epidemiology affecting the spatial heterogeneity in disease occurrence (which villages had been affected by sleeping sickness) as well as prevalence (how intense was the transmission within affected areas) which are confounded in a one-step approach.

Forward stepwise addition beginning with the null model (no explanatory variables) was used in the model fitting. At each step the variable resulting in the greatest reduction in deviance was selected. A Chi-squared likelihood ratio test was used to compare models, and additional explanatory variables were accepted only if this test was significant and the covariate was significant within the model. Any variables that did not retain significance in subsequent steps were removed from the model. The stepwise addition of plausible interaction terms was then carried out in the same manner after the variables were centred (variable mean was subtracted from each value).

The sensitivity (the percentage of case villages which are correctly classified as such by the regression equation) and specificity (the percentage of control villages which are correctly classified as such by the regression equation) of the fitted model were calculated with the *ROCR* package in the R statistical software using the predicted and observed values (Metz 1978; Brooker *et al.* 2002b). A variety of cut-off points (the value of the predicted probability of occurrence above which a location would be defined as a case village) were used for the calculation of sensitivity and

specificity. These were then plotted against the cut-off points to show the change in sensitivity and specificity at different cut-offs. The cut-off point where the sensitivity and specificity crossed was selected as a suitable threshold for the classification of case and non-case villages: this point maximises both the specificity and the sensitivity of the classification of locations. A receiver operator characteristic (ROC) curve was plotted using the *epi* package in the R statistical software. This plots the sensitivity against the false positive fraction (false positive fraction = 1 – specificity) with varying cut-off points and illustrates the compromise between sensitivity and specificity for different threshold values (Metz 1978; Brooker *et al.* 2002b). The area under the ROC curve (AUC) was calculated to give a measure of the overall accuracy of the model in classifying villages. An AUC of between 0.5 and 0.7 would indicate a model with poor discriminatory power, an AUC of between 0.7 and 0.9 would indicate a reasonable discrimination between cases and controls and an AUC of over 0.9 would indicate a very good discriminatory accuracy (Pearce and Ferrier 2000; Brooker *et al.* 2002b).

A 10-fold cross-validation (where predicted values are compared with observed values) was performed with ten random sub-divisions of the dataset using the *DAAG* package in the R statistical software. Each of the random sub-divisions was removed from the dataset in turn, and the model was then re-fit using the data which remained. Predictions were made for the observations which had been removed, and the comparison of fitted and observed values was used to calculate an overall measure of predictive accuracy.

The resulting regression equation (probability of occurrence as a function of the explanatory variables) was used to predict probability of occurrence of sleeping sickness across a grid with an area of 30,000 km^2 (including the study region) and a 1.1 km cell size (this was the minimum spatial resolution from the covariate datasets). All villages in the study area lying within an area of high predicted probability of occurrence (probability of occurrence above the selected cut-off value) were extracted for use in the second step of the analysis.

The Pearson residuals were calculated (the difference between the fitted and observed values divided by the standard deviation) to give an indication of the level of additional variation in the observations which was not explained by the covariates in the model (McCullagh 1989). Biased regression parameters and underestimated standard errors may occur if the covariates used in a model do not explain all of the spatial variation in the observations (Thomson *et al.* 1999; Boyd *et al.* 2005; Pfeiffer *et al.* 2008b). To assess the amount of residual spatial autocorrelation, a residual variogram was plotted (with the Pearson residuals) using the *gstat* package in the R statistical software. The variogram is used to illustrate the degree of similarity between observations (in this case the degree of similarity between the Pearson residuals) at different distances from one another, and can give an indication of the presence of additional spatially correlated variation which has not been explained by the covariates used.

The outcome variable for the second step of the two-step regression was defined as prevalence of sleeping sickness (number of cases divided by village population). Data from all villages within areas of high predicted probability of occurrence were included in the model, including those with no reported cases (i.e. a reported prevalence of zero). Forwards stepwise addition was used in the model fitting procedure, as for the first step. For this section of the analysis, the distance to health centre variable was forced into the model (regardless of its significance) to ensure that access to health care was controlled for in the final results. The fitted model was used to predict the prevalence of *T. b. rhodesiense* sleeping sickness across the same area as was used in the first step.

A scatter plot was created with observed prevalence against fitted prevalence (fitted using the second regression model for villages with a high probability of occurrence). The correlation coefficient was also calculated using the *stats* package in the R statistical software to give an indication of the degree of association between predicted and observed prevalence values. The prediction errors (difference between observed and predicted prevalence per 100 population) were calculated for each village. The median errors and mean absolute errors were then calculated to illustrate

the error distribution and to give measures of prediction bias (median) and predictive accuracy (mean absolute error) respectively. The Pearson residuals were also calculated and a residual variogram was plotted as in the first step.

3.3.2.2. One-step analysis of prevalence using all villages

For the one-step analysis (second analysis, Figure 3.1), the same methodology was used as the second step of the two-step regression, using prevalence data from all villages rather than a subset. Diagnostics were carried out as discussed in the second step of the two-step regression analysis (Section 3.3.2.1): a scatter plot of observed versus predicted prevalence values and calculation of the correlation coefficient; the median and mean absolute errors and the residual variogram.

3.4. Results

A total of 690 villages within Kaberamaido and Dokolo districts were geo-referenced. Two villages were not geo-referenced due to logistical difficulties, and 18 villages that had separated recently into two were merged for the purpose of the analysis. A total of 52 patient records could not be matched to any of the known villages in the study area and so were excluded from the analysis. This was most likely due to inaccuracies in the recording of patient details in the hospital records. The total number of cases used in the study was 302. The distribution of villages, along with the village prevalence of sleeping sickness using data from 2004 – 2006, is illustrated in Figure 3.2. The results of the exploratory data analyses are not shown; the variable selection procedure was based on observed relationships from the exploratory analysis and current knowledge regarding the spatial epidemiology of sleeping sickness.

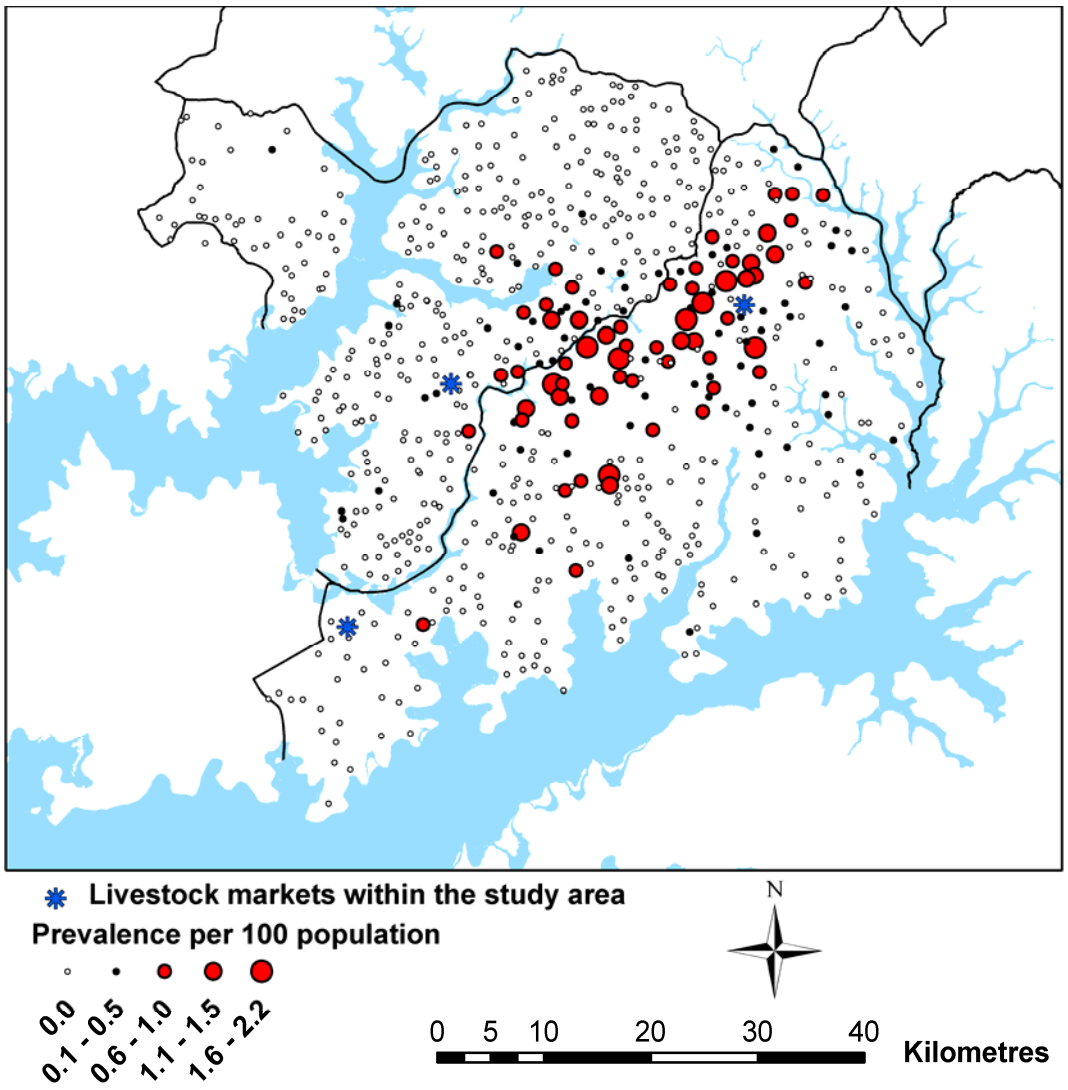


Figure 3.2: Village level period prevalence of sleeping sickness in Kaberamaido and Dokolo districts, 2004-2006. District boundaries are shown as black lines.

3.4.1. Two-step regression analysis of sleeping sickness suitability and prevalence

Four covariates were found to influence significantly the occurrence of sleeping sickness across the study area ($p < 0.05$) as shown in Table 3.2. Occurrence of sleeping sickness was negatively correlated with distance to the closest livestock market, with a 21% reduction in odds of disease for every kilometre increase in distance, when accounting for the additional variables. This was found to interact (the effect of one variable on odds of disease changes in relation to the effect of

another variable) with maximum NDVI, which also demonstrated a negative correlation with sleeping sickness occurrence. In addition, occurrence was positively correlated with minimum LST and negatively correlated with distance to the closest health centre.

Variable	**Odds ratio (95% CI)**	**p-value**
Intercept	7.42E-6	<0.0001
	(3.39 E-8 – 0.002)	
Distance to livestock market	0.79	<0.0001
	(0.75 – 0.84)	
Maximum NDVI	9.56 E-7	0.03
	(2.64 E-12 – 0.35)	
Minimum LST	2.10	0.0001
	(1.44 - 3.04)	
Distance to health centre	0.84	0.002
	(0.74 - 0.94)	
Distance to market * Max NDVI	32.46	0.003
	(3.34 - 315.52)	

Table 3.2: Results of the first model from the two-step regression analysis, using a binary response variable and all villages.

For prediction purposes, the selected probability cut-off point for the prediction of areas suitable for transmission was 0.2, and model diagnostics indicated that the model provided a reasonable fit to the data, and reliable predictions (AUC: 0.87, 10-fold cross-validation estimate of accuracy: 85%). The predicted suitability for transmission across the study area using the specified model is illustrated in Figure 3.3. The residual variogram indicates that some residual spatial autocorrelation remains after accounting for covariate effects (Figure 3.4).

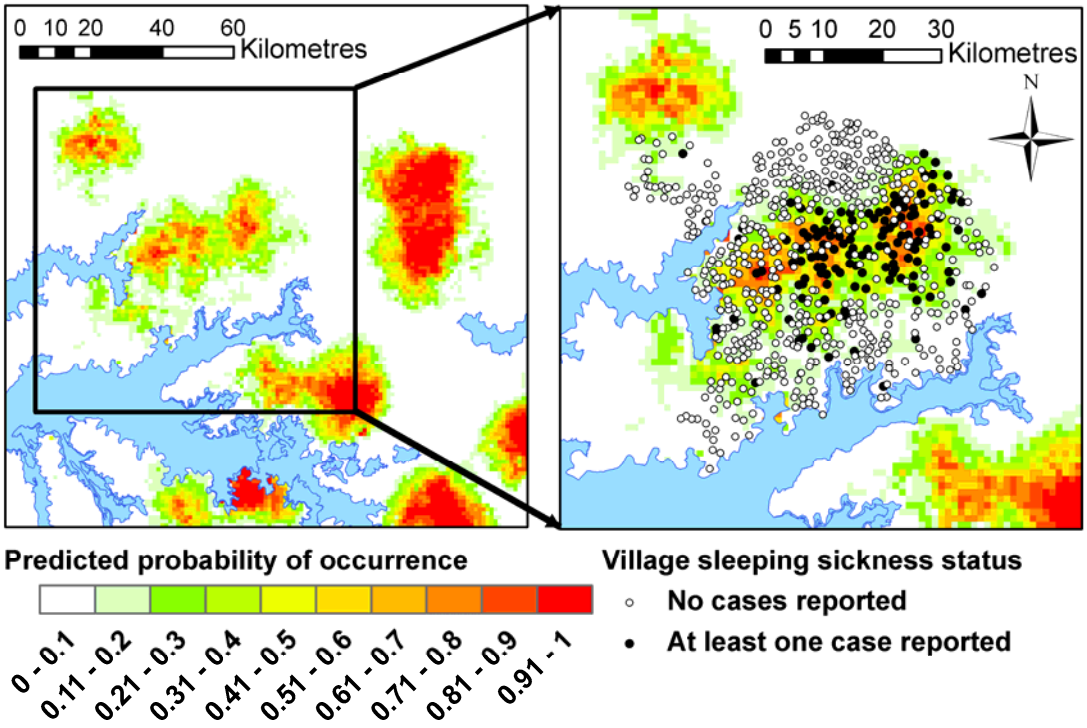


Figure 3.3: Predicted probability of sleeping sickness occurrence from the first step of the two-step analysis. White and pale green indicate areas with low predicted probability of occurrence. Black circles indicate case villages and white circles represent non-case villages within the study area.

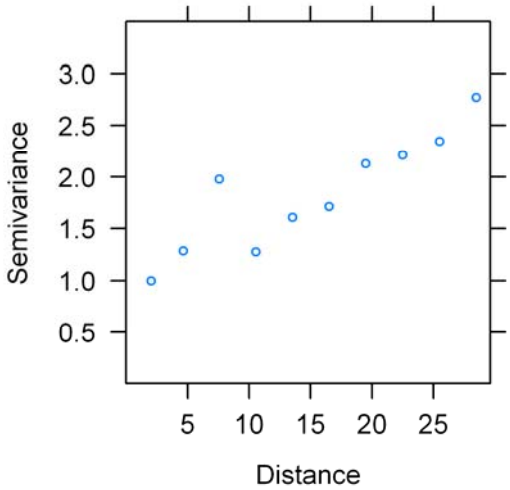


Figure 3.4: Residual variogram for the first step of the two step analysis (using Pearson residuals).

The prediction was used to create a mask over the study area; all areas with a predicted probability of occurrence less than 0.2 were excluded. 279 villages lay within the area defined as having a high probability of occurrence. However, seven

of those villages had no population data and so were excluded from the remaining analysis leaving 272 villages. The results from the second (prevalence) model are shown in Table 3.3.

Sleeping sickness prevalence was significantly correlated with nine variables in addition to distance to the closest health centre which was negatively correlated and of borderline significance (p=0.05; variable forced into the model). Prevalence was negatively correlated with distance to the closest livestock market with every additional kilometre resulting in a 20% decrease in odds of disease. This was shown to interact with distance to the closest area of woodland, which in turn showed a positive correlation with prevalence. In addition, sleeping sickness prevalence was negatively correlated with distance to the closest area of bush and maximum NDVI and positively correlated with NDVI phase of annual cycle, NDVI annual amplitude, LST phase of annual cycle, LST annual amplitude and minimum LST.

Variable	Odds ratio (95% CI)	p-value
Intercept	1.72 E-8	0.0001
	(1.78 E-12 – 0.0002)	
Distance to health centre[1]	0.92	0.05
	(0.85 – 1.00)	
Distance to livestock market	0.80	< 0.0001
	(0.77 – 0.83)	
NDVI phase of annual cycle	3.46	0.0008
	(1.67 – 7.14)	
NDVI annual amplitude	2.18 E+11	<0.0001
	(1.85 E+6 – 2.59 E+16)	
LST phase of annual cycle	1.27	<0.0001
	(1.13 – 1.43)	
Distance to woodland	1.15	0.18
	(0.95 – 1.40)	
Distance to bush	0.93	0.0007
	(0.90 – 0.97)	
Maximum NDVI	3.50 E-5	0.01
	(1.46 E-8 – 0.08)	
LST annual amplitude	1.27	0.009
	(1.07 – 1.52)	
Minimum LST	1.46	0.004
	(1.13 – 1.89)	
Distance to livestock market *	0.91	0.002
Distance to woodland	*(0.86 – 0.97)*	

Table 3.3: Results of the second step from the two-step regression analysis, using prevalence response variable and a subset of villages. [1]Forced into the model.

The two-step regression analysis resulted in a correlation between observed and predicted prevalence of 0.57 (a value of 1 indicates perfect correlation and 0 no correlation). The model had a small tendency to over-predict prevalence with a median error of 0.05% (error calculations are based on prevalence per 100 population and so are expressed as a percentage). The mean absolute error for the predicted prevalence per 100 population was 0.24%. The predicted prevalence from the two-step analysis is shown in Figure 3.5. The scatter plot of predicted prevalence against observed prevalence (Figure 3.6) highlights the tendency for over-prediction of prevalence, particularly in villages with an observed prevalence of zero. Variogram

analysis of the Pearson residuals (Figure 3.7) indicates the absence of residual spatial autocorrelation following the second step of the two-step regression analysis.

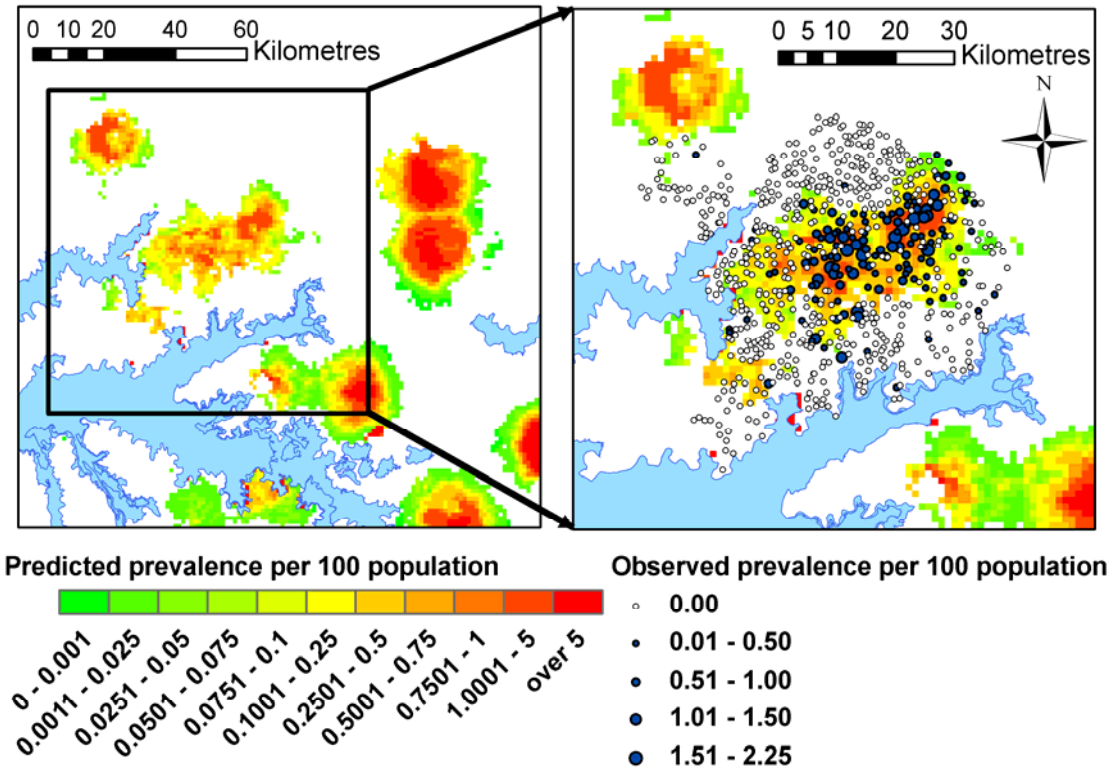


Figure 3.5: Predicted prevalence of sleeping sickness from the second step of the two-step analysis. White indicates areas predicted to be unsuitable for transmission. Blue circles indicate case villages and white circles represent control villages within the study area.

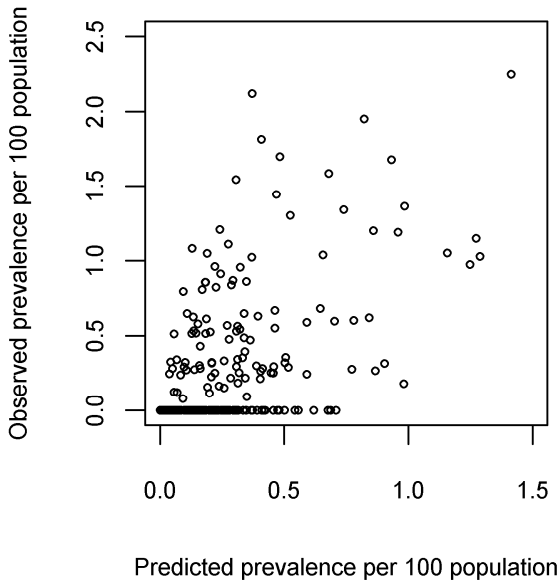


Figure 3.6: Scatter plot of observed prevalence versus predicted prevalence (per 100 population) using the two-step analysis.

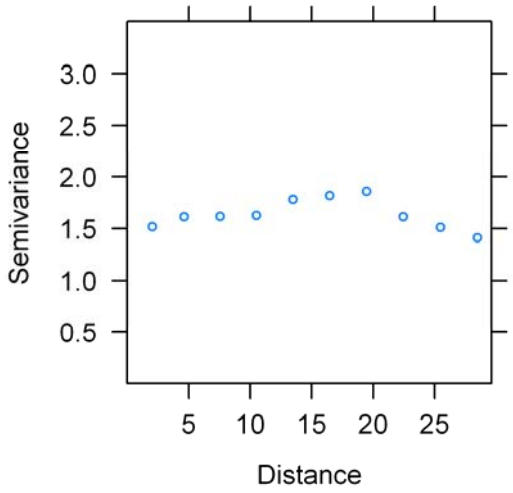


Figure 3.7: Residual variogram for the second step of the two step analysis (using Pearson residuals).

3.4.2. One-step regression analysis of prevalence using all villages

The prevalence of sleeping sickness was shown to be significantly associated with nine variables across the study area using the one-step regression (Table 3.4). Sleeping sickness prevalence was negatively correlated with distance to the closest livestock market, with a 21% reduction in odds of disease for every kilometre

increase in distance. This was shown to interact significantly with both NDVI phase of annual cycle and distance to the closest area of woodland, both of which were also negatively correlated with prevalence. Additionally, prevalence was negatively correlated with maximum NDVI, mean LST and distance to the closest health centre. Sleeping sickness prevalence was positively correlated with minimum LST, LST phase of annual cycle and LST annual amplitude.

Variable	Odds ratio (95% CI)	p-value
Intercept	0.32 *(0.0005 – 200.1)*	0.003
Distance to livestock market	0.79 *(0.76 –0.82)*	<0.001
Maximum NDVI	2.6E-06 *(3.59 E-9 – 0.002)*	0.0001
Minimum LST	2.05 *(1.62 – 2.60)*	<0.0001
LST phase of annual cycle	1.26 *(1.12 – 1.42)*	<0.0001
LST annual amplitude	1.75 *(1.36 – 2.26)*	<0.001
Mean LST	0.56 *(0.39 – 0.81)*	0.003
Distance to woodland	0.96 *(0.76 – 1.22)*	0.76
Distance to health centre	0.87 *(0.80 – 0.94)*	<0.0001
NDVI phase of annual cycle	0.98 *(0.42 – 2.33)*	0.97
Distance to market * NDVI phase of annual cycle	0.84 *(0.74 – 0.94)*	0.002
Distance to market * distance to woodland	0.95 *(0.91 – 0.99)*	0.01

Table 3.4: Results of one-step regression analysis using prevalence outcome variable and all villages

The correlation between fitted and observed prevalence values was 0.58 indicating a modest linear association. The model was slightly biased with a small tendency to over-predict prevalence (median error = 0.02%) and the mean absolute error was 0.13%. Figure 3.8 illustrates the predicted prevalence across the study area using the final prevalence model. The scatter plot of predicted prevalence against observed prevalence values (Figure 3.9) illustrates that many of the errors are associated with over-prediction for villages with an observed prevalence of zero. From the residual variogram (Figure 3.10) the presence of spatial autocorrelation in the residuals can be observed.

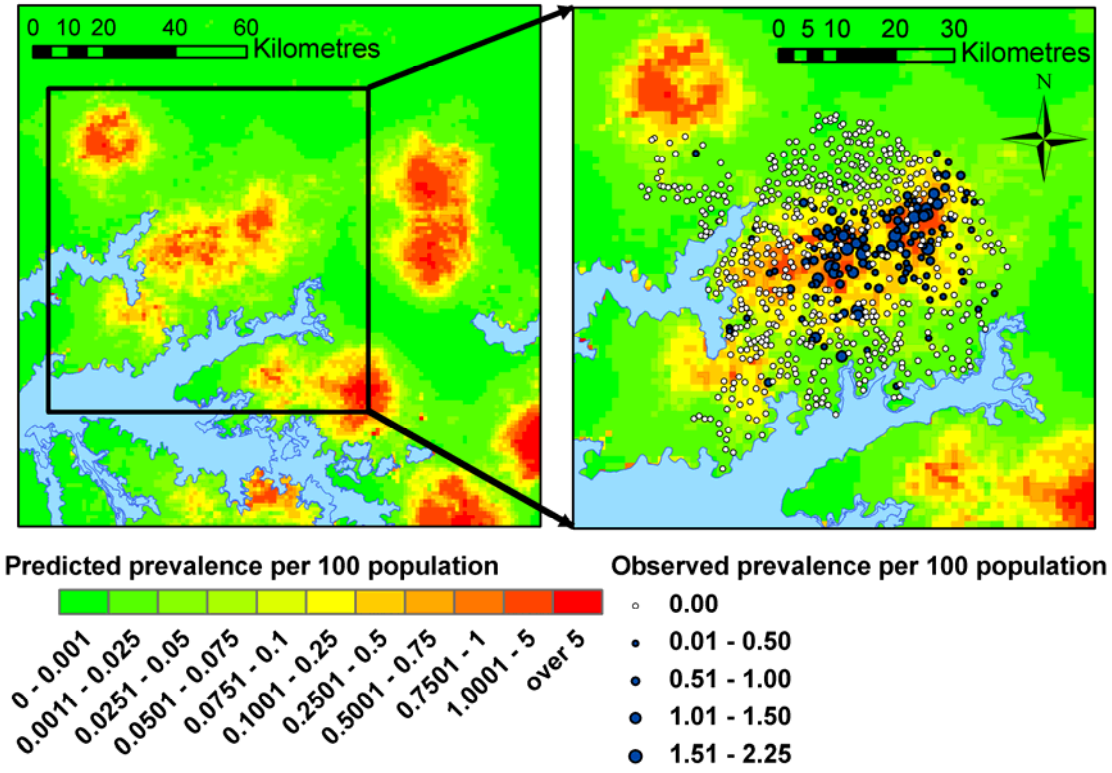


Figure 3.8: Predicted prevalence of sleeping sickness from one-step regression analysis.

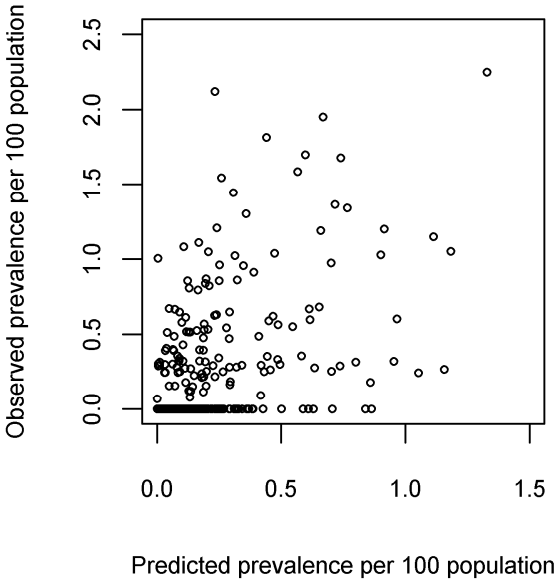


Figure 3.9: Scatter plot of observed prevalence versus predicted prevalence (per 100 population) using the one-step analysis.

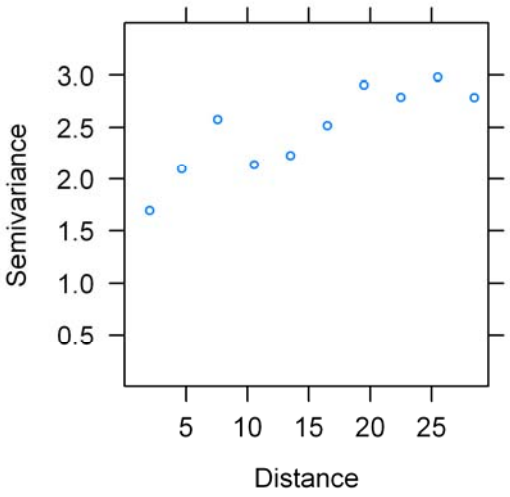


Figure 3.10: Residual variogram for the one step analysis (using Pearson residuals).

To allow a direct comparison of the predictive accuracy of the two methodologies, the one-step model was used to calculate predicted prevalence for the villages with high predicted probabilities of occurrence from the two-step analysis (i.e. excluding areas with a predicted probability of occurrence of less than 0.2). The correlation between predicted and observed prevalence was 0.50, lower than that for the two-step regression method (0.57). Again, the model was shown to have a tendency to over-predict prevalence, with a median error of 0.05% (calculated using prevalence per 100 population). The mean absolute error was 0.24%, equal to the mean absolute error from the two-step regression methodology.

3.5. Discussion

The analysis discussed in this chapter examined the relationships between Rhodesian sleeping sickness and several environmental, climatic and social factors in two newly affected districts, Kaberamaido and Dokolo. The application of a two-step regression approach for the prediction of sleeping sickness prevalence in a newly affected area of Uganda allowed the investigation of factors influencing the occurrence and prevalence of sleeping sickness separately, and overall resulted in slight increase in predictive accuracy when compared to a one-step analysis in areas with high predicted probability of occurrence. Each of the models has illustrated an increased

risk of sleeping sickness in villages closer to livestock markets than in villages further away, suggesting that the persistent spread of Rhodesian sleeping sickness in Uganda may have resulted from the continued movement of untreated cattle.

The two-step regression model gave a small increase in predictive accuracy in comparison with the one-step analysis with a correlation between fitted and observed prevalence values of 0.57 for the two-step regression and 0.50 for the one-step regression analysis (when considering only areas with a high predicted probability of occurrence). Both models tended to predict higher prevalence than was observed, particularly in villages of zero prevalence, with a median error of 0.05% for both models. The mean absolute error was also equal for the two methods (0.24%). The difference in predicted prevalence of sleeping sickness from the two methods was small over the majority of the prediction area, with divergences mainly occurring in areas of high predicted prevalence outside of the study area (see Figure 3.11).

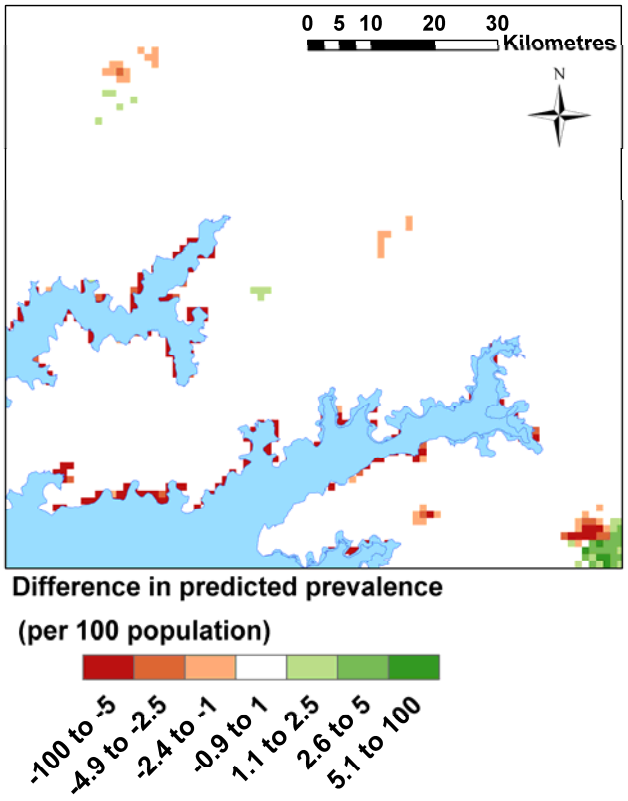


Figure 3.11: Difference in predicted prevalence between two-step and one-step analyses.

Analysis of the residual variation (after accounting for the covariate effects) indicated that there was some spatial autocorrelation in the residuals from the one-step regression and the probability of occurrence analysis (first step of the two-step regression analysis). For the two-step regression, the probability of occurrence regression was carried out partially to provide a mask over areas with low predicted probability of occurrence to enable the focusing of the prevalence analysis, and so the small amount of spatial autocorrelation in the residuals is not seen as problematic as it would have a negligible effect on the final prevalence model. However, for the one-step regression, the small amount of spatial autocorrelation in the residuals may lead to inflated statistical significance for the covariates. When comparing the residual variogram from the one-step regression (Figure 3.10) with that from the second step of the two-step logistic regression analysis (Figure 3.7), it can be seen that the two-step regression analysis resulted in a smaller amount of residual spatial autocorrelation. However, it is likely that this is due to the exclusion of areas with an observed prevalence of zero by using the mask created during the first step of the analysis. A variety of different methods are available to deal with spatial autocorrelation in regression analysis, such as model-based geostatistics (Diggle *et al.* 1998). A model-based geostatistics approach was used to extend the analysis described within this chapter and to assess any increase in the predictive accuracy by dealing with the residual spatial autocorrelation, as discussed in Chapter 4.

Distance to the closest livestock market was an important predictor in the one-step regression and in both steps of the two-step regression, with decreasing odds of infection at increasing distances (a decrease in odds of infection of approximately 20% for every additional km). Previous research has indicated the introduction of sleeping sickness to a previously unaffected area via the movement of untreated, infected livestock (Fèvre *et al.* 2001a). These results suggest that despite reinforced policy regarding the treatment of livestock for trypanosomes prior to movement from endemic areas (Wendo 2002), the subsequent spread of sleeping sickness into Kaberamaido and Dokolo may have been facilitated by the movement of infected cattle through one or more of the local livestock markets. The work presented does not, however, differentiate between the three different livestock markets operating

within the study area. To further characterise the source of *T. b. rhodesiense* within the study area, additional analyses could investigate the significance of each livestock market individually on the spatial distribution of Rhodesian sleeping sickness. By comparing these models it may be possible to identify a single market as the initial source of infection.

The main cattle trading routes within this part of Uganda run from *T. b. rhodesiense* endemic areas in the south east, through the study area and neighbouring districts, to the *T. b. gambiense* endemic areas in the far north west of Uganda and towards southern Sudan (R. Selby, pers. com.). This increases the risk of overlap of the two subspecies, particularly if the regulations regarding the treatment of cattle being moved from *T. b. rhodesiense* endemic areas are not strictly enforced. The stringent implementation of these regulations should be a priority for the Ugandan Government and tsetse control efforts may be more efficiently targeted to areas surrounding livestock markets. These interventions may help prevent the further spread of Rhodesian sleeping sickness and the establishment of transmission in previously unaffected areas as occurred in Soroti district in the late 1990s and Kaberamaido and Dokolo districts in 2004.

There were only two health centres trained and equipped to diagnose and treat sleeping sickness serving the study population during the study period. It has been shown previously that geographical accessibility to treatment facilities can affect the observed spatial distribution of sleeping sickness, with smaller numbers of cases reported from areas which are further from the treatment centres (Odiit *et al.* 2004a). However, an added complication arises within the study area; following the detection of a number of cases in Kaberamaido district in 2004, appropriate training and equipment were provided to one hospital within the area. The facility was selected based on a number of criteria, including its location within the affected area, thus, creating problems for the separation of the effects of differential utilisation of the sleeping sickness treatment centre, from the purposeful siting of the treatment centre. Moreover, this facility is close to one of the major livestock markets in the study area (7.5 km away) making their separate influences on observed prevalence difficult to

distinguish. Distance to the closest health centre (of any type) was used in the analyses to account for the effects of health care accessibility without using the potentially misleading variable, distance to the sleeping sickness treatment centre. Distance to the closest health centre was a significant factor in each model (although it was forced into the model in the second step of the two-step regression model), with decreasing prevalence observed at increasing distances. This suggests a confounding relationship due to accessibility of health services as has been previously reported (Odiit *et al.* 2004a).

Each of the regression models (the one-step regression model and each step of the two-step regression models) also included maximum NDVI (negative association) and minimum LST (positive association) as significant predictors. These are likely to relate to the habitat and environmental requirements of the tsetse fly vector of disease. Other variables that were significantly correlated with sleeping sickness prevalence and/or occurrence included NDVI phase of annual variation, NDVI annual amplitude, LST phase of annual variation, LST annual amplitude, mean LST, distance to the closest area of woodland and distance to the closest area of bush. These additional variables may also be linked to the suitability of an area for the tsetse fly vector (due to their preferred habitat and also climatic requirements), and so will influence the intensity of transmission and observed prevalence of sleeping sickness.

From these and previous findings (Fèvre *et al.* 2001a), it is thought to be likely that the movement of *T. b. rhodesiense* infected livestock from endemic areas through livestock markets within the study area occurs periodically. A complex interaction of factors is involved in the establishment of transmission following such an occurrence. In addition to the environmental, climatic and social variables which significantly influenced the distribution and prevalence of Rhodesian sleeping sickness in the current analysis, tsetse and livestock densities, human-cattle-tsetse contact and also to a large degree, chance, may play roles. This analysis has been extended by conducting a spatial analysis in which a generalised linear geostatistical model is used in a Bayesian framework to account explicitly for spatial

autocorrelation and incorporate uncertainty in input data and model parameters, as discussed in Chapter 4. This modelling approach allows the more robust assessments of covariate effects, providing a further test of the hypothesis that the parasite was introduced to Kaberamaido and Dokolo districts via the movement of untreated livestock from *T. b. rhodesiense* endemic areas.

Chapter 4. Investigating the spatial distribution of *T. b. rhodesiense* in a newly affected area of Uganda using Bayesian geostatistics

"After a certain high level of technical skill is achieved, science and art tend to coalesce in aesthetics, plasticity, and form. The greatest scientists are always artists as well."

Albert Einstein (1879 - 1955)

4.1. Introduction

Many statistical methods commonly used in epidemiology are not strictly applicable to spatial datasets. Classical statistical methods (such as logistic regression analysis) assume observations are independent (or, more specifically, model residuals are not correlated). These assumptions are frequently violated when using spatial data due to the phenomenon of spatial autocorrelation (McCullagh 1989), where observations separated by small distances are more alike than observations separated by larger distances. When assessing the effects of spatially varying factors, it is important to examine any residual spatial autocorrelation. If the variables used capture all (or most) of the spatial variation in the observations so that there is no remaining spatial autocorrelation in the residuals, then the analysis should not be prone to any problems (Boyd *et al.* 2005; Pfeiffer *et al.* 2008b). However, if the covariates do not entirely explain the spatial patterns, then the model results may be inaccurate. The regression parameters are likely to be biased and standard errors underestimated, potentially resulting in falsely narrow confidence intervals and an overestimation of the significance of covariates (Legendre 1993; Thomson *et al.* 1999; Boyd *et al.* 2005). The spurious parameter and uncertainty estimates can ultimately lead to misinterpretation of the relationships between observations and explanatory variables (Boyd *et al.* 2005).

The use of non-spatial logistic regression methods, as discussed in Chapter 3, have provided an in-depth exploration of the social, environmental and climatic factors influencing the spatial distribution of Rhodesian sleeping sickness within two of the most recently affected districts in Uganda. A large number of variables were significantly associated with sleeping sickness prevalence; the final one-step regression model of village level prevalence contained nine covariates, plus two interaction terms, complicating our understanding of the relationships and hindering the biological interpretation of covariate effects. This analysis utilised non-spatial logistic regression methods and, thus, did not account explicitly for the spatial structure of the data and potential spatial autocorrelation in observations. An examination of the residual variogram following the logistic regression analyses

indicated spatial autocorrelation in the residuals from the one-step prevalence analysis (Section 3.4.2), and from the first step of the two-step regression analysis (Section 3.4.1). The residual autocorrelation from the first step of the two-step regression (analysis of sleeping sickness occurrence) was not considered problematic as the resulting predictions from this model were used as a mask for the subsequent analysis only. However, the residual autocorrelation in the one-step regression analysis indicates that the explanatory variables used in the analysis did not explain fully the observed pattern of prevalence. In addition, the utilisation of non-spatial regression may have resulted in biased parameter estimates, underestimated standard errors and inflated significance of covariate effects as discussed above (Legendre 1993; Boyd *et al.* 2005).

By dealing with the residual spatial autocorrelation in the dataset (as used in Chapter 3), more accurate parameter and significance estimates can be obtained which can clarify and aid the interpretation of exposure-disease relationships. The significance of distance to the closest livestock market in the non-spatial logistic regression analyses suggested the implication of livestock movements in the spread of Rhodesian sleeping sickness into the study area. By extending the previous analysis (one-step logistic regression, as discussed in Chapter 3) with the explicit incorporation of spatial autocorrelation, more robust parameter estimates and significance values may be obtained, which can add substantial support to (or disprove) the hypothesis regarding the source of *T. b. rhodesiense* within the study area.

4.1.1. Geostatistical methods

Geostatistical methods allow spatial autocorrelation to be explicitly incorporated in the analytical framework, thus, alleviating the problems discussed above. Many of these methods are adaptations of classical statistical methods and can be applied readily to spatial datasets. There are many available spatial regression methods of varying complexity which can allow the assessment of covariate effects and spatial prediction, while accounting for the underlying spatial structure of the data. These

methods can give a more accurate representation of the environmental or biological processes taking place.

Classical geostatistical methods assume a continuous distribution, which makes them inappropriate for applications where response variables are bounded, such as binary or proportion data (De Oliveira *et al.* 1997; Palacios and Steel 2006; Pilz and Spock 2008). Epidemiological variables are often non-continuous; for example disease occurrence or prevalence data, in which case GLMs are the most appropriate regression models. Several methods are available for the geostatistical analysis of non-continuous data, including transformations of the response variable (De Oliveira *et al.* 1997) and model-based geostatistics (Christensen and Ribeiro Jr 2002), each possessing different attributes. Further discussion of the benefits of model-based geostatistics is provided in Section 4.1.1.1.

4.1.1.1. Model-based geostatistics

Model-based geostatistics is a set of formal statistical methods under an explicitly assumed spatial stochastic model (Diggle *et al.* 1998). A generalised linear geostatistical model allows the spatial analysis of non-continuous response variables using a GLM framework by incorporating a stochastic spatial effect to model the residual spatial autocorrelation (Diggle *et al.* 1998; Christensen and Ribeiro Jr 2002). This framework links multivariate regression methods (including GLM to deal with non-continuous data) for the assessment of covariate effects, with geostatistical methods to model the spatial structure. This also allows the prediction of the underlying spatial process and the visualisation of predicted values in areas for which no observations are available (kriging). The assessment of covariate effects can be used to gain a greater understanding of factors contributing to the observed spatial distribution in the same manner as with the use of non-spatial multivariate regression (as discussed in Chapter 3, Section 3.1.2). As an adjunct to the robust estimation of parameters, hypothesis testing can be carried out, for example to look at the significance of specific explanatory variables on the observed distribution.

Likelihood based methods of inference for generalised linear geostatistical models are computationally very intensive and, thus, Bayesian inference is the preferred method of implementation (Diggle and Ribeiro Jr 2007). In the Bayesian inference of model parameters, realisations from the posterior and predictive distributions are generated using Markov Chain Monte Carlo (MCMC) methods (Gelman *et al.* 2004; Diggle and Ribeiro Jr 2007). MCMC methods involve the construction of a Markov chain (a mathematical representation of a random process where future values are conditionally independent of past values, and depend only on the present value), with the desired probability distribution at its equilibrium (the stationary posterior distribution which the chain will converge to following a suitable number of iterations). Samples are then drawn from the equilibrium distribution and summarised to provide parameter estimates, quantiles and other measures of the distribution (Gelman *et al.* 2004). Bayesian inference of generalised linear geostatistical models has been used for a variety of different applications in recent years, including several epidemiological problems (Diggle *et al.* 2002; Stevenson *et al.* 2005; Kazembe *et al.* 2006; Diggle *et al.* 2007; Craig *et al.* 2007; Hay *et al.* 2009).

Bayesian inference for a generalised linear geostatistical model as described in the model-based geostatistics framework has two key benefits when compared with frequentist approaches such as those used in Chapter 3. Firstly, the explicit incorporation of spatial autocorrelation in the analysis enables more accurate estimation of covariate effects and their significance (Boyd *et al.* 2005; Clements *et al.* 2006c), allowing the generation of strengthened support for hypotheses with regards to the implication of specific factors in the observed spatial distribution. Secondly, the use of a fully Bayesian implementation allows the incorporation of uncertainty at each step of the modelling process (Diggle *et al.* 1998; Best *et al.* 2005). These benefits make this type of approach suitable for the analysis of disease distribution data, where both the covariate effects and the predictive distributions are of interest. In addition, the incorporation of uncertainty in model outputs can assist the interpretation of results or add strength to evidence used for the targeting of control measures or future data collections.

4.1.1.2. Case study: malaria in Malawi

A model-based geostatistics framework was used by Kazembe *et al* (2006) for the analysis of malaria point-referenced prevalence in Malawian children. The significance of several environmental covariates including elevation, temperature, rainfall and potential evapotranspiration was assessed; firstly using univariate non-spatial logistic regression models and secondly using univariate spatial logistic models (generalised linear geostatistical models with the logistic link function). Following the univariate analyses, parameter estimation was carried out using a full Bayesian implementation of the generalised linear geostatistical model, incorporating covariates as determined to be significant using the univariate screening. In addition, prediction was carried out to provide a risk surface across the country, based on covariate effects and the autocorrelation structure. The final spatial model highlighted the significance of elevation, maximum temperature and potential evapotranspiration, with the predictive map showing high risk of malaria in lower elevation regions on the lake shore and lower risks in the highlands (Kazembe *et al.* 2006). The application of model-based geostatistics in this and other similar studies demonstrates the utility of these methods in epidemiological analysis; the results presented by Kazembe *et al* (2006) can be used to assist in the future planning and monitoring of malaria control activities.

4.1.1.3. Geostatistical analysis of sleeping sickness: previous studies

Despite recent advances in geostatistical methods, particularly for their use with non-continuous distributions, their implementation for the spatial analysis of sleeping sickness or animal trypanosomiasis distributions has not yet been undertaken. As with other vector-borne diseases, sleeping sickness and animal trypanosomiasis distributions (and also tsetse distributions) display spatial autocorrelation and the use of classical statistical methods without explicitly accounting for spatial structure may not give an accurate representation of covariate effects and their significance.

Regardless of the large amount of literature regarding tsetse modelling and mapping, such as Rogers *et al* (1996) and Hendrickx *et al* (2001b), the published applications

of geostatistical methods to tsetse distributions are thus far sparse. Kitron *et al* (1996) and Odulaja *et al* (2001) explicitly determine the existence of spatial autocorrelation in tsetse fly trap catches using Moran's *I* index, with Kitron *et al* (1996) going on to assess covariate effects on tsetse densities having used a spatial filtering method to remove the effects of spatial autocorrelation. Results indicated that spatial autocorrelation accounted for a large amount of the observed correlation between fly densities and Landsat thematic mapper (TM) Band 7. However, when the spatial effects were removed via filtering, the correlation remained significant for some of the time periods analysed (Kitron *et al.* 1996). More recently, geostatistical methods including kriging were used for the identification of areas with increased tsetse densities to allow the implementation of appropriate control measures (Sciarretta *et al.* 2005).

A different approach, discriminant analysis using categorical outcome variables (i.e. present or absent), has been used in several tsetse (Rogers *et al.* 1996; Robinson *et al.* 1997a; Hendrickx *et al.* 2001b) and animal trypanosomiasis (Hendrickx *et al.* 2000) mapping exercises. These types of methods include the construction of a set of linear functions of the predictor variables (discriminant functions) based on observed data. The discriminant functions are then applied to locations with an unknown outcome to predict the outcome class (i.e. present or absent) based on the values of predictor variables at those locations. Discriminant analysis differs from logistic regression methods (and, thus, the generalised linear geostatistical model) in the assumption that both the outcome and the explanatory variables should be normally distributed (Press and Wilson 1978; Lachenbruch and Goldstein 1979; Pfeiffer *et al.* 2008b); an assumption which may not always be met in epidemiological studies. In addition, spatial autocorrelation in the observations is not explicitly accounted for in standard discriminant analysis methodologies.

4.2. Specific objectives

Following on from previous work in Chapter 3, a generalised linear geostatistical model was used in a Bayesian framework to account explicitly for spatial

autocorrelation and incorporate uncertainty in the input data and model parameters. The previous analysis (Chapter 3) provided an in-depth exploration of the relationships between sleeping sickness occurrence and prevalence, but did not account explicitly for spatial autocorrelation in the data. This approach may result in inaccurate parameter estimates and inflated significance of covariate effects. The current analysis provides a more rigorous analytical approach, incorporating spatial autocorrelation explicitly, potentially resulting in more accurate parameter and significance estimates and increased predictive accuracy. Overall, the aim of this chapter was to investigate the utility of model based geostatistics for sleeping sickness prevalence and to provide a more accurate assessment of the environmental, climatic and social determinants of the spatial distribution of Rhodesian sleeping sickness prevalence within the two recently affected districts. Specifically, the validity of the hypothesis that the movement of untreated, infected livestock resulted in the introduction of *T. b. rhodesiense* to the study area was assessed, given a more robust parameter estimation and appropriate assessment of covariate effects.

4.3. Methods

4.3.1. Data

The analysis within this chapter used the same dataset as described in Chapter 3, Section 3.3.1; all sleeping sickness cases residing within Kaberamaido and Dokolo districts, linked to a geo-referenced database of all villages within the study area. Village level prevalence was calculated for all villages using population data from the Ugandan Bureau of Statistics (Uganda Bureau of Statistics 2008). Further information on the data acquisition can also be found in Chapter 2, Sections 2.4.2.1 and 2.4.3.1.

The values of each of the covariates listed in Table 4.1 were extracted for each village in the study area. These variables were found to be significant during the one-step logistic regression analysis of village level prevalence described in Chapter 3, Section 3.4.2. Further details on the provenance of these data can be found in Chapter 2 (Section 2.4.4).

Distance to closest livestock market	LST phase of annual cycle
Distance to closest health centre	NDVI phase of annual cycle
Distance to woodland	Mean LST
Maximum NDVI	LST annual amplitude
Minimum LST	

Table 4.1: Significant variables from previous logistic regression analysis of sleeping sickness prevalence

4.3.2. Statistical analysis

Non-spatial logistic regression methods were used to identify a set of environmental, climatic and social variables that were significantly correlated with sleeping sickness prevalence, as presented in Chapter 3. The covariates from the final one-step logistic regression model were used as the starting point for the current analysis (see Table 4.1 for the covariates included). The spatial variation in prevalence of sleeping sickness within Kaberamaido and Dokolo districts was modelled using model-based geostatistics (Diggle *et al.* 1998; Diggle and Ribeiro Jr 2007) with a spatial GLM and Bayesian inference of model parameters (Diggle and Ribeiro Jr 2007).

4.3.2.1. Model specification

The total number of sleeping sickness cases Y_i within village i was modelled as a conditionally independent binomial variable, $Y_i \sim Bin(n_i, p_i)$, where n_i is the total village population and p_i is the underlying population prevalence of sleeping sickness at location i. The method is an extension of a GLM using the logit link function, incorporating a stochastic spatial effect $S(\mathbf{x})$, where $\mathbf{x}$ is a vector of spatial locations, as follows:

$$Log\{p_i/[1-p_i]\} = \beta_0 + \beta_1 v1_i ... + \beta_k vk_i + S_i$$

where β_0 is the intercept term and β_1 to β_k are the regression coefficients relating to covariates $v1$ to vk.

The stochastic spatial component, $S(\mathbf{x})$, is modelled as a zero mean Gaussian process with variance σ^2 and autocorrelation function $\rho(d_{ij}, \theta)$, where $d_{ij} = x_i - x_j$ measures the Euclidean distance between x_i and x_j ; $\theta = \{\phi, \sigma^2, \tau^2\}$; ϕ is the range (effectively the maximum distance at which there is spatial autocorrelation between observations) and τ^2 is the relative nugget. This serves to model the spatial variation in the residuals after accounting for the covariates, $v1$ to vk .

The model parameters were estimated using a Bayesian framework with an MCMC algorithm in the package *geoRglm* (Christensen and Ribeiro Jr 2002) in the R statistical software (R Development Core Team 2006). Priors were selected for each parameter to represent prior knowledge of their distributions. Non-informative, uniform priors were selected for the regression parameters, β_k and the variance, σ^2 in the absence of prior knowledge. This allows the observed data to have the greatest influence on posterior distributions without being constrained by the choice of prior and can also improve MCMC convergence (Diggle *et al.* 2002; Clements *et al.* 2006b). The range parameter, ϕ, was fixed and optimised by minimising the mean squared error due to potential problems with the estimation of both ϕ and σ^2 (Zhang 2004). This acted as a compromise between statistical rigor and computational ease: the range parameter would ideally be estimated along with all other parameters, but previous trials indicated problems with the mixing and convergence of MCMC chains when both ϕ and σ^2 were being estimated. A Matérn correlation function was used (Diggle *et al.* 2002) with a discrete order (smoothness parameter) of 0.5. This equates to the use of an exponential correlation function; $\rho(d_{ij}) = \exp(-d_{ij} / \phi)$. The relative nugget parameter, τ^2, was fixed at 1 after inspection of the residual variogram.

4.3.2.2. Univariate parameter estimation

Due to convergence and mixing problems when including all significant covariates from the logistic regression model (as listed in Table 4.1), each of the explanatory

variables was examined independently using the model based geostatistics framework described. The MCMC algorithms were tuned to give an acceptance rate of approximately 60%, and the fixed prior for ϕ was optimised for each explanatory variable using several iterations to obtain a minimised mean squared error.

The univariate spatial models were run for 2,000,000 iterations, with the first 1,000,000 discarded and every 100^{th} iteration thereafter stored to assess the significance of each explanatory variable. Convergence and mixing of the MCMC algorithms were judged based on traceplots and autocorrelation plots for each model parameter to ensure that convergence had been reached, the chains had mixed adequately and autocorrelation amongst the samples was minimal. The mean parameter values from the posterior distribution and their 95% credible intervals (CrIs) were calculated and exponentiated to provide ORs and their respective uncertainty measures. Only those covariates which were significantly associated with sleeping sickness prevalence (i.e. the 95% CrI for the OR did not include the value 1) were selected for the multivariate spatial regression model.

4.3.2.3. Multivariate parameter estimation

An initial run of the multivariate model was carried out following the optimisation of ϕ and tuning of the MCMC algorithm as described in Section 4.3.2.2. The regression parameters and 95% CrI were inspected. Any covariates which were non-significant in the multivariate model (95% CrI for the OR crosses one) were discarded from the final model.

The fixed ϕ value was again optimised for the final multivariate model and the MCMC algorithm tuned to achieve an acceptance rate of approximately 60%. Following a burn-in of 1,000,000 iterations, the chain was run for a further 5,000,000 iterations, with every 1000^{th} iteration thereafter stored, resulting in a total of 5,000 samples from the posterior distributions. The regression parameters and 95% CrIs were obtained from the model and exponentiated as above.

4.3.2.4. Spatial predictions

A 2 km spatial resolution prediction grid was created for the study area, containing covariate values at each grid cell. Samples from the predictive distribution for each prediction location were generated using the MCMC algorithm, given the explanatory variables at the locations. The posterior medians and lower and upper 95% CrI limits were extracted from the predictive distributions to give predicted prevalence and uncertainty estimates at all locations. The predictions were exported to ArcMap 9.1 for illustrative purposes.

A scatter plot of predicted prevalence versus observed prevalence was created to illustrate the relationship between the model predictions and the observations. The correlation between fitted and observed prevalence was calculated using the *stats* package in the R statistical software. In addition, the mean error, median error and absolute mean error (calculated using prevalence per 100 population, therefore expressed as percentages) were calculated based on the difference between observed and predicted prevalence at each location, to give an indication of the prediction bias (mean and median error) and precision (absolute mean error). The Pearson residuals were calculated and plotted against the fitted prevalence values (McCullagh 1989). In addition, the residual variogram was plotted to examine any residual spatial autocorrelation, and this was compared with the residual variogram from the non-spatial logistic regression model.

4.4. Results

4.4.1. Univariate parameter estimation

From the univariate spatial regression, five variables which were significantly correlated with sleeping sickness prevalence using non-spatial logistic regression did not retain statistical significance (see Table 4.2 for ORs and 95% CrI). Four covariates retained their significance in the spatial regression analysis. Increasing distance from the closest livestock market had a protective effect in terms of sleeping sickness prevalence (OR = 0.81, 95% CrI = 0.76 to 0.86), with a 19% decrease in the

odds of sleeping sickness for every 1 km of distance. Increasing distance to the closest health centre also demonstrated a protective effect (OR = 0.85, 95% CrI = 0.76 to 0.95). Areas with higher minimum LST and larger NDVI phase of annual cycle had significantly increased odds of sleeping sickness (OR = 1.57, 95% CrI = 1.11 to 2.29 and OR = 3.25, 95% CrI = 1.02 to 10.07 respectively).

	Odds Ratio	**95% CrI**
Distance to closest livestock market	0.81	0.76 to 0.86*
Distance to closest health centre	0.85	0.76 to 0.95*
Maximum NDVI	$1.06E^{-5}$	$2.75E^{-12}$ to 36.60
Minimum LST	1.57	1.11 to 2.29*
LST phase of annual cycle	1.19	0.91 to 1.51
Distance to woodland	1.19	0.95 to 1.46
NDVI phase of annual cycle	3.25	1.02 to 10.07*
Mean LST	0.90	0.59 to 1.38
LST annual amplitude	0.86	0.66 to 1.13

Table 4.2: Odds ratios and 95% CrI from Bayesian univariate regression analysis. * indicates significance at the 95% level.

4.4.2. Multivariate parameter estimation

The NDVI phase of annual cycle covariate did not retain statistical significance when included along with the other three significant covariates in the multivariate spatial regression (OR = 1.73, 95% CrI = 0.61 to 5.00), and so was omitted from the final multivariate model.

The remaining three covariates retained statistical significance at the 95% level in the multivariate regression (see Table 4.3). Both increasing distance to the closest livestock market and increasing distance to the closest health centre had protective effects (OR = 0.83, 95% CrI = 0.78 to 0.88 and OR = 0.88, 95% CrI = 0.79 to 0.97 respectively). Additionally, areas with a higher minimum LST had increased odds of

sleeping sickness (OR = 1.49, 95% CrI = 1.09 to 2.10). The variance (σ^2) was estimated to be 1.17 (95% CrI = 0.74 to 1.75).

The posterior distributions for all parameters were normally distributed; although for σ^2 there was a slight positive skew (see Figure 4.1 for posterior distribution curves for the final model parameters). Traceplots (see Figure 4.2) and autocorrelation plots (see Figure 4.3) for model parameters were examined to assess the mixing and convergence of the MCMC algorithms and each appeared to have reached convergence during the burn-in period and to be mixing well. Autocorrelation amongst samples was minimal.

	Odds ratio	**95% CrI**
Intercept	$9.02E^{-6}$	$4.77E^{-8}$ to 0.001
Distance to closest livestock market	0.83	0.78 to 0.88
Distance to closest health centre	0.88	0.79 to 0.97
Minimum LST	1.49	1.09 to 2.10
σ^2 (variance)	*1.17**	*0.74 to 1.75*

Table 4.3: Odds ratios and 95% CrI from Bayesian multivariate regression analysis. * Indicates variance value rather than OR.

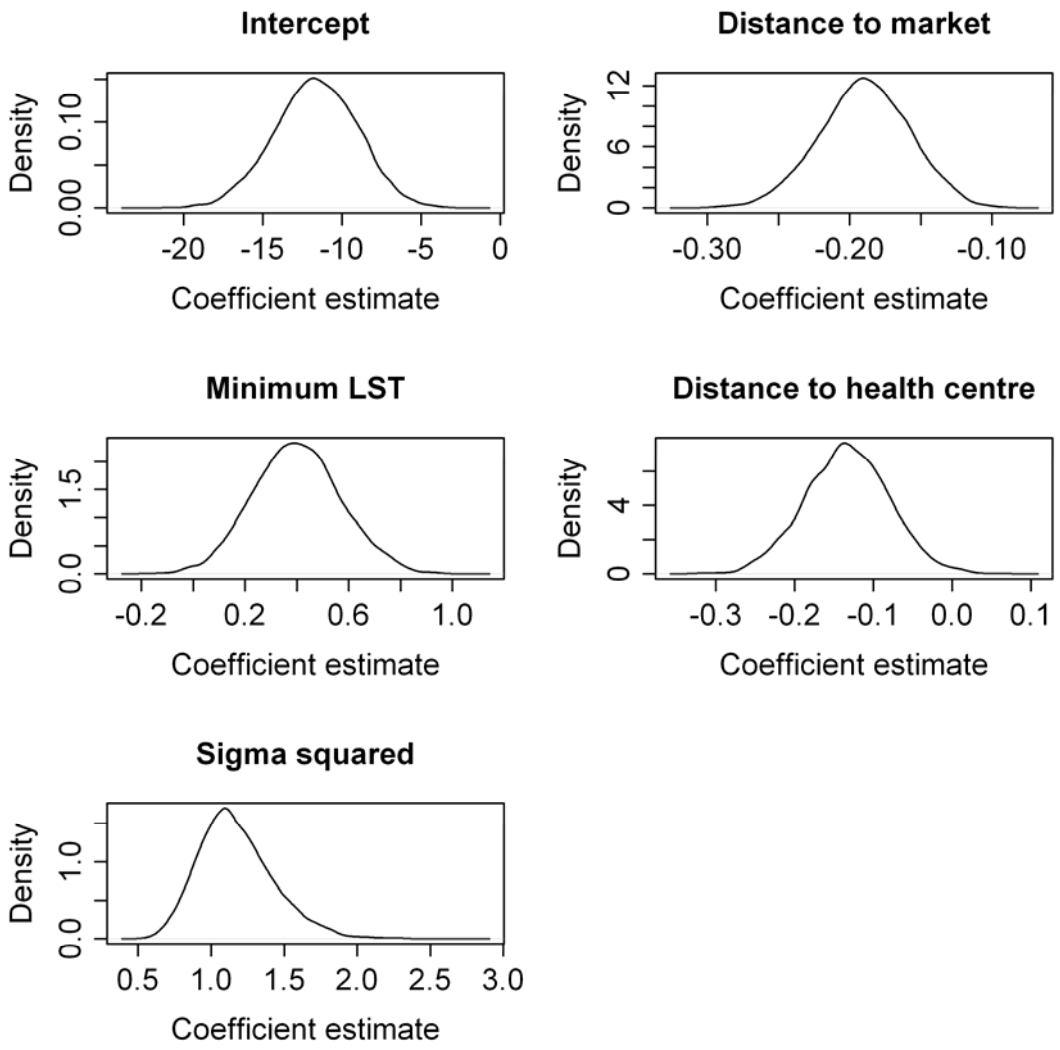


Figure 4.1: Posterior distributions for model parameters

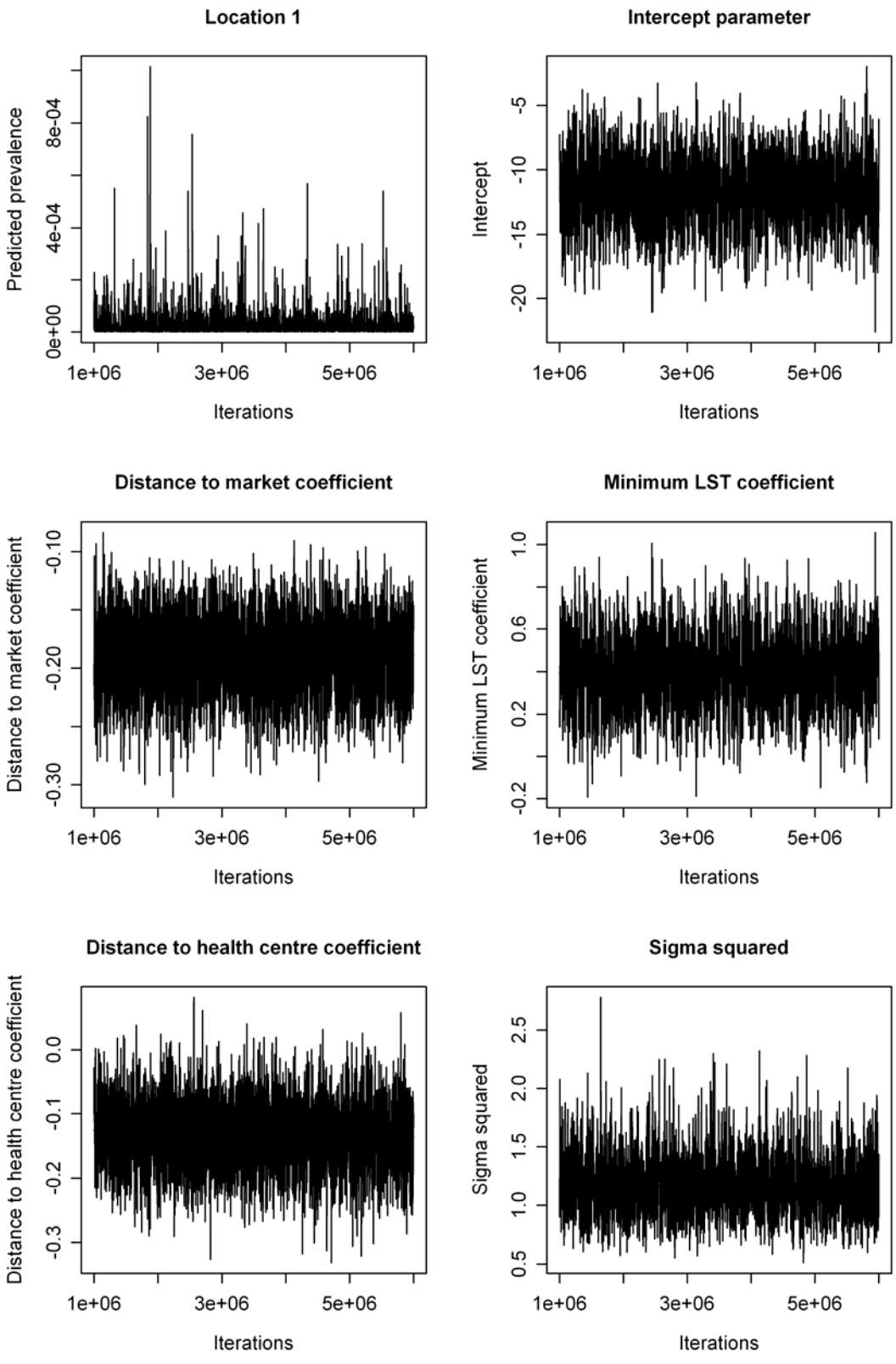


Figure 4.2: Traceplots of MCMC output for each parameter.

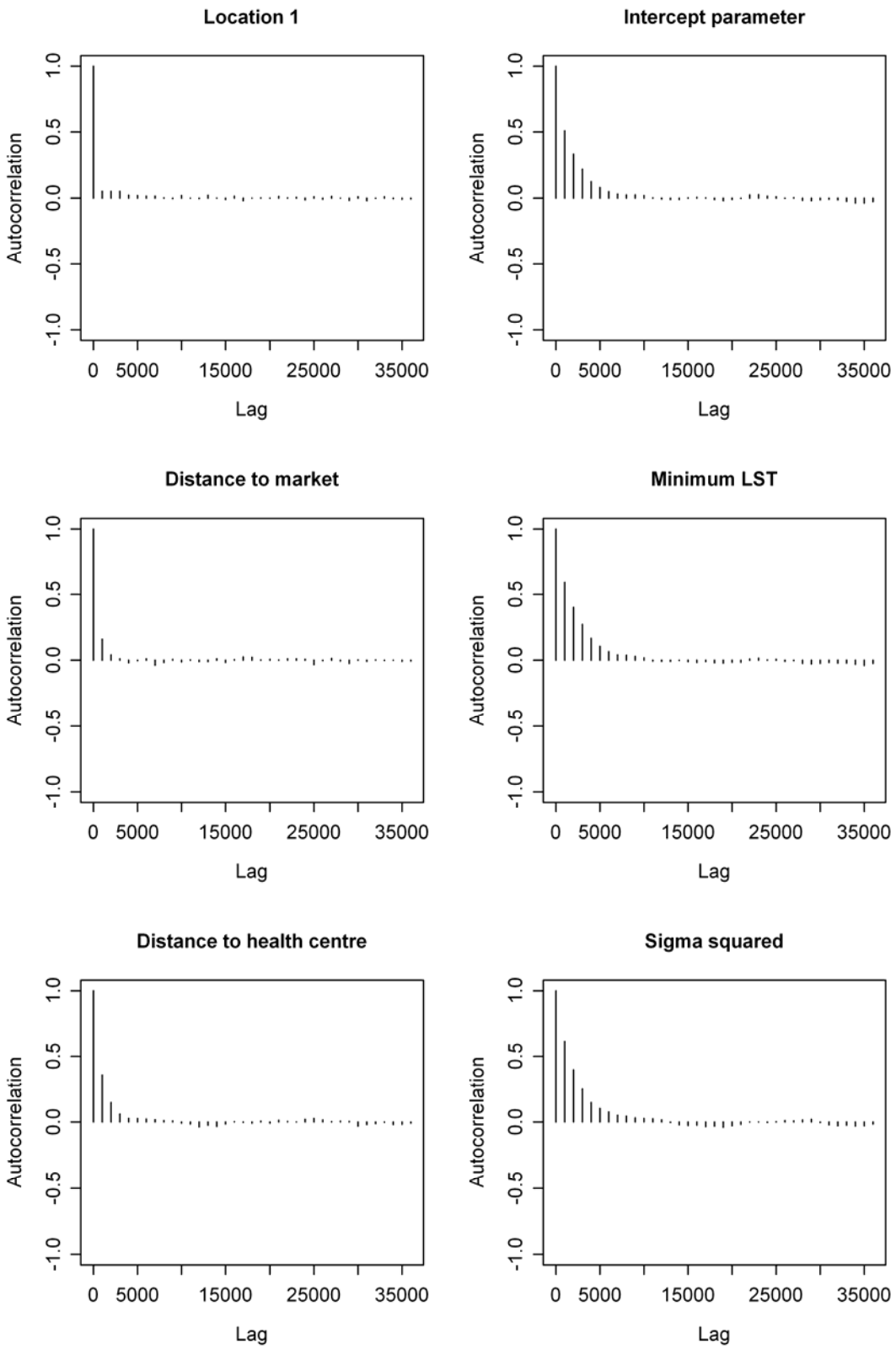


Figure 4.3: Autocorrelation plots of MCMC output for each parameter.

4.4.3. Spatial predictions

The predicted prevalence surface from the final spatial model is illustrated in Figure 4.4a and is overlaid with observed village prevalence data in Figure 4.4b. Figures 4.4c and 4.4d illustrate the lower and upper 95% credible limits for the prediction. The area of highest predicted prevalence within the study area corresponds with the majority of high prevalence villages. Several potential high prevalence areas out with the study area correspond to areas surrounding livestock markets, with the effect of distance to the closest health centre and minimum LST also accounted for. The areas with largest predicted prevalence also have the largest 95% CrIs, which is due to the greater variability of observed village level prevalence in the high prevalence areas (i.e. villages with high prevalence are interspersed with zero prevalence villages within the high prevalence area).

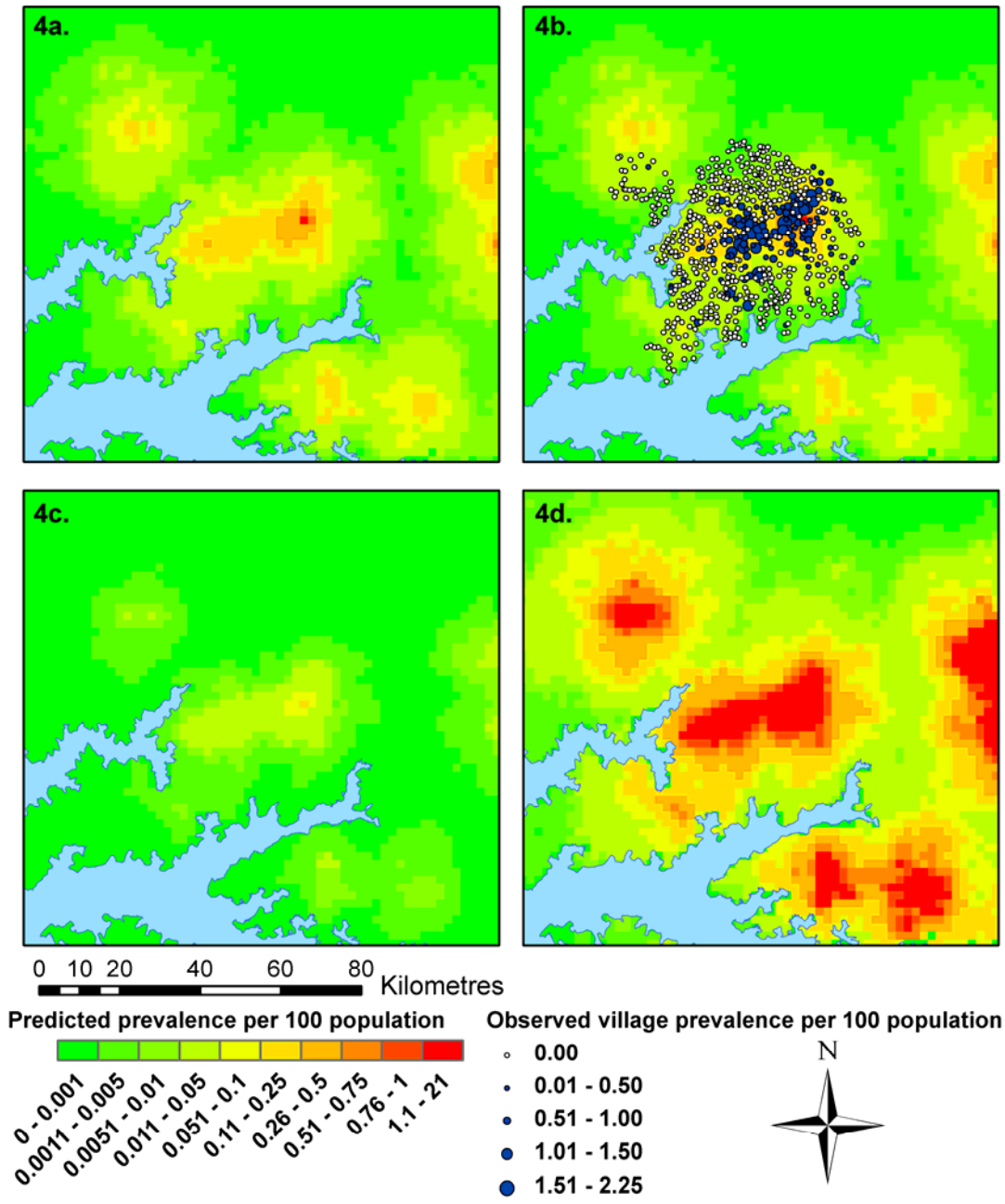


Figure 4.4: Predicted prevalence of sleeping sickness per 100 population from final spatial model (4.4a), observed village prevalence (4.4b) and lower (4.4c) and upper (4.4d) 95% credible limits.

A plot of predicted prevalence versus observed prevalence (Figure 4.5) shows a tendency to under-predict the prevalence in high prevalence villages and over-predict in zero prevalence villages, with an overall correlation between observed and fitted prevalence of 0.95. The predicted prevalence (expressed as prevalence per 100 population, therefore a percentage) had a mean error of -0.00094%, a median error of 0.018% and an absolute mean error of 0.064%. A plot of the Pearson residuals versus

fitted prevalence (Figure 4.6) highlights the propensity for over-predicting in zero prevalence villages and under-predicting in high prevalence villages.

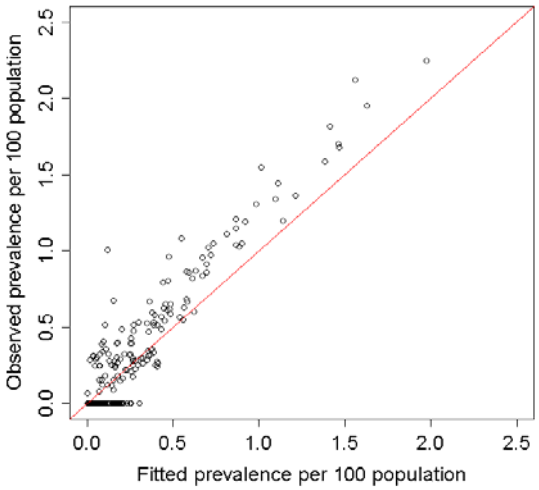


Figure 4.5: Fitted village prevalence versus observed village prevalence.

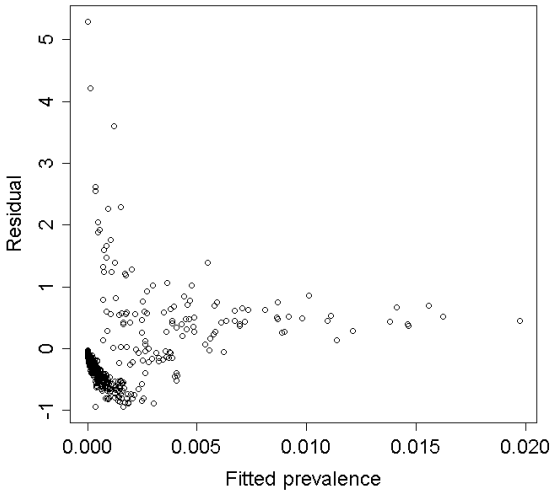


Figure 4.6: Fitted village prevalence versus Pearson residuals

The empirical variogram of the Pearson's residuals from the non-spatial model discussed in Chapter 3 (Figure 4.7a) indicates the presence of some unexplained spatial variation in the residuals. The residual spatial autocorrelation from the spatial model (Figure 4.7b) gives a flatter variogram, with a smaller amount of residual variation than the non-spatial model, indicating that the spatial model has accounted for a larger amount of the spatially correlated variation in the prevalence data. Overall, the diagnostics show that although the spatial model results in less residual variation and greater correlation between observed and predicted prevalence, there is still some residual spatial variation in sleeping sickness prevalence within the study area which is not being accounted for; in particular, several zero-prevalence villages have higher predicted prevalence than was observed.

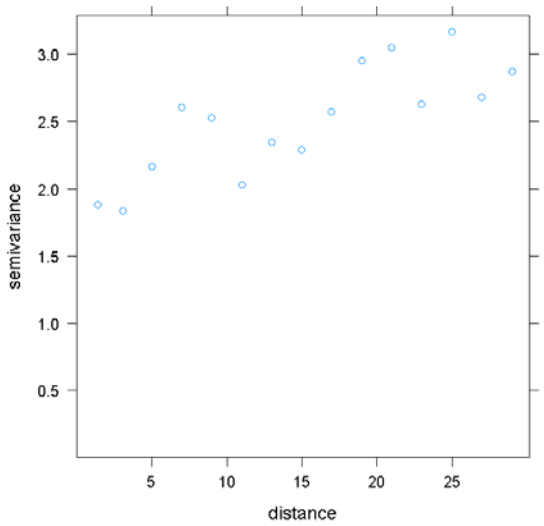


Figure 4.7a: Residual variogram using Pearson residuals from the non-spatial model

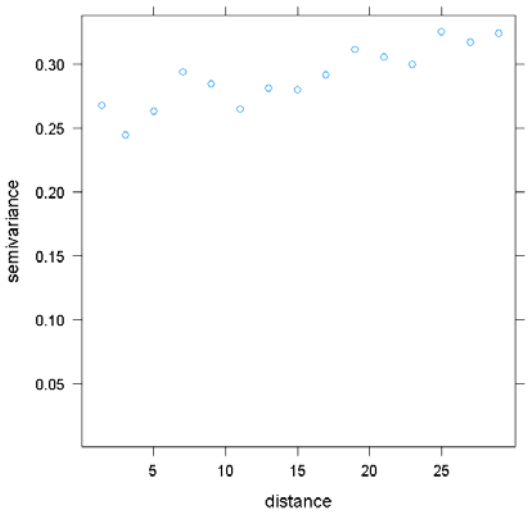


Figure 4.7b: Residual variogram using Pearson residuals from the spatial model

4.5. Discussion

The results presented in this chapter extend an initial (non-spatial) analysis as discussed in Chapter 3. A spatial analysis was conducted, in which a generalised linear geostatistical model was applied with Bayesian implementation, as described by Diggle *et al* (1998). This method explicitly accounts for spatial autocorrelation and incorporates uncertainty in the input data and model parameters. This approach allowed a more robust assessment of covariate effects, thereby strengthening the hypothesis that Rhodesian sleeping sickness was introduced into Kaberamaido and Dokolo districts via the movement of infected livestock. In addition, the significant relationships between sleeping sickness prevalence and environmental, climatic and social factors detected using the non-spatial regression in Chapter 3 have been clarified.

As a starting point for the spatial model, it would have been preferable to include all covariates from the final non-spatial logistic regression model. Any covariates which were found to have become non-significant when accounting for residual spatial autocorrelation would then be removed prior to the final fitting of the model. However, when including all covariates (as listed in Table 4.1), problems with the

convergence and mixing of the MCMC algorithms were encountered. For any application of an MCMC algorithm it must be ensured that the Markov chain has been run for a large enough number of iterations to converge to the required posterior distribution and is mixing well throughout the density distribution: this means that the chain will be able to thoroughly sample the posterior distribution (Gelman *et al.* 2004; Lynch 2007). It is likely that the problematic MCMC performance with the full multivariate model was due to correlation and redundancy amongst some of the covariates and potentially also difficulties in estimating a large number of parameters at the same time. In addition, Zhang (2004) found that ϕ and σ^2 cannot be estimated consistently, potentially creating problems with parameter estimation and MCMC convergence and mixing. The problems encountered with the mixing and convergence of the Markov chains were prevented by modelling each covariate independently and fixing the range parameter, ϕ (after optimisation by minimising the mean squared error).

Following on from the non-spatial logistic regression methods discussed in Chapter 3, five of the covariates that retained significance in the multivariate one-step model of sleeping sickness prevalence did not retain statistical significance using the Bayesian implementation of a spatial logistic regression model. This indicates that the non-spatial model may have inflated the significance of covariates and produced inaccurate parameter estimates. In addition, this may signify the presence of redundancy in the covariates used in the final non-spatial logistic regression model. A further covariate (NDVI phase of annual cycle) did not retain statistical significance when accounting for the effects of distance to the closest livestock market, distance to the closest health centre and minimum LST in the multivariate spatial regression.

The final spatial model included three covariate effects: distance to the closest livestock market, distance to the closest health centre and minimum LST. These results, using a more robust assessment of covariate effects, provide considerable strength to the hypothesis that the movement of infected, untreated livestock from endemic areas resulted in the introduction of *T. b. rhodesiense* to Kaberamaido and

Dokolo districts. Previous research has established that the introduction of Rhodesian sleeping sickness transmission within Soroti district (which neighbours the study area) was due to the movements of untreated cattle from endemic areas through a local livestock market (Fèvre *et al.* 2001a). The results discussed here, supported by the findings discussed in Chapter 3, strongly indicate a similar occurrence in Kaberamaido and Dokolo districts; *T. b. rhodesiense* is likely to have been introduced to the area via the continued movement of untreated livestock, despite the introduction of a law requiring the treatment of livestock from endemic areas prior to sale (Wendo 2002). The significance of distance to the closest health centre in the spatial regression model strengthens the evidence of a confounding effect created by the level of access to health care services, as noted in Chapter 3 and in previous studies (Odiit *et al.* 2004b; Odiit *et al.* 2006). Individuals living in more remote areas and further from health care services may be less likely to access treatment and, thus, be diagnosed with and treated for sleeping sickness.

Minimum LST was observed to be a risk factor for sleeping sickness, with higher prevalence in areas with higher minimum LST. Minimum LST is calculated using measurements of radiance modified by the atmosphere in several spectral wavebands and varies depending on climate and landcover properties (e.g. amount of vegetation, urbanisation or soil moisture) (Dash *et al.* 2002). The size of the study area (approximately 60 km by 60 km) suggests that the observed correlations are more likely to be due to the heterogeneous landcover profile and soil and vegetation moisture contents than to climatic variability across the two districts, although the precise interpretation of this predictor is not clear. Further work is planned to disentangle the effects of climate and landcover by encompassing a larger study area; the proposed research will investigate the temporal dynamics of the distribution of sleeping sickness in relation to climatic, environmental and social covariates (including temperature, rainfall and landcover classes).

From the presented analysis, as well as the results discussed in Chapter 3, it can be seen that distance to the closest livestock market is the factor with the largest effect on the distribution of Rhodesian sleeping sickness in Kaberamaido and Dokolo

districts. The predictive maps highlight a number of areas of elevated risk (higher predicted prevalence) which correspond to the locations of livestock markets. From these results it appears likely that within Kaberamaido and Dokolo districts during the study period, the disease had not yet fully dispersed into its preferred environmental niche, as the observed distribution seems to be constrained primarily by proximity to the cattle market. It is possible that reintroduction of the parasite at one or more of the local livestock markets throughout the study period, rather than the hypothesised single introduction at the beginning of the study period, has contributed to the observed constrained distribution. It would be expected that over time, the correlation between sleeping sickness prevalence and environmental factors would increase as the disease has a chance to disperse away from the hypothesised site of introduction into more suitable habitats.

When the performance of the spatial regression model is compared with that of the non-spatial model (one-step model of prevalence) from Chapter 3, the predictions from the spatial model are seen to be more accurate. The correlation between observed and fitted prevalence for the non-spatial model was 0.58, compared with a correlation of 0.95 for the spatial model. The absolute mean error for the non-spatial model was 0.13%; double that of the spatial model (0.064%). Despite the increase in accuracy gained by modelling the residual spatial autocorrelation after accounting for covariate effects, there was still a tendency to over-predict in zero prevalence villages and under-predict in high prevalence villages. The over-prediction in zero prevalence villages indicates the presence of extra-binomial variation (greater variability in the observations than can explained by the model), whereby additional unmeasured factors may be influencing the spatial heterogeneity of sleeping sickness occurrence within small areas. From the observed prevalence it can be seen that within the main 'focus' of infection there are many zero prevalence villages interspersed amongst high prevalence villages, which are not explained adequately by the spatial regression model. The estimates of model uncertainty (95% CrIs) also highlight this, with larger predictive uncertainty in the areas with higher predicted prevalence as can be seen in Figures 4.4c and 4.4d. This non-constant variance in the error is known as heteroscedasticity. Future research as described above aims to deal

with these issues by utilising a wider range of covariate datasets, with finer spatial resolutions.

The research presented in this chapter illustrates the importance of spatial autocorrelation in epidemiological variables; the use of non-spatial logistic regression analysis resulted in a model with a large number of covariates, complicating the interpretation of their effects. The use of a generalised linear geostatistical modelling framework, which models the residual autocorrelation after accounting for covariate effects, gave more precise and less biased parameter and significance estimates, with only three covariates retaining significance in the final model. The Bayesian implementation of the method allowed the incorporation of uncertainty in each of the model parameters from the posterior distributions and from the definition of a random variable. By carrying out the spatial-regression analysis, the quantified relationships between sleeping sickness prevalence and significant covariates can be more confidently described and interpreted. The predictive accuracy was also increased by using the spatial regression when compared to the non-spatial logistic regression analysis. These results strengthen the evidence in support of the hypothesis generated by the analysis discussed in Chapter 3; that the movement of untreated, infected livestock from endemic areas resulted in the introduction of Rhodesian sleeping sickness to the study area.

Chapter 5. The temporal landscape epidemiology of *Trypanosoma brucei rhodesiense* in Soroti district

“Science is facts; just as houses are made of stones, so is science made of facts; but a pile of stones is not a house and a collection of facts is not necessarily science."

Jules Henri Poincaré (1854 - 1912)

5.1. Introduction

5.1.1. Landscape epidemiology

The environmental landscape within an area is a significant factor in the spatial distribution of many different disease vectors, reservoirs, intermediate hosts and parasites and, thus, also the spatial distribution of a variety of diseases, including sleeping sickness. Populations of vectors, reservoir or intermediate hosts (including insects and small mammals) are often linked explicitly to land cover due to specific habitat and environmental requirements for their development and survival. In addition, the survival and reproduction of many parasite species depends upon external environmental conditions. These links between land cover; vector, reservoir or intermediate host distributions and disease distributions can be quantified and described to allow a greater understanding of the relationships and highlight areas with potentially higher risks of vector or reservoir presence, disease transmission, or both (Kitron 1998; Curran *et al.* 2000; Ostfeld *et al.* 2005). The study of land cover and other environmental landscape features in relation to distributions of vectors, reservoirs or intermediates hosts and human or animal disease is commonly known as landscape epidemiology, after the phrase was coined in the 1960s (Pavlovskii 1966).

5.1.1.1. Case study: alveolar echinococcosis in China

Following initial work by Craig *et al* (2000) and Danson *et al* (2003), which highlighted the correlation between grassland distributions and human alveolar echinococcosis cases within China, landscape analysis was used to quantify the relationships between landscape composition and human disease (Danson *et al.* 2004). The transmission cycle of the causative agent, *Echinococcus multilocularis,* normally involves an intermediate small mammal host (i.e. rodents) and a definitive canid host (i.e. fox or dog) and is dependent on the availability of suitable habitat for the intermediate host, as well as the environmental conditions necessary for egg survival. Human infection occurs via the accidental ingestion of eggs from direct contact with an infected canid or indirect environmental transmission. Thus, the

landscape directly surrounding a village is thought to be fundamental to the risk of alveolar echinococcosis within the village population, due to the zoonotic transmission cycle (Craig *et al.* 2000).

To further quantify the relationships between human alveolar echinococcosis and landscape composition, the total area of each land cover type (from a classification of Landsat TM imagery) was extracted from within a range of buffer sizes around each village, and expressed as the proportion of the total area within the buffer zone. Several buffer sizes ranging from 500 m to 3.5 km were used to identify the scale at which land cover composition is most significant for disease transmission to humans. A positive correlation was found between human infection and the area of land surrounding the village which was forest, grassland or shrubland, and a negative correlation with the area that was agricultural land. Variations in these correlations were observed at each of the different spatial scales (Danson *et al.* 2004). In this case, landscape epidemiology emphasized the dependence of the distribution of human alveolar echinococcosis on the peri-domestic landscape composition at different spatial scales.

The human alveolar echinococcosis research highlights important aspects of landscape epidemiology; there are many parallels between the transmission of human alveolar echinococcosis and that of Rhodesian sleeping sickness. The research specifically accounts for interaction between village inhabitants and the landscape within and outside of the village at varying spatial scales. Transmission of sleeping sickness normally occurs outside of the village of residence and is dependent upon contact between humans, livestock (as the primary reservoir of the disease) and the tsetse vector species (Wyatt *et al.* 1985; Okia *et al.* 1994). Due to the environmental and habitat requirements of the tsetse vector, the interactions between individuals and the landscape directly surrounding their village of residence are important in determining sleeping sickness risk. Therefore, similar methods as those used by Danson *et al* (2004) may be of particular value to the study of sleeping sickness and the disease's landscape epidemiology.

Similar landscape epidemiology methods have been used for a variety of other epidemiological applications. A sample of such studies have focussed on the landscape epidemiology of malaria (Beck *et al.* 1994), Lyme disease (Maupin *et al.* 1991; Dister *et al.* 1997) and West Nile virus (Brownstein *et al.* 2003).

5.1.1.2. The landscape epidemiology of sleeping sickness: previous studies

Within the study areas the predominant species of tsetse vector is *Glossina fuscipes fuscipes*, a member of the *palpalis* group that is restricted to riverine vegetation habitats (patches of vegetation on the banks of rivers, lakes or wetlands) (Food and Agricultural Organization of the United Nations 1982; Leak 1999). Due to this association with particular types of land cover, sleeping sickness (or tsetse) distributions can be correlated with landscape information that captures the distribution of potential tsetse habitats.

Despite the reliance of tsetse populations and, thus, sleeping sickness distributions on the availability of suitable habitat, there have been relatively few studies examining the distribution of sleeping sickness in relation to land cover. A recent study by Odiit *et al* (2006) as discussed in Chapter 3, Section 3.1.2.4, used a classified Landsat image to assess the significance of land cover to the distribution of Rhodesian sleeping sickness cases in an area of eastern Uganda. This study demonstrated an increased risk of sleeping sickness in areas close to 'long vegetation swamp' habitats. Further work by Zoller *et al* (2008) revealed that homes between 500 m and 3 km from areas of wetland have a significantly increased risk of Rhodesian sleeping sickness. A small number of other studies have examined tsetse populations and risk of sleeping sickness in relation to the presence of particular crop types (such as coffee or cocoa) (Fournet *et al.* 1999; Fournet *et al.* 2000), and the level of human land use and disturbance of vegetation (Mahama *et al.* 2005; Bouyer *et al.* 2006). These studies were conducted in western Africa (where Gambian sleeping sickness occurs and the tsetse species responsible for sleeping sickness transmission differ from those predominant in Uganda) and therefore the observed relationships may differ from those in Uganda.

Since 1987 Rhodesian sleeping sickness has spread within Uganda and established endemnicity in eight previously unaffected districts (Matovu 1982; Okiria 1985; Fèvre *et al.* 2001a; Fèvre *et al.* 2005; Picozzi *et al.* 2005). The successful invasion of a new area by a species requires several steps: the introduction of the species to that area via long distance movements; the establishment of transmission following the introduction; an increase in the abundance of the species within the area of introduction and the subsequent spatial dispersion of the species via short distance movements (Peterson 2008; Schreiber and Lloyd-Smith 2009). In the case of Rhodesian sleeping sickness, the introduction of the parasite into new areas has been brought about via the movement of infected, untreated livestock from endemic areas (Fèvre *et al.* 2001a; Fèvre *et al.* 2005). Following transmission establishment and an increase in prevalence (which is most likely to have occurred within the livestock reservoir prior to the occurrence of human cases), dispersion away from the point of introduction has been observed (Fèvre *et al.* 2001a). The dispersal of species following an introduction can increase the probability of a successful invasion: the movement of the species is guided by environmental cues into higher quality habitats. The fitness of the organism in terms of survival and reproduction will be enhanced in the more suitable habitats and, thus, successful persistence in the new environment is more likely (Schreiber and Lloyd-Smith 2009). As discussed above, the influence of environmental and climatic factors, including proximity to specific types of land cover (specifically areas of wetland), on the spatial distribution of tsetse vectors and Rhodesian sleeping sickness is well known, although few studies have made use of these relationships to further the understanding of sleeping sickness distributions. The influence of such relationships on the spatial dispersal of sleeping sickness, following its introduction to a previously unaffected area, has not yet been investigated, despite the potential practical applications of such knowledge for targeted disease control activities.

5.1.2. Sources of land cover data

Landscape epidemiology investigations necessitate the availability of up-to-date, accurate land cover maps along with knowledge of vector requirements to enable the

identification of areas with suitable habitat for the vector species and, thus, the identification of areas with a higher disease risk. However, access to detailed, accurate and up-to-date land cover maps is often limited, particularly in resource poor settings and field surveys to obtain such information can be extremely costly and time consuming. By using remotely-sensed images from Earth observing satellites, information about Earth's biosphere, oceans, land surface and atmosphere can be obtained in a timely and standardised manner for use in epidemiological research (Hay *et al.* 1997; Curran *et al.* 2000; Goetz *et al.* 2000). Ground-based surveys and other traditional methods for the production of land cover maps can be complemented by the classification of remotely sensed images, thus, allowing the production of land cover maps for any area in the world, at multiple time points and at a variety of spatial scales (Anderson *et al.* 1976). Remote sensing imagery covering the entire globe is available from an array of satellite sensors, each with different attributes, allowing the application of remotely sensed data in many different areas and for a variety of purposes (Hay *et al.* 1997). Land cover maps can be created for remote, rural and difficult to reach locations with a minimum of ground-based surveying required, giving an affordable alternative to the more resource intensive ground-based methods.

A frequently used source of remotely sensed data for use in land cover classifications is the Landsat programme. The first Landsat satellite was launched in 1972, and there have been a further five launched since, each with a different combination of sensors. The currently orbiting satellite, Landsat 7, carries the Enhanced Thematic Mapper Plus (ETM+) sensor onboard which collects data in seven bands from the blue, green, red, near-infrared, mid-infrared (2 bands) and thermal infrared sections of the electromagnetic spectrum as well as a panchromatic band at a finer spatial resolution (further information on the electromagnetic spectrum is provided in Section 5.1.2.1) (Lillesand *et al.* 2004c; Jensen 2007b). The ETM+ sensor obtains images of the entire surface of Earth every 16 days (Landsat 1-3 obtains images every 18 days, Landsat 4-7 every 16 days) and the archived Landsat images from Landsat 1 through to Landsat 7 provide a comprehensive catalogue of images of Earth from 1972 to the current time. Since the launch of the TM sensor on Landsat 4 in 1982, the same

spectral bands have been continually collected (bands 1 to 7; see Table 5.1 for information on the spectral bands) to allow for greater continuity in data collection. The panchromatic band was subsequently added on the Landsat 7 satellite in 1999 (Jensen 2007b).

Band	Spectral range (microns)	Spatial resolution (metres)	Spectral characteristic
1	0.45 to 0.52	30	Blue-green
2	0.52 to 0.60	30	Green
3	0.63 to 0.69	30	Red
4	0.76 to 0.90	30	Near infrared
5	1.55 to 1.75	30	Mid infrared
6	10.4 to 12.5	60	Thermal infrared
7	2.08 to 2.35	30	Mid infrared
Panchromatic	0.50 to 0.90	15	Blue-green to near infrared

Table 5.1: Spectral bands collected by the ETM+ sensor onboard the Landsat 7 satellite.

The recent move to allow free access to the entire Landsat archive and all new Landsat scenes has greatly increased the opportunities for land cover classification and the application of remotely sensed images to a variety of research problems. The relatively fine temporal (16 days) and spatial (30 m) resolutions of Landsat TM and ETM+ images makes them ideal candidates for the production of land cover classifications at the regional or countrywide level. There are many examples of land cover classifications using images obtained by Landsat sensors in the literature, with a variety of different epidemiological applications including sleeping sickness (Odiit *et al.* 2006), malaria (Pope *et al.* 1994), Lyme disease (Glass *et al.* 1995) and also the alveolar echinococcosis study discussed in Section 5.1.1.1 (Danson *et al.* 2004). A recent review found that Landsat was the second most frequently used satellite in published epidemiological work, with very few studies utilising expensive fine spatial resolution imagery (Herbreteau *et al.* 2007).

5.1.2.1. Land cover classification methods

The type of energy measured and recorded in remote sensing is electromagnetic radiation; a form of energy made up of oscillating electric and magnetic fields, which are perpendicular to each other, and to the direction of propagation. For the purposes of remote sensing, the source of electromagnetic radiation is usually the Sun. Different types of electromagnetic radiation include visible light, radio waves, infrared and x-rays, each of which has a different wavelength (distance over which the wave repeats) (Jensen 2007a). The electromagnetic spectrum includes all possible wavelengths of electromagnetic radiation, from radio waves (the longest wavelengths; 10^3 m) through visible light, which makes up only a small proportion of the entire electromagnetic spectrum (intermediate wavelengths; 0.5 x 10^{-6} m), to gamma rays (the shortest wavelengths; 10^{-12} m). Different types of surface have different spectral properties: they absorb, transmit and reflect electromagnetic radiation of different wavelengths in different proportions. Surfaces which appear green to the human eye are actually reflecting the green wavelengths more than the red and blue wavelengths (Lillesand *et al.* 2004a; Jensen 2007a).

Based on these properties, different types of land cover have different spectral profiles (the proportion of light that they reflect in a range of wavelengths). For example, areas of land with dense, green vegetation will have high reflectance in the green wavelengths, and low reflectance in the red and blue wavelengths (Campbell 2002a). Earth observation satellites contain sensors which record information on the amount of electromagnetic radiation reflected by the Earth's surface in particular wavelengths (Jensen 2005a), and it is using this information in comparison with the reflectance patterns in areas of known land cover types that a land cover classification map can be produced from a remotely sensed image (Jensen 2005b).

The extraction of meaningful information from remotely sensed images relies on an array of statistical algorithms, each with different characteristics and suitable for diverse types of applications. Land cover classification aims to assign each pixel within the image (or each object within the image, as described below) to a particular

land cover class, dependent on its spectral characteristics in comparison with the characteristics of other pixels (or objects) of known land cover type (Campbell 2002c).

5.1.2.1.1. Supervised versus unsupervised classifications

A priori knowledge regarding the land cover classes occurring at defined locations or within defined areas (training sites) from various sources can be used in supervised classification methods. The spectral properties of each of the training sites is assessed, and every pixel (or object) within the image is assigned to the land cover class to which it is most similar. In contrast, unsupervised classification methods do not use any existing knowledge regarding the spectral attributes of land cover classes. For this type of classification, algorithms group pixels in the image into homogeneous clusters based on the spectral responses in several wavelengths. Unsupervised classification takes advantage of the naturally occurring homogeneity within land cover classes, and following the clustering of pixels the user will re-label (and may also combine) the identified groups into land cover classes relevant to the proposed application. Many different classification algorithms, both supervised and unsupervised, have been developed. Hybrid classifications may also be used which link together an unsupervised and a supervised classification, the first to obtain groups of spectrally similar pixels, and the second to match these groups to *a priori* land cover information, resulting in a land cover classification.

5.1.2.1.2. Pixel-based versus object-based classification

The traditional approach to land cover classification is pixel-based (pixels are the smallest discrete units in a remotely sensed image), where every pixel within an image is assigned to a land cover class. However, the use of single pixels in the classification algorithm discards much of the relevant information that is inherently present within an image. When looking at an image, the human eye automatically groups areas of similar colour together, and using previous experience and knowledge these observed groups can provide far more meaningful information than

looking individually at each pixel. Object-based classification methods utilise a segmentation step prior to classification to group pixels together based on their spectral similarity and proximity to one another, which mimics the automatic grouping of information by eye. The segmentation algorithm aims to extract objects from the image with minimised internal heterogeneity (Blaschke *et al.* 2000; Benz *et al.* 2004). The use of objects in the classification procedure (as opposed to pixels) provides a wealth of additional information which is not available in pixel-based analyses. Pixel-based classification is usually constrained to using the reflectance values for each band in the image, but when using image objects (groups of similar pixels), additional attributes such as object mean or standard deviation of reflectance values, object shape parameters, band ratios, texture or contextual information regarding relationships to neighbouring objects can be used in the classification procedure (Blaschke and Strobl 2001; Benz *et al.* 2004). By using a combination of different types of attributes in the classification procedure, it is possible to differentiate between classes that are spectrally very similar and, therefore, may be problematic when using traditional pixel-based methods.

The final classification result is also based on objects rather than pixels, which more closely relates to patterns of 'real-world' objects (i.e. an area of forest or a village) than a pixel-based classification and is, therefore, arguably a more intuitive process (Benz *et al.* 2004). In addition, pixel-based classifications can suffer from 'noise' in the classification (when individual pixels are assigned to a different class from surrounding pixels). Due to the individual pixels being classified independently of one another, with spatial autocorrelation in land cover not normally being accounted for, pixel-based classifications can result in a 'salt-and-pepper' effect, whereas using an object-based method will produce a far smoother classification will less noise in the final product.

Although object-based classification is a relative newcomer to the portfolio of classification methods, it has been used for several different applications, including the mapping of urban-sprawl (Jacquin *et al.* 2008), ecological habitat identification (Jobin *et al.* 2008), urban biodiversity assessments (Mathieu *et al.* 2007), tree

mortality (Guo *et al.* 2007) and landscape epidemiology (Koch *et al.* 2007), in some cases demonstrating an increased accuracy when compared to pixel-based methods (Flanders *et al.* 2003; Koch *et al.* 2007; Guo *et al.* 2007).

5.2. Specific objectives

The aims of this chapter were to produce a land cover classification for Soroti district, Uganda using Landsat images from 2001 and to analyse the distribution of sleeping sickness within this area in relation to specific land cover classes and other relevant covariates. In particular, a classification containing land cover classes of importance as potential tsetse habitats, such as seasonally flooding grassland, was required to allow subsequent quantification of the relationships between these land cover classes and sleeping sickness occurrence. Temporal changes in the observed relationships were studied to provide some understanding of the factors influencing the dispersion of sleeping sickness following its introduction (via the movements of infected cattle from endemic areas) into a previously unaffected area (Fèvre *et al.* 2001a). The landscape features of potential high transmission areas were also examined over time to give a greater understanding of high transmission risk areas and the temporal dynamics of disease transmission.

5.3. Methods

5.3.1. Land cover classification

5.3.1.1. Image processing

Three Landsat ETM+ images (path 171, row 59) were selected for use in the analysis from 27th January, 17th April and 27th November 2001, corresponding to the dry season, beginning of the long rains and end of the short rains respectively (see Figure 5.1). These three seasons were selected to give the optimal differentiation between land cover classes based on seasonal information; Figure 5.2 shows the monthly average rainfall for Soroti district during 2001 to highlight the seasonal precipitation

pattern within the study area. The images were all level 1 terrain (corrected) products (level 1T), which have undergone radiometric and geometric correction.

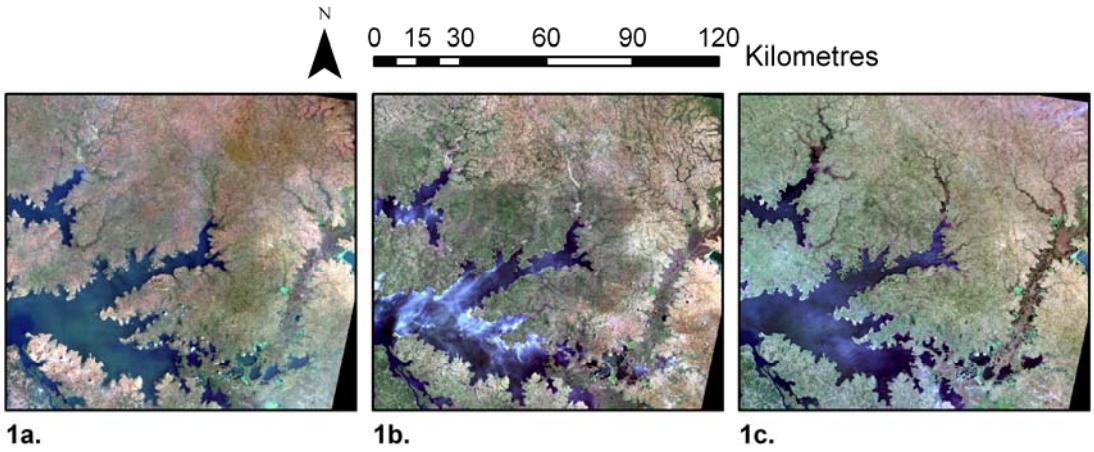


Figure 5.1: Thumbnail images of the three Landsat ETM+ images used in the land cover classification process. January (1a), April (1b) and November (1c). Displayed as true colour composites.

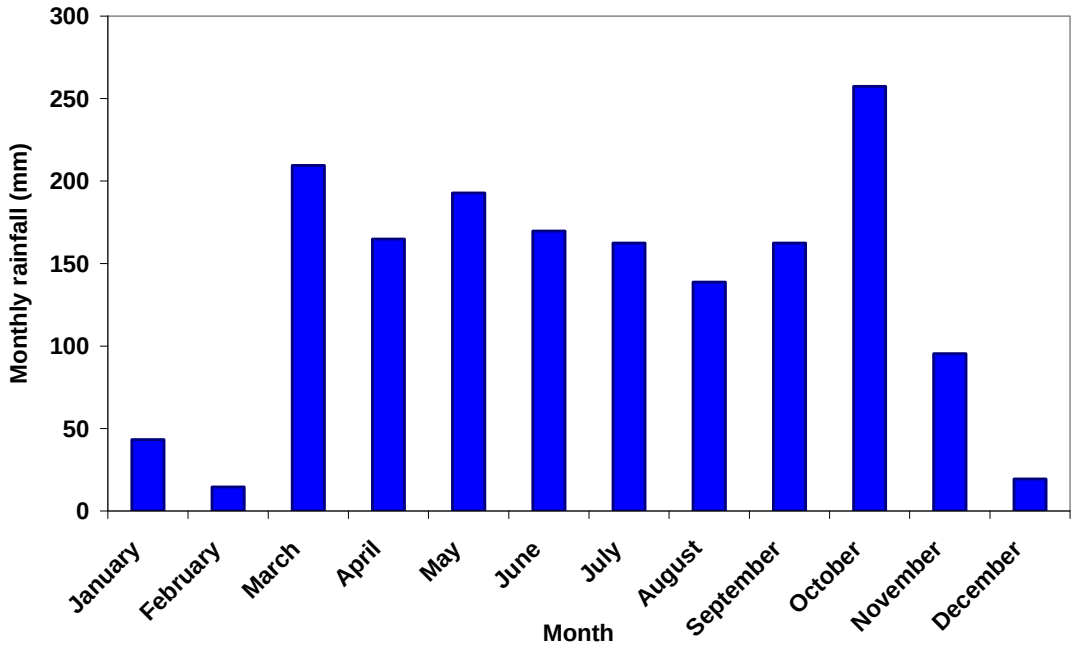


Figure 5.2: Monthly accumulated rainfall for Soroti district, 2001. District averages extracted from Tropical Rainfall Measuring Mission (TRMM) product 3B43 (V6) (National Aeronautics and Space Administration and Japan Aerospace Exploration Agency 2001).

Atmospheric correction for each of the Landsat images was carried out using ATCOR-2 (Atmospheric & Topographic Correction for Small FOV Satellite Images;

ReSe Applications Schläpfer, Switzerland; (Richter 2009)). A thin to moderately thick haze removal mask was used to remove the effects of haze and clouds from the April image (no haze removal masks were used on the January or November images). The aerosol type was set as rural and the water vapour category as tropical. Further information regarding the settings used in the atmospheric correction is provided in Appendix B. During atmospheric correction, the image units were converted from digital numbers to reflectance. NDVI was calculated for each image using the red and near-infrared wavebands of a Landsat ETM+ image using the following formula: NDVI = (near-infrared – red)/(near-infrared + red) in the eCognition software Version 4.2 (Definiens, Munich, Germany) (Tucker 1979).

5.3.1.2. Training data

Ground-based training data were collected by Dr Eric M Fèvre in 2001 for a variety of land cover classes in Soroti district. A handheld GPS was used to georeference polygons for each of the selected land cover classes shown in Table 5.2. Additional training areas were selected using high spatial resolution Quick Bird imagery available in Google Earth (Google™, Mountain View, CA) for land cover classes which were easily identifiable, were not likely to change over time and were also clearly identifiable on the Landsat images (e.g. seasonally flooding grassland). Training polygons were defined and stored in the WGS-84 datum, with a Universal Transverse Mercator (UTM) 36N projection. A random 20% sample of the training polygons was selected for use in the classification validation, with the remaining 80% being used to train the classification (see Figure 5.3 for the spatial distribution of training and testing data in the study area).

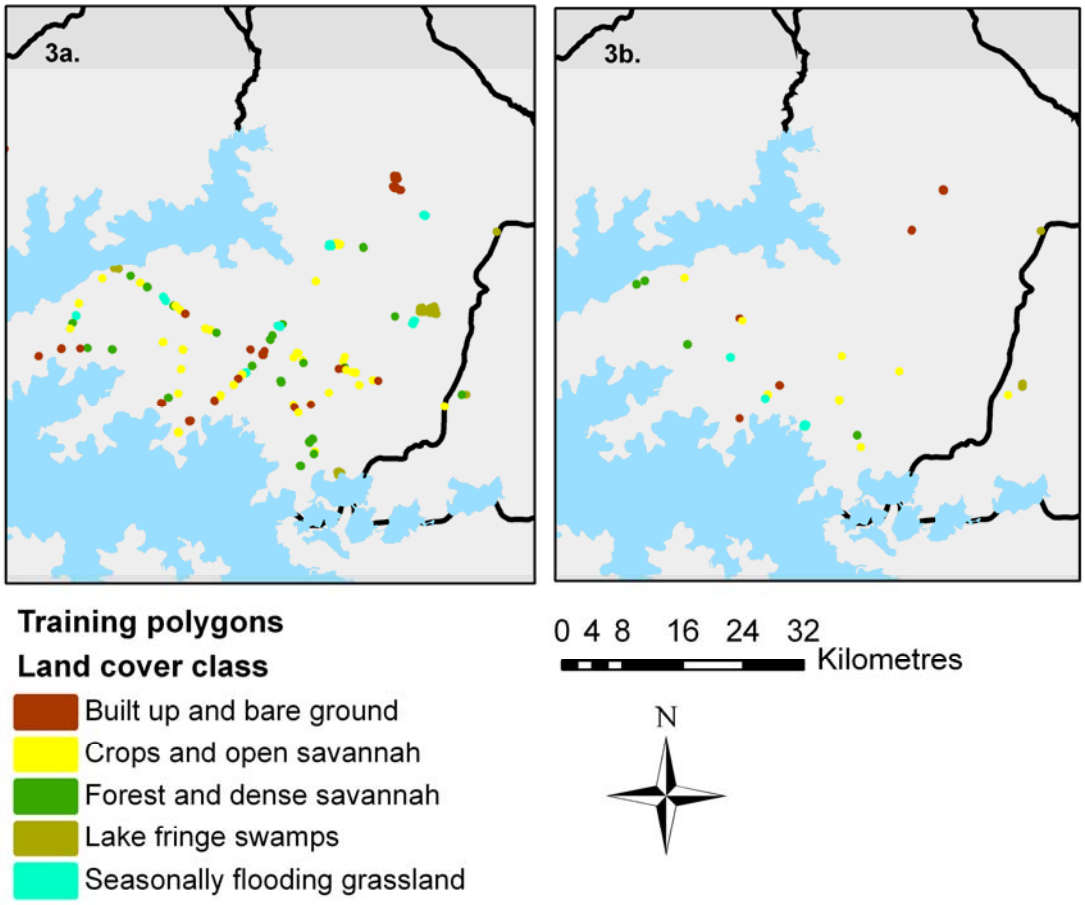


Figure 5.3: Spatial distribution of training (5.3a) and testing (5.3b) polygons within Soroti district (black lines represent district boundaries). Please note the sizes of polygons are not accurate as they have been enlarged to allow visualisation.

Initial land cover class	Description	Final class
Maize Cassava Banana and coffee Ground nuts Mille Peas Pineapple Rice Sim sim	Fields of specific crops	Crops, agricultural and open savannah
Ploughed fields	Ploughed fields with no growth	
Open savannah	Grassland with occasional trees/bushes (not seasonally flooding)	
Dense savannah	Grassland with dense trees/bushes (not seasonally flooding)	Woodland and dense savannah
Deciduous woodland	Patches of deciduous woodland	
Evergreen woodland	Patches of pine woodland	
Riverine woodland	Patches of deciduous woodland along rivers	
Rural built up	Rural towns/villages with high density of buildings	Built up and bare ground
Urban	Major town with high density of roads, roofs and covered markets	
Grass lawns	Large school playing fields (dry grass)	
Bare ground	Bare murrum or mud	
Open water	Areas of open water (lake or wide river)	Open water
Lake fringe swamps	Lake edges with a high density of papyrus, water hyacinth and water lilies	Lake fringe swamps
Seasonally flooded grassland	Savannah which floods during the wet season, with occasional trees/bushes	Seasonally flooded grassland

Table 5.2: Initial land cover classes which were collected during ground-truthing along with final class for land cover classification.

5.3.1.3. Supervised object-based classification

Feature space plots using the reflectance in each spectral band were used to visualise the separability of classes. Based on these plots, the class descriptions and knowledge of the study area and land cover types, several of the initial land cover classes were aggregated as detailed in Table 5.2. The aggregation was used to ensure the accurate classification of classes of importance to tsetse habitat and, thus, sleeping sickness transmission (particularly seasonally flooding grassland, woodland and dense savannah). The final classes used in the classification are shown in Table 5.2.

Segmentation and classification of the images was carried out using the eCognition software. Segmentation of the April image was used to produce homogeneous objects for classification using a scale parameter of 5 (determines the size of the resulting objects). The homogeneity criteria included 90% emphasis on spectral homogeneity and 10% shape (compactness and smoothness), and 50% for both compactness (optimises for the compactness of resulting objects; a perfectly compact object is a square) and smoothness (optimises for the smoothness of object borders). Two layers were created for the classification using these segmentation criteria to allow a hierarchical classification. The segmentation criteria used were based on trial-and-error to achieve objects with the desired size, homogeneity and shape for the application as recommended by Definiens (Baatz *et al.* 2000).

The classification was performed in two steps. An initial classification of open water; built up and bare ground; dry vegetation types and wet vegetation types (see Figure 5.4, Level 1) was carried out using the first object layer. This was followed by the classification of sub-classes woodland and dense savannah; crops and open savannah; seasonally flooding grassland and lake-fringe swamps (see Figure 5.4, Level 2) on the second object layer.

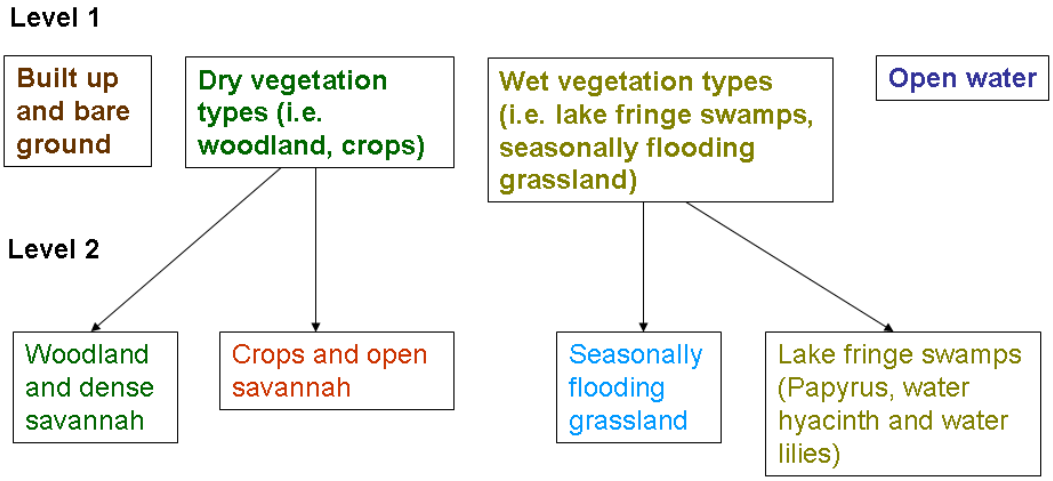


Figure 5.4: Diagram of land cover class hierarchy showing super-classes (level 1) and sub-classes (level 2).

Open water was classified using the object mean of band 4 (April image) as water absorbs nearly all the light within this band (Campbell 2002b). The open water in the image could be observed clearly by eye using band 4, and was differentiated from all other classes using a threshold of 5.5% (any object with a mean of less than or equal to 5.5% reflectance for band 4 was classified as open water).

Built up and bare ground, wet vegetation types and dry vegetation types were differentiated using both the April and the November images with a nearest neighbour classification algorithm. The object mean of band 3 from the April image and bands 2, 3, 5 and 7 from the November image as well as the object's length divided by width, and the difference in NDVI from January to November were used. Landsat ETM+ band 2 is useful for the identification of healthy vegetation, band 3 for the discrimination of different vegetation types, band 5 for observing vegetation and soil moisture content and band 7 for the identification of built up areas and bare ground (Lillesand *et al.* 2004c). The difference in NDVI between January and November was used to show the amount of change in vegetation cover and, thus, assist in the differentiation of vegetation types.

Following the first classification (Level 1), a second classification was carried out on the second object layer to further differentiate two of the classes into sub-classes: dry vegetation types and wet vegetation types. The sub-classification of dry vegetation

types into woodland and dense savannah, and crops and open savannah was carried out only on objects for which the super-object had been classified as dry vegetation types in Level 1, and likewise the sub-classification of wet vegetation types into seasonally flooding grassland and lake-fringe swamps was carried out only on the sub-objects of super-objects which were classified as wet vegetation types in Level 1. The classes open water and built up and bare ground remained as in the Level 1 classification.

For the sub-classification of dry vegetation types the object mean for bands 2, 3, 4 and 5 from the April image were used, along with the difference in NDVI from January to April, January to November and April to November. As the April image was from the beginning of the long rains and, thus, the beginning of the growing season this image can enhance the differentiation of the two dry-vegetation sub-classes. Each of the bands selected for this classification can be related to healthy vegetation, biomass, plant type or vegetation moisture content (Lillesand *et al.* 2004c). In addition, the differences in NDVI between the three images allow the identification of the two classes based on differences in their growing seasons and growth rates. The wet-vegetation types were sub-classified using the mean of bands 4 and 5 from the November image which are useful for the identification of vegetation and soil moisture content, biomass, plant vigour and water (Lillesand *et al.* 2004c). As November is just after the end of the short rains, the area should be wetter in the November image than the January or April images and, thus, this image should allow a more accurate differentiation of seasonally flooding grassland and lake-fringe swamps.

Using the pixels within the testing polygons (a random 20% sample of the original training polygons), the land cover classification was validated by creating an error matrix and calculating producer and user accuracies for each class, plus an overall accuracy value. The error matrix is a method of visualising actual versus predicted occurrences of a class and can allow the easy identification of a classification which is repeatedly confusing two (or more) classes. The formulae for the calculation of producer and user accuracies are given below (Lillesand *et al.* 2004b).

Producer accuracy:

- number correctly classified as class *x* / number of reference pixels for class *x*

User accuracy:

- number correctly classified as class *x* / total number classified as class *x*

Overall accuracy:

- total number correctly classified / total number of reference pixels

A pixel-based classification was also performed using maximum likelihood classification of the April Landsat ETM+ image, implemented in ENVI Version 4.3 (ITT Industries, Boulder, Colorado). This was carried out to allow a brief comparison with the object-based classification method and to ensure the highest possible accuracy was obtained in the final classification.

5.3.2. Sleeping sickness case control analysis

5.3.2.1. Data

Previously collected sleeping sickness case records for Soroti district from December 1998 to November 2002 were used for all of the analysis in this chapter, along with matched control data as detailed in Chapter 2, Sections 2.4.2.2 and 2.4.3.2. Part of these data has previously been published by Fèvre *et al* (Fèvre *et al.* 2001a). All sleeping sickness cases recorded as having been actively detected were excluded from the analysis to avoid the introduction of spatial bias in the distribution of cases.

The study design used (matched case-control) relies on the selection of representative controls from the study population (hospital-going individuals in which the cases of sleeping sickness have occurred, within the catchment of Serere Health Centre). To ensure that the controls adequately represented the entire population from which the sleeping sickness cases came, they were matched to the cases (using a one-to-one matching design) on age group, sex and month of admission. This prevents bias in the selection of controls, which can adversely affect the results of an analysis (Woodward 2004). A one-to-one matching design was used as there were not enough

suitable controls to use a one-to-many design. Having used a matched case-control design, it is vital that this matching is taken into account during the analysis of the dataset, as discussed in Section 5.3.2.2 (Woodward 2004).

The geo-referenced case-control data were visualised using ArcMap 9.1 along with the final land cover classification described in Section 5.3.1.3. Circular buffers of 1 km and 3 km radii were created around each village centroid. The two selected radii were based on previous work illustrating the significance of the proportion of buffer zones with radii between 500 m and 3.5 km which intersected areas of wetland, in terms of the distribution of sleeping sickness. The greatest significance was demonstrated with radii of 800 to 900 m and 2.5 to 3 km (Zoller *et al.* 2008). Additionally, a recent study of daily mobility patterns within Uganda found the average distance of daily short-distance trips (e.g. to work or to fetch water) for village residents (the predominant population group in the study area) to range from approximately 2 km for low income households to 4 km for high income households (Bryceson *et al.* 2003). The smaller buffer zone (1 km radius) was selected to encompass the majority of the village area, and the larger buffer zone (3 km radius) was selected to include the areas immediately surrounding the village, and should be representative of the average distance village inhabitants walk from their village on a daily basis (for activities such as fetching water or firewood and watering livestock). The total number of pixels of each land cover class was calculated within the circular buffers and expressed as a percentage, thus, giving a quantitative measure of the landscape within and surrounding the case and control villages.

The population density (expressed as hundreds of people per square kilometre) (Oak Ridge National Laboratory 2006), elevation (in metres) (US Geological survey 2006), night-time lights of the world (Defence Meteorological Satellite Program 2004) and predicted suitability (as a percentage) for *G. f. fuscipes* (the predominant vector species within the study area) (Wint and Rogers 2000) for each case and control were extracted. Additional information on each of these data can be found in Chapter 2, Section 2.4.4. The distance between each village and Brookes Corner livestock market, which has previously been implicated in the introduction of

Rhodesian sleeping sickness to the study area (Fèvre *et al.* 2001a), was calculated and expressed in kilometres.

5.3.2.2. Exploratory analysis

The dataset was divided annually to allow the separate analysis of each year and illustrate any temporal patterns in the observed relationships following the introduction of the disease to the area in 1998. As the study period began in December 1998 and ended in November 2002, the annual periods for analysis ran from December to November. For clarity, the four annual periods will be referred to as the first (December 1998 to November 1999), second (December 1999 to November 2000), third (December 2000 to November 2001) and fourth (December 2001 to November 2002) annual periods.

Relationships between the dichotomous dependent variable of sleeping sickness status (case or control) and each of the independent variables (percentage of 1 km and 3 km buffers intersected by each land cover class, population density, elevation, night-time lights, predicted suitability for *G. fuscipes* and distance to Brookes Corner livestock market) for each of the four annual periods were examined to identify spatial risk factors and temporal changes in the observed relationships. As mentioned in Section 5.3.2.1, the analysis of matched case-control data must take into special consideration the matching within the data: unconditional logistic regression analysis (as used in Chapter 3) of matched data will result in biased parameter estimates (Breslow and Day 1980; Woodward 2004). Conditional logistic regression, a variant of logistic regression, can be used for matched study designs with a binary or proportional response variable. By specifying within the analysis which 'stratum' each case and control belongs to (i.e. the matched pair), the regression is carried out within each stratum (Breslow and Day 1980; Hosmer and Lemeshow 1989b). Unadjusted odds ratios for each of the independent variables were calculated using conditional logistic regression for each of the four annual periods separately. The *Survival* package in the R statistical software, which maximises the exact conditional likelihood, was used to implement the conditional logistic regression analysis. The

resulting odds ratios were plotted over time using Microsoft® Excel software (Microsoft Corporation, Redmond, Washington) to highlight any temporal changes in the correlations.

5.3.2.3. Multivariate analysis

Multivariate conditional logistic regression analysis was carried out separately for each of the four annual time periods, using forwards stepwise methodologies, beginning with the null model in each time period. At each step in the model fitting procedure the variable which resulted in the largest decrease in deviance was selected. Models were compared using Chi-squared likelihood ratio tests, and variables were only retained in the model if this test was significant with more than 95% confidence (p-value < 0.05), and the variable was significant within the model (p-value < 0.05). Any variables which lost significance in subsequent steps (significance of the variable in the model, or the Chi-squared likelihood ratio test for deletion of the variable gave a p-value ≥ 0.05) were removed from the model. The stepwise addition of interaction terms between the significant variables was then carried out to assess any effect modification. R^2 values were calculated to assess the proportion of the variability in the response variable which was described by the explanatory variables included in the model (the maximum R^2 value for a conditional logistic regression is 0.5).

5.3.3. Identification and characterisation of potential high transmission zones

5.3.3.1. Identification of potential high transmission zones

Areas of potential high sleeping sickness transmission were identified using the spatial distribution of cases and controls for each of the four annual periods and knowledge regarding the average distances travelled on a daily basis by residents of rural Ugandan villages. The majority of sleeping sickness infections are acquired outside of the village of residence while individuals are undertaking activities such as watering livestock or collecting water or firewood (Wyatt *et al.* 1985; Okia *et al.*

1994). In areas with a higher observed number of sleeping sickness cases, it can be hypothesised that there are factors within that area promoting a higher intensity of transmission, such as elements of the landscape and interactions between tsetse, humans and livestock. Utilising information on the average daily short-trip distances travelled by inhabitants of rural Ugandan villages (approximately 3 km (Bryceson *et al.* 2003)) and previous findings by Zoller *et al* (2008) which showed an increased risk of sleeping sickness in homes between 500 m and 3 km away from areas of wetland, it can be assumed that areas of sleeping sickness transmission are normally within 3 km of the village centroid. If there are several villages with a large number of observed cases in close proximity it may be possible to identify areas of potential high transmission.

Circular buffers of 3 km radius, as described in Section 5.3.2.1, were used to represent the daily areas of mobility (referred to as 'daily activity areas') for cases and controls. Areas where a large number of the daily activity areas of sleeping sickness cases overlap were identified as potential high transmission zones. These are areas which are likely to be visited frequently by sleeping sickness cases resident in several surrounding villages, and may constitute areas with landscape features that promote a high level of interaction between tsetse, livestock reservoirs (mainly cattle) and humans, thus, encouraging a high intensity of sleeping sickness transmission.

Similar, more traditional methods such as KDE (where a weighted average is calculated within a spatially moving window (Diggle 2000)) do not fully capture the desired effect. Weighted KDE produces a higher intensity at village centres where a lot of cases have occurred and lower intensity in areas outside villages where transmission is more likely to occur, due to the weighting used; more weight is generally given to values in the centre of the moving window than those on the periphery. These methods are used to give a smooth representation of the intensity of a point process (Pfeiffer *et al.* 2008a). As sleeping sickness transmission normally occurs outside of the village, transmission zones may not be accurately represented by village centroids and therefore weighted KDE will not adequately identify areas

which may be considered as high transmission zones. The methods presented here give results comparable to those which may be obtained using an un-weighted kernel density smoothing algorithm (a flat kernel, where equal weight is given to all values within the moving window).

A spatial grid was created over the entire study area, with a 500 m by 500 m cell size. The number of daily activity areas (3 km buffers) for cases which overlapped within each grid cell was calculated using ArcMap to give the number of intersecting daily activity areas. This was done for each of the four annual periods. Grid cells in the 99th percentile based on count values (the 1% of grid cells with the highest number of overlapping buffers) were defined as being areas with a high number of overlapping daily activity areas (these will be referred to as 'high overlap areas'). The analysis was also carried out for the control daily activity areas to allow a comparison between areas with a high number of overlapping case and control daily activity areas and to highlight differences in landscape makeup in the high overlap areas for cases and controls over the four year period. The high overlap areas for both cases and controls over the study period were visualised using ArcMap.

5.3.3.2. Characterisation of potential high transmission zones

The areas of different land cover classes within each of the high overlap areas were calculated and expressed as a proportion of the overall high overlap area for each of the four time periods, for both cases and controls. In addition, the average elevation within the high overlap areas was extracted for analysis. The *z*-test for two proportions was used to assess the significance of the difference in proportions of the various land cover types within high overlap areas for cases and controls. One-tailed tests were used for land cover types for which a clearly defined difference was expected (built up and bare ground – lower proportion in case high overlap areas than control; crops and open savannah – lower proportion in case high overlap areas and seasonally flooding grassland – higher proportion in case high overlap areas). A two-tailed test was used for woodland and dense savannah due to the indistinct relationships observed. Open water and lake-fringe swamps were not included in this

analysis due to the predominance of high overlap areas with no intersections with these land cover types. Additionally, a *t*-test was used to assess the significance of the difference between mean elevations for case and control high overlap areas during each of the annual periods.

The landscape profiles (the proportions of each land cover type within the high overlap areas) for each annual period were illustrated using percentage component bar charts created in Microsoft® Excel, for both the case and control high overlap areas. This was carried out to highlight any changes in the landscape within the high overlap areas over time, and to allow a comparison between the case and control high overlap areas. A line graph was also used to demonstrate changes in the mean elevation within the case and control high overlap areas over time. The average population density (Oak Ridge National Laboratory 2006) within the case and control high overlap areas was calculated to allow a comparison between the high overlap area locations and the distribution of the general population.

5.4. Results

5.4.1. Supervised object-based classification

The initial classification of Level 1 classes (open water; built up and bare ground; wet vegetation types and dry vegetation types) is shown in Figure 5.5. The overall accuracy from the Level 1 classification was 88.9% based on pixels within the testing polygons, with user accuracies for all classes above 80% and producer accuracies for all classes above 70%. The error matrix for the classes is shown in Table 5.3.

The level 2 classification is shown in Figure 5.6 (built up and bare ground; open water; crops and open savannah; woodland and dense savannah; lake-fringe swamps and seasonally flooding grassland). The overall accuracy for the level 2 classification was 86%, with producer and user accuracies of over 70% for all classes except crops and open savannah (producer accuracy = 69%, user accuracy = 47%). The error matrix for the level 2 classification is shown in Table 5.4. The class crops and open

savannah was not thought to be a significant tsetse habitat within the study area and so the lower accuracy for this class was not problematic, although it may have resulted in lower proportions of the other classes of relevance as potential tsetse habitat. The spatial distribution of testing polygons incorrectly classified as crops and open savannah is demonstrated in Figure 5.7. For comparison, the utilisation of a maximum likelihood pixel-based classification for the same classes gave an overall accuracy of 81%, with producer and user accuracies above 68% for all classes except crops and open savannah (producer accuracy = 47%, user accuracy = 31%).

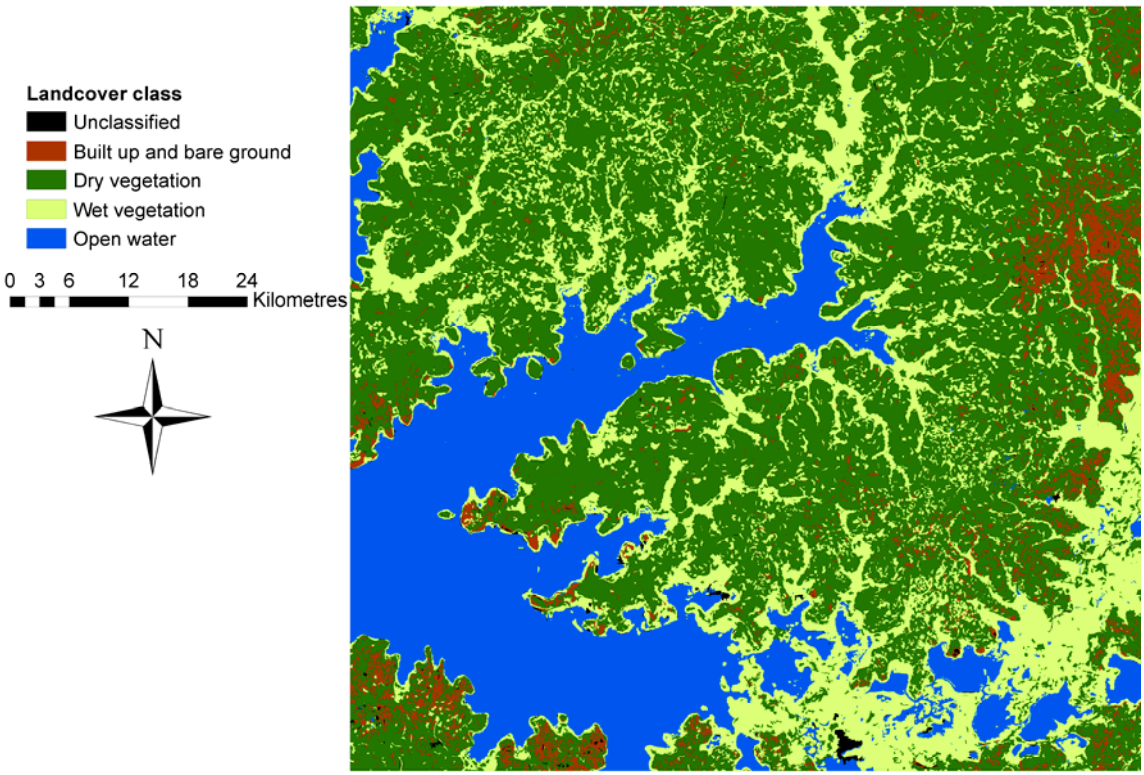


Figure 5.5: Level 1 land cover classification using supervised object-based classification.

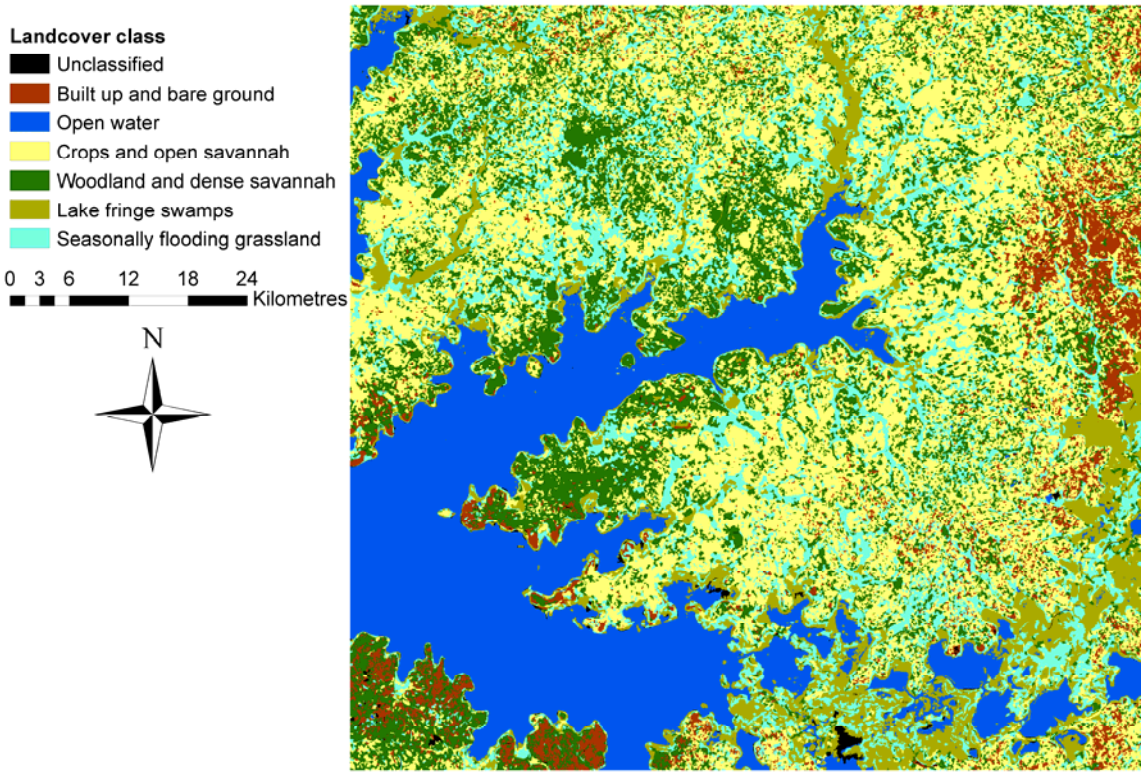


Figure 5.6: Level 2 land cover classification using supervised object-based classification.

		Actual land cover class				
		Built up and bare ground	Dry vegetation types	Wet vegetation types	Total	User accuracy
Predicted land cover class	Built up and bare ground	48	3	0	**51**	94.1%
	Dry vegetation types	11	274	0	**285**	96.1%
	Wet vegetation types	0	50	254	**304**	83.6%
	Unclassified	8	0	0	**8**	
	Total	**67**	**327**	**254**		
	Producer accuracy	71.6%	83.8%	100%		

Table 5.3: Error matrix for Level 1 classification.

Predicted land cover class	Actual land cover class: Built up and bare ground	Crops and open savannah	Woodland and dense savannah	Lake fringe swamps	Seasonally flooding grassland	Total	User accuracy
Built up and bare ground	56	3	0	0	0	**59**	94.9%
Crops and open savannah	9	27	22	0	0	**58**	46.6%
Woodland and dense savannah	2	3	222	0	0	**227**	97.8%
Lake fringe swamps	0	6	30	111	3	**150**	74.0%
Seasonally flooding grassland	0	0	14	0	140	**154**	90.9%
Total	**67**	**39**	**288**	**111**	**143**		
Producer accuracy	83.6%	69.2%	77.1%	100%	97.9%		

Table 5.4: Error matrix for Level 2 classification.

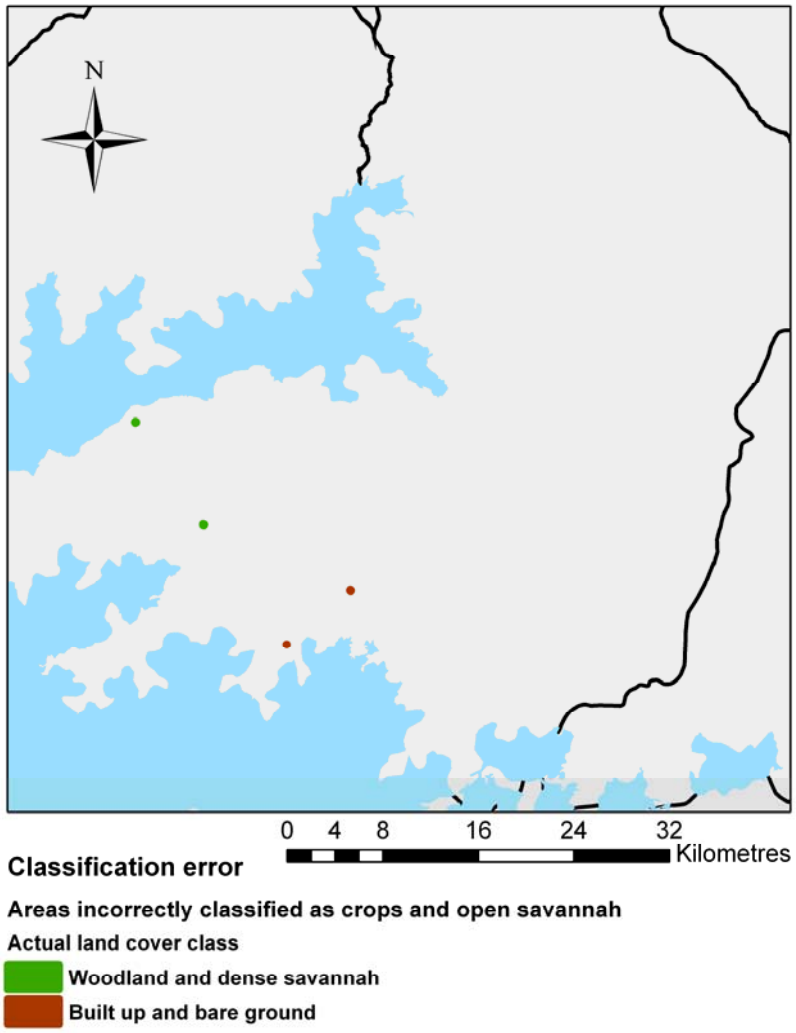


Figure 5.7: Spatial distribution of testing polygons misclassified as crops and open savannah. Please note the sizes of polygons are not accurate, but have been enlarged to allow visualisation.

Within the study area (a subset of the entire classified image; covering all case and control villages with a 3 km buffer surrounding them), the predominant land cover classes were crops and open savannah (31.4% of the study area), followed by open water and woodland and dense savannah (both 17.9%). Seasonally flooding grassland accounted for 14.3%, lake-fringe swamps 11% and built up and bare ground was the least common land cover class (7.5%).

5.4.2. Exploratory analysis

A total of 258 sleeping sickness cases resident within Soroti district were diagnosed at Serere hospital during the study period (December 1998 to November 2002). Eighteen of these cases were detected during active surveillance activities and, thus, were excluded from the analysis. In addition, suitable matched controls could not be identified for seven of the cases, and the unmatched cases were excluded from the analysis. A total of 58 cases were analysed in the first annual period, 52 in the second annual period, 44 in the third annual period and 79 in the fourth annual period (each with the same numbers of matched controls; see Table 5.5).

	Time period	**Number of unmatched cases**	**Number of matched cases**
First annual period	December 1998 to November 1999	2	58
Second annual period	December 1999 to November 2000	4	52
Third annual period	December 2000 to November 2001	1	44
Fourth annual period	December 2001 to November 2002	0	79
Total		7	233

Table 5.5: Number of matched and unmatched cases for each annual period (excluding cases detected during active surveillance).

The spatial distribution of cases within the study area varied through the study period (Figures 5.8a to 5.8d). Between December 1998 and November 1999, the cases were primarily located in close proximity to Brookes Corner livestock market (Figure 5.8a). The parasite has previously been demonstrated to have been introduced to the study area via the trade of untreated cattle at this market (Fèvre *et al.* 2001a). In contrast, the controls were evenly distributed across the study area, with no apparent clustering. Over the subsequent years, the distribution of controls remained evenly dispersed across the

study area with no directional movement. The spatial distribution of cases changed over the study period, with a directional movement away from the livestock market (Figures 5.8b, 5.8c and 5.8d) as has been previously reported (Fèvre *et al.* 2001a).

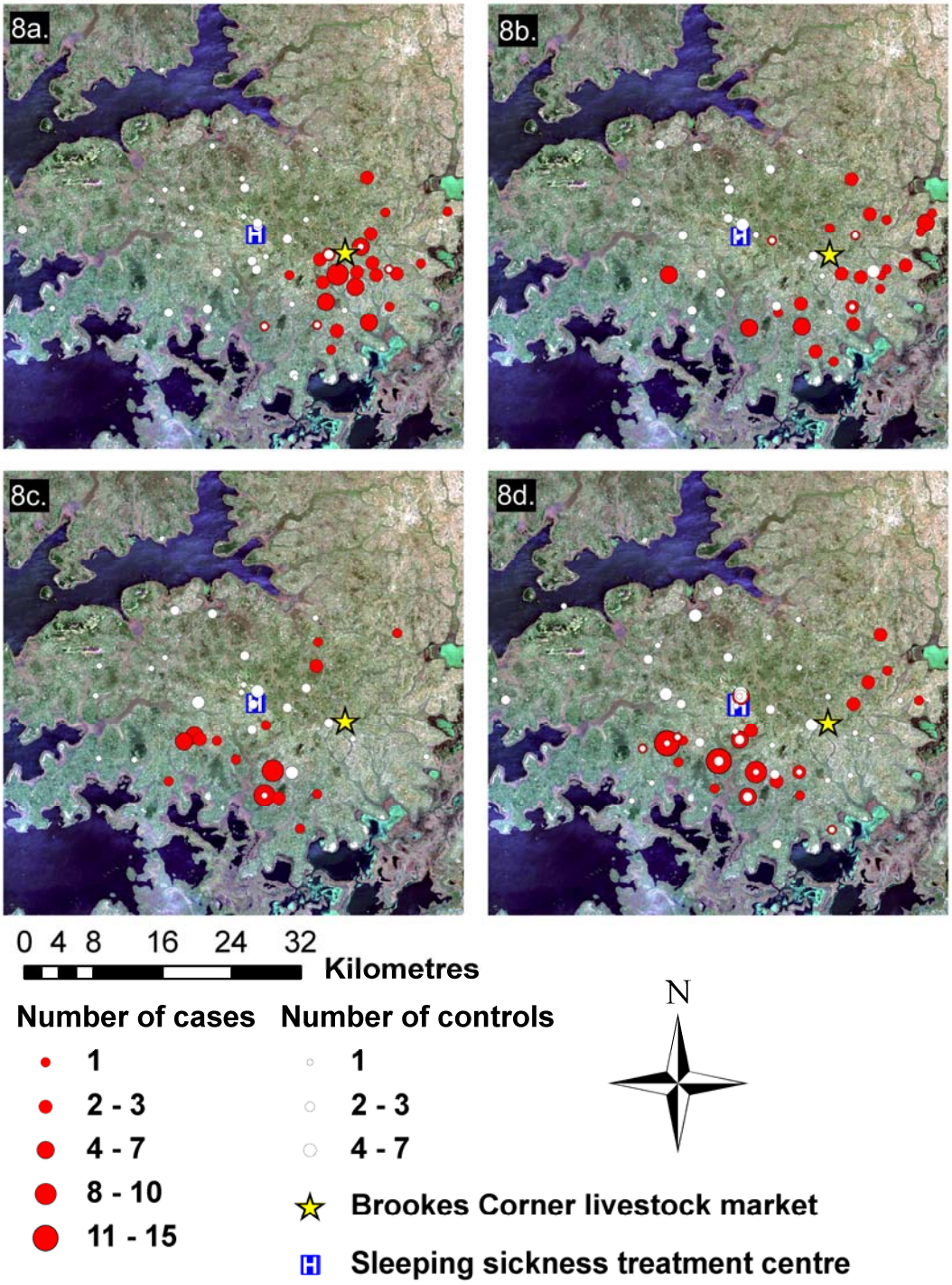


Figure 5.8: True colour Landsat ETM+ composite of study area, showing distribution of cases and controls (as counts for each location) in the first (5.8a), second (5.8b), third (5.8c) and fourth (5.8d) annual periods, with Serere hospital and Brookes Corner livestock market.

The occurrence of sleeping sickness was significantly correlated with five variables across the study period; temporal trends were observed in the odds ratios for several of the covariates (see Tables 5.6 and 5.7, and Figures 5.9 a - i). Increasing population density demonstrated a protective effect in each of the four time periods (OR = 0.95, 0.91, 0.90 and 0.93 for the first to fourth annual periods respectively although this was not statistically significant in the third annual period; see Table 5.6), with no apparent change in the association over time (see Figure 5.9a). Similarly, night-time lights had a significantly protective effect during each annual period (OR = 0.10, 0.06, 0.44 and 0.44 for the first, second, third and fourth annual periods respectively; see Table 5.6) with no clear temporal change (see Figure 5.9b).

In the first annual period, areas with a higher predicted suitability for *G. f. fuscipes* had a significantly decreased odds of sleeping sickness occurrence (OR = 0.95), but in the second, third and fourth annual periods the correlation reversed with significantly increased odds of sleeping sickness occurrence in areas predicted to be more suitable for the *G. f. fuscipes* vector (OR = 1.04, 1.09 and 1.07 respectively; see Table 5.6 and Figure 5.9c). A similar temporal pattern was observed for elevation, with an increased odds of sleeping sickness in higher elevation areas in the first annual period (OR = 1.02), and decreased odds in the second, third and fourth annual periods (OR = 0.99, 0.96 and 0.98 respectively although this was not statistically significant in the second annual period; see Table 5.6 and Figure 5.9d). A protective effect was also seen at increasing distances from Brookes Corner livestock market during all four annual periods (OR = 0.69, 0.91, 0.97 and 0.96 for the first, second, third and fourth annual periods respectively), although this correlation loses statistical significance in the third and fourth annual periods as the strength of the association decreases (see Table 5.6 and Figure 5.9e).

With respect to land cover in the areas surrounding case and control villages, significant correlations were seen between sleeping sickness occurrence and the proportions of both seasonally flooding grassland and woodland and dense savannah within 1 km and 3 km buffer zones. In the first annual period, villages with a higher proportion of seasonally

flooding grassland in the surrounding areas had a lower odds of sleeping sickness occurrence, but in the second, third and fourth annual periods the relationship reversed with a higher odds of sleeping sickness occurrence in villages with a higher proportion of seasonally flooding grassland in the surrounding areas (see Table 5.7 and Figures 5.9f and 5.9h). The correlation with seasonally flooding grassland had higher significance using a buffer of 3 km radius as opposed to 1 km (OR = 0.94, 1.01, 1.11 and 1.08 for the first, second, third and fourth annual period respectively using a 3 km buffer, although this was not statistically significant in the second annual period). The presence of a higher proportion of the woodland and dense savannah land cover class within the 1 km and 3 km buffer zones demonstrated a protective effect, with higher significance for the 1 km radius buffers than the 3 km radius buffers (OR = 0.98, 0.93, 0.91 and 0.92 for the first, second, third and fourth annual periods respectively using a 1 km buffer, although this was not statistically significant for the first annual period; see Table 5.7 and Figures 5.9g and 5.9i).

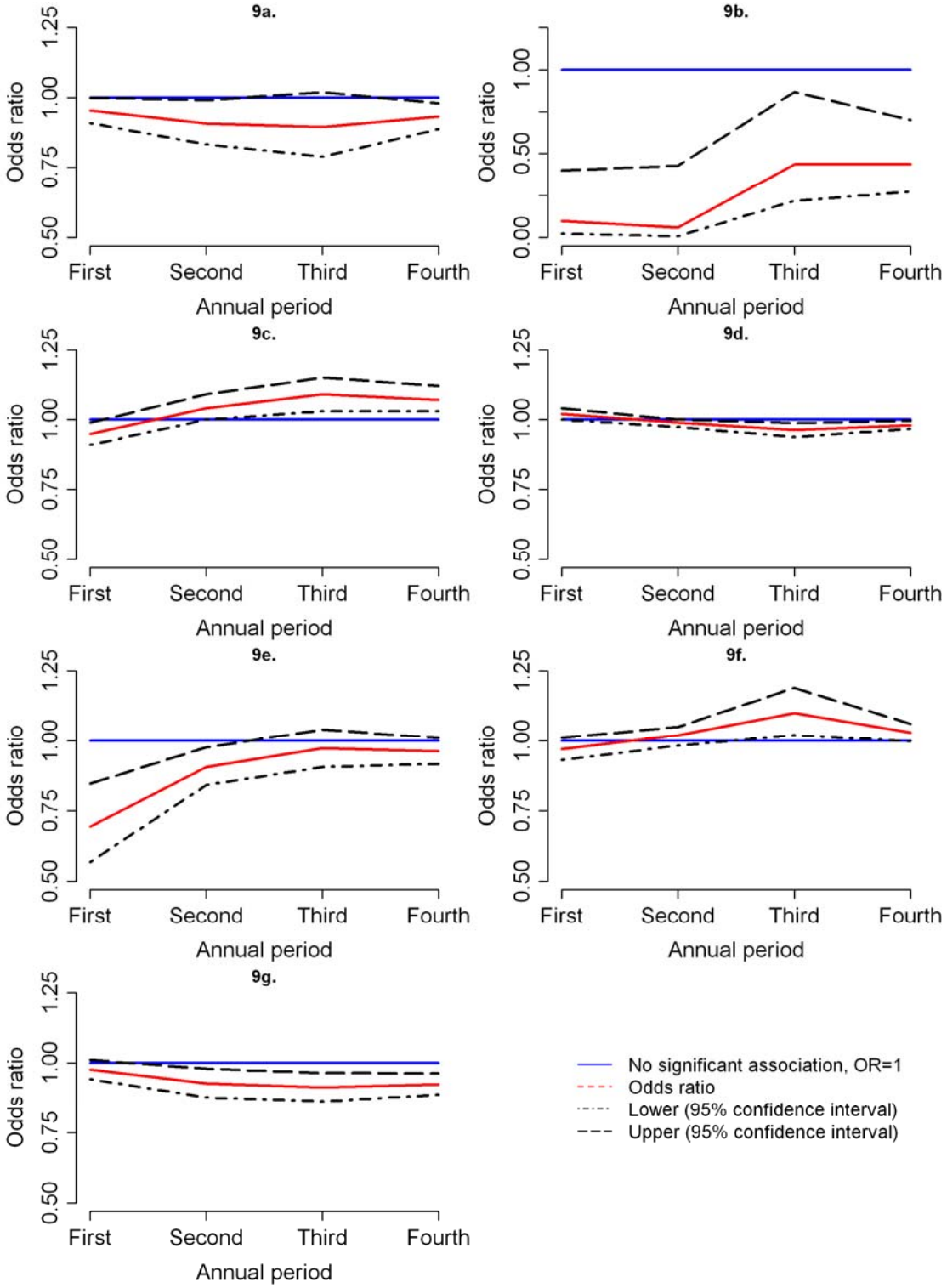
No correlation was observed between occurrence of sleeping sickness and the percent coverage within 1 km or 3 km radius buffers of the open water or lake-fringe swamp land cover classes (data not shown). The percent coverage of the built up or bare ground and crops or open savannah classes within 1 km and 3 km buffer areas showed unclear patterns of correlation with sleeping sickness occurrence across the four annual periods (data not shown). In addition, these classes were not thought to represent high risk environments in terms of sleeping sickness transmission (based on the known habitat preferences of the tsetse vector within the study area) and so the eight variables related to these land cover classes were not used in the multivariate analysis.

		Dec 1998 - Nov 1999	Dec 1999 - Nov 2000	Dec 2000 - Nov 2001	Dec 2001 - Nov 2002
Population density	Odds ratio	0.95	0.91	0.90	0.93
	(95% CI)	*(0.91 to 1.00)*	*(0.83 to 0.99)*	*(0.79 to 1.02)*	*(0.89 to 0.98)*
	p-value	0.05	0.03	0.09	0.006
Nighttime lights	Odds ratio	0.10	0.06	0.44	0.44
	(95% CI)	*(0.02 to 0.40)*	*(0.008 to 0.43)*	*(0.22 to 0.87)*	*(0.27 to 0.70)*
	p-value	<0.005	0.005	0.02	<0.005
Predicted suitability for *G. fuscipes*	Odds ratio	0.95	1.04	1.09	1.07
	(95% CI)	*(0.91 to 0.99)*	*(1.00 to 1.09)*	*(1.03 to 1.15)*	*(1.03 to 1.12)*
	p-value	0.01	0.05	<0.005	<0.005
Elevation	Odds ratio	1.02	0.99	0.96	0.98
	(95% CI)	*(1.00 to 1.04)*	*(0.97 to 1.00)*	*(0.94 to 0.99)*	*(0.97 to 0.99)*
	p-value	0.03	0.15	<0.005	0.008
Distance to livestock market	Odds ratio	0.69	0.91	0.97	0.96
	(95% CI)	*(0.57 to 0.85)*	*(0.84 to 0.98)*	*(0.91 to 1.04)*	*(0.92 to 1.01)*
	p-value	<0.005	0.009	0.44	0.11

Table 5.6: Unadjusted odds ratios for population density, nighttime lights, predicted *G. fuscipes* suitability, elevation and distance to Brookes Corner livestock market during the four annual periods.

			Dec 1998 - Nov 1999	***Dec 1999 - Nov 2000***	***Dec 2000 - Nov 2001***	***Dec 2001 - Nov 2002***
Percent coverage within 1km buffer	Seasonally flooding grassland	Odds ratio	0.97	1.02	1.10	1.03
		(95% CI)	*(0.93 to 1.01)*	*(0.98 to 1.05)*	*(1.02 to 1.19)*	*(1.00 to 1.06)*
		p-value	0.16	0.35	0.02	0.07
	Woodland and dense savannah	Odds ratio	0.98	0.93	0.91	0.92
		(95% CI)	*(0.94 to 1.01)*	*(0.88 to 0.98)*	*(0.86 to 0.96)*	*(0.89 to 0.96)*
		p-value	0.16	0.008	<0.005	<0.005
Percent coverage within 3 km buffer	Seasonally flooding grassland	Odds ratio	0.94	1.01	1.11	1.08
		(95% CI)	*(0.89 to 0.99)*	*(0.97 to 1.06)*	*(1.02 to 1.21)*	*(1.02 to 1.14)*
		p-value	0.02	0.69	0.02	0.007
	Woodland and dense savannah	Odds ratio	0.98	0.91	0.95	0.88
		(95% CI)	*(0.92 to 1.03)*	*(0.83 to 1.01)*	*(0.88 to 1.03)*	*(0.81 to 0.96)*
		p-value	0.38	0.07	0.21	<0.005

Table 5.7: Unadjusted odds ratios for land cover proportions within 1km and 3 km buffer zones during the four annual periods.


9a.
9b.
9c.
9d.
9e.
9f.
9g.
Odds ratio
Annual period
First
Second
Third
Fourth
0.50
0.75
1.00
1.25
0.00
0.50
No significant association, OR=1
Odds ratio
Lower (95% confidence interval)
Upper (95% confidence interval)

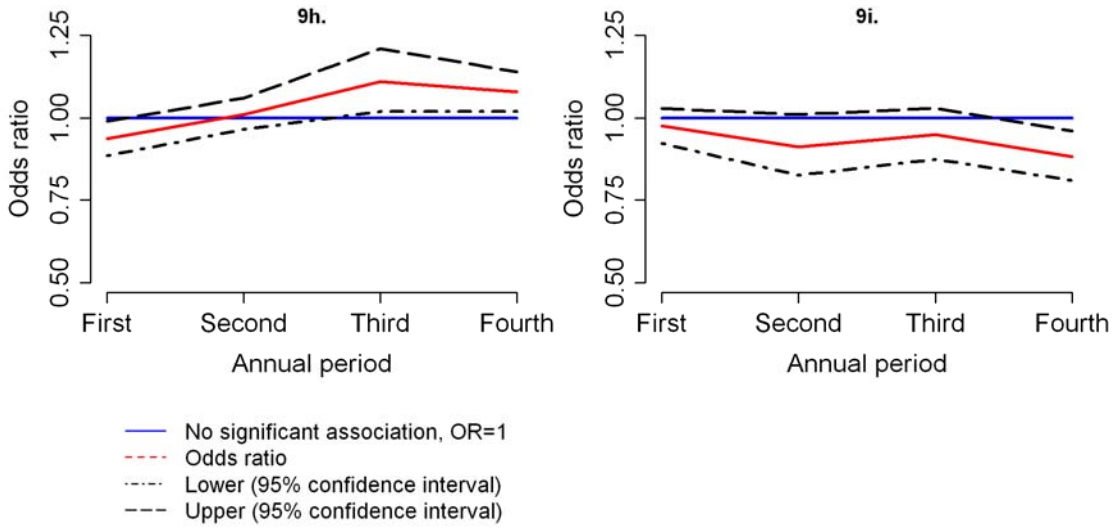


Figure 5.9: Temporal trend in odds ratios for population density (5.9a), nighttime lights (5.9b), predicted suitability for *G. fuscipes fuscipes* (5.9c), elevation (5.9d), distance to the livestock market (5.9e), proportion of land within 1 km that was seasonally flooding (5.9f), proportion of land within 1 km that was woodland and dense savannah (5.9g), proportion of land within 3 km that was seasonally flooding (5.9h) and proportion of land within 3 km that was woodland and dense savannah (5.9i). Note the y-axes for all plots except for 5.9b begin at 0.5 rather than 0.

5.4.3. Multivariate analysis

As mentioned in Section 5.4.2, the proportions of open water and lake-fringe swamps were not used in the multivariate analysis. The following covariates were also excluded from the multivariate analysis: nighttime lights (due to the association with population density) and predicted suitability for *G. f. fuscipes* (this variable is an estimated output from a predictive model which encompasses the effects of several different covariates).

Forward stepwise conditional logistic regression for each of the four annual periods resulted in three different final regression models (see Table 5.8). The model for the first and second annual periods contained distance to Brookes Corner livestock market (OR = 0.48, $p < 0.005$ for the first annual period, OR = 0.79, $p < 0.005$ for the second annual

period), which had a protective effect at greater distances, although this correlation was smaller in the second annual period than the first, and elevation (OR = 0.91, p = 0.02 for the first annual period, OR = 0.95, p < 0.005 for the second annual period) which demonstrated a protective effect in areas with higher elevation. The final model for the third annual period contained the proportion of woodland and dense savannah within a 1 km radius buffer (OR = 0.88, p < 0.005) which showed a protective effect in areas with a higher proportion of the land cover class, and the proportion of seasonally flooding grassland within a 3 km buffer (OR = 1.18, p = 0.01) which resulted in an increased odds of sleeping sickness in areas with a higher proportion of flooding grassland. For the fourth annual period, the final model contained the proportion of woodland and dense savannah within a 1 km radius buffer (OR = 0.90, p < 0.005) which again showed a protective effect in areas with a higher proportion of the class, and distance to the livestock market (OR = 0.90, p = 0.01), which resulted in lower odds of sleeping sickness occurrence in areas at a greater distance from the market (this effect was smaller in the fourth annual period than in the first or second annual periods).

The R^2 for each model differed, with the largest R^2 for the first annual period (0.41 compared with 0.21, 0.26 and 0.18 for the second, third and fourth annual periods respectively), indicating that the model for the first annual period explains a larger amount of the spatial variation in the data than the subsequent time period's models.

Annual period	Variable	Odds ratio *(95% CI interval)*	*p* - value
Dec 1998 – Nov 1999	Distance to market	0.48 *(0.30 to 0.77)*	<0.005
	Elevation	0.91 *(0.84 to 0.99)*	0.02
	R^2 = 0.41, AUC = 0.99		
Dec 1999 – Nov 2000	Distance to market	0.79 *(0.68 to 0.90)*	<0.005
	Elevation	0.95 *(0.93 to 0.98)*	<0.005
	R^2 = 0.21, AUC = 0.86		
Dec 2000 – Nov 2001	Proportion woodland (1km)	0.88 *(0.81 to 0.95)*	<0.005
	Proportion flooding (3 km)	1.18 (1.04 to 1.33)	0.01
	R^2 = 0.26, AUC = 0.90		
Dec 2001 – Nov 2002	Proportion woodland (1km)	0.90 (0.85 to 0.95)	<0.005
	Distance to market	0.90 (0.83 to 0.98)	0.01
	R^2 = 0.18, AUC = 0.84		

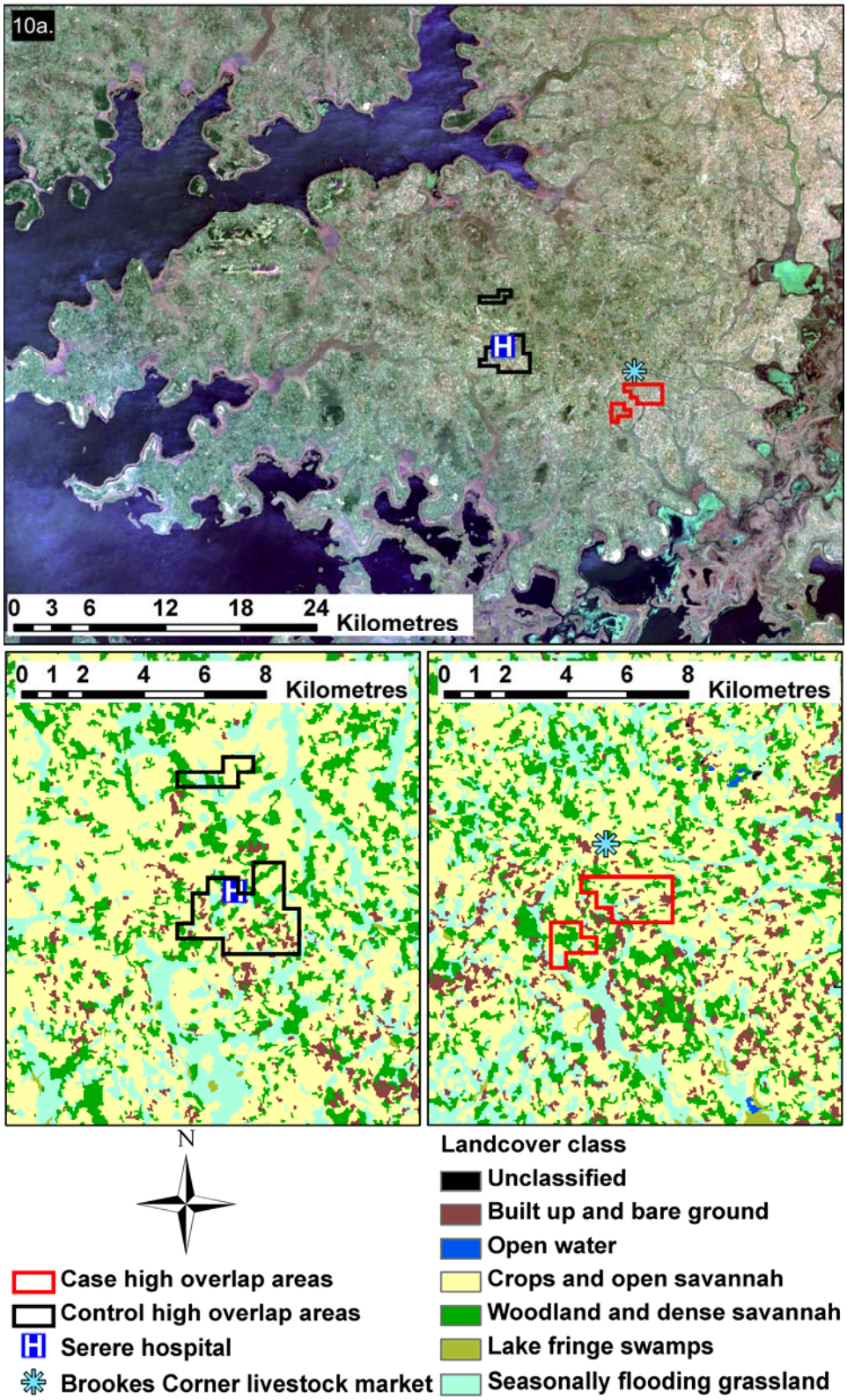
Table 5.8: Multivariate models for each of the four annual periods.

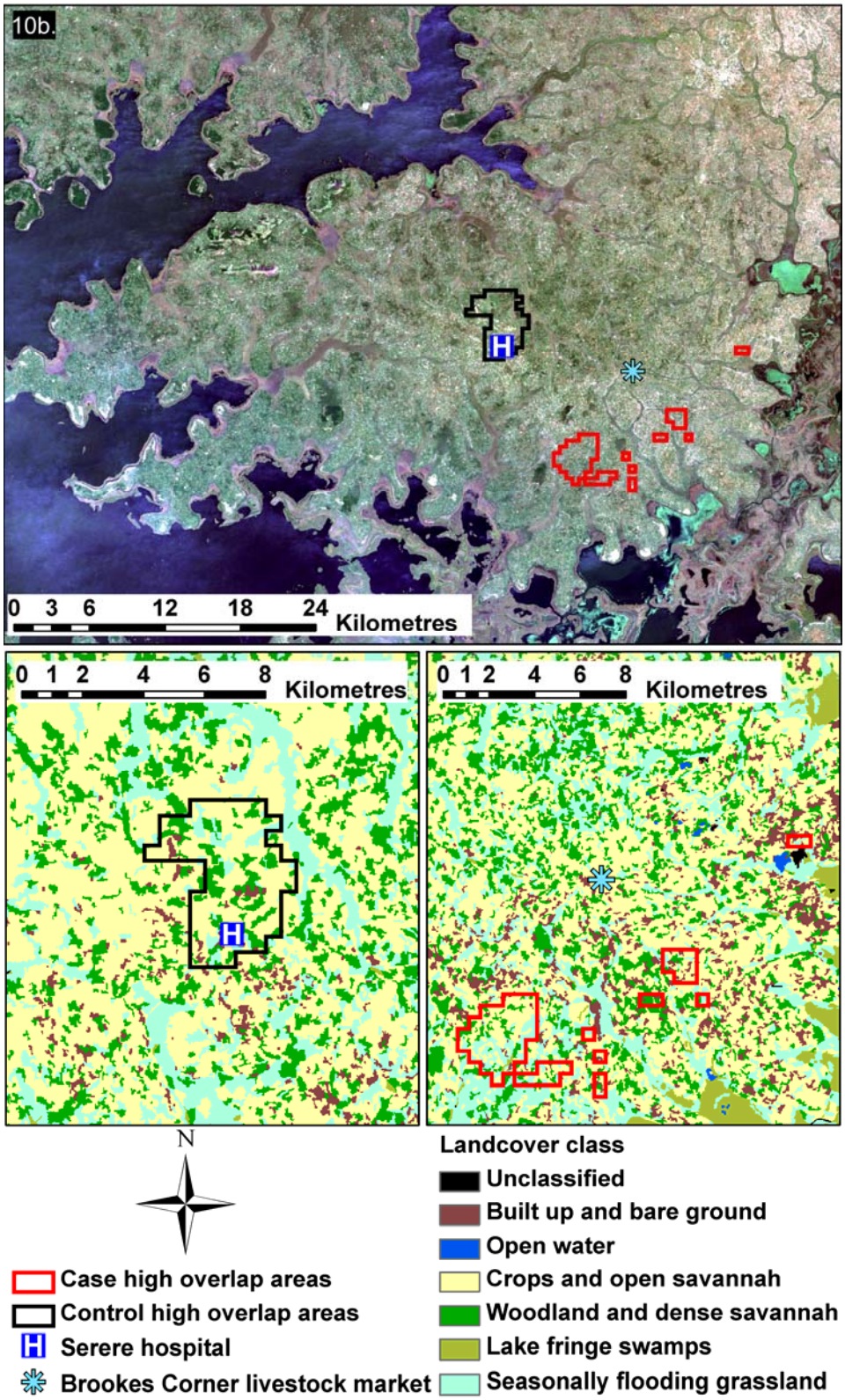
5.4.4. Identification and characterisation of potential high transmission zones

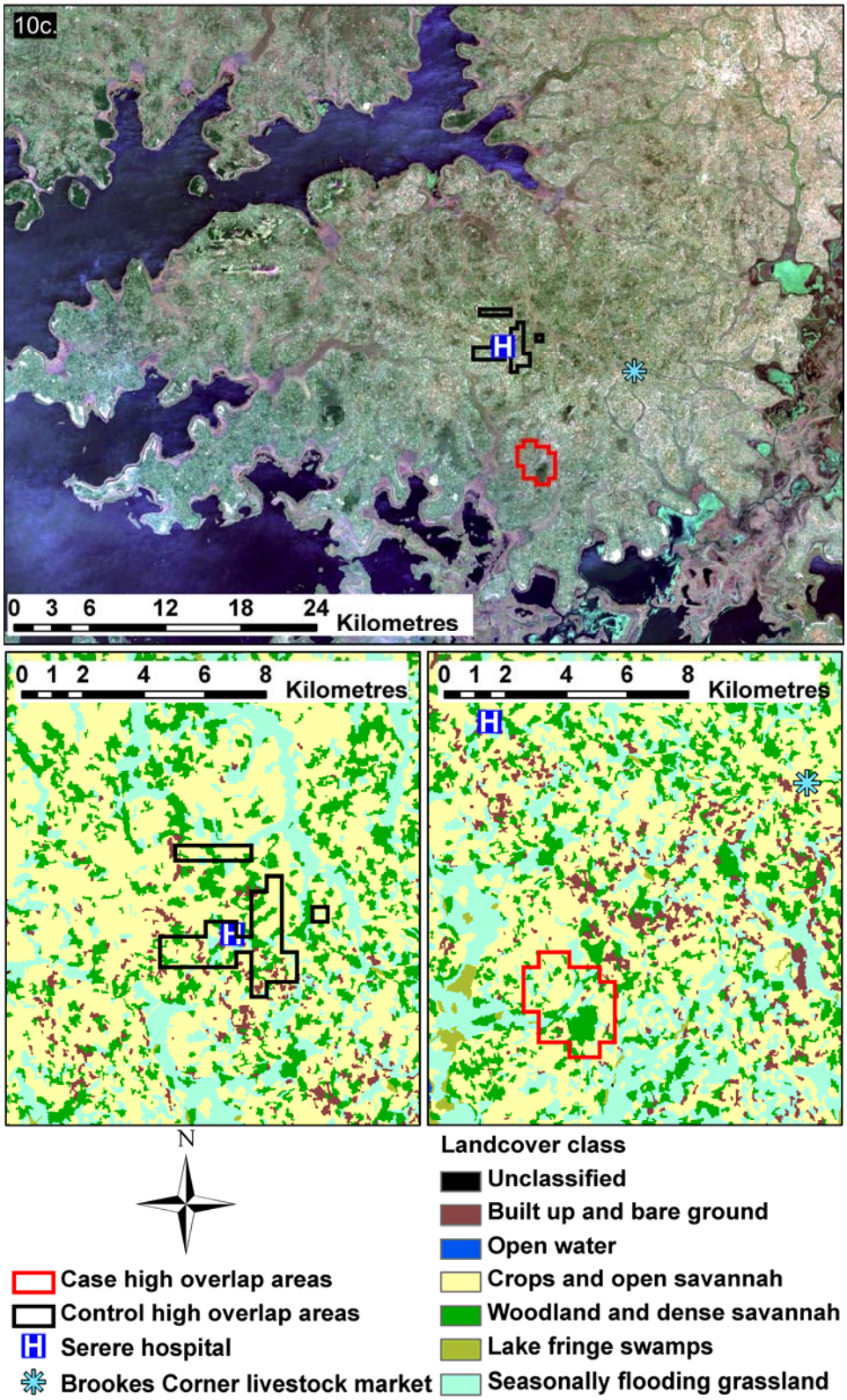
5.4.4.1. Identification of potential high transmission zones

Several high overlap areas were identified using 3 km buffer zones surrounding cases and controls for each of the four annual periods (see Figures 5.10 a - d for maps highlighting the case and control high overlap areas). The high overlap areas for the controls were distinct from those of the cases. For each time period, high overlap areas for controls were located within close proximity to the sleeping sickness treatment centre (Serere Hospital), and no directional trend was evident across the four annual periods. The high overlap areas for cases (potential high transmission zones) in the first annual period were located close to Brookes Corner livestock market, which has

previously been identified as the original source of the disease within the study area (Fèvre *et al.* 2001a). Over the study period, the high overlap areas for the cases moved from the original location (close to Brookes Corner livestock market) in a south-westerly direction.


10a.
0 3 6 12 18 24 Kilometres
0 1 2 4 6 8 Kilometres
0 1 2 4 6 8 Kilometres
N
Case high overlap areas
Control high overlap areas
Serere hospital
Brookes Corner livestock market
Landcover class
Unclassified
Built up and bare ground
Open water
Crops and open savannah
Woodland and dense savannah
Lake fringe swamps
Seasonally flooding grassland


10b.
0 3 6 12 18 24 Kilometres
0 1 2 4 6 8 Kilometres
0 1 2 4 6 8 Kilometres
N
Case high overlap areas
Control high overlap areas
Serere hospital
Brookes Corner livestock market
Landcover class
Unclassified
Built up and bare ground
Open water
Crops and open savannah
Woodland and dense savannah
Lake fringe swamps
Seasonally flooding grassland


10c.
0 3 6 12 18 24 Kilometres
0 1 2 4 6 8 Kilometres
0 1 2 4 6 8 Kilometres
N
Case high overlap areas
Control high overlap areas
Serere hospital
Brookes Corner livestock market
Landcover class
Unclassified
Built up and bare ground
Open water
Crops and open savannah
Woodland and dense savannah
Lake fringe swamps
Seasonally flooding grassland

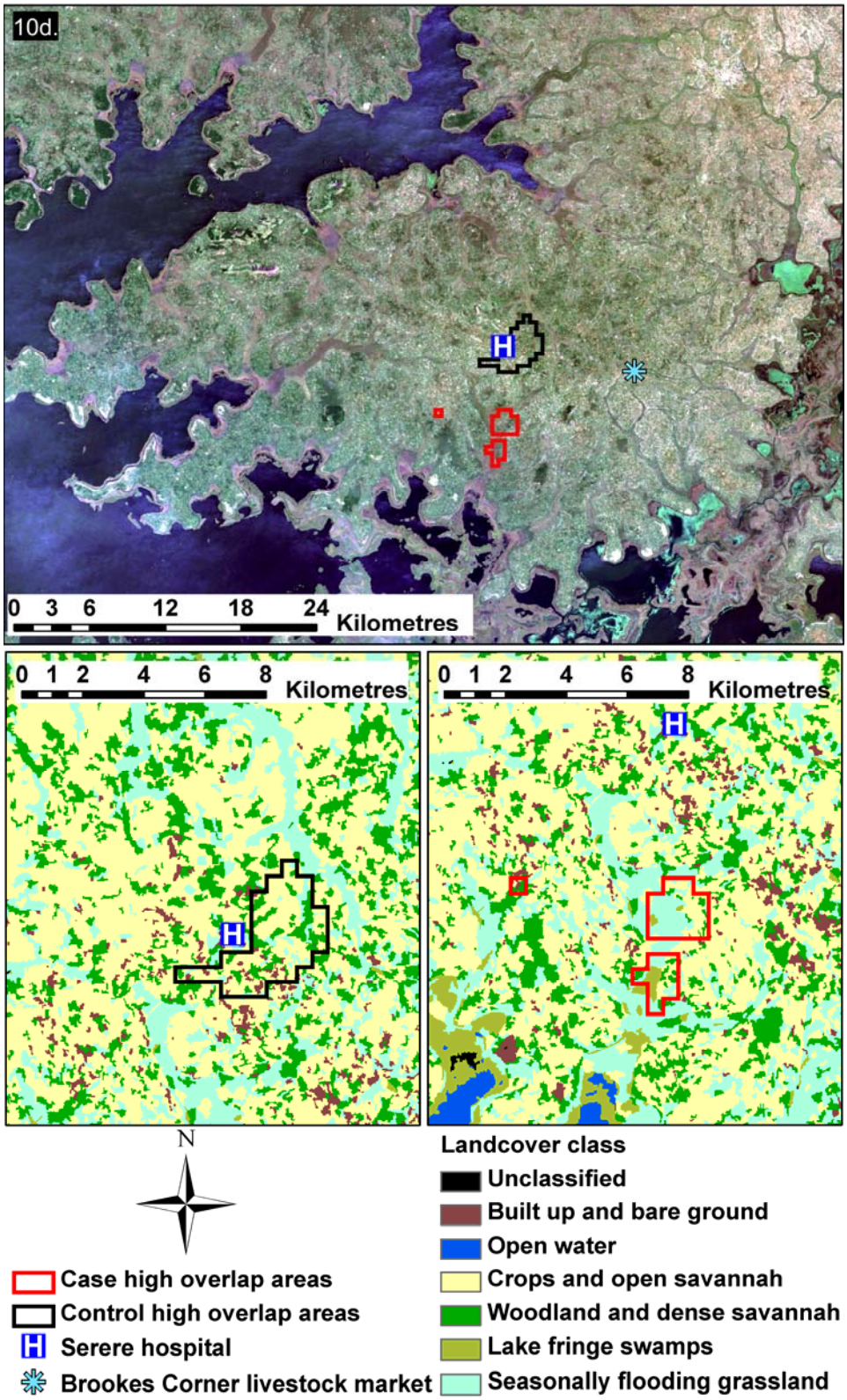

10d.
0 3 6 12 18 24 Kilometres
0 1 2 4 6 8 Kilometres
0 1 2 4 6 8 Kilometres
N
Case high overlap areas
Control high overlap areas
Serere hospital
Brookes Corner livestock market
Landcover class
Unclassified
Built up and bare ground
Open water
Crops and open savannah
Woodland and dense savannah
Lake fringe swamps
Seasonally flooding grassland

Figure 5.10 (previous four pages): True colour Landsat ETM+ composite of study area, highlighting case (red) and control (black) high overlap areas in the first (5.10a), second (5.10b), third (5.10c) and fourth (5.10d) annual periods, also with close up images of high overlap areas with landcover classes. Note the spatial scales for the close up images in 5.10b are not the same.

5.4.4.2. Characterisation of potential high transmission zones

A comparison of the landscape characteristics for the first annual period shows that the proportion of built up and bare ground, crops and open savannah and seasonally flooding grassland are not significantly different between the high overlap areas for cases and controls (see Table 5.9), although the case high overlap areas have a significantly lower proportion of woodland and dense savannah than control high overlap areas ($p < 0.001$). The same significant difference was seen in the second and final annual periods ($p < 0.001$), although in the third period the difference was reversed and the proportion of woodland and dense savannah was higher for the high overlap areas of cases than controls ($p < 0.001$).

From the second annual period onwards, the proportion of high overlap areas classified as crops or open savannah was significantly lower for cases than controls ($p < 0.001$ for each annual period), and the proportion classified as seasonally flooding grassland was significantly higher for cases than controls ($p < 0.001$ for each annual period). The proportion of built up and bare ground was significantly lower in high overlap areas for cases than controls in the two final annual periods ($p < 0.001$ for the third and fourth annual periods). Additionally, the mean elevations within the high overlap zones were significantly lower for cases than controls for the second, third and fourth annual periods ($p < 0.001$ for each of these three periods; see Table 5.9 and Figure 5.11).

From Table 5.9 and Figure 5.12a, a changing landscape within the potential high transmission zones (the high overlap areas for cases) can be observed over the study period. The proportion high overlap areas that were classified as crops and open savannah decreased from 62% to 34%; the proportion classified as built up and bare

ground decreased from 10% to 1%, and the proportion classified as seasonally flooding grassland increased from 10% to 41%. In comparison, the landscape seen in high overlap areas for controls remained relatively constant over time, with a predominance of the crops and open savannah land cover class (see Figure 5.12b). The average elevation within the high overlap zones for cases also demonstrated a temporal trend, with a decreasing mean elevation over time (decreasing from 1098 m during the first annual period, to 1048 m in the fourth annual period; see Figure 5.11), while the average elevation for high overlap areas for the controls remained relatively constant over the study period.

Annual period		Proportion of high overlap area classified as:				Mean elevation (metres)
		Built up & bare ground	Crops & open savannah	Seasonally flooding grassland	Woodland & dense savannah (2-tailed)	
Dec 1998 - Nov 1999	Cases	10.48%	61.05%	10.09%	18.38%	1098.27
	Controls	10.19%	55.21%	12.15%	22.44%	1095.37
	Difference	0.29%	5.84%	-2.06%	-4.07%	2.90
	p-value	0.72	>0.999	>0.999	<0.001	>0.999
Dec 1999 - Nov 2000	Cases	6.96%	48.86%	26.38%	17.13%	1066.31
	Controls	4.54%	59.61%	11.04%	24.81%	1100.10
	Difference	2.42%	-10.76%	15.34%	-7.67%	-33.79
	p-value	>0.999	<0.001	<0.001	<0.001	<0.001
Dec 2000 - Nov 2001	Cases	1.54%	52.66%	17.14%	28.66%	1058.35
	Controls	7.88%	59.56%	9.53%	23.03%	1099.74
	Difference	-6.34%	-6.90%	7.61%	5.64%	-41.39
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Dec 2001 - Nov 2002	Cases	0.90%	33.68%	41.60%	14.06%	1047.82
	Controls	5.81%	60.84%	10.02%	23.34%	1093.86
	Difference	-4.91%	-27.15%	31.58%	-9.27%	-46.04
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001

Table 5.9: Land cover profiles of case and control high overlap areas, and mean elevation for the four annual periods. *P*-values shown represent the significance of the difference in proportion (for land cover classes) from a *z*-test, or difference in mean (for elevation) from a *t*-test. The alternative hypotheses for the proportion of built up and bare ground and crops and open savannah classes and mean elevation were that the proportion or mean for case high overlap areas would be lower than for control high overlap areas (one-tailed test). The alternative hypothesis for the seasonally flooding grassland class was that the proportion would be higher for case high overlap areas than control high overlap areas (one-tailed test). The alternative hypothesis for the woodland and dense savannah class was that the difference in proportions would not be zero (two-tailed test).

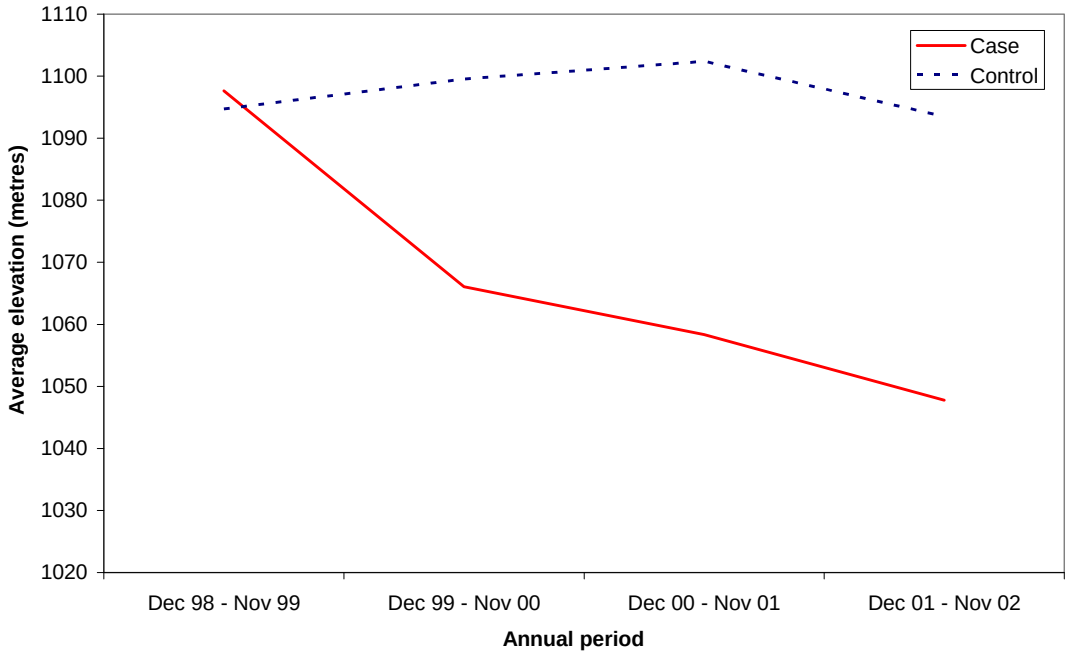
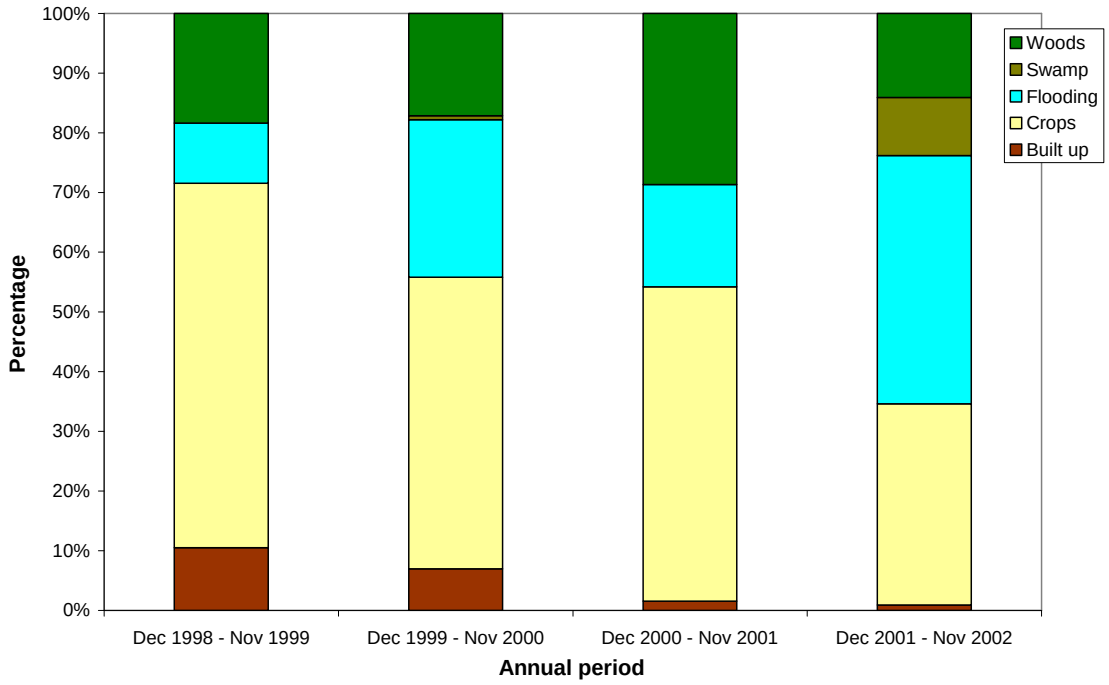


Figure 5.11: Average elevation within case and control high overlap areas, over the four annual periods.

12a.



12b.

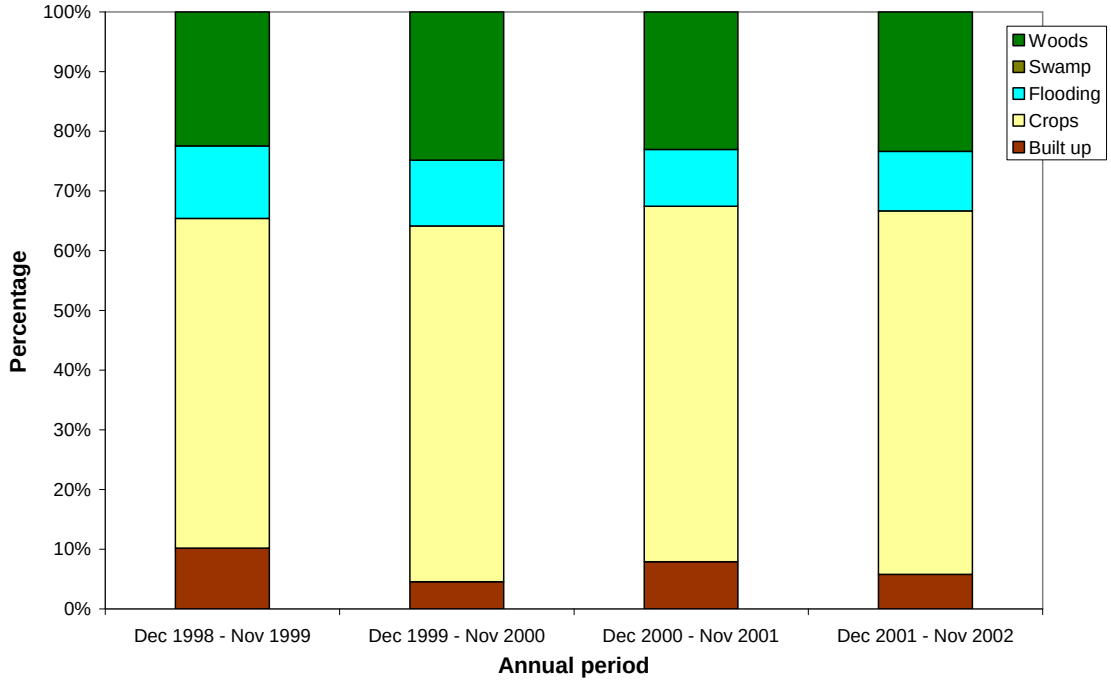


Figure 5.12: Land cover profile within high overlap area for cases (5.12a) and controls (5.12b), over the four annual periods.

The average population density within the high overlap areas for cases and controls during each annual period is shown in Table 5.10. Within the case high overlap areas, the average population density decreased over the study period (249 per km^2,

250.7 per km^2, 192.9 per km^2 and 190.7 per km^2 for the first, second third and fourth annual periods respectively). In contrast, the control high overlap areas show no temporal trend in their average population densities (155.2 per km^2, 650.6 per km^2, 781.4 per km^2 and 189.2 per km^2 for the first, second, third and fourth annual periods respectively). The average population densities for the first and fourth annual periods were slightly higher for the case high overlap areas than the control high overlap areas. The converse is true for the second and third annual periods, with considerably higher population densities (more than double) in the control high overlap areas than the case high overlap areas.

	Case	**Control**
Dec 1998 - Nov 1999	249.0	155.2
Dec 1999 - Nov 2000	250.7	650.6
Dec 2000 - Nov 2001	192.9	781.4
Dec 2001 - Nov 2002	190.7	189.2

Table 5.10: Average population density (number per km^2) within case and control high overlap areas, over the four annual periods.

5.5. Discussion

The use of a land cover classification alongside sleeping sickness case-control data was used to investigate the temporal landscape epidemiology of Rhodesian sleeping sickness, following the introduction of the disease to a previously unaffected area. Previous work has demonstrated the introduction of the parasite to Soroti district via livestock trading at Brookes Corner livestock market and the subsequent dispersion away from the market over time (Fèvre *et al.* 2001a). The current study furthers this work by illustrating the temporal movement of the disease following its introduction to the district, providing an insight into the ecology of Rhodesian sleeping sickness dispersal. The results presented may contribute towards the implementation of evidence-based control measures to prevent the further spread of Rhodesian sleeping sickness, or to prevent the establishment of transmission if a further introduction occurs in an area not currently affected.

The object-based land cover classification gave an overall accuracy of 86%. The classification had a higher accuracy than a similar utilisation of pixel-based methods which resulted in an overall accuracy of 81%. Producer and user accuracies were also higher for all classes using the object-based method (data not shown). The differentiation between classes using object-based classification was high despite the low separability of classes based on the spectral values of pixels alone (data not shown). Producer and users accuracies for all classes were high (over 70%), with the exception of the class crops and open savannah (producer accuracy = 69.2%, user accuracy = 46.6%). The lower user accuracy for this class was due to the misclassification of built up and bare ground, and forest and dense savannah, and represents a small classification bias towards the classification of crops and open savannah. This misclassification may have resulted in lower proportions of other classes which are of importance to the spatial distribution of sleeping sickness within the study area. However, the land cover class which demonstrated the clearest relationship to the spatial distribution of sleeping sickness throughout the study period, seasonally flooding grassland, did not suffer from large amounts of misclassification (producer accuracy = 97.9%, user accuracy = 90.9%).

The classification illustrated a predominance of agricultural land and open savannah within the study area, with 31.25% of the area classified as crops and open savannah. Areas of cropland and open savannah were thought to be restrictive for tsetse populations due to the disturbance of potential resting sites by human activities (Reid *et al.* 2000; Bouyer *et al.* 2006) and a lack of water which is required for the vector's survival. Almost 15% of the area was classified as seasonally flooding grassland which, based on the habitat requirements of the predominant vector species within the study area, *G. f. fuscipes*, was hypothesised to be a suitable vector habitat and, thus, potentially implicated in sleeping sickness transmission (Odiit *et al.* 2006; Zoller *et al.* 2008).

In line with previous studies, increasing population density was shown to have a protective effect in all four annual periods (Reid *et al.* 2000; Odiit *et al.* 2006).

Similarly, increasing nighttime lights, which are implicitly linked to population density (the population density dataset utilises nighttime lights data) and are also indirectly associated with poverty (Noor *et al.* 2008b), demonstrated a protective effect. These two variables illustrate the preference of tsetse fly vectors for areas with lower population densities and, thus, less human disturbance of vector habitats (Reid *et al.* 2000). Furthermore, higher predicted suitability for the vector *G. fuscipes* was shown to be a risk factor in the second, third and fourth annual periods indicating that the spatial dispersal within Soroti district may have been limited by habitat suitability.

Previous research has illustrated the statistically significant spatial clustering of sleeping sickness cases around Brookes Corner livestock market at the beginning of the study period, and described the subsequent dispersion of the cluster away from the livestock market over time (Fèvre *et al.* 2001a). The current research has also illustrated this phenomenon; the strength of the association between sleeping sickness occurrence and distance to Brookes Corner livestock market decreased as the disease dispersed away from the market and the correlation lost statistical significance in the third and fourth annual periods. In comparison, the distribution of the controls remained relatively static, supporting the hypothesis that external factors were influencing the movement of the disease over time.

When considering associations between the distribution of sleeping sickness and landscape features over time, a temporal trend is apparent. Initially, increased elevation was associated with higher odds of infection, but over time the odds ratio decreased, with increased elevation having a protective effect. A similar temporal trend was observed in the odds ratios for the proportion of land within 1 km and 3 km that was woodland or dense savannah (decreasing odds ratios over time, with higher proportions giving a protective effect in the later annual periods). The reverse of this relationship was true for the proportion of seasonally flooding grassland within 1 km or 3 km, with an initial protective effect in areas with a higher proportion of seasonally flooding grassland. In the second, third and fourth annual periods, areas with a higher proportion of seasonally flooding grassland in 1 km and

3 km buffers had higher odds of sleeping sickness. These trends signify the involvement of these landscape features in the temporal dispersion of sleeping sickness away from the point of introduction. The areas into which the disease dispersed may represent higher quality habitats (lower elevation, higher proportions of seasonally flooding grassland and lower proportions of woodland and dense savannah) which can more readily support vector populations and promote sleeping sickness transmission. The observed relationships with elevation may be related to the proportion of seasonally flooding grassland, as lower lying areas are more likely to become flooded during the wet season.

From the multivariate analysis, it can be seen that the most significant factors influencing the spatial distribution of sleeping sickness cases in the first two years following its introduction are distance to the livestock market and elevation, both of which have a protective effect when controlling for the other (lower odds of infection further from the livestock market and at higher elevations). Over time, the cases occur further from the livestock market and the significance of the surrounding landscape increases, with proportion of woodland and dense savannah within 1 km and seasonally flooding grassland within 3 km having the largest significance in the third annual period, and the proportion of woodland within 1 km along with distance to the livestock market included in the final model for the fourth annual period. The inclusion of distance to the livestock market in the multivariate model for the fourth annual period provides conflicting evidence from the univariate analysis, which illustrated the apparent dispersion of cases away from the livestock market over the study period.

In addition to the changing relationships observed in the multivariate regression analysis, the R^2 values indicate that a lower proportion of the spatial variation is being described by the regression models in later years. In the first annual period, the distribution was strongly dependent on proximity to Brookes Corner livestock market, which is demonstrated by a small odds ratio (OR = 0.48, $p < 0.005$). The correlation decreased in the subsequent annual periods (OR = 0.79 for the second annual period, non-significant in the third annual period and OR = 0.9 for the third

annual period) as the cases began dispersing further from the market. The R^2 values decreased from 0.41 in the first annual period, to 0.21, 0.26 and 0.18 in the second, third and fourth annual periods respectively, indicating that the first model describes the largest amount of the observed variation and the fourth model explains the smallest amount. The smaller R^2 values in the later annual periods may represent a more complex interaction of factors influencing the spatial distribution of the disease as the cases dispersed outwards from the livestock market, some of which were not included in the current analysis (such as tsetse distribution, livestock population densities and social factors).

Previous studies have found underreporting of sleeping sickness in Uganda to be a serious issue (Odiit *et al.* 2004b; Odiit *et al.* 2005). Access to diagnostic facilities can adversely affect the results of spatial analyses; individuals residing closer to the treatment centre are more likely to travel there for diagnosis and treatment than cases further away, thus, resulting in a spatial bias in the observed cases (Odiit *et al.* 2004a; Odiit *et al.* 2006). The matched case-control design in this study, taking controls specifically from Serere Hospital's patient records ensures that the controls were selected from the same population as the cases, thus, avoiding spatial bias.

Using the 3 km buffer zones around case villages, several potential high transmission zones were identified. These areas may be frequented by the residents of a number of surrounding villages (i.e. to water and graze livestock or collect firewood), and due to their particular landscapes and environmental conditions, may promote a high level of interaction between tsetse vectors, livestock and humans. The conditions present within these potential high transmission zones may be implicated in the observed increased occurrence of sleeping sickness in the surrounding villages. When comparing the location of high overlap zones of cases and controls a distinct difference can be observed, with the high overlap areas for cases moving gradually over the study period, but the high overlap zones for controls remaining relatively static and close to Serere hospital (as was expected).

The temporal changes in the potential high transmission zones over time further supports the hypothesis that the disease dispersed outwards from the original source of infection (the livestock market) into areas which were more suitable for the tsetse vector, and which promote interaction between the vector, the reservoir hosts and humans. The landscape within the potential high transmission zones in the second, third and fourth annual periods demonstrated higher proportions of seasonally flooding grassland, and lower proportions of crops and open savannah and built up and bare ground than the first annual period and also when compared with the high overlap zones for controls. Additionally, the elevation was lower in the latter annual periods than the first annual period, and when compared with the high overlap areas for controls. These landscapes may constitute areas with a higher risk of *T. b. rhodesiense* transmission due their suitability for vector populations, combined with an increased amount of interaction between humans, livestock and tsetse. The population density is low over the majority of the study area (the study area is predominantly rural) with a small number of areas with higher population densities (corresponding to the main towns). The high overlap areas for the controls in the second and third annual periods included the main town of Serere, whereas the high overlap areas for the cases throughout the study period encompass rural areas with low population densities. This demonstrates that the selection of potential high transmission zones using 3 km buffers is not influenced solely by the population distribution within the study area.

The probability of invasion by a species increases with increasing introduction challenge (a measure of the number of individuals of a species introduced to an area in which they are not already present); several separate introductions, each with few parasites may increase the chance of a successful invasion when compared with a single introduction attempt with many parasites (Lockwood *et al.* 2005). The relatively constant movement of livestock from endemic areas in the south east of Uganda into the recently affected areas may have increased the introduction challenge, with many infected cattle passing through the local livestock markets across a period of months or years, thus, increasing the likelihood of the establishment of transmission. The continued movement of livestock from *T. b.*

rhodesiense endemic areas of south east Uganda towards *T. b. gambiense* endemic areas in the north west of Uganda increases the risk of subsequent *T. b. rhodesiense* introductions into areas not currently affected and the probability of a successful invasion. An introduced species can persist at low levels in an area which may be pertained as unsuitable for a period of time prior to dispersion (Schreiber and Lloyd-Smith 2009). Due to the sparseness of livestock prevalence data, it is not possible to know how long the parasite had been present within the area in livestock populations, prior to the presentation of human cases. However, the results discussed here give some indications of the factors influencing the disease's dispersal and may assist in the focussing of disease control measures. To prevent further spread of the disease, periodic surveillance of cattle populations in areas surrounding potential points of introduction (livestock markets) could be implemented where feasible, to ensure the early detection of, and response to, any future introductions. Should a further introduction be detected, intensive control efforts may be required in areas surrounding the point of introduction and also in areas with higher quality habitats, to prevent the establishment of transmission and more widespread dispersal.

In this chapter, the dispersal of Rhodesian sleeping sickness into higher quality habitats over time has been illustrated, with decreasing spatial dependency on the point of introduction and an increasingly complex interaction of landscape factors influencing its distribution. From each element of this analysis, it can be seen that the landscapes more conducive to sleeping sickness transmission have a higher proportion of seasonally flooding grassland and lower proportions of the other land cover classes. In addition, lower elevation was seen to be a risk factor; presumably in correlation with seasonally flooding grassland (lower areas are more likely to flood during the rainy season). Low population density was also a risk factor, with higher odds of infection in areas with lower population density as has been shown previously (Odiit *et al.* 2006). However, the spatial epidemiology of sleeping sickness is guided by a very complex combination of factors, only some of which have been considered here. Finer spatial resolution satellite sensor imagery (for land cover classification) may be beneficial to aid a more accurate identification of high risk transmission zones and, thus, enable the more efficient targeting of control

efforts, although this is a more costly option. An increase in the spatial resolution of the land cover data would also necessitate increased spatial resolution patient geo-referencing (i.e. geo-referencing patients to their household rather than their village of residence). Further research is planned to build upon these findings using spatio-temporal modelling of an extended spatially-referenced Rhodesian sleeping sickness dataset. The planned research will provide further insights into the dynamic spatial distribution of the disease and aims to predict areas at highest risk of sleeping sickness in the future, based on a combination of environmental, social and climatic variables, including land cover.

Chapter 6. General discussion

“We can only see a short distance ahead, but we can see plenty there that needs to be done.”

Alan Turing (1912 – 1954)

The research discussed within this thesis focused on the spatial distribution of Rhodesian sleeping sickness within several recently affected districts of Uganda, with the primary aims of identifying contributory factors for the continued spread of the disease, explaining spatial heterogeneity in disease prevalence and investigating the dispersal ecology of the disease following its introduction to a new area. Using spatially referenced sleeping sickness and covariate data (as discussed in Chapter 2), three different regression methods (non-spatial [Chapter 3] and spatial [Chapter 4] logistic regression) were applied and compared for the assessment of environmental, climatic and social factors influencing the spatial distribution of Rhodesian sleeping sickness. Additionally, predictive maps were produced to highlight areas with the potential for high prevalence, based on significant covariates. Using a temporally stratified analysis, the spatial dispersion of Rhodesian sleeping sickness, following its introduction to a new area was investigated in relation to environmental factors (Chapter 5). Potential high transmission zones were also identified and characterised, providing evidence of the landscape features which promote transmission of the disease.

6.1. Sleeping sickness distributions

The analysis discussed in Chapter 3 suggested that the spread of *T. b. rhodesiense* into Kaberamaido and Dokolo districts may have been brought about via livestock movements. The extension of this analysis (Chapter 4) provided strengthened evidence in support of this hypothesis; despite the introduction of regulations requiring the treatment of all cattle from *T. b. rhodesiense* endemic areas prior to sale at livestock markets (Wendo 2002), the spread of the disease into Kaberamaido and Dokolo districts is likely to have resulted from the continued movement of untreated, infected livestock. The predictive maps in Chapters 3 and 4 illustrate that the areas directly surrounding livestock markets have the highest predicted prevalence. This emphasises the significance of livestock movements for the importation of the parasite to new areas and highlights the spatial constraint that the livestock market exerts on the initial distribution of the disease, as has been demonstrated previously (Fèvre *et al.* 2001a; Fèvre *et al.* 2005).

Results from a recent survey tracking livestock movements in Uganda suggest that a large amount of livestock traffic travels on a weekly basis from *T. b. rhodesiense* endemic areas in the south east of Uganda to *T. b. gambiense* endemic areas in the north west of Uganda, in some cases entering southern Sudan (via illegal smuggling). In addition, it was found that the implementation of cattle treatment regulations was incomplete due to a combination of factors including: inability or unwillingness to pay for the treatment; illegal trading outside of official livestock markets; lack of drugs and poor market facilities (R. Selby, pers. com.). The continuing flow of livestock from *T. b. rhodesiense* to *T. b. gambiense* endemic areas, with the percentage of cattle treated with trypanocidal drugs remaining below 100%, clearly results in a sustained risk of further spread of the disease into new areas, with the potential for an overlap of the two forms of the disease.

The findings in Chapters 3, 4 and 5 illustrate the spatial constraints on sleeping sickness distributions due to a complex interaction of factors: livestock movements, climatic factors and land cover all directly or indirectly influence the observed spatial distribution of Rhodesian sleeping sickness. This supports previous research which has linked tsetse, sleeping sickness and animal trypanosomiasis distributions with climatic and environmental factors (Rogers *et al.* 1996; Wint and Rogers 2000; Hendrickx *et al.* 2000; Hendrickx *et al.* 2001b; Odiit *et al.* 2006). Proximity to the closest livestock market (the most likely point of introduction) was a strong constraining factor, but as shown in Chapter 5 this relationship weakened over time in Soroti district, giving rise to a more complex combination of influencing factors. This can be related to the dispersion of the disease outwards from the point of introduction over time, guided by environmental cues, into areas more suitable for the establishment and maintenance of transmission (Schreiber and Lloyd-Smith 2009).

Accessibility to health care services was found to exert a confounding effect on the observed distribution of disease, as has been previously reported (Odiit *et al.* 2006). It has been estimated that almost 40% of all sleeping sickness cases in an outbreak in eastern Uganda were not diagnosed or treated and, therefore, were not included in

sleeping sickness surveillance data (Odiit *et al.* 2005). This indicates the likelihood that Rhodesian sleeping sickness is present within the study areas at a higher prevalence than observed and that sleeping sickness transmission may also be occurring outside of the observed distribution, in areas more remote from the available diagnostic and treatment facilities. Unfortunately the very low prevalence of infection coupled with the unavailability of easy to use, rapid and cheap diagnostic tests makes active screening for Rhodesian sleeping sickness outside of areas known to be affected an inefficient exercise. The use of regular livestock sampling as a form of sentinel surveillance for *T. b. rhodesiense* may be a more realistic option to gain a more complete picture of the distribution of the parasite. However, the presence of *T. b. rhodesiense* in livestock does not always equate to the occurrence of sleeping sickness in humans.

6.2. Methods for the spatial analysis of sleeping sickness data

The application of both non-spatial (logistic regression) and spatial (model-based geostatistics) methods to the analysis of sleeping sickness distributions has offered a unique comparison of these methods. The spatial logistic regression method applied in Chapter 4 provided a more robust method for the assessment of covariate effects and prediction of disease prevalence while accounting for spatial autocorrelation and resulted in increased predictive accuracy when compared with the non-spatial methods used in Chapter 3. The application of model-based geostatistics to infectious disease distributions has become more widespread in recent years, with several studies demonstrating its utility (Diggle *et al.* 2002; Diggle *et al.* 2007; Hay *et al.* 2009). It seems clear that future research should focus methodologically on the further application of explicitly spatial regression methods with Bayesian implementation to ensure consistent high standards. The approach used in Chapter 4 may also be extended as discussed in Section 6.5.

6.3. Study limitations

The presence of an animal reservoir of disease complicates issues regarding the surveillance of *T. b. rhodesiense*. Tsetse feed preferentially on animals, and Rhodesian sleeping sickness cases are normally the result of sporadic infections outside of the normal animal-tsetse-animal transmission cycle (Onyango *et al.* 1966; Waiswa *et al.* 2003; Waiswa *et al.* 2006). Therefore, prevalence in the animal reservoir is likely to be higher than in humans (Hide *et al.* 1996; Welburn *et al.* 2001b; Fèvre *et al.* 2005), with human infections occurring only where a high prevalence exists in reservoir populations, or where landscape features promote an increased level of interaction between reservoir animals, tsetse and humans. The analyses discussed within the thesis do not explicitly account for the distribution of livestock hosts or of *T. b. rhodesiense* prevalence within animals, both of which may influence significantly the spatial distribution of human sleeping sickness. Data regarding livestock populations within Uganda are available from the national livestock census. The most recent census was carried out in 2008; these data were not available for inclusion in the analyses. The previous livestock census in Uganda was carried out in 1992 and may now be considered as out of date. A comparison of livestock census data with data collected during the implementation of the Stamping Out Sleeping Sickness campaign indicated considerable inaccuracies in the available livestock population estimates (S. Welburn, pers. com.). In addition, the availability of livestock *T. b. rhodesiense* prevalence data is spatially and temporally irregular, resulting in problems with its utilisation. The future incorporation of data capturing the distribution and density of livestock populations, along with livestock prevalence of *T. b. rhodesiense* could enhance the analyses discussed in Chapters 3, 4 and 5 by integrating an additional part of the Rhodesian sleeping sickness transmission cycle. Additional temporal dynamics could be investigated by using time-series data to illustrate the seasonality of animal *T. b. rhodesiense* infections in relation to human sleeping sickness. An extension of the analysis discussed in Chapter 4 may allow the integration of such data in future analyses, as discussed in Section 6.5.

The identification of local livestock markets as the likely point of introduction in the study areas previously (Fèvre *et al.* 2001b) and in the current analyses (Chapters 3 and 4) illustrates the importance of livestock trading and ownership in the spatial epidemiology of Rhodesian sleeping sickness. However, the discussed analysis has not taken into account livestock trading networks other than via the location of livestock markets. An integrated approach utilising further information on the livestock trading network (e.g. volume of livestock coming into each market from *T. b. rhodesiense* endemic areas per week), or a dynamic network model representing the livestock trading network may further our understanding of the spread of Rhodesian sleeping sickness. In particular, such analysis could provide an indication of which areas currently free of *T. b. rhodesiense* have the highest risk of an introduction based on livestock imports.

Although the distribution of sleeping sickness depends on the availability of competent vector populations, the availability of vector distribution or abundance data within Uganda is scarce. The study areas are located in a 'boundary' area on the periphery of the predicted range of *G. fuscipes*, where the predicted suitability is predominantly less than 50% (Wint and Rogers 2000). Within these areas, the spatial distribution of tsetse is spatially heterogeneous, but currently there is little observational data available to provide accurate information on the distribution or abundance of vectors. Such information would provide invaluable input for analyses of the spatial distribution of sleeping sickness. The predicted distribution of *G. f. fuscipes* was used in exploratory analyses, but was excluded from the multivariate regression models as the variable is a model output itself, and the variables used for its creation may be correlated with the variables selected for use in the analyses discussed in Chapters 3, 4 and 5, resulting in redundancy. A quantification of the tsetse population within the study areas in terms of spatial distribution and abundance, using trapping methods, would provide important information relating to sleeping sickness transmission. Additional laboratory analysis of the captured tsetse flies could be used to establish *T. b. rhodesiense* infection rates (or infections rates for other trypanosome species), and to identify the source of blood meals (e.g. human, livestock, wild animals). The results of this type of tsetse analysis could

serve to enhance the spatial analyses discussed in the thesis by providing further data relating to the transmission cycle of *T. b. rhodesiense*.

In terms of spatial heterogeneity, the spatial scale of such heterogeneity is of importance for spatial analyses. The distribution of sleeping sickness within affected areas is extremely patchy, with some villages within an area of high prevalence having an observed prevalence of zero. The predictions resulting from the regression models demonstrated a reasonable ability to delineate areas of high prevalence, although within these areas there was a tendency to over-predict prevalence in villages with an observed prevalence of zero. One possible reason for this lack of accuracy within high prevalence areas is the spatial scale of the covariates used. The fine scale spatial heterogeneity in sleeping sickness prevalence within the affected area may be dictated by features smaller than the spatial scale of the covariates, such that these features may not be detected adequately. Further research utilising datasets with a finer spatial scale (e.g. Moderate-resolution Imaging Spectroradiometer data with a 500 m spatial resolution (Savtchenko *et al.* 2003)) may be beneficial to increase the accuracy of predictions within areas of high prevalence. However, to accompany an increase in the spatial resolution of covariate data, finer spatial scale geo-referencing of sleeping sickness patient data may be necessary (i.e. geo-referencing patient's homes rather than village of residence). The use of multiple spatial scales may also be of assistance by providing a fuller understanding of the different factors influencing sleeping sickness distributions at different spatial scales (Marshall 1991). Multiple scale studies have been used frequently in ecological research (Levin 1992) and may have applicability for a variety of environmentally constrained infectious diseases.

Further social factors exerting an influence on the spatial heterogeneity of Rhodesian sleeping sickness include mobility patterns, occupation, livestock ownership, land use patterns (as opposed to simply land cover) and other behavioural factors (Wyatt *et al.* 1985; Okia *et al.* 1994; Fournet *et al.* 2000). Human behaviour and mobility control the level of contact between an individual, tsetse vectors and reservoir hosts, thus, altering the likelihood of infection. Such

factors could be integrated in future analysis by including individual level covariates (e.g. occupation or frequent travel outside the village of residence) as an extension of the analyses in Chapter 3, 4 and 5 (section 5.3.2), or by incorporating information on land use patterns, daily movement patterns or inter-village footpath networks to assist in the identification of potential high transmission zones (Chapter 5, section 5.3.3). By extending the current analyses in such a manner, the interaction between humans and a spatially heterogeneous landscape, due to behaviour and mobility patterns, can be accounted for.

6.4. Implications for disease control

The most effective method for the prevention of Rhodesian sleeping sickness spread within Uganda is to prevent the introduction of the parasite to any areas where it does not currently exist (Mack *et al.* 2000). The findings discussed in the thesis indicate that the enforcement of cattle treatment regulations at livestock markets should be a priority for the Ugandan Government to prevent further long distance movements. Methods preventing the movement of animals potentially infected with specific diseases into areas where the disease is not currently endemic are used widely; for example the use of quarantine in the United Kingdom aims to prevent the importation of diseases such as Rabies. However, it may not be feasible to ensure complete adherence to livestock treatment policies for various reasons, including the illegal trade of livestock outside of authorised markets. The persistent risk of further spread leads to continued concern of a potential future overlap with the Gambian form of the disease. The areas thought to be at highest risk of an introduction of Rhodesian sleeping sickness (the districts currently within the buffer zone directly between areas of Gambian sleeping sickness and Rhodesian sleeping sickness transmission) should be targeted for the provision of diagnostic capabilities, along with awareness raising among health care workers to ensure the timely detection of any further spread.

The information gained on the invasion dynamics of Rhodesian sleeping sickness has provided invaluable information for the planning and targeting of control

activities to prevent the successful invasion of further areas, should another introduction occur. The implementation of regular surveillance in areas with an increased risk of introduction (i.e. livestock markets in currently unaffected areas) can allow the early detection of an introduction; this is vital to allow the timely application of control measures for the prevention or limitation of disease dispersion. Parasitological surveys of livestock may offer an ideal opportunity for the detection of *T. b. rhodesiense* transmission in an area before the presentation of human sleeping sickness cases.

Following the detection of an introduction, two primary options exist regarding the targeting of control efforts to prevent the establishment and dispersion of the disease within an area. Areas surrounding the point of introduction (most likely to be a livestock market) may be targeted to prevent the establishment of transmission within the locality, from whence the disease can disperse outwards over time. Alternatively, areas which are defined as being high quality habitats may be targeted; these are areas into which the organisms are mostly likely to disperse and will support an increased reproductive capacity. If the disease reaches such areas, the invasion will be more likely to succeed (Schreiber and Lloyd-Smith 2009). Further data regarding the dispersion of *T. b. rhodesiense* in animal populations over time would be beneficial for the planning of such interventions.

6.5. Future research priorities

Despite *T. b. rhodesiense* prevalence in livestock being higher than in humans, livestock prevalence data are sparse and the large scale livestock surveys necessary to build a comprehensive picture of the spatial distribution of *T. b. rhodesiense* would be extremely time demanding and expensive. The recent development of joint disease mapping methods may provide a realistic option for the incorporation of animal trypanosomiasis prevalence data. Joint disease mapping (also called multivariate modelling) can be used to model the distributions of two or more related diseases. These methods use information gained from the joint distribution of the two (or more) diseases to increase the accuracy of estimates of covariate

effects and predictions (Langford *et al.* 1999; Assuncao and Castro 2004; Held *et al.* 2005; Downing *et al.* 2008). The spatial distribution of *T. b. rhodesiense* in humans is dependent on the distribution of the parasite in livestock populations. Thus, *T. b. rhodesiense* prevalence data in livestock may provide additional information to the modelling process. Outputs of joint disease mapping would be the estimation of covariate effects and spatial predictions for both of the diseases (human and animal infections). The application of joint disease modelling techniques to the distributions of several cancer types has demonstrated its utility and indicated an increased predictive precision when compared to the modelling of cancer types separately (Assuncao and Castro 2004; Held *et al.* 2005; Downing *et al.* 2008).

The incorporation of an explicitly temporal element in the spatial modelling would also enhance future research, by providing a more comprehensive analysis of the dynamic spatial patterns of sleeping sickness. By linking together the dynamic distribution of sleeping sickness with the environmental and climatic constraints and knowledge of livestock trading networks, it is possible to build a more complete picture of the spatio-temporal dynamics of sleeping sickness distributions. The integration of a livestock movement network model would allow an assessment of the probability of an introduction of Rhodesian sleeping sickness to areas not currently affected, and the likelihood of establishment of transmission if such an introduction was to take place. Ideally, this would be supplemented by the modelling of Gambian sleeping sickness if appropriate data become available. Further research aims include the production of sleeping sickness risk maps covering areas endemic for Rhodesian sleeping sickness and Gambian sleeping sickness, as well as the buffer zone between the two; a quantification of the risk of overlap of the two forms of sleeping sickness and cartographic representations of overlap risk to allow the most effective targeting of control efforts.

The incorporation of information relating more directly to the transmission cycle of *T. b. rhodesiense* in livestock, tsetse and human populations could also provide additional insights into the heterogeneous spatial distribution of disease. An investigation of the possible use of agent-based modelling for sleeping sickness

transmission within Uganda, as has previously been used for the simulation of Gambian sleeping sickness spread (Muller *et al.* 2004), is planned. If the data requirements can be achieved, this type of method may be used to allow the incorporation of the mobility patterns of tsetse, livestock and human populations, in relation to heterogeneous landscape features (i.e. land cover and elevation). This type of analysis will explicitly include many key parameters and relationships in the transmission cycle of *T. b. rhodesiense* and, thus, may give a more accurate representation of the spatio-temporal dynamics of sleeping sickness distributions (Perez and Dragicevic 2009). By integrating the outputs from geostatistical modelling with the outputs from a systems-based approach, it should be possible to build a broader picture of the spatial heterogeneity in *T. b. rhodesiense* transmission and prevalence, with the ultimate aim of providing cartographic representations of risk that can be used to improve the spatial targeting of control measures by the Ugandan Ministry of Health.

Chapter 7. Conclusions

“Begin at the beginning…and go on till you come to the end: then stop.”

Lewis Carroll (1832 - 1898)

Said by the King to the White Rabbit in Alice in Wonderland.

The application of non-spatial and spatial regression analyses and a temporally stratified analysis has illuminated various features of the spatial epidemiology of Rhodesian sleeping sickness in recently affected areas of Uganda. These findings will provide valuable information for the planning of further sleeping sickness control strategies within Uganda to prevent the further spread of Rhodesian sleeping sickness and, ultimately, prevent an overlap between the Rhodesian and Gambian forms of the disease. The key findings from the research discussed within the thesis are summarised below:

- The spread of Rhodesian sleeping sickness into Kaberamaido and Dokolo districts is likely to have resulted from the continued trade and movement of untreated, infected livestock from endemic areas.
- The enforcement of cattle treatment regulations within markets should be a priority to prevent further spread of Rhodesian sleeping sickness within Uganda.
- Until cattle treatment regulations can be implemented more completely, areas surrounding livestock markets will have the highest risk of disease introduction and transmission. Targeting of *T. b. rhodesiense* surveillance and control activities to these areas may be beneficial to ensure early detection of an introduction and the control of the parasite prior to more widespread dispersion.
- Proximity to the point of introduction is a predominant constraining factor on the distribution of Rhodesian sleeping sickness, although the strength of the association decreases over time along with an increasing influence exerted by environmental and climatic factors.
- Accessibility of health care services acts as a confounding factor, with fewer sleeping sickness cases diagnosed in villages more remote from the closest health care facility.
- The dispersion of Rhodesian sleeping sickness following its introduction to a previously unaffected area is guided by environmental factors, including land cover, into more suitable habitats.

- The application of model-based geostatistics, which accounts for spatial autocorrelation, provides increased predictive accuracy and more precise estimates of covariate effects when compared to non-spatial methods.

References

Adams, E. B. 1907, *Report of the sleeping sickness commission of the Royal Society*.

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. 1976, *A land use and land cover classification system for use with remote sensor data*, United States Department of the Interior, Washington, USA, 964.

Apted, F. I. C. 1970, "Clinical manifestations and diagnosis of sleeping sickness," *In The African Trypanosomiases*, H. W. e. a. Mulligan, ed., London: George Allen and Unwin, pp. 661-683.

Assuncao, R.M. & Castro, M.S.M. 2004. Multiple cancer sites incidence rates estimation using a multivariate Bayesian model. *International Journal of Epidemiology*, 33, (3) 508-516

Azene, G., Guyatt, H., Brooker, S., Hall, A., Bundy, D.A.P., Adjei, S., Allan, E., Cobbinah, J., Kihamia, C., Nkungu, D.R., Kiwelu, V.M., Magingo, J., Yona, E., & Issae, W. 1999. The cost of large-scale school health programmes which deliver anthelmintics to children in Ghana and Tanzania. *Acta Tropica*, 73, (2) 183-204

Baatz, M., Benz, U., Dehghani, S., Heynen, M., Höltje, A., Hofmann, P., Lingenfelder, I., Mimler, M., Sohlbach, M., Weber, M., & Willhauck, G. 2000, *eCognition professional user guide 4*, Definiens Imaging GmbH, Munich, Germany.

Barrett, M.P., Burchmore, R.J.S., Stich, A., Lazzari, J.O., Frasch, A.C., Cazzulo, J.J., & Krishna, S. 2003. The trypanosomiases. *Lancet*, 362, (9394) 1469-1480

Beck, L.R., Rodriguez, M.H., Dister, S.W., Rodriguez, A.D., Rejmankova, E., Ulloa, A., Meza, R.A., Roberts, D.R., Paris, J.F., Spanner, M.A., Washino, R.K., Hacker, C., & Legters, L.J. 1994. Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *American Journal of Tropical Medicine and Hygiene*, 51, (3) 271-280

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58, (3-4) 239-258

Bergquist, N.R. 2001. Vector-borne parasitic diseases: new trends in data collection and risk assessment. *Acta Tropica*, 79, (1) 13-20

Berrang-Ford, L. 2007. Civil conflict and sleeping sickness in Africa in general and Uganda in particular. *Conflict and Health*, 1, (1) 6

Berrang-Ford, L., Berke, O., Abdelrahman, L., Waltner-Toews, D., & McDermott, J. 2006. Spatial analysis of sleeping sickness, southeastern Uganda, 1970-2003. *Emerging Infectious Diseases*, 12, (5) 813-820

Best, N., Richardson, S., & Thomson, A. 2005. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, (1) 35-59

Blaschke, T., Lang, S., Lorup, E., Strobl, J., & Zeil, P. Object-oriented image processing in an intergrated GIS/remote sensing environment and perspectives for environmental applications, *In International Symposium Computer Science for Environmental Protection of the Deutsche Gesellschaft für Informatik*, A. Cremers & K. Greve, eds., Marburg: Metropolis, pp. 555-570.

Blaschke, T. & Strobl, J. 2001. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. *GIS Zeitschrift für Geoinformationssysteme*, 6, 12-17

Borst, P., Rudenko, G., Taylor, M.C., Blundell, P.A., vanLeeuwen, F., Bitter, W., Cross, M., & McCulloch, R. 1996. Antigenic variation in trypanosomes. *Archives of Medical Research*, 27, (3) 379-388

Bouyer, J., Guerrini, L., Desquesnes, M., De La Rocque, S., & Cuisance, D. 2006. Mapping African animal trypanosomosis risk from the sky. *Veterinary Research*, 37, (5) 633-645

Boyd, D.S., Foody, G.M., & Curran, P.J. 1999. The relationship between the biomass of Cameroonian tropical forests and radiation reflected in middle infrared wavelengths (3.0-5.0 mu m). *International Journal of Remote Sensing*, 20, (5) 1017-1023

Boyd, H.A., Flanders, W.D., Addiss, D.G., & Waller, L.A. 2005. Residual spatial correlation between geographically referenced observations - A Bayesian hierarchical modeling approach. *Epidemiology*, 16, (4) 532-541

Breslow, N. E. & Day, N. E. 1980, "Conditional logistic regression for matched sets," *In Statistical Methods in Cancer Research Volume 1: The Analysis of Case-Control Studies*, First ed. W. Davis, ed., Lyon, France: IARC Scientific Publications, pp. 247-279.

Brooker, S. 2002. Schistosomes, snails and satellites. *Acta Tropica*, 82, (2) 207-214

Brooker, S. 2007. Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101, (1) 1-8

Brooker, S., Beasley, M., Ndinarotan, M., Madjiouroum, E.M., Baboguel, M., Djenguinabe, E., Hay, S.I., & Bundy, D.A.P. 2002a. Use of remote sensing and a geographical information system in a national helminth control programme in Chad. *Bulletin of the World Health Organization*, 80, (10) 783-789

Brooker, S., Hay, S.I., & Bundy, D.A.P. 2002b. Tools from ecology: useful for evaluating infection risk models? *Trends in Parasitology*, 18, (2) 70-74

Brooker, S., Hay, S.I., Issae, W., Hall, A., Kihamia, C.M., Lwambo, N.J.S., Wint, W., Rogers, D.J., & Bundy, D.A.P. 2001. Predicting the distribution of urinary schistosomiasis in Tanzania using satellite sensor data. *Tropical Medicine & International Health*, 6, (12) 998-1007

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

Brown, J.H., Mehlman, D.W., & Stevens, G.C. 1995. Spatial variation in abundance. *Ecology*, 76, (7) 2028-2043

Brownstein, J.S., Rosen, H., Purdy, D., Miller, J.R., Merlino, M., Mostashari, F., & Fish, D. 2003. Spatial analysis of West Nile Virus: Rapid risk assessment of an introduced Vector-Borne Zoonosis (vol 2, pg 157, 2002). *Vector-Borne and Zoonotic Diseases*, 3, (3) 155

Bruce, D. 1895. *Tsetse fly disease or Nagana* London, Harrison and Sons.

Bruce, D., Hamerton, A.E., Bateman, H.R., & Mackie, F.P. 1910. Experiments to ascertain if cattle may act as a reservoir of the virus of sleeping sickness (*Trypanosoma gambiense*). *Proceedings of the Royal Society of London.Series B, Containing Papers of a Biological Character*, 82, (558) 480-484

Bryceson, D.F., Mbara, T.C., & Maunder, D. 2003. Livelihoods, daily mobility and poverty in sub-saharan Africa. *Transport Reviews*, 23, (2) 177-196

Bukachi, S.A., Wandibba, S., & Nyamongo, I.K. 2009. The treatment pathways followed by cases of human African trypanosomiasis in western Kenya and eastern Uganda. *Annals of Tropical Medicine and Parasitology*, 103, (3) 211-220

Burri, C., Stich, A., & Brun, R. 2004, "Current chemotherapy of human African trypanosomiasis," *In The Trypanosomiases*, I. Maudlin, P. H. Holmes, & M. A. Miles, eds., Wallingford: CABI publishing, pp. 403-419.

Buscher, P. & Lejon, V. 2004, "Diagnosis of human African trypanosomaisis," *In The Trypanosomiases*, I. Maudlin, P. H. Holmes, & M. A. Miles, eds., Wallingford: CABI Publishers, pp. 203-218.

Campbell, J. B. 2002a, "Electromagnetic radiation," *In Introduction to Remote Sensing*, Third ed. New York: Taylor and Francis, pp. 25-52.

Campbell, J. B. 2002b, "Hydrospheric sciences," *In Introduction to Remote Sensing*, Third ed. New York: The Guildford Press, pp. 523-553.

Campbell, J. B. 2002c, "Image classification," *In Introduction to Remote Sensing*, Third ed. New York: Taylor and Francis, pp. 319-360.

Cattand, P. & De Raadt, P. 1991. Laboratory diagnosis of trypanosomiasis. *Clinics in Laboratory Medicine*, 11, (4) 899-908

Cattand, P., Jannin, J., & Lucas, P. 2001. Sleeping sickness surveillance: an essential step towards elimination. *Tropical Medicine & International Health*, 6, (5) 348-361

Ceva Sante Animal, University of Edinburgh, University of Makerere, & Industri Kapital. Stamp Out Sleeping Sickness in Uganda. 2006.

Christensen, O.S. & Ribeiro Jr, P.J. 2002. geoRglm: A package for generalised linear spatial models. *R News*, 2, (2) 26-28

Clements, A.C.A., Brooker, S., Nyandindi, U., Fenwick, A., & Blair, L. 2008. Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. *International Journal for Parasitology*, 38, (3-4) 401-415

Clements, A.C.A., Lwambo, N.J.S., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi, S., Webster, J.P., Fenwick, A., & Brooker, S. 2006a. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Tropical Medicine & International Health*, 11, (4) 490-503

Clements, A.C.A., Moyeed, R., & Brooker, S. 2006b. Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. *Parasitology*, 133, 711-719

Clements, A.C.A., Moyeed, R., & Brooker, S. 2006c. Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. *Parasitology*, 133, 711-719

Cook, J.H. 1901. Notes on cases of "sleeping sickness" occurring in the Uganda Protectorate. *Journal of Tropical Medicine*, 4, 236-239

Courtin, F., Jamonneau, V., Oke, E., Coulibaly, B., Oswald, Y., Dupont, S., Cuny, G., Doumenge, J.P., & Solano, P. 2005. Towards understanding the presence/absence of human African trypanosomosis in a focus of Cote d'Ivoire: a spatial analysis of the pathogenic system. *International Journal of Health Geographics*, 4, (1) 27

Craig, M.H., Sharp, B.L., Mabaso, M.L.H., & Kleinschmidt, I. 2007. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. *International Journal of Health Geographics*, 6,

Craig, P.S., Giraudoux, P., Shi, D., Bartholomot, B., Barnish, G., Delattre, P., Quere, J.P., Harraga, S., Bao, G., Wang, Y., Lu, F., Ito, A., & Vuitton, D.A. 2000. An epidemiological and ecological study of human alveolar echinococcosis transmission in south Gansu, China. *Acta Tropica*, 77, (2) 167-177

Curran, P.J., Atkinson, P.M., Foody, G.M., & Milton, E.J. 2000. Linking remote sensing, land cover and disease. *Advances in Parasitology*, 47, 37-80

Danson, F.M., Craig, P.S., Man, W., Shi, D.H., & Giraudoux, P. 2004. Landscape dynamics and risk modeling of human alveolar echinococcosis. *Photogrammetric Engineering and Remote Sensing*, 70, (3) 359-366

Danson, F.M., Graham, A.J., Pleydell, D.R.J., Campos-Ponce, M., Giraudoux, P., & Craig, P.S. 2003. Multi-scale spatial analysis of human alveolar echinococcosis risk in China. *Parasitology*, 127, S133-S141

Dash, P., Gottsche, F.M., Olesen, F.S., & Fischer, H. 2002. Land surface temperature and emissivity estimation from passive sensor data: theory and practice-current trends. *International Journal of Remote Sensing*, 23, (13) 2563-2594

De La Rocque, S., Michel, V., Plazanet, D., & Pin, R. 2004. Remote sensing and epidemiology: examples of applications for two vector-borne diseases. *Comparative Immunology Microbiology and Infectious Diseases*, 27, (5) 331-341

De Oliveira, V., Kedem, B., & Short, D.A. 1997. Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92, (440) 1422-1433

Defence Meteorological Satellite Program 2004, *Nighttime lights of the world*, National Geophysical Data Centre, Boulder, USA.

Department of Communicable Disease Surveillance and Response 2000, *WHO report on Global Surveillance of Epidemic-Prone Infectious Diseases*, World Health Organization.

Diggle, P., Moyeed, R., Rowlingson, B., & Thomson, M. 2002. Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 51, 493-506

Diggle, P. J. 2000, "Disease mapping and it's relationship to cluster detection," *In Spatial Epidemiology Methods and Applications*, P. Elliott et al., eds., Oxford: Oxford University Press, pp. 87-103.

Diggle, P.J. & Ribeiro Jr, P.J. 2007. *Model-based Geostatistics* New York, Springer.

Diggle, P.J., Tawn, J.A., & Moyeed, R.A. 1998. Model-based geostatistics. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 47, 299-326

Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J.H., Boussinesq, M., & Molyneux, D.H. 2007. Spatial modelling and the prediction of *Loa loa* risk: Decision making under uncertainty. *Annals of Tropical Medicine and Parasitology*, 101, (6) 499-509

Dister, S.W., Fish, D., Bros, S.M., Frank, D.H., & Wood, B.L. 1997. Landscape characterization of peridomestic risk for Lyme disease using satellite imagery. *American Journal of Tropical Medicine and Hygiene*, 57, (6) 687-692

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., & Worley, B.A. 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, 66, (7) 849-857

Downing, A., Forman, D., Gilthorpe, M.S., Edwards, K.L., & Manda, S.O.M. 2008. Joint disease mapping using six cancers in the Yorkshire region of England. *International Journal of Health Geographics*, 7,

Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. 2000, "Spatial epidemiology: methods and applications," *In Spatial Epidemiology: Methods and Applications*, P. Elliott et al., eds., Oxford: Oxford University Press, pp. 3-14.

Elliott, P. & Wartenberg, D. 2004. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112, (9) 998-1006

Fèvre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C., & Woolhouse, M.E.J. 2001a. The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda. *Lancet*, 358, (9282) 625-628

Fèvre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C., & Woolhouse, M.E.J. 2001b. The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda. *Lancet*, 358, (9282) 625-628

Fèvre, E.M., Coleman, P.G., Welburn, S.C., & Maudlin, I. 2004. Reanalyzing the 1900-1920 sleeping sickness epidemic in Uganda. *Emerg.Infect.Dis.*, 10, (4) 567-573

Fèvre, E.M., Picozzi, K., Fyfe, J., Waiswa, C., Odiit, M., Coleman, P.G., & Welburn, S.C. 2005. A burgeoning epidemic of sleeping sickness in Uganda. *Lancet*, 366, (9487) 745-747

Fèvre, E.M., von Wissmann, B., Welburn, S.C., & Lutumba, P. 2008. The burden of human African trypanosomiasis. *PLoS Neglected Tropical Diseases*, 2, (12)

Flanders, D., Hall-Beyer, M., & Pereverzoff, J. 2003. Preliminary evaluation of eCognition object-based software for cut block delineation and feature extraction. *Canadian Journal of Remote Sensing*, 29, (4) 441-452

Food and Agricultural Organization of the United Nations 1982, *Ecology and behaviour of tsetse*, Food and Agricultural Organization of the United Nations, Rome.

Forde, R.M. 1902. The discovery of the human trypanosoma. *British Medical Journal*, 2, 1741

Forest Department 2002, *National Biomass Study*, Ministry of Water Lands and Environment, Kampala.

Fountain Publishers 2007. *Uganda districts information handbook*, Expanded 2007-2008 ed. Kampala, Fountain Publishers.

Fournet, F., Kone, A., Traore, S., & Hervouet, J.P. 1999. Heterogeneity in the risk of sleeping sickness in coffee and cocoa commercial plantations in Ivory Coast. *Medical and Veterinary Entomology*, 13, (3) 333-335

Fournet, F., Traore, S., Prost, A., Cadot, E., & Hervouet, J.P. 2000. Impact of the development of agricultural land on the transmission of sleeping sickness in Daloa, Cote d'Ivoire. *Annals of Tropical Medicine and Parasitology*, 94, (2) 113-121

Fox, J. 1997, "What is regression analysis?," *In Applied Regression Analysis, Linear Models, and Related Methods*, Thousand Oaks: Sage Publications Inc, pp. 15-32.

Gatrell, A.C., Bailey, T.C., Diggle, P.J., & Rowlingson, B.S. 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21, (1) 256-274

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2004, "Posterior simulation," *In Bayesian Data Analysis*, Second ed. Boca Raton, FL: Chapman and Hall/CRC, pp. 283-310.

Gemade, E.I.I., Jiya, J.Y., Nwoke, B.E.B., Ogunba, E.O., Edeghere, H., Akoh, J.I., & Omojola, A. 1998. Human onchocerciasis: current assessment of the disease burden in Nigeria by rapid epidemiological mapping. *Annals of Tropical Medicine and Parasitology*, 92, S79-S83

Gilbert, M., Jenner, C., Pender, J., Rogers, D. J., Slingenbergh, J. H. W., & Wint, W. The development and use of the Programme Against African Trypanosomiasis Information System (PAATIS), *In The Jubilee Meeting of the International Council for Trypanosomiasis Research and Control*.

Glass, G.E., Schwartz, B.S., Morgan, J.M., Johnson, D.T., Noy, P.M., & Israel, E. 1995. Environmental risk factors for lyme disease identified with geographic information systems. *American Journal of Public Health*, 85, (7) 944-948

Godfrey, D.G. & Fairbairn, H. 1958. Sections cut through a chancre developing in a human volunteer previously exposed to the bite of an infected tsetse. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 52, 21-22

Goetz, S.J., Prince, S.D., & Small, J. 2000. Advances in satellite remote sensing of environmental variables for epidemiological applications. *Advances in Parasitology*, 47, 289-307

Graham, A.J., Atkinson, P.M., & Danson, F.M. 2004. Spatial analysis for epidemiology. *Acta Tropica*, 91, (3) 219-225

Greenwood, B.M. & Whittle, H.C. 1980. The pathogenesis of sleeping sickness. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 74, (6) 716-725

Guisan, A., Edwards, T.C., & Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157, (2-3) 89-100

Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135, (2-3) 147-186

Guo, Q.H., Kelly, M., Gong, P., & Liu, D.S. 2007. An object-based classification approach in mapping tree mortality using high spatial resolution imagery. *GIScience and Remote Sensing*, 44, (1) 24-47

Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S., Smith, D.L., Moyeed, R.A., & Snow, R.W. 2009. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine*, 6, (3)

Hay, S.I., Packer, M.J., & Rogers, D.J. 1997. The impact of remote sensing on the study and control of invertebrate intermediate hosts and vectors for disease. *International Journal of Remote Sensing*, 18, (14) 2899-2930

Hay, S.I., Tatem, A.J., Graham, A.J., Goetz, S.J., & Rogers, D.J. 2006. Global environmental data for mapping infectious disease distribution. *Advances in Parasitology*, 62, 37-77

Hay, S.I., Tucker, C.J., Rogers, D.J., & Packer, M.J. 1996. Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. *Annals of Tropical Medicine and Parasitology*, 90, (1) 1-19

Held, L., Natario, I., Fenton, S.E., Rue, H., & Becker, N. 2005. Towards joint disease mapping. *Statistical Methods in Medical Research*, 14, (1) 61-82

Hendrickx, G., De La Rocque, S., Reid, R., & Wint, W. 2001a. Spatial trypanosomosis management: from data-layers to decision making. *Trends in Parasitology*, 17, (1) 35-41

Hendrickx, G., Napala, A., Dao, B., Batawui, D., De Deken, R., Vermeilen, A., & Slingenbergh, J.H.W. 1999a. A systematic approach to area-wide tsetse distribution and abundance maps. *Bulletin of Entomological Research*, 89, (3) 231-244

Hendrickx, G., Napala, A., Dao, B., Batawui, K., Bastiaensen, P., De Deken, R., Vermeilen, A., Vercruysse, J., & Slingenbergh, J.H.W. 1999b. The area-wide epidemiology of bovine trypanosomosis and its impact on mixed farming in subhumid West Africa; a case study in Togo. *Veterinary Parasitology*, 84, (1-2) 13-31

Hendrickx, G., Napala, A., Slingenbergh, J.H.W., De Deken, R., & Rogers, D.J. 2001b. A contribution towards simplifying area-wide tsetse surveys using medium resolution meteorological satellite data. *Bulletin of Entomological Research*, 91, (5) 333-346

Hendrickx, G., Napala, A., Slingenbergh, J.H.W., De Deken, R., Vercruysse, J., & Rogers, D.J. 2000. The spatial pattern of trypanosomosis prevalence predicted with the aid of satellite imagery. *Parasitology*, 120, 121-134

Herbreteau, V., Salem, G., Souris, M., Hugot, J.P., & Gonzalez, J.P. 2007. Thirty years of use and improvement of remote sensing, applied to epidemiology: From early promises to lasting frustration. *Health & Place*, 13, (2) 400-403

Hide, G., Tait, A., Maudlin, I., & Welburn, S.C. 1996. The origins, dynamics and generation of *Trypanosoma brucei rhodesiense* epidemics in East Africa. *Parasitology Today*, 12, (2) 50-55

Hosmer, D.W. & Lemeshow, S. 1989a. *Applied Logistic Regression* New York, John Wiley.

Hosmer, D. W. & Lemeshow, S. 1989b, "Logistic regression for matched case-control studies," *In Applied Logistic Regression*, New York: John Wiley, pp. 223-259.

Hutchinson, O.C., Fèvre, E.M., Carrington, M., & Welburn, S.C. 2003. Lessons learned from the emergence of a new *Trypanosoma brucei rhodesiense* sleeping sickness focus in Uganda. *Lancet Infectious Diseases*, 3, (1) 42-45

International Lake Environment Committee. World Lakes Database: Lake Kyoga. 2008. Shiga, Japan, International Lake Environment Committee. 8-10-2009.

Jacquin, A., Misakova, L., & Gay, M. 2008. A hybrid object-based classification approach for mapping urban sprawl in periurban environment. *Landscape and Urban Planning*, 84, (2) 152-165

Jensen, J. R. 2005a, "Remote sensing and digital image processing," *In Introductory Digital Image Processing. A Remote Sensing Perspective*, Third ed. K. C. Clarke, ed., Upper Saddle River: Pearson Prentice Hall, pp. 1-34.

Jensen, J. R. 2005b, "Thematic information extraction: pattern recognition," *In Introductory Digital Image Processing. A Remote Sensing Perspective*, Third ed. K. C. Clarke, ed., Upper Saddle River: Pearson Prentice Hall, pp. 337-406.

Jensen, J. R. 2007a, "Electromagnetic radiation principles," *In Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd ed. K. C. Clarke, ed., Upper Saddle River, NJ: Pearson Prentice Hall, pp. 37-60.

Jensen, J. R. 2007b, "Multispectral remote sensing systems," *In Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd ed. K. C. Clarke, ed., Upper Saddle River, NJ: Pearson Prentice Hall, pp. 193-248.

Jobin, B., Labrecque, S., Grenier, M., & Falardeau, G. 2008. Object-based classification as an alternative approach to the traditional pixel-based classification to identify potential habitat of the Grasshopper Sparrow. *Environmental Management*, 41, (1) 20-31

Kabasa, J.D. 2007. Public-private partnership works to stamp out sleeping sickness in Uganda. *Trends in Parasitology*, 23, (5) 191-192

Kabatereine, N.B., Brooker, S., Tukahebwa, E.M., Kazibwe, F., & Onapa, A.W. 2004. Epidemiology and geography of *Schistosoma mansoni* in Uganda: implications for planning control. *Tropical Medicine & International Health*, 9, (3) 372-380

Kazembe, L., Kleinschmidt, I., Holtz, T., & Sharp, B. 2006. Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. *International Journal of Health Geographics*, 5, (1) 41

Kitron, U. 1998. Landscape ecology and epidemiology of vector-borne diseases: Tools for spatial analysis. *Journal of Medical Entomology*, 35, (4) 435-445

Kitron, U. 2000. Risk maps: Transmission and burden of vector borne diseases. *Parasitology Today*, 16, (8) 324-325

Kitron, U., Otieno, L.H., Hungerford, L.L., Odulaja, A., Brigham, W.U., Okello, O.O., Joselyn, M., MohamedAhmed, M.M., & Cook, E. 1996. Spatial analysis of the distribution of tsetse flies in the Lambwe Valley, Kenya, using Landsat TM satellite imagery and GIS. *Journal of Animal Ecology*, 65, (3) 371-380

Klugman, J., Neyapti, B., & Stewart, F. 1999. *Conflict and Growth in Africa. Volume 2: Kenya, Tanzania and Uganda* Paris, Development Centre of the Organisation for Economic Co-operation and Development.

Koch, D.E., Mohler, R.L., & Goodin, D.G. 2007. Stratifying land use/land cover for spatial analysis of disease ecology and risk: an example using object-based classification techniques. *Geospatial Health*, 2, (1) 15-28

Koerner, T., Deraadt, P., & Maudlin, I. 1995. The 1901 Uganda sleeping sickness epidemic revisited - a case of mistaken identity. *Parasitology Today*, 11, (8) 303-306

Kulldorff, M. & Nagarwalla, N. 1995. Spatial disease clusters - detection and inference. *Statistics in Medicine*, 14, (8) 799-810

Lachenbruch, P.A. & Goldstein, M. 1979. Discriminant analysis. *Biometrics*, 35, (1) 69-85

Lancien, J., Muguwa, J., Lannes, C., & Bouvier, J.B. 1990. Tsetse and human trypanosomiasis challenge in south eastern Uganda. *Insect Science and Its Application*, 11, (3) 411-416

Langford, I.H., Leyland, A.H., Rasbash, J., & Goldstein, H. 1999. Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, (2) 253-268

Langlands, B. W. 1967, *The sleeping sickness epidemic of Uganda, 1900 - 1920: a study in historical geography. Occasional Paper No. 1.*, Makerere University College, Kampala.

Leak, S.G.A. 1999. *Tsetse Biology and Ecology* Guildford, CABI Publishing.

Legendre, P. 1993. Spatial autocorrelation - trouble or new paradigm. *Ecology*, 74, (6) 1659-1673

Lehmann, A., Overton, J.M., & Austin, M.P. 2002. Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, 11, (12) 2085-2092

Levin, S.A. 1992. The problem of pattern and scale in ecology. *Ecology*, 73, (6) 1943-1967

Lilienfeld, D. E. & Stolley, P. D. 1994, "Introduction to epidemiology," *In Foundations of Epidemiology*, 3rd ed. New York: Oxford University Press, pp. 1-56.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. 2004a, "Concepts and foundations of remote sensing," *In Remote Sensing and Image Interpretation*, 5th ed. Hoboken, NJ: John Wiley and Sons, pp. 1-57.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. 2004b, "Digital image processing," *In Remote Sensing and Image Interpretation*, Fifth ed. Hoboken: John Wiley and Sons, pp. 491-637.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. 2004c, "Earth resource satellites operating in the optical spectrum," *In Remote Sensing and Image Interpretation*, Fifth ed. Hoboken: John Wiley and Sons, pp. 397-490.

Lockwood, J.L., Cassey, P., & Blackburn, T. 2005. The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, 20, (5) 223-228

Lynch, S. M. 2007, "Evaluating Markov chain Monte Carlo (MCMC) algorithms and model fit," *In Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, New York: Springer, pp. 131-163.

Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M., & Bazzaz, F.A. 2000. Biotic invasions: Causes, epidemiology, global consequences, and control. *Ecological Applications*, 10, (3) 689-710

MacKichan, I.W. 1944. Rhodesian sleeping sickness in Eastern Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 38, (1) 49-60

Mahama, C.I., Kone, A., De La Rocque, S., De Deken, R., & Geerts, S. 2005. Prediction of the distribution of *Glossina tachinoides* (Diptera : Glossinidae) in the Volta basin of northern Ghana. *Bulletin of Entomological Research*, 95, (1) 63-67

Marshall, R.J. 1991. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 154, 421-441

Mathieu, R., Aryal, J., & Chong, A.K. 2007. Object-based classification of ikonos imagery for mapping large-scale vegetation communities in urban areas. *Sensors*, 7, (11) 2860-2880

Matovu, F.S. 1982. Rhodesian sleeping sickness in South-Eastern Uganda: (the present problems). *East African Medical Journal*, 59, (6) 390-393

Maupin, G.O., Fish, D., Zultowsky, J., Campos, E.G., & Piesman, J. 1991. Landscape ecology of Lyme disease in a residential area of Westchester County, New York. *American Journal of Epidemiology*, 133, (11) 1105-1113

Mbulamberi, D.B. 1989. Possible causes leading to an epidemic outbreak of sleeping sickness - facts and hypotheses. *Annales de la Societe Belge de Medecine Tropicale*, 69, 173-179

McCullagh, P. N. J. A. 1989, "An outline of generalized linear models," *In Generalized Linear Models*, 2 ed. London: Chapman and Hall, pp. 21-47.

Metz, C.E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, (4) 283-298

Mladenoff, D.J., Sickley, T.A., & Wydeven, A.P. 1999. Predicting gray wolf landscape recolonization: Logistic regression models vs. new field data. *Ecological Applications*, 9, (1) 37-44

Moisen, G.G., Edwards, T.C., & Osborne, P.E. 2006. Further advances in predicting species distributions. *Ecological Modelling*, 199, (2) 129-131

Moore, A. & Richer, M. 2001. Re-emergence of epidemic sleeping sickness in southern Sudan. *Tropical Medicine & International Health*, 6, (5) 342-347

Moore, A., Richer, M., Enrile, M., Losio, E., Roberts, J., & Levy, D. 1999. Resurgence of sleeping sickness in Tambura County, Sudan. *American Journal of Tropical Medicine and Hygiene*, 61, (2) 315-318

Morris, K.R.S. 1962. The epidemiology of sleeping sickness in east Africa. V. Epidemics on the Albert Nile. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 56, (4) 316-338

Muller, G., Grebaut, P., & Gouteux, J.P. 2004. An agent-based model of sleeping sickness: simulation trials of a forest focus in southern Cameroon. *Comptes Rendus Biologies*, 327, (1) 1-11

NASA Landsat Programme. Landsat ETM+ Scene. p171r059[L1G]. 2004. Sioux Falls, USA, USGS. 27-11-2001.

National Aeronautics and Space Administration & Japan Aerospace Exploration Agency. Tropical Rainfall Measuring Mission product 3B43 (V6). 2001. National Aeronautics and Space Administration.

Newsom, S.W.B. 2006. Pioneers in infection control: John Snow, Henry Whitehead, the Broad Street pump, and the beginnings of geographical epidemiology. *Journal of Hospital Infection*, 64, (3) 210-216

Noma, M., Nwoke, B.E.B., Nutall, I., Tambala, P.A., Enyong, P., Namsenmo, A., Remme, J., Amazigo, U.V., Kale, O.O., & Seketeli, A. 2002. Rapid epidemiological mapping of onchocerciasis (REMO): its application by the African Programme for Onchocerciasis Control (APOC). *Annals of Tropical Medicine and Parasitology*, 96, 29-39

Noor, A.M., Clements, A.C.A., Gething, P.W., Moloney, G., Borle, M., Shewchuk, T., Hay, S.I., & Snow, R.W. 2008a. Spatial prediction of *Plasmodium falciparum* prevalence in Somalia. *Malaria Journal*, 7, (159)

Noor, A., Alegana, V., Gething, P., Tatem, A., & Snow, R. 2008b. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics*, 6, (1) 5

Oak Ridge National Laboratory. Landscan 2006 Global Population Dataset. 2006. Oak Ridge National Laboratory. 23-4-2008.

Odiit, M., Bessell, P.R., Fèvre, E.M., Robinson, T., Kinoti, J., Coleman, P.G., Welburn, S.C., McDermott, J., & Woolhouse, M.E.J. 2006. Using remote sensing and geographic information systems to identify villages at high risk for rhodesiense sleeping sickness in Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100, (4) 354-362

Odiit, M., Coleman, P.G., Liu, W.C., McDermott, J.J., Fèvre, E.M., Welburn, S.C., & Woolhouse, M.E.J. 2005. Quantifying the level of under-detection of *Trypanosoma brucei rhodesiense* sleeping sickness cases. *Tropical Medicine & International Health*, 10, (9) 840-849

Odiit, M., Coleman, P.G., McDermott, J.J., Fèvre, E.M., Welburn, S.C., & Woolhouse, M.E.J. 2004a. Spatial and temporal risk factors for the early detection of *Trypanosoma brucei rhodesiense* sleeping sickness patients in Tororo and Busia districts, Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98, (10) 569-576

Odiit, M., Kansiime, F., & Enyaru, J.C.K. 1997. Duration of symptoms and case fatality of sleeping sickness caused by *Trypanosoma brucei rhodesiense* in Tororo, Uganda. *East African Medical Journal*, 74, (12) 792-795

Odiit, M., Shaw, A., Welburn, S.C., Fèvre, E.M., Coleman, P.G., & McDermott, J.J. 2004b. Assessing the patterns of health-seeking behaviour and awareness among sleeping-sickness patients in eastern Uganda. *Annals of Tropical Medicine and Parasitology*, 98, (4) 339-348

Odulaja, A., Baumgartner, J., Mihok, S., & bu-Zinid, I.M. 2001. Spatial and temporal distribution of tsetse fly trap catches at Nguruman, southwest Kenya. *Bulletin of Entomological Research*, 91, (3) 213-220

Okia, M., Mbulamberi, D.B., & Demuynck, A. 1994. Risk factors assessment for *Trypanosoma brucei rhodesiense* sleeping sickness acquisition in SE Uganda - a case-control study. *Annales de la Societe Belge de Medecine Tropicale*, 74, (2) 105-112

Okiria, R. 1985. The prevalence of human trypanosomiasis in Uganda, 1970 to 1983. *East African Medical Journal*, 62, (11) 813-816

Onyango, R.J., Van Hoeve, K., & De Raadt, P. 1966. The epidemiology of *Trypanosoma rhodesiense* sleeping sickness in alego location, Central Nyanza, Kenya I. Evidence that cattle may act as reservoir hosts of trypanosomes infective to man. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 60, (2) 175-182

Ostfeld, R.S., Glass, G.E., & Keesing, F. 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in Ecology & Evolution*, 20, (6) 328-336

Palacios, M.B. & Steel, M.F.J. 2006. Non-gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101, (474) 604-618

Pavlovskii, E.N. 1966. *Natural nidality of transmissible diseases, with special reference to the landscape epidemiology of zooanthroponoses* Urbana, Illinois, University of Illinois Press.

Pearce, J. & Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133, (3) 225-245

Pepin, J. & Meda, H.A. 2001. The epidemiology and control of human African trypanosomiasis. *Advances in Parasitology*, 49, 71-132

Pepin, J. & Milord, F. 1994. The treatment of human African trypanosomiasis. *Advances in Parasitology*, 33, 1-47

Perez, L. & Dragicevic, S. 2009. An agent-based approach for modeling dynamics of contagious disease spread. *International Journal of Health Geographics*, 8,

Peterson, A.T. 2008. Biogeography of diseases: a framework for analysis. *Naturwissenschaften*, 95, (6) 483-491

Pfeiffer, D., Robinson, T., Stevenson, M., Stevens K., Rogers, D., & Clements, A. 2008a, "Spatial variation in risk," *In Spatial Analysis in Epidemiology*, Oxford: Oxford University Press, pp. 67-80.

Pfeiffer, D., Robinson, T., Stevenson, M., Stevens, K., Rogers, D., & Clements, A. 2008b, "Identifying factors associated with the spatial distribution of disease," *In Spatial Analysis in Epidemiology*, New York: Oxford University Press, pp. 81-109.

Picozzi, K., Fèvre, E.M., Odiit, M., Carrington, M., Eisler, M.C., Maudlin, I., & Welburn, S.C. 2005. Sleeping sickness in Uganda: a thin line between two fatal diseases. *British Medical Journal*, 331, (7527) 1238-1241

Pilz, J. & Spock, G. 2008. Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, 22, (5) 621-632

Pope, K.O., Rejmankova, E., Savage, H.M., Rredondo-Jimenez, J.I., Rodriguez, M.H., & Roberts, D.R. 1994. Remote sensing of tropical wetlands for malaria control in Chiapas, Mexico. *Ecological Applications*, 4, (1) 81-90

Press, S.J. & Wilson, S. 1978. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73, (364) 699-705

R Development Core Team. R: A language and environment for statistical computing. 2006. Vienna, Austria, R Foundation for Statistical Computing.

Reid, R.S., Kruska, R.L., Deichmann, U., Thornton, P.K., & Leak, S.G.A. 2000. Human population growth and the extinction of the tsetse fly. *Agriculture Ecosystems & Environment*, 77, (3) 227-236

Richter, R. 2009, *Atmospheric/topographic correction for satellite imagery: ATCOR-2/3 user guide, Version 7.0*, DLR German Aerospace Centre, Wessling, Germany.

Robays, J., Kadima, A.E., Lutumba, P., Bilenge, C.M.M., Mesu, V.K.B.K., De Deken, R., Makabuza, J., Deguerry, M., Van der Stuyft, P., & Boelaert, M. 2004. Human African trypanosomiasis amongst urban residents in Kinshasa: a case-control study. *Tropical Medicine & International Health*, 9, (8) 869-875

Robinson, T., Rogers, D., & Williams, B. 1997a. Mapping tsetse habitat suitability in the common fly belt of Southern Africa using multivariate analysis of climate and remotely sensed vegetation data. *Medical and Veterinary Entomology*, 11, (3) 235-245

Robinson, T., Rogers, D., & Williams, B. 1997b. Univariate analysis of tsetse habitat in the common fly belt of Southern Africa using climate and remotely sensed vegetation data. *Medical and Veterinary Entomology*, 11, (3) 223-234

Robinson, T.P. 1998. Geographic information systems and the selection of priority areas for control of tsetse-transmitted trypanosomiasis in Africa. *Parasitology Today*, 14, (11) 457-461

Robinson, T.P., Harris, R.S., Hopkins, J.S., & Williams, B.G. 2002. An example of decision support for trypanosomiasis control using a geographical information system in eastern Zambia. *International Journal of Geographical Information Science*, 16, (4) 345-360

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

Rogers, D.J. 2000. Satellites, space, time and the African trypanosomiases. *Advances in Parasitology*, 47, 129-171

Rogers, D.J., Hay, S.I., & Packer, M.J. 1996. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology*, 90, (3) 225-241

Rogers, D.J. & Randolph, S.E. 1986. Distribution and abundance of tsetse flies (*Glossina* spp). *Journal of Animal Ecology*, 55, (3) 1007-1025

Rogers, D.J. & Randolph, S.E. 1991. Mortality rates and population density of tsetse flies correlated with satellite imagery. *Nature*, 351, (6329) 739-741

Rogers, D.J. & Randolph, S.E. 1993. Distribution of Tsetse and Ticks in Africa - Past, Present and Future. *Parasitology Today*, 9, (7) 266-271

Rogers, D.J. & Williams, B.G. 1993. Monitoring trypanosomiasis in space and time. *Parasitology*, 106, S77-S92

Rushton, G. 2003. Public health, GIS, and spatial analytic tools. *Annual Review of Public Health*, 24, 43-56

Savtchenko, A., Ouzounov, D., Gopalan, A., Yuan, D., Nickless, D., & Ostrenga, D. 2003. MODIS data from Terra and Aqua satellites. *unknown*

Schreiber, S.J. & Lloyd-Smith, J.O. 2009. Invasion dynamics in spatially heterogeneous environments. *American Naturalist*, 174, (4) 490-505

Sciarretta, A., Girma, M., Tikubet, G., Belayehun, L., Ballo, S., & Baumgartner, J. 2005. Development of an adaptive tsetse population management scheme for the Luke community, Ethiopia. *Journal of Medical Entomology*, 42, (6) 1006-1019

Stephens, J.W.W. & Fantham, H.B. 1910. On the peculiar morphology of a trypanosome from a case of sleeping sickness and the possibility of its being a new species (*T. rhodesiense*). *Annals of Tropical Medicine and Parasitology*, 4, (343) 350

Stevenson, M.A., Benard, H., Bolger, P., & Morris, R.S. 2005. Spatial epidemiology of the Asian honey bee mite (Varroa destructor) in the North Island of New Zealand. *Preventive Veterinary Medicine*, 71, (3-4) 241-252

Symeonakis, E., Robinson, T., & Drake, N. 2007. GIS and multiple-criteria evaluation for the optimisation of tsetse fly eradication programmes. *Environmental Monitoring and Assessment*, 124, (1-3) 89-103

Thomson, M.C. & Connor, S.J. 2000. Environmental information systems for the control of arthropod vectors of disease. *Medical and Veterinary Entomology*, 14, (3) 227-244

Thomson, M.C., Connor, S.J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., & Greenwood, B. 1999. Predicting malaria infection in Gambian

children from satellite data and bed net use surveys: The importance of spatial correlation in the interpretation of results. *American Journal of Tropical Medicine and Hygiene*, 61, (1) 2-8

Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8, 127-150

Uganda Bureau of Statistics 2008, *Uganda National Census, 2002*, Uganda Bureau of Statistics, Kampala.

US Geological Survey 2006, *"Finished" 3-arc second SRTM Format Documentation*, USGS.

US Geological survey. Global Digital Elevation Model. 2006. U. S. Geological Survey. 8-11-2007.

Waiswa, C., Olaho-Mukani, W., & Katunguka-Rwakishaya, E. 2003. Domestic animals as reservoirs for sleeping sickness in three endemic foci in south-eastern Uganda. *Annals of Tropical Medicine and Parasitology*, 97, (2) 149-155

Waiswa, C., Picozzi, K., Katunguka-Rwakishaya, E., Olaho-Mukani, W., Musoke, R.A., & Welburn, S.C. 2006. *Glossina fuscipes fuscipes* in the trypanosomiasis endemic areas of south eastern Uganda: Apparent density, trypanosome infection rates and host feeding preferences. *Acta Tropica*, 99, (1) 23-29

Welburn, S.C., Coleman, P.G., Maudlin, I., Fèvre, E.M., Odiit, M., & Eisler, M.C. 2006. Crisis, what crisis? Control of Rhodesian sleeping sickness. *Trends in Parasitology*, 22, (3) 123-128

Welburn, S.C., Fèvre, E.M., Coleman, P.G., Odiit, M., & Maudlin, I. 2001a. Sleeping sickness: a tale of two disease. *Trends in Parasitology*, 17, (1) 19-24

Welburn, S.C., Picozzi, K., Fèvre, E.M., Coleman, P.G., Odiit, M., Carrington, M., & Maudlin, I. 2001b. Identification of human-infective trypanosomes in animal reservoir of sleeping sickness in Uganda by means of serum-resistance-associated (SRA) gene. *Lancet*, 358, (9298) 2017-2019

Wendo, C. 2002. Uganda revises cattle treatment to protect humans from sleeping sickness. *Lancet*, 359, (9302) 239

Wint, W. & Rogers, D. J. 2000, *Predicted distributions of tsetse in Africa*, Food and Agriculture Organisation of the United Nations, Rome.

Woodward, M. 2004, "Case-control studies," *In Epidemiology Study Design and Data Analysis*, Second ed. Boca Raton: CRC Press, pp. 273-334.

World Health Organization 2006a. Human African Trypanosomiasis (sleeping sickness): epidemiological update. *Weekly Epidemiological Record*, 81, (8) 71-80

World Health Organization 2006b, *Human African Trypanosomiasis disease distributions and population at risk. Draft document under revision*, WHO, Geneva.

World Health Organization 2008, *The global burden of disease: 2004 update*, WHO, Geneva.

Wyatt, G.B., Boatin, B.A., & Wurapa, F.K. 1985. Risk factors associated with the acquisition of sleeping sickness in Northeast Zambia - a case-control study. *Annals of Tropical Medicine and Parasitology*, 79, (4) 385-392

Yu, P., Habtemariam, T., Oryang, D., Obasa, M., Nganwa, D., & Robnett, V. 1995. Integration of temporal and spatial models for examining the epidemiology of African trypansomiasis. *Preventive Veterinary Medicine*, 24, (2) 83-95

Zhang, H. 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99, (465) 250-261

Zoller, T., Fèvre, E.M., Welburn, S.C., Odiit, M., & Coleman, P.G. 2008. Analysis of risk factors for *T. brucei rhodesiense* sleeping sickness within villages in south-east Uganda. *BMC Infectious Diseases*, 8, (1) 88

Appendix A: Full structure of trypanosomiasis database

	Prefix	Meaning
Objects	tbl	Table
	tlkp	Lookup table
Fields	anm	Auto number
	num	Number
	str	Text
	dtm	Date/Time
	mem	Memo
Keys	PK	Primary key
	FK	Foreign key

Table A1: Naming conventions used in trypanosomiasis database

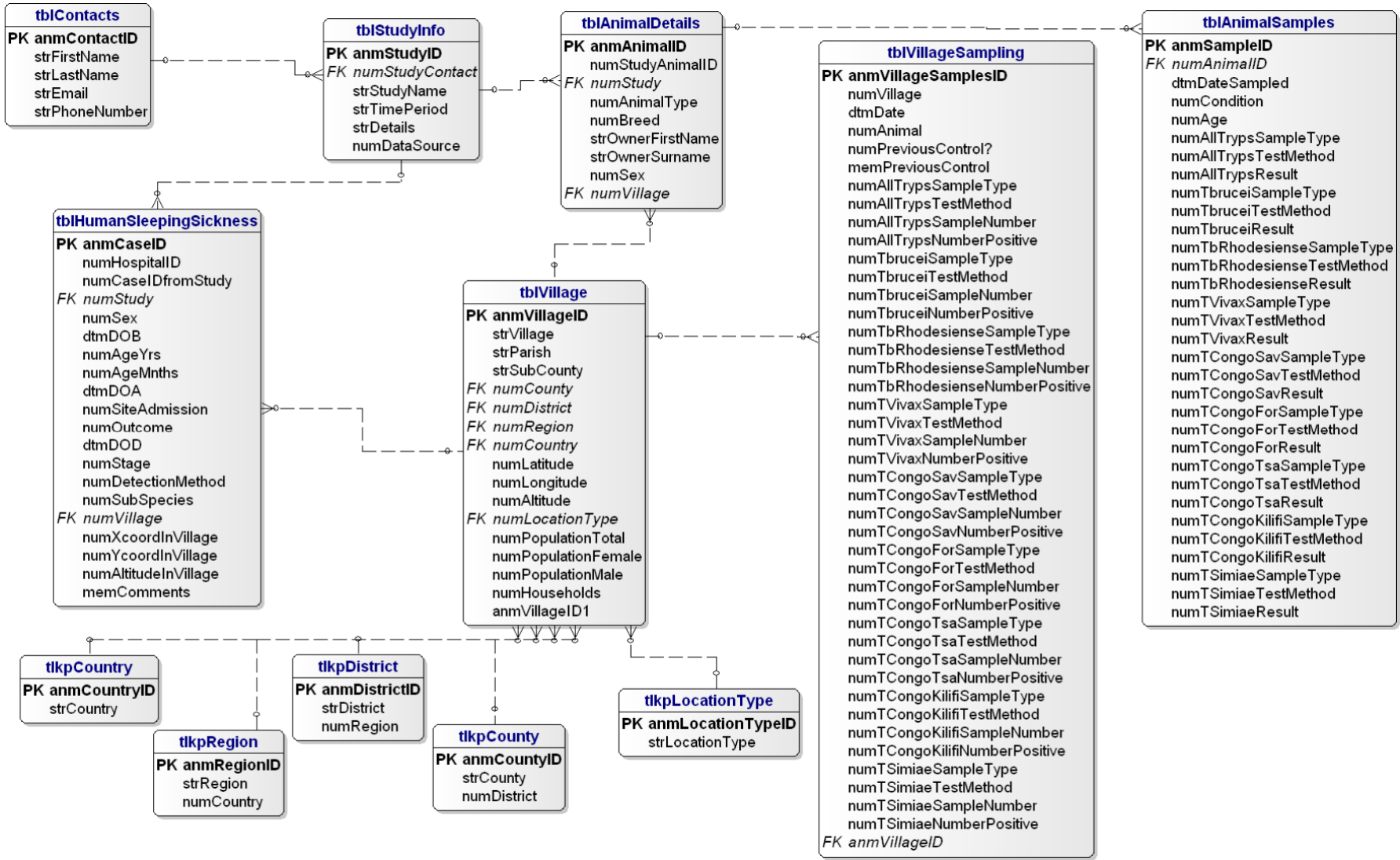

tblContacts
PK anmContactID
strFirstName
strLastName
strEmail
strPhoneNumber
tblStudyInfo
PK anmStudyID
FK numStudyContact
strStudyName
strTimePeriod
strDetails
numDataSource
tblAnimalDetails
PK anmAnimalID
numStudyAnimalID
FK numStudy
numAnimalType
numBreed
strOwnerFirstName
strOwnerSurname
numSex
FK numVillage
tblAnimalSamples
PK anmSampleID
FK numAnimalID
dtmDateSampled
numCondition
numAge
numAllTrypsSampleType
numAllTrypsTestMethod
numAllTrypsResult
numTbruceiSampleType
numTbruceiTestMethod
numTbruceiResult
numTbRhodesienseSampleType
numTbRhodesienseTestMethod
numTbRhodesienseResult
numTVivaxSampleType
numTVivaxTestMethod
numTVivaxResult
numTCongoSavSampleType
numTCongoSavTestMethod
numTCongoSavResult
numTCongoForSampleType
numTCongoForTestMethod
numTCongoForResult
numTCongoTsaSampleType
numTCongoTsaTestMethod
numTCongoTsaResult
numTCongoKilifiSampleType
numTCongoKilifiTestMethod
numTCongoKilifiResult
numTSimiaeSampleType
numTSimiaeTestMethod
numTSimiaeResult
tblVillageSampling
PK anmVillageSamplesID
numVillage
dtmDate
numAnimal
numPreviousControl?
memPreviousControl
numAllTrypsSampleType
numAllTrypsTestMethod
numAllTrypsSampleNumber
numAllTrypsNumberPositive
numTbruceiSampleType
numTbruceiTestMethod
numTbruceiSampleNumber
numTbruceiNumberPositive
numTbRhodesienseSampleType
numTbRhodesienseTestMethod
numTbRhodesienseSampleNumber
numTbRhodesienseNumberPositive
numTVivaxSampleType
numTVivaxTestMethod
numTVivaxSampleNumber
numTVivaxNumberPositive
numTCongoSavSampleType
numTCongoSavTestMethod
numTCongoSavSampleNumber
numTCongoSavNumberPositive
numTCongoForSampleType
numTCongoForTestMethod
numTCongoForSampleNumber
numTCongoForNumberPositive
numTCongoTsaSampleType
numTCongoTsaTestMethod
numTCongoTsaSampleNumber
numTCongoTsaNumberPositive
numTCongoKilifiSampleType
numTCongoKilifiTestMethod
numTCongoKilifiSampleNumber
numTCongoKilifiNumberPositive
numTSimiaeSampleType
numTSimiaeTestMethod
numTSimiaeSampleNumber
numTSimiaeNumberPositive
FK anmVillageID
tblHumanSleepingSickness
PK anmCaseID
numHospitalID
numCaseIDfromStudy
FK numStudy
numSex
dtmDOB
numAgeYrs
numAgeMnths
dtmDOA
numSiteAdmission
numOutcome
dtmDOD
numStage
numDetectionMethod
numSubSpecies
FK numVillage
numXcoordInVillage
numYcoordInVillage
numAltitudeInVillage
memComments
tblVillage
PK anmVillageID
strVillage
strParish
strSubCounty
FK numCounty
FK numDistrict
FK numRegion
FK numCountry
numLatitude
numLongitude
numAltitude
FK numLocationType
numPopulationTotal
numPopulationFemale
numPopulationMale
numHouseholds
anmVillageID1
tlkpCountry
PK anmCountryID
strCountry
tlkpDistrict
PK anmDistrictID
strDistrict
numRegion
tlkpLocationType
PK anmLocationTypeID
strLocationType
tlkpRegion
PK anmRegionID
strRegion
numCountry
tlkpCounty
PK anmCountyID
strCounty
numDistrict

Figure A1 (previous page): Entity relationship diagram showing data tables and relationships in the trypanosomiasis database (some lookup tables are excluded, Table A3 shows the full list of lookup tables with field descriptions).

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
tblAnimalDetails	**Details of individual animals from which samples were taken**	anmAnimalID	Unique ID for animal	
		numStudyAnimalID	ID used for animal in the study of origin	
		numStudy	Study from which the data was taken	tblStudyInfo.anmStudyID
		numAnimalType	Type of animal	tlkpAnimalType.anmAnimalTypeID
		numBreed	Breed of animal, if recorded	tlkpBreed.anmBreedID
		strOwnerFirstName	First name of animals owner, if recorded	
		strOwnerSurname	Surname of animals owner, if recorded	
		numSex	Gender of the animal (male, female, unknown)	tlkpGender.anmGenderID
		numVillage	Village in which the animal's owner is resident	tblVillage.anmVillageID
tblAnimalSamples	**Details of samples taken from individual animals, and the results (May be multiple samples for each animal)**	anmSampleID	Unique ID for sample	
		numAnimalID	ID of animal	tblAnimalDetails.anmAnimalID
		dtmDateSampled	Date the sample was taken on	
		numCondition	Condition score for the animal (L-, L, L+, M-, M, M+, F-, F, F+, Unknown)	tlkpCondition.anmConditionID
		numAge	Age group of the animal (calf, juvenile, adult, unknown)	tlkpAge.anmAgeID
		numAllTrypsSampleType	Type of sample taken for testing for *Trypanosoma* spp (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
		numAllTrypsTestMethod	Test method used to detect *Trypanosoma* spp (general microscopy, thin film microscopy, thick film microscopy, buffy coat microscopy, ITS PCR, other)	tlkpAllTrypsTestMethod.anmAllTrypsMethodID
		numAllTrypsResult	Result of test for *Trypanosoma* spp (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTbruceiSampleType	Type of sample taken for *T. brucei* testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTbruceiTestMethod	Test method used to detect *T. brucei* (TBR primers, PLC, other, microscopy)	tlkpBruceiTestMethod.anmBruceiTestID
		numTbruceiResult	Result of test for *T. brucei* (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTbRhodesienseSampleType	Type of sample taken for *T. b. rhodesiense* testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTbRhodesienseTestMethod	Test method used to detect *T. b. rhodesiense* (SRA, other)	tlkpRhodesienseTestMethod.anmRhodesienseTestID
		numTbRhodesienseResult	Result of test for *T. b. rhodesiense* (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTVivaxSampleType	Type of sample taken for *T. vivax* testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTVivaxTestMethod	Test method used to detect *T.*	tlkpVivaxTestMethod.anmVivaxTestID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
			vivax (TVW A/B, TWJ, wet blood film)	
		numTVivaxResult	Result of test for *T. vivax* (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTCongoSavSampleType	Type of sample taken for *T. congolense* savannah testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoSavTestMethod	Test method used to detect *T. congolense* savannah (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoSavResult	Result of test for *T. congolense* savannah (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTCongoForSampleType	Type of sample taken for *T. congolense* forest testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoForTestMethod	Test method used to detect *T. congolense* forest (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoForResult	Result of test for *T. congolense* forest (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTCongoTsaSampleType	Type of sample taken for *T. congolense* tsavo testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
		numTCongoTsaTestMethod	Test method used to detect *T. congolense* tsavo (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoTsaResult	Result of test for *T. congolense* tsavo (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTCongoKilifiSampleType	Type of sample taken for *T. congolense* kilifi testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoKilifiTestMethod	Test method used to detect *T. congolense* kilifi (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoKilifiResult	Result of test for *T. congolense* kilifi (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
		numTSimiaeSampleType	Type of sample taken for *T. simiae* testing (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTSimiaeTestMethod	Test method used to detect *T. simiae* (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTSimiaeResult	Result of test for *T. simiae* (positive, negative, unknown, not tested)	tlkpTestResult.anmTestResultID
tblContacts	**Contact person for the study, with phone and email**	anmContactID	Unique ID for contact person	
		strFirstName	First name of contact person	
		strLastName	Last name of contact person	
		strEmail	Email address of contact	
		strPhoneNumber	Phone number of contact person	

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
tblHumanSleepingSickness	**Details of human sleeping sickness cases**	anmCaseID	Unique ID for case	
		numHospitalID	ID taken from the hospital record	
		numCaseIDfromStudy	Case ID from the original data set, if one is available	
		numStudy	Study during which the data was collected	tblStudyInfo.anmStudyID
		numSex	Gender of the patient (male, female, unknown)	tlkpGender.anmGenderID
		dtmDOB	Date of birth, if available	
		numAgeYrs	If no date of birth is available, the age of the patient, in years	
		numAgeMnths	If no date of birth is available and the patient is under 1, the age of the patient, in months	
		dtmDOA	Date of admission to the hospital	
		numSiteAdmission	Hospital site the patient was treated at	tlkpSite.anmSiteID
		numOutcome	Outcome of infection (improved, died, unknown)	tlkpOutcome.anmOutcomeID
		dtmDOD	Date of discharge or death	
		numStage	Stage of infection (early, late, not staged, unknown)	tlkpStage.anmStageID
		numDetectionMethod	Method of detection of case (active surveillance, passive surveillance, population screening, unknown)	tlkpDetectionMethod.anmDetectionMethodID
		numSubSpecies	Sub species of parasite (not necessarily lab confirmed - based on location) (*T. b. rhodesiense*, *T. b. gambiense*)	tlkpSubSpecies.anmSubSpID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
		numVillage	Village of origin of the patient	tblVillage.anmVillageID
		numXcoordInVillage	X coordinate of the patient's homestead within the village, if available (in WGS 84)	
		numYcoordInVillage	Y coordinate of the patient's homestead within the village, if available (in WGS 84)	
		numAltitudeInVillage	Altitude of the patients homestead within the village, if available	
		memComments	Any additional comments regarding the patient	
tblStudyInfo	**Basic information regarding the study**	anmStudyID	Unique ID for the study	
		numStudyContact	Contact person for the study	tblContacts.anmContactID
		strStudyName	Name of the study	
		strTimePeriod	Study time period	
		strDetails	Brief description of the study	
		numDataSource	Original source of the data (hospital records, population screening, cattle sampling – cattle market, cattle sampling – village, cattle sampling – homestead)	tlkpDataSource.anmDataSourceID
tblVillage	**Location information for each village included in the database**	anmVillageID	Unique ID for the village	
		strVillage	Village (or sampling site) name	
		strParish	Parish the village is in	
		strSubCounty	Sub county the village is in	
		numCounty	County the village is in	tlkpCounty.anmCountyID
		numDistrict	District the village is in	tlkpDistrict.anmDistrictID
		numRegion	Region the village is in	tlkpRegion.anmRegionID
		numCountry	Country the village is in	tlkpCountry.anmCountryID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
		numLatitude	X coordinate of the village (if available) in WGS84	
		numLongitude	Y coordinate of the village (if available) in WGS84	
		numAltitude	Altitude of the village (if available)	
		numLocationType	Type of location (village centre, cattle market, homestead, abbatoir, sampling site)	tlkpLocationType.anmLocationTypeID
		numPopulationTotal	Total population of the village (if known)	
		numPopulationFemale	Female population of the village (if known)	
		numPopulationMale	Male population of the village (if known)	
		numHouseholds	Number of households in the village (if known)	
tblVillageSampling	**Village level sampling results (May be multiple results for each village)**	anmVillageSamplesID	Unique ID for village sampling	
		numVillage	Village that the sampling was carried out in	tblVillage.anmVillageID
		dtmDate	Date the sampling was carried out	
		numAnimal	Animal species tested	tlkpAnimalType.anmAnimalTypeID
		memPreviousControl?	Details of any previous control measures used in the area	
		numAllTrypsSampleType	Type of sample taken to test for *Trypanosoma* spp (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numAllTrypsTestMethod	Method used to test for *Trypanosoma* spp (general microscopy, thin film	tlkpAllTrypsTestMethod.anmAllTrypsMethodID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
			microscopy, thick film microscopy, buffy coat microscopy, ITS PCR, other)	
		numAllTrypsSampleNumber	Number of samples tested for *Trypanosoma* spp	
		numAllTrypsNumberPositive	Number positive for *Trypanosoma* spp	
		numTbruceiSampleType	Type of sample taken to look for *T. brucei* (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTbruceiTestMethod	Method used to test for *T. brucei* (TBR primers, PLC, other, microscopy)	tlkpBruceiTestMethod.anmBruceiTestID
		numTbruceiSampleNumber	Number of samples tested for *T. brucei*	
		numTbruceiNumberPositive	Number positive for *T. brucei*	
		numTbRhodesienseSampleType	Type of samples taken to look for *T. b. rhodesiense* (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTbRhodesienseTestMethod	Method used to test for *T. b. rhodesiense* (SRA, other)	tlkpRhodesienseTestMethod.anmRhodesienseTestID
		numTbRhodesienseSampleNumber	Number of samples tested for *T. b. rhodesiense*	
		numTbRhodesienseNumberPositive	Number positive for *T. b. rhodesiense*	
		numTVivaxSampleType	Type of sample taken to look for *T. vivax* (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTVivaxTestMethod	Method used to test for *T. vivax*	tlkpVivaxTestMethod.anmVivaxTestID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
			(TVW A/B, TWJ, wet blood film)	
		numTVivaxSampleNumber	Number of samples tested for *T. vivax*	
		numTVivaxNumberPositive	Number positive for *T. vivax*	
		numTCongoSavSampleType	Type of sample taken to look for *T. congolense* Savannah (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoSavTestMethod	Method used to test for *T. congolense* Savannah (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoSavSampleNumber	Number of samples tested for *T. congolense* Savannah	
		numTCongoSavNumberPositive	Number positive for *T. congolense* Savannah	
		numTCongoForSampleType	Type of sample taken to look for *T. congolense* Forest (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoForTestMethod	Method used to test for *T. congolense* Forest (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoForSampleNumber	Number of samples tested for *T. congolense* Forest	
		numTCongoForNumberPositive	Number positive for *T. congolense* Forest	
		numTCongoTsaSampleType	Type of sample taken to look for *T. congolens*e Tsavo (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID

Table name	Table description	Field names	Field descriptions (and options where appropriate)	Links to
		numTCongoTsaTestMethod	Method used to test for *T. congolense* Tsavo (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoTsaSampleNumber	Number of samples tested for *T. congolense* Tsavo	
		numTCongoTsaNumberPositive	Number positive for *T. congolense* Tsavo	
		numTCongoKilifiSampleType	Type of sample taken to look for *T. congolense* Kilifi (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTCongoKilifiTestMethod	Method used to test for *T. congolense* Kilifi (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTCongoKilifiSampleNumber	Number of samples tested for *T. congolense* Kilifi	
		numTCongoKilifiNumberPositive	Number positive for *T. congolense* Kilifi	
		numTSimiaeSampleType	Type of sample taken to look for *T. simiae* (blood, CSF, DNA, FTA card, isocode cards, filter paper, none)	tlkpSampleMethod.anmSampleTypeID
		numTSimiaeTestMethod	Method used to test for *T. simiae* (PCR, other, microscopy)	tlkpCongo&SimiaeTestMethod.anmCongoMethodID
		numTSimiaeSampleNumber	Number of samples tested for *T. simiae*	
		numTSimiaeNumberPositive	Number positive for *T. simiae*	

Table A2 (previous 10 pages): Main data storage tables used in the trypanosomiasis database, including field descriptions and relationships with other tables. Field options are included only for fields with a strictly defined number of options.

Table name	Table description	Field names	Field descriptions	Links to
tlkpAge	**Lookup table of animal age groups**	anmAgeID	Unique ID for age group	
		strAge	Age group of animal	
tlkpAllTrypsTestMethod	**Lookup table of test methods for *Trypanosoma* spp**	anmAllTrypsMethodID	Unique ID for testing method	
		strAllTrypsMethod	Test method used to detect *Trypanosoma* spp (not species specific)	
tlkpAnimalType	**Lookup table of animal types**	anmAnimalTypeID	Unique ID for animal type	
		strAnimalType	Type of animal	
tlkpBreed	**Lookup table of animal breeds**	anmBreedID	Unique ID for animal breed	
		strBreed	Breed of animal	
tlkpBruceiTestMethod	**Lookup table of test methods for *T. brucei s.l.***	anmBruceiTestID	Unique ID for testing method	
		strTestMethod	Test method used to detect *T. brucei s.l.*	
tlkpCondition	**Lookup table of condition scores for animals**	anmConditionID	Unique ID for condition score	
		strCondition	Condition of animal	
tlkpCongo&SimiaeTestMethod	**Lookup table for *T. congolense* and *T. simiae* test methods**	anmCongoMethodID	ID for *T. congolense* and *T. simiae* test methods	
		strCongoTestMethod	Test method for *T. congolense* or *T. simiae*	
tlkpCountry	**Lookup table for country**	anmCountryID	Unique ID for country	
		strCountry	Country from which data was collected	
tlkpCounty	**Lookup table for county**	anmCountyID	Unique ID for county	
		strCounty	County	
		numDistrict	District the county is in	tlkpDistrict.anmDistrictID
tlkpDataSource	**Lookup for the source of data**	anmDataSourceID	Unique ID for data source	
		strDataSource	Source of original data	

Table name	Table description	Field names	Field descriptions	Links to
tlkpDetectionMethod	**Lookup table for method of detection (human SS)**	anmDetectionMethodID	Unique ID for detection method	
		strDetectionMethod	Method of detection of case	
tlkpDistrict	**Lookup table for district**	anmDistrictID	Unique ID for district	
		strDistrict	District	
		numRegion	Region the district is in	tlkpRegion.anmRegionID
tlkpGambienseTestMethod	**Lookup table of test methods for *T. b. gambiense***	anmGambienseTestID	ID for *T. b. gambiense* test methods	
		strGambienseTestMethod	Method for testing for *T. b. gambiense*	
tlkpGender	**Lookup table for gender**	anmGenderID	Unique ID for gender	
		strGender	Gender	
tlkpLocationType	**Lookup table for location type**	anmLocationTypeID	Unique ID for location type	
		strLocationType	Location type	
tlkpOutcome	**Lookup table for outcome of infection**	anmOutcomeID	Unique ID for Outcome	
		strOutcome	Outcome of infection	
tlkpRegion	**Lookup table for Region**	anmRegionID	Unique ID for region	
		strRegion	Region	
		numCountry	Country the region is in	tlkpCountry.anmCountryID
tlkpRhodesienseTestMethod	**Lookup table for *T. b. rhodesiense* test method**	anmRhodesienseTestID	Unique ID for *T. b. rhodesiense* test method	
		strRhodesienseTestMethod	Test method for *T. b. rhodesiense*	
tlkpSampleMethod	**Lookup table for sample methods**	anmSampleTypeID	Unique ID for sample type	
		strSampleType	Type of sample taken	

Table name	Table description	Field names	Field descriptions	Links to
		strSampleDetails	Further details of sample	
tlkpSite	**Lookup table for sleeping sickness clinic site**	anmSiteID	Unique ID for hospital site	
		strSite	Site of admission	
tlkpStage	**Lookup table for stage of infection**	anmStageID	Unique ID for stage of infection	
		strStage	Stage of disease	
tlkpSubSpecies	**Lookup table for subspecies of *T. brucei* (human infective)**	anmSubSpID	Unique ID for subspecies	
		strSubSpecies	Subspecies of parasite	
tlkpTestResult	**Lookup table for test result**	anmTestResultID	Unique ID for test result	
		strTestResult	Test result	
tlkpVivaxTestMethod	**Lookup table for *T. vivax* test methods**	anmVivaxTestID	Unique ID for the *T. vivax* test method	
		strVivaxTestMethod	Test method for *T. vivax*	

Table A3 (previous 3 pages): Lookup tables used in the trypanosomiasis database, including field descriptions and relationships with other tables.

Appendix B: ATCOR-2 settings used for atmospheric correction of Landsat ETM+ images

Image acquisition date	**27/01/2001**
Image acquisition time (GTM)	07:51:13
Solar elevation	52.6
Solar azimuth	123.6
Solar zenith	37.4
Sensor tilt angle	0
Sensor view azimuth angle	0
Scene centre	1.44600 N, 33.02080 E
Water vapour category	Tropical
Aerosol type	Rural
Visibility estimate	13 km
Adjacency range	0.03 km (same as pixel size)

Image acquisition date	**17/04/2001**
Image acquisition time (GTM)	07:51:03
Solar elevation	59.67
Solar azimuth	71.38
Solar zenith	30.3
Sensor tilt angle	0
Sensor view azimuth angle	0
Scene centre	1.44600 N, 33.02080 E
Water vapour category	Tropical
Aerosol type	Rural
Visibility estimate	30 km
Adjacency range	0.03 km (same as pixel size)

Image acquisition date	**27/11/2001**
Image acquisition time (GTM)	07:49:29
Solar elevation	55.6
Solar azimuth	132.5
Solar zenith	34.5
Sensor tilt angle	0
Sensor view azimuth angle	0
Scene centre	1.44600 N, 33.02080 E
Water vapour category	Tropical
Aerosol type	Rural
Visibility estimate	40 km
Adjacency range	0.03 km (same as pixel size)

Appendix C: Spatial predictions of Rhodesian Human African Trypanosomiasis (sleeping sickness) prevalence in Kaberamaido and Dokolo, two newly affected districts of Uganda.

Batchelor, N, Atkinson, P M, Gething, P W, Picozzi, K, Fevre, E M, Kakembo, A, and Welburn, S. Spatial predictions of Rhodesian Human African Trypanosomiasis prevalence in Kaberamaido and Dokolo, two newly affected districts of Uganda, *PLoS Neglected Tropical Diseases* 2009 (30); e563.