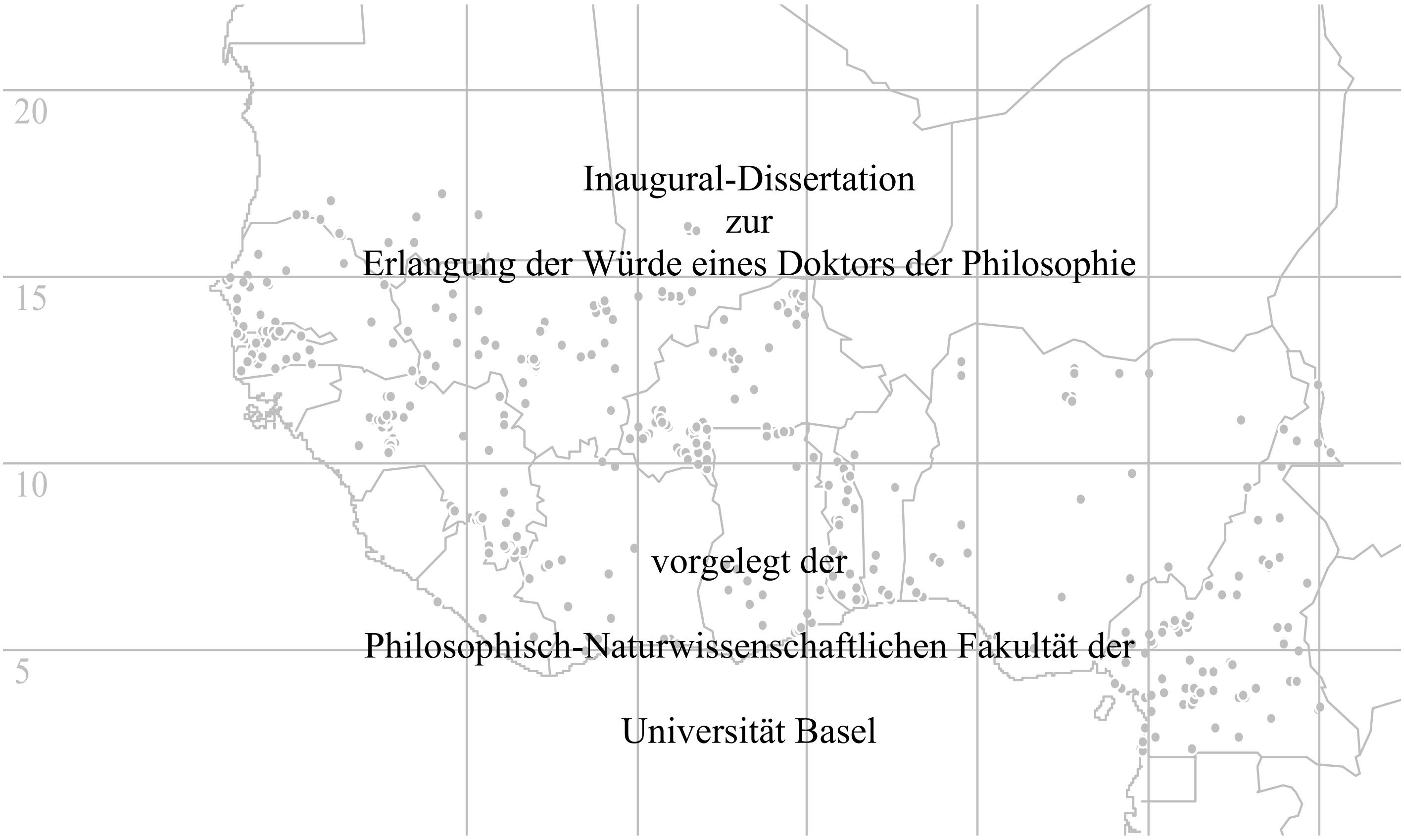
Spatial statistical analysis, modelling and mapping of malaria in Africa

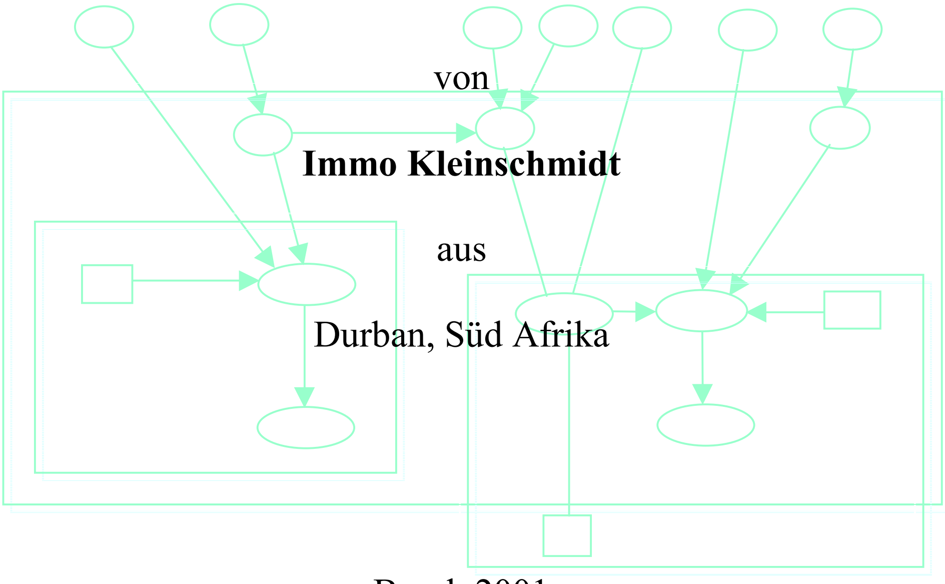


Inaugural-Dissertation
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät der

Universität Basel



von

Immo Kleinschmidt

aus

Durban, Süd Afrika

Basel, 2001

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. M. Tanner, Dr. T. Smith, Prof. Dr. M. Weiss und Dr. P. Vounatsou

Basel, den 3. Juli 2001

Prof. Dr. Andreas Zuberbühler, Dekan.

Dedicated to
to Mags, Lissy, Joe and Tusha
and to Nelson Mandela

Table of Contents

Acknowledgments

This thesis would not have been possible without significant contributions and assistance from a number of people. First and foremost I would like to thank my supervisor Dr. Tom Smith, who came up with the proposal for me to do this PhD a few years ago at a MARA meeting in Durban. I am indebted to him for many ideas and levelheaded suggestions and above all for many discussions which benefited from his uniquely simultaneous insights into malaria epidemiology and statistical modelling. I would also like to acknowledge Tom's enduring sense of humour that provided many of the lighter moments that helped to keep me going.

Likewise I am highly indebted to my co-supervisor, Dr Penelope Vounatsou, who taught me what I needed to know about Bayesian statistics and MCMC methods, and who always managed to come up with a new idea if a particular method did not seem to work. Her perseverance in pursuing a particular line of analysis and her insight into spatial methods has had a major impact on some studies in this thesis and I am very grateful for this.

I am extremely grateful to both Tom and Penelope for their generosity in making financial arrangements that funded my stay in Basel on three extended visits and paid for some of my flights from Durban to Basel. The support of the Swiss Tropical Institute is hereby gratefully acknowledged as well as partial funding of this work by the Swiss National Science Foundation (grant SNF 3200-057165.99).

Furthermore I would like to thank the head of the Swiss Tropical Institute, Professor Marcel Tanner, for his overall guidance and for making it possible for this PhD to be done at the Institute. I would also like to thank my *Koreferent* Professor Mitchell Weiss for his interest in my project, his hospitality in the Department of Epidemiology and Public Health and for many informal chats.

Much of the work in this thesis would not have been possible without the input I have received from Dr Brian Sharp, director of the National Malaria Research Programme of the Medical Research Council in Durban. Brian's particular insights into malaria in

areas of seasonal and unstable malaria, his insights into the Kwa Zulu Natal malaria data, and his constant encouragement made a major contribution through many discussions around the stone table outside the MRC offices in Durban. In addition he taught me some of the very basics of entomology.

Of central importance to the early chapters of this thesis is the contribution of Professor Peter Clarke, at that time head of Statistics and Biometry at the University of Natal in Pietermaritzburg. Peter mentored me into combined applications in the fields of spatial statistics, geostatistics and generalised linear mixed models. His contribution in terms of encouragement, novel ideas and attention to detail has helped shape much of the work that resulted in chapters 2, 3 and 4. I am also indebted to his successor, Prof Linda Haines, also of University of Natal, for many discussions on Bayesian modelling, and for filling the whiteboard in my office with equations, inspirations and illustrations.

Further thanks are due to my supervisor at the South African Medical Research Council, Dr Carl Lombard, for his encouragement and his generosity in being accommodating regarding my visits to Basel and other demands relating to doing a PhD at the same time as working. I also wish to express my thanks to my colleagues in the Biostatistics Unit in Durban for their understanding, and particular thanks to Salo Pillay for sorting out my references.

My sincere thanks also to the MARA principal investigators, Marlies Craig, and Dr Abraham Mnzava, for facilitating this work, for support, and part-funding.

I am very thankful to Dr Christian Lengeler and his family for their generous hospitality in Basel, and to Christian for his support and encouragement of the collaborative arrangements within the MARA.

Many thanks to Ms Cornelia Nauman at the Swiss Tropical Institute, for sorting out so many administrative issues and for the highly efficient manner in which she made all the practical arrangements for my visits. Sincere thanks also go to Ms Christine Walliser for her pleasant manner in sorting out many practical matters during my visits to Basel.

Many thanks also to fellow students at STI for their friendship and hospitality, in particular to Dr Ivo Müller, Dr Hassan Mshinda, Armin Gemperli, Sebastien Gagneux, Lucy Ochola and Owusu-Agyei. Very special thanks to Armin for his help with the *Zusammenfassung*.

In a wider context I would like to acknowledge the democratic changes in South Africa that have made it possible for South African scientists to make a contribution in internationally collaborative projects. Before the overthrow of apartheid it would have been unthinkable for an African collaboration such as the MARA project to include South African institutions and researchers. It is all too easy for South Africans to take for granted the freedoms we now enjoy, which were only attained through struggle, and extreme sacrifice by many of our compatriots.

On a very personal note, I acknowledge above all with deep gratitude the support, love, tolerance and selfless devotion of my wife Mags throughout the period of this PhD, and Lissy and Joe for their love and forbearance that sustained me throughout.

Summary

Estimates of the disease burden due to malaria in Africa show that the toll it is exacting in terms of loss of life, episodes of serious illness, and impediment to economic development is enormous. In many areas the situation has become worse due to failing drugs, failing insecticides, failing health systems, large scale population movements and possibly due to co-infection with HIV. On the other hand, recent studies have shown that widespread use of insecticide treated bed nets has the potential for making substantial inroads into this disease burden, particularly in areas of high endemicity.

Recording the geographical distribution of any major disease forms an important basis for locating appropriate interventions for its control and a means to monitoring their effectiveness. It also provides a possibility for identifying ecological factors with which the disease may be associated.

The objective of this thesis was to produce evidence-based maps of malaria prevalence and incidence by means of spatial statistical modelling; to evaluate and advance the application of methodology in the analysis of spatially correlated disease data; and to undertake detailed analysis of malaria incidence for one particular area in order to establish underlying patterns of malaria risk over space and time and in relation to population, climatic and environmental factors. Altogether six individual studies were carried out, which modelled malaria distribution at three different levels of scale. These levels and their locations, were: regional level in sub-Saharan West Africa, country level in Mali and district level in Ubombo and Ngwavuma in KwaZulu Natal, South Africa. In the case of the regional and country maps, the malariometric measure was parasite prevalence in children, obtained from the MARA database. In the case of the district-level analysis, routinely recorded small area malaria incidence data were used, which were obtained from the provincial malaria control programme. Three of the studies modelled malaria distribution over space and time.

There are well-documented difficulties with the mapping of raw disease rates, since such maps will be dominated by sampling variability and analyses based on them will be flawed due to the lack of independence in the rates. Spatial statistical methods can be used to overcome these difficulties, but these have rarely been applied in the context of malaria distribution modelling. In this thesis two such approaches were employed: 1) classical geo-statistical methods, based on variograms and generalised linear mixed models, and 2) autoregressive models in a Bayesian context using Markov Chain Monte Carlo (MCMC) methods. Some minor adaptations of the methods have been suggested.

The main findings of the studies carried out in this thesis were:

- Both classical geostatistical and autoregressive MCMC methods are feasible for modelling malaria distribution and advantages and limitations of each method have to be weighed up in a particular context. The development of extensions to the MCMC spatial modelling approach to cater for point referenced (as opposed to areal) spatial data will make this method more generally applicable. The ability to adequately reflect the effects of random errors comprehensively in the resulting map estimates is an important advantage of the Bayesian modelling approach.
- It is feasible to produce evidence-based maps of transmission intensity, which are a refinement of expert opinion maps, from parasite ratio surveys.
- Malariometric measures of transmission intensity (and their proxies) are often highly correlated in space as well as in time and this must be taken into account in any modelling, particularly at the short range scales.
- Due to strong spatial heterogeneity it is difficult to model malaria transmission intensity without leaving considerable unexplained, residual variation, which may be spatially correlated. It is therefore unsatisfactory to map model predictions directly. One method of overcoming this problem is to produce a map of kriged (interpolated) model residuals, and to add these to model predictions which can then be mapped. In large heterogeneous regions, models should be derived within ecological zones, and special smoothing methods

should be employed in boundary areas between these zones, rather than attempting to derive a single unified distribution model for the whole region.

- Spatial variation in malaria transmission intensity is significantly associated with basic climatic factors in areas of endemic stable malaria and in areas of epidemic unstable malaria, but the relationship is usually not straightforward. However, an association between temporal variation in malaria transmission intensity and variation in weather, whilst plausible, could not be proven in the data that were analysed.
- Sharp increases in malaria caseloads in Kwa Zulu Natal appear to originate mainly from areas of previously low incidence, whilst high incidence areas have partly stabilized. This suggests a geographical expansion of malarious areas, and the acquisition of clinical tolerance to disease in some individuals in high incidence areas. The finding that adults in high transmission sub-regions of the province experience lower incidence rates than teenagers, supports the hypothesis of clinical immunity to infection in these relatively high incidence areas. Children under five in the same area, experience the lowest incidence rates compared to other age groups, possibly as a result of being more adequately protected by vector control measures than older children and adults.
- In areas of unstable fluctuating malaria transmission intensity, incidence in individual localities is highly correlated to incidence at the same locality in previous seasons.

One of the maps (West Africa) that were produced in this thesis has already been put to use in malaria control. The findings relating to Kwa Zulu Natal will be presented directly to the provincial malaria control programme. Two of the six studies have been published, three have been submitted for publication and one is being prepared for submission, to ensure widespread dissemination of the findings.

A number of future research questions arise out of this work. These are, amongst others:

- Methodological development of Bayesian spatial modelling software, particularly to accommodate point referenced spatial data.

- Further analysis using the MARA database to produce endemicity maps of other regions in Africa.
- Prospective studies should be undertaken to assess the relationship between malaria and weather changes in epidemic prone areas, with a view to further exploring the feasibility of epidemic forecasting systems.
- Further investigation of factors that influence the acquisition of clinical immunity in adults in areas of moderate transmission intensity; investigation whether this is confirmed in similar areas elsewhere (e.g. Namibia, Botswana), and whether it is supported by age specific differences in case-fatality rates.

Zusammenfassung

Schätzungen der Malariabelastung in Afrika zeigen, dass diese Krankheit eine sehr hohe Sterberate und eine enorme Anzahl schwerer Erkrankungen verursacht, sowie ein beachtliches Hindernis für die wirtschafliche Entwicklung darstellt. In vielen Teilen des Kontinents hat sich die Situation wegen Fehlmedikation verschlechtert, sowie durch unwirksame Insektenbekämpfungsmittel, mangelhafte Gesundheitsdienste, grosse Bevölkerungsumsiedlungen und möglicherweise durch Koinfektion mit HIV. Demgegenüber haben neuere Studien gezeigt, dass die flächendeckende Nutzung von insektizidbehandelten Mückennetzen das Potential haben, grosse Erfolge gegen die Belastung durch Malaria zu erzielen, besonders in Gegenden mit hoher Endemizität.

Das Aufzeichnen der geographischen Ausbreitung einer Krankheit stellt eine wichtige Basis dar, um entsprechende Kontrollinterventionen zu lokalisieren, und um die Effektivität solcher Interventionen zu überwachen. Weiter dient es der Möglichkeit Umweltfaktoren zu identifizieren, mit der die Krankheit verbunden sein könnte.

Das Ziel der vorliegenden Dissertation war: Mittels räumlicher statistischer Modellierung Karten zu erstellen, welche die Prävalenz und das Auftreten von Malaria dokumentieren; die Anwendung von Methoden in der Analyse räumlich korrelierter Daten auszuwerten und zu verbessern; und eine detaillierte Analyse von Malariameldungen einer bestimmten Gegend durchzuführen, damit zugrundeliegende räumliche und zeitliche Tendenzen von Malaria Erkrankungsrisiken aufgezeigt und mit Bevölkerungs-, Klima- und Umweltfaktoren in Zusammenhang gestellt werden können. Im Ganzen wurden sechs verschiedene Studien durchgeführt, welche die Malariaausbreitung in drei verschiedenen Masstäben modellierten. Diese drei Masstäbe waren: Regionaler Masstab in West Afrika südlich der Sahara, Landesmasstab in Mali, und Distriktmasstab in Ubombo und Ngwavuma in KwaZulu Natal, Süd Afrika. Bei den Regional- und Landeskarten wurde das Vorkommen von Malariaparasiten bei Kindern als Malariaindikator benutzt, welches in der MARA Datenbank enthalten ist. Im Falle der Distriktanalyse wurden routinemässig gemessene Malaria Inzidenzdaten benutzt, die vom Malaria-Kontrolldienst der

Provinz Kwa Zulu Natal zugestellt wurden. Drei der Studien modellierten Malaria Ausbreitung in der räumlichen sowie in der zeitlichen Dimension.

Es gibt gutdokumentierte Schwierigkeiten die auftreten, wenn rohe Krankheitsraten auf Landkarten übertragen werden, da solche Landkarten überwiegend von Stichprobenvariabilität dominiert werden, und jegliche Analyse, die darauf beruht, wegen der nicht vorhandenen Unabhängigkeit der Daten, fälschliche Resultate aufweist. Räumliche statistische Methodik, welche zu diesem Zweck entwickelt wurde, kann solche Schwierigkeiten überwinden, wurde aber bisher selten im Zusammenhang mit Malaria Ausbreitungsmodellierung verwendet. In der vorliegenden Dissertation wurden zwei solche Ansätze angewandt: erstens klassische geostatistische Methoden, die auf Variogrammen und verallgemeinerten linearen gemischten Modellen beruhen, und zweitens autoregressive Modelle, die in einem bayesianischen Kontext Markov Chain Monte Carlo (MCMC) Methoden anwenden. Ferner werden geringfügige Abweichungen der Methodik vorgeschlagen.

Die wichtigsten Ergebnisse der Untersuchungen in dieser Doktorarbeit waren:

- Klassische geostatistische sowie autoregressive MCMC Methoden können erfolgreich zur Modellierung der Malariaausbreitung angewandt werden - ihre jeweiligen Vor- und Nachteile müssen im einzelnen Fall abgewogen werden. Die Weiterentwicklung des MCMC Ansatzes zur räumlichen Modellierung punktueller, im Gegensatz zu flächigen Daten, wird diese Methodik allgemeiner anwendtbar machen. Die Fähigkeit der bayesianischen Methodik die Effekte von Stichprobenfehlern in den sich ergebenden Kartenschätzungen zu reflektieren, ist ein wichtiger Vorteil dieses Ansatzes.
- Es ist durchaus möglich, mittels Erhebungen über Malariaparasiten Prävalenz, empirische Karten der Übertragungsintensitäten zu erstellen, die eine Verfeinerung der Expertenkarten darstellen.
- Messungen der Malariaübertragungsintensität sind oft räumlich sowie zeitlich stark korreliert. Diese Korrelation muss bei jeglicher Modellierung in Betracht gezogen werden, besonders bei kürzeren Distanzen.

- Wegen starker räumlicher Heterogenität ist es schwierig Malaria Ausbreitungsmodelle zu entwickeln, bei denen nicht beachtliche unerklärte Residualvariation zurückbleibt, welche räumlich korreliert sein kann. Es ist deshalb nicht zufriedenstellend, Modellvohersagen direkt auf Karten zu übertragen. Stattdessen kann eine Karte von gekrigden Modelresiduen erstellt werden, um diese dann zu den Modellvorhersagen zu addieren - diese addierten Werte können schliesslich auf Karten übertragen werden. In heterogenen Gebieten sollten Modelle in der Regel innerhalb von ökologischen Zonen erstellt werden, und spezielle Glättungsmethoden sollten in den Grenzgebieten zwischen diesen Zonen durchgeführt werden, statt zu versuchen ein einziges, ganzumfassendes Verbreitungsmodel abzuleiten.
- Die Räumliche Variation der Malaria Übertragunsintensität ist erheblich assoziert mit grundlegenden klimatischen Faktoren in Gegenden von endemischer, stabiler Malaria Übertragung, sowie in Gegenden von unstabiler epidemischer Übertragung, aber der Zusammenhang ist meistens nicht einfacher Natur. Eine Assoziation von zeitlicher Veränderung von Malaria Übertragungsintensität, und zeitlicher Veränderungen des Wetters, ist zwar plausibel, konnte aber nicht nachgewiesen werden in den Datensätzen die analysiert wurden.
- Der starke Anstieg der Anzahl Malariafälle in Kwa Zulu Natal scheint hauptsächlich aus Gegenden zu stammen, wo vorher nur geringes Auftreten der Krankheit vorhanden war, während es sich in Gegenden mit zuvor hohen Auftretensraten teilweise stabilisierte. Dieser Umstand deutet auf eine räumliche Ausbreitung von Malaria Gegenden hin, sowie das Erwerben einer klinischen Toleranz bei manchen Bewohnern der Gegenden mit bisher hoher Malaria Inzidenz. Die Feststellung, dass Erwachsene in Gebieten höherer Übertragungsintensität einem niedrigerem Auftreten von Malariaepisoden unterliegen als Teenagern, unterstützt die Hypothese klinischer Immunität in diesen Regionen. In diesen Orten, erleben Kinder unter fünf Jahren weniger Episoden von Malariaerkrankung als irgend eine andere Altersgruppe, möglicherweise infolge von Mückenbekämpfungsmassnahmen, welche diese Altersgruppe vorwiegend schützen.

- In Gegenden unstabiler, fluktuierender Malaria Übertragungsintensität, ist die Inzidenz einzelner Teilgebiete stark korreliert mit der vorjahres Inzidenz desselben Teilgebietes.

Eine der Karten, die in dieser Dissertation erstellt wurde (West Afrika, Kapitel 4), ist bereits in der Malariabekämpfung benutzt worden. Die Ergebnisse, die sich mit Kwa Zulu Natal befassen, werden dem örtlichen Malaria-Kontrolldienst direkt vorgetragen. Zwei der sechs Studien dieser Dissertation sind bereits publiziert worden, drei weitere sind zur Publikation eingesandt, und ein weiteres wird zur Publikation vorbereitet, um diese Ergebnisse so weit wie möglich zu verbreiten.

Einige weitere Forschungsthemen ergeben sich aus dieser Arbeit. Diese sind unter anderem:

- Methodische Weiterentwickelung von bayesianischer räumlicher modellierungs Software damit diese punkt-bezogene räumliche Daten verarbeiten können.
- Weiterbearbeitung der MARA Datenbank um Malaria Endemizitätskarten für andere Regionen Afrikas zu erstellen.
- Prospektive Studien sollten unternommen werden, um den Zusammenhang zwischen Malaria und Wetterveränderung zu bewerten, in Gegenden welche Malaria Epidemien unterliegen, um weitere Möglichkeiten eines Malaria Vorherrsagesystems zu beurteilen.
- Weitere Untersuchung von Faktoren, die Erwerb von klinischer Immunität bei Erwachsenen in Gegenden mässiger Übertragungsintensität beeinflussen. Untersuchungen ob dieses Phänomen sich in anderen ähnlichen Gegenden wiederholt (z.B. Namibia, Botswana), und ob es durch Unterschiede in Fatalitätsraten bestätigt wird.

Abbreviations

AEZ	Agro-ecological zone
AR(1)	First order autoregressive
ARTEMIS	African real time environmental monitoring using imaging satellites
AVHRR	Advanced very high resolution radiometer
CAR	Conditional autoregressive
CCD	Cold cloud duration
EA	Enumeration area
EIR	entomological inoculation rate
EPD	Expected predictive deviance
ESHAW	Ecosystem health analytic workshop
FAO	Food and agricultural organisation
GLM	Generalised linear model
GLMM	Generalised linear mixed model
GIS	Geographic information system
GPS	global positioning systems
HRR	High resolution radiometer
ITM/ITBN	Insecticide treated material/bednet
LRS	Likelihood ratio statistic
LST	Land surface temperature
MARA	Mapping malaria risk in Africa
MCMC	Markov Chain Monte Carlo
NOAA	National Oceanographic and Atmospheric Administration
NDVI	normalised difference vegetation index
PEN	Penalty
RS	Remote sensing
SD	Standard deviation

List of tables

List of figures

Chapter 1

Introduction: The epidemiology of malaria distribution

The burden of malaria in Africa

In areas of stable endemic malaria transmission in sub-Saharan Africa it has been estimated that in 1995 about 1 million deaths were directly attributable to malaria infection (Snow *et al*. 1999). Of these deaths, three-quarters were in children below the age of 5 years. In the same population, it is estimated that about 200 million clinical attacks of malaria occurred in the same year. In areas of unstable or epidemic prone malaria in southern Africa (“fringe areas”), about 2000 deaths and 200,000 clinical episodes occurred that were due to malaria and that were not prevented despite malaria control measures in these areas. According to a World Bank report of 1993, malaria accounts for an estimated 35 million disability –adjusted life years (DALYs) per year lost in Africa due to ill-health and premature death (World Bank, 1993).

The discovery of an interactive effect between HIV infection and malaria morbidity (Whitworth *et al.* 2000; Chandramohan and Greenwood 1998; Verhoef et al. 1999) exacerbates the potential for devastating health consequences in populations with large numbers of individuals who are co-infected. In resource-poor countries in Africa, malaria prevention and treatment consume large proportions of health budgets, and since it poses a threat to indigenous populations as well as visitors, it acts as a deterrent to tourism and foreign investment in these countries. Malaria therefore not only affects the health status of Africa’s population, but also has far-reaching economic consequences inhibiting economic development (Wernsdorfer and Wernsdorfer 1988). The impact of malaria on the region has been recognized by the convening of the first African summit of heads of state on malaria in Abudja, Nigeria in April 2000. A report to the summit meeting calls, amongst other things, for more research on trends in incidence and prevalence, epidemic outbreaks and clinical epidemiology (Sachs 2000). A better understanding of the distribution of malaria has been identified as an important tool in its control (Snow *et al*. 1996). More accurate maps make it possible for interventions to be mounted which are appropriate to the disease profile which characterises particular levels of endemicity, for clinical trials and evaluations of new approaches to be located correctly, and for planners of irrigation and other development schemes to take cognisance of the potential effects of these schemes on malaria transmission intensities.

Transmission of malaria

Malaria is caused by the parasite of genus *Plasmodium*. The four species of *Plasmodium* are *P. falciparum*, *P.malariae*, *P.ovale* and *P.vivax*. In Africa the predominant species of the disease causing-parasite is *P. falciparum*. Infection of the human host occurs when a person is bitten by a female *Anopheles* mosquito which has previously become infected. The parasite, called sporozoite at this stage of its cycle, enters the human body via the saliva of the mosquito which is injected into the blood. The parasites multiply in the liver, and re-invade the blood via red blood cells as merozoites. These develop into a stage known as the trophozoite, which is the one visible in blood films, and subsequently divide by the process of schizogony to produce further merozoites, which invade non-infected blood-cells. Some of the merozoites develop into new trophozoites whilst others develop into male micro- or female macrogametocytes. Uninfected *Anopheles* mosquitoes become infected if they feed on a person with mature gametocytes in their peripheral blood. In the mosquito, the microgametozytes exflagellate into gametes before fertilising the macrogametocytes, thereby forming zygotes. The zygote changes into an ookinete and then into an oocyst, which is found in the mid-gut wall of the mosquito. Large numbers of sporozoites are formed within the oocyst. The rate of development of sporozoites in the oocyst is temperature dependent. The sporozoites leave the oocyst to invade the mosquito's salivary glands, from where they can infect another human host when the mosquito takes a blood meal. The incubation period of the parasite in the vector takes 13 days to complete at 24˚ C. for *P.falciparum*. The vector will only become infective if it survives this *sporogonic* cycle (Gilles and Warrell 1993, chapter 2).

Malaria as a disease is therefore closely bound to conditions which favour the survival of the anopheles mosquito in the form of habitat and breeding sites and which favour the life cycle of the parasite in terms of suitable temperatures. In the absence of any human intervention these conditions are predominantly determined by climatic and environmental factors.

The most important vectors of malaria in Africa are members of the *An. gambiae* complex and *An. funestus*. Identification of the distribution of particular species is important since malaria vector control measures may have to take account of behavioural differences between species to be effective (Coetzee *et al*. 2000; Gillies and De Meillon 1968). For example, indoor biting and indoor resting habits (endophagy and endophily respectively), make mosquitoes more susceptible to control by residual insecticide on interior walls of houses, and to other insecticide treated materials such as bednets.

Five species of the *An. gambiae* complex are vectors of malaria. The two species which are the most efficient vectors of malaria parasites, *An. gambiae sensu stricto* and *An. arabiensis*, are also the most widely distributed throughout most of sub-Saharan Africa. They often occur together, but *An. arabiensis* predominates in drier areas, whilst *An gambiae* predominates in more humid areas. *An gambiae* generally has a higher vectorial capacity than any of the other species, in part due to it being highly anthropophilic. It is also mainly endophagic and endophilic, making it amenable to control by indoor house-spraying of residual insecticide, at least in areas of moderate transmission intensity. *An. arabiensis,* on the other hand, is partly zoophagic and mainly exophagic and exophilic. It is generally considered a less efficient vector of malaria than *An gambiae*, but it is nevertheless the principal malaria vector in many areas (White 1974). *A. bwambae* is found only in the Semliki forest area in Uganda. It is partially endophagic and partially endophilic. The two saltwater species of the *An. gambiae* complex are *An. melas* and *An. merus* which are found in West Africa and in East Africa respectively. *An. merus* is exophilic and mainly zoophagic, whereas *An. melas* displays a more mixed resting and biting behaviour. *An. funestus* of the *An funestus* group, the other major vector of malaria in many parts of tropical and sub-tropical Africa (Armah *et al.* 1997; Gillies and De Meillon, 1968) bites humans; it is exophagic and endophilic. Since it breeds mainly in permanent water bodies, it is associated with all-year as opposed to seasonal malaria transmission (Sharp *et al*. 2000).

One of the main environmental factors affecting malaria transmission is temperature. The effect of an increase in temperature on the parasite is to shorten the sporogony cycle and hence to accelerate transmission. The duration of sporogony can be calculated by the formula $n=T/(t-t_{min})$ where n=duration of sporogony in days, t= average temperature in ˚C, and for *P.falciparum* T =105 and t_{min} =16˚ C. Below 16˚ C parasite development ceases. Rising temperature also increases transmission by increasesing the frequency with which the vector takes blood meals, which increases the growth rate of vector populations through a shortening of the generation time. The optimal range of temperature for most vectors lies between 20 and 30˚ C. Higher temperatures reduce the longevity of adult vectors, and hence fewer of them will survive the sporogony cycle to become infective. There are thus upper and lower thresholds outside which malaria transmission is very inefficient or impossible. The dependence of malaria transmission on temperature is indirectly expressed in the Macdonald model which formulates the dependence of the basic reproduction rate of malaria in terms of the daily survival probability of the vector and the length of the incubation period (Bruce-Chwatt 1980, pp. 149-159; Molineaux 1988, pp. 923).

Increasing rainfall and vegetation density generally have a favourable impact on malaria transmission through the provision of breeding sites and habitat for the vector. However, the differing breeding habits of different species of Anopheles, complicate the relationship between rainfall and malaria transmission. Flooding, for example, may flush out larvae pools and lead to a temporary reduction in vector populations. Forest vegetation may inhibit *An. gambiae* because of the lack of sunlight. Nevertheless, insufficient annual rainfall, or seasonal rainfall, constitutes a distinct limitation to malaria transmission in areas where temperature is not a limiting factor. Rainfall of about 80mm per month for at least five months of the year has been identified as a minimum requirement for stable transmission to occur (Craig *et al.* 1999).

Clinical manifestations

Clinical malaria manifests itself in its mild form as a febrile illness associated with other non-specific symptoms (Bruce-Chwatt 1980, ch.3). The first clinical signs will only appear after the incubation period, which varies between nine and fourteen days for *falciparum* malaria. Clinical diagnosis is usually confirmed by a blood test, involving microscopic evidence of parasites in the blood, or by rapid diagnostic kit (Craig and Sharp 1997). However, in endemic countries infected individuals are often asymptomatic, so that parasitological evidence does not necessarily prove that the symptoms are due to malaria in a particular patient (Bruce-Chwatt 1980, pp. 35-51; Snow *et al*. 1997).

Severe life threatening malaria is usually due to *P.falciparum* malaria. In non-endemic areas cerebral malaria is the sequel that often sets in after the initial general symptoms. In such areas death due to malaria in both children and adults is usually due to cerebral malaria. In highly endemic areas severe malaria affects mainly young children, and women during pregnancy. In such areas infants may enjoy a period of inherited immunity of up to 6 months. As this declines, clinical attacks become more severe, and often take the form of severe anaemia which is responsible for most deaths due to malaria in these areas. Depending on the intensity of exposure to the parasite, these children develop relative tolerance to malaria infection in their first few years of life. As a result of this older children and adults usually exhibit mild, non life-threatening clinical symptoms, if any.

Malaria control

In areas of high transmission intensity the use of insecticide treated bednets (ITBNs) and materials has become recognized as an effective means of malaria vector control for reducing mortality and severe morbidity in young children and pregnant mothers (Binka 1997; Abdulla *et al.* 2001). In an integrated strategy these would be used in conjunction with rapid and effective algorithms for diagnosis and the availability of efficient and affordable drugs for case management.

In areas of low transmission intensities (particularly in southern Africa), house spraying with residual insecticide (for example pyrethroids, or DDT) has been widely used as an effective means of vector control, coupled with definitive diagnosis and treatment towards parasitological cure (Sharp *et al.* 2000). More recently, this has been complemented with the use of ITBNs in specific areas (Mnzava *et al.* 1999).

Malaria parasite control in most parts of Africa, including the malaria “fringe” areas in the south, has been affected by large scale parasite resistance to the cheap anti-malarial drugs such as chloroquine and increasingly to sulphadoxine/pyrimethamine (SP). In KwaZulu Natal in South Africa this has necessitated a recent decision to introduce combination therapy including artemisinin in place of previously used SP.

Malaria distribution data and measures of transmission intensity

For modelling malaria transmission intensity, the measure of choice is the entomological inoculation rate (EIR), which is the number of infective bites per person per year, since it is a direct measure of exposure to which individuals are subjected. Unfortunately this is not widely available. Other potential measures would be the vectorial capacity, man-biting rate, parasite ratio and incidence rates. Irrespective of the merits and de-merits of these measures, the only one that is widely available for the whole continent is the parasite ratio or prevalence of infection. This is obtained by a random survey of individuals who are tested for the presence of parasites in their blood. The results of thousands of these surveys taken over time across the length and breadth of malarious areas in Africa, have been consolidated in the MARA database (MARA/ARMA Collaboration 1998). Due to the effects of partial immunity in endemic malaria areas, surveys that include older children and adults do not give a reliable measure of potential infection rates. For this reason only surveys (or components of surveys) restricted to children under 10 years of age have been included in analyses for the purpose of malaria distribution modelling. A general problem with such surveys is that they are predominantly located in areas of high transmission intensity, leading to an under-representation of populations living in low transmission environments.

It has been shown that parasite ratios are reasonably well correlated with EIR (Beier *et al*. 1999). For this reason the parasite ratio is an acceptable proxy for transmission intensity. It needs to be remembered, however, that the parasite ratio is dependent on the age-group of children being surveyed, and to some extent on season. If the main objective of modelling is to predict malaria risk in broad categories, then the parasite ratio is the most practical measure due to its abundant availability.

Another proxy of transmission intensity that is fairly widely available in southern Africa is parasitologically confirmed disease incidence. Incidence data generally are biased due to the fact that they may reflect patient access to health services rather than true morbidity, and they are dependent on good denominator data being available at the same level of aggregation as the case data. In the northern most magisterial districts of KwaZulu Natal a surveillance system is used which is believed to identify the vast majority of cases, since active case finding supplements the passively reported cases, as part of a malaria control strategy that seeks to identify and treat every infected individual. Reasonably good population data are also available for this area. Incidence data for this population are unique in that they have been recorded over many years. Since malaria in the area is seasonal and highly variable over space and time, the data present an unequalled opportunity to investigate the relationship between climatic variability and malaria incidence in a mainly non-immune population and to explore the potential of epidemic prediction using satellite derived meteorological data.

This thesis therefore used both parasite ratios and malaria incidence data to undertake spatial statistical analysis of malaria distribution. In chapters 2 and 4 parasite ratios are used to model the relationship between malaria and climatic factors in order to produce prediction maps of prevalence of infection. Chapters 3, 5, 6 and 7 use incidence data to analyse spatial and temporal variation in incidence and to investigate relationships between climate and malaria at a small area level by using spatial and spatial-temporal models.

There have been previous projects to map the distribution of malaria in Africa. These have ranged from expert opinion maps (Molineaux 1988), to suitability maps (Craig *et al.* 1999), to maps for a single country that have used parasite ratios (Thomson *et al.* 1999). Whilst this thesis is not attempting to produce a detailed empirically derived risk map for the whole continent, it attempts to show approaches using modern statistical methods that are suitable at different levels of scale ranging from regional to sub-district maps.

Spatial statistical modelling and mapping of malaria

There is a wide range of approaches to spatial analysis and modelling in the statistical and Geographic Information Systems (GIS) literature. Many of these approaches have been recently developed in response to the interest in spatial processing and presentation of data, and the opportunities that have been opened up through the collection of small area data and the development of GIS technology and software. However, the idea of spatial analysis to solve epidemiological problems goes back to the very beginnings of epidemiological research (Snow J 1855).

Statistical approaches to spatial analysis have in common the concept of correlation or non-independence of spatial data. This can be a problem that needs to be taken into account when analysing such data since the degrees of freedom tend to be exaggerated, or it can be usefully exploited, for example in stabilising small counts of cases in small areas by borrowing strength from neighbouring areas. Sometimes the mere existence of significant spatial correlation is a statistical result of interest in itself (Walter 1994). Results of spatial statistical modelling are estimated quantities (parameters) that are intended to quantify the true underlying magnitudes in a map and their uncertainty rather than the mere mapping of recorded data that are subject to sampling error. The role of GIS in such analysis is twofold: (a) to pre-process the data, for example by extracting values, or calculating distance or proximity, and (b) to post-process the results, for example by plotting estimated area effects in a map. The essential core of such spatial analysis is however, stochastic and uses statistical programs that take account of the random nature of the processes involved. Modelling approaches that are based purely on GIS techniques tend not to deal with the random nature of processes explicitly and hence produce point estimates of processed quantities for individual pixels in a map.

In this thesis spatial statistical analysis was performed, with GIS employed as a pre- and post-processing tool, but with statistical software used for the main analysis. Two distinctly different approaches to spatial statistical modelling have been followed, without attempting to make direct comparisons between the two. In chapters 2, 3 and 4, geostatistical approaches in conjunction with generalised linear mixed models

(GLMM) have been followed, whereas in chapters 5 and 6 hierarchical fully Bayesian methods using Markov Chain Monte Carlo modelling was used.

Geostatistical, or variogram approaches have occasionally been applied to disease mapping (Carrat and Valleron 1992; Oliver *et al.* 1992). In these the method of "ordinary kriging" is used as a means of interpolating disease prevalence or incidence across a map, based on observed values at known grid locations. A variogram is used to model spatial dependence in the observed data. Classical kriging is based on the assumption that the response is a continuous variable, that its underlying value is constant across the map (stationarity) and that the covariance between two points is entirely a function of distance between them. Details are given in ch. 2. In this thesis this method has not been used directly, since these assumptions are generally not satisfied in malaria distribution data. Instead, kriging has been applied to residuals (which do satisfy the assumptions) in order to improve map estimates obtained from a regression model (ch. 2). In chapters 3 and 4 variograms are used to estimate the co-variance matrix of the GLMM which is used to analyse the relationship between the disease, and climatic and other factors. This approach requires software that allows a spatial model to be used to define the covariance matrix. Regression coefficients are estimated using residual maximum likelihood methods (Littell *et al.* 1996). The method lends itself well to data consisting of observations that represent points.

Hierarchical fully Bayesian methods using MCMC sampling (Gelfand and Smith, 1990) have been widely applied to disease mapping and *ecological* regression analysis in recent years (see Wakefield *et al.* 2000 for an overview). In this approach the correlation between neighbouring areas is modelled via conditional autoregressive (CAR) priors. Such methods have been developed for data in which the response represents an areal unit as well as for data representing points. However, readily available statistical software using these methods is currently restricted to area based spatial data which limits its application to malaria distribution data, which are generally point referenced, with the exception of the reporting system that is available in South Africa. Virtually all applications of Bayesian disease mapping methods in the literature are in the context of rare diseases such as rare cancers in developed countries of Europe and North America. Vector borne diseases in tropical countries differ in that the disease is often not rare and in that the spatial correlation is often much stronger due to the links with climatic and environmental factors. The quality of both disease data and age-sex specific population data is also generally of a lower standard than is the case for example with cancer registration data in first world countries. In chapters 5 and 6 of this thesis these methods were applied to malaria incidence data thereby representing an evaluation of this methodology to the tropical disease setting.

Currently the only “off the shelf” software that is available for this type of analysis is WinBUGS (WinBUGS 2000) and this was used in this thesis. In chapter 5 the simple spatial model without co-variates was extended to a spatial-temporal model by adding a linear temporal term with spatial smoothing of the rate of change of incidence. In chapter 6 a spatio-temporal model using first order autoregressive effects was used to investigate the effects of rainfall and temperature on malaria incidence at different points in time. The methodological details are given in the respective chapters.

Overall aim

This thesis sets out to estimate malaria prevalence and incidence at map locations or areal units by means of spatial statistical modelling; to determine factors that are associated with spatial and temporal heterogeneity of malaria transmission intensity and to evaluate the potential of using remote sensed meteorological satellite data for explaining and hence predicting variation in malaria incidence at small area level. It does so by applying state of the art methodology in the spatial analysis of correlated disease data and thereby evaluates the potential of this methodology to vector borne disease and other tropical disease data in general. It also attempts to document the time trend of malaria incidence in an area of unstable malaria and to suggest some reasons why malaria incidence has increased so unevenly in this area.

Chapter 2

A spatial statistical approach to malaria mapping

Kleinschmidt I[1], Bagayoko M[2], Clarke GPY[3], Craig M[1], Le Sueur D[1].

[1] Medical Research Council (South Africa), 771 Umbilo Road, Congella, Durban 4001, South Africa.

[2] Malaria Research and Training Center DEAP/FMPOS,
Universite du Mali, Bamako, Mali

[3] Department of Statistics and Biometry, University of Natal, Pietermaritzburg, South Africa.

This article has been published in *International Journal of Epidemiology* (2000), 29(2): 355-361.

Summary

Good maps of malaria risk have long been recognised as an important tool for malaria control. The production of such maps relies on modelling to predict the risk for most of the map, with actual observations of malaria prevalence usually only known at a limited number of specific locations. Estimation is complicated by the fact that there is often local variation of risk that cannot be accounted for by the known co-variates and because data points of measured malaria prevalence are not evenly or randomly spread across the area to be mapped. We describe, by way of an example, a simple two stage procedure for producing maps of predicted risk: we use logistic regression modelling to determine approximate risk on a larger scale and we employ geo-statistical ('kriging') approaches to improve prediction at a local level.

Malaria prevalence in children under 10 was modeled using climatic, population and topographic variables as potential predictors. After the regression analysis, spatial dependence of the model residuals was investigated. Kriging on the residuals was used to model local variation in malaria risk over and above that which is predicted by the regression model. The results of the method are illustrated by a map showing the improvement of risk prediction brought about by the second stage. The advantages and shortcomings of this approach are discussed in the context of the need for further development of methodology and software.

Keywords: malaria risk, disease maps, geo-statistics, spatial analysis, kriging, climatic factors.

Introduction

Malaria is a major cause of morbidity and mortality in Africa, and is a leading cause of death especially amongst children, in many African countries (Snow *et al*. 1999; Binka, 1997). The MARA/AMRA project (MARA/AMRA Collaboration, 1998) has been set up recently to collate sources of data on malaria, and to model and map malaria risk across the continent. Accurate maps of malaria have been recognised as an important tool in the hands of control programme managers (Snow *et al*. 1996; Kitron *et al*. 1994). This paper describes the statistical methods used to produce a map of malaria risk for Mali and discusses the methodological issues that are raised. A companion paper discusses in detail the substantive aspects of the results of this work and its policy implications (Bagayoko M, Kleinschmidt I, Sogoba N, Craig M, le Seur D, Toure YTT. Mapping malaria risk in Mali. (*in preparation*)).

The production of malaria maps relies on modelling to predict the risk for most of the map, with actual observations of malaria prevalence usually only known at a limited number of specific locations. Accurate prediction of risk is dependant on knowledge of a number of environmental and climatic factors that are related to malaria transmission (Craig *et al*. 1999; Snow *et al*. 1998; Beck *et al*. 1994). However, the estimation is complicated by the fact that there is often local variation of risk that cannot easily be accounted for by the known co-variates. A further complication arises from the fact that data points of measured malaria prevalence are not evenly or randomly spread across a country, but are often closely clustered in areas of high risk. Any modelling of risk has to take account of spatial autocorrelation of the data, and allow for local deviation from predictions that are based on the known climatic covariates

In this project a two-stage procedure was followed: (1) generalised linear regression modelling was applied to determine approximate risk on a larger scale by identifying important climatic and environmental determinants and (2) the geo-statistical kriging method was used to improve prediction at a local level.

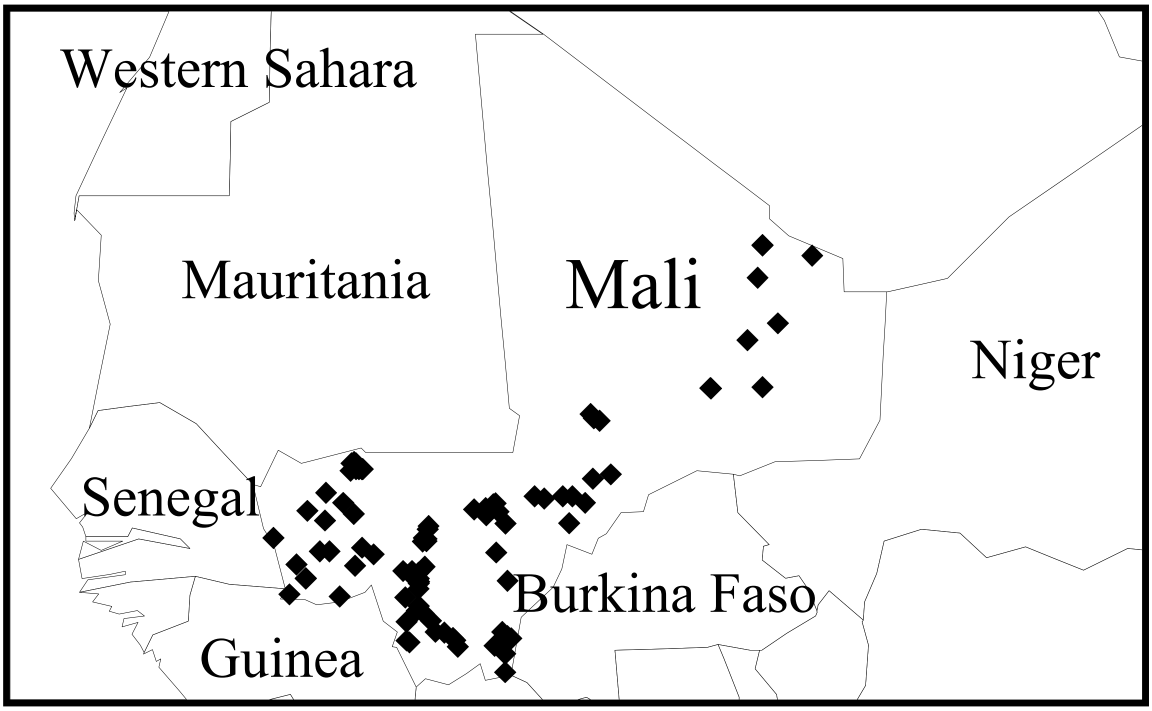
Data collection and data preparation

Malaria prevalence data were collated from surveys of childhood populations in Mali since 1960. Altogether 101 such surveys were identified yielding suitable estimates of malaria prevalence. The surveys represent historical data whose screening for inclusion in the MARA/AMRA database has been documented elsewhere. (MARA/AMRA Collaboration, 1998) For example surveys carried out amongst non-representative samples of respondents were excluded. Similarly, surveys conducted during known malaria epidemics were also excluded. In the absence of large scale intervention or climatic change it was assumed that malaria endemicity in Mali has remained reasonably stable. All the surveys were carried out in a confined locality so that the survey results collectively could be regarded as a cross-section of point referenced malaria prevalence observations.

For each survey the total sample size and number of individuals testing positive was known. The geographical co-ordinates of each survey were established using paper maps, electronic maps and global positioning systems. The distribution of surveys across Mali was uneven, with higher concentrations of surveys in more densely populated areas and in areas where malaria risk was perceived to be high. The location of each survey is shown in fig. 2.1.

For each of the survey co-ordinates long term climatic averages, normalised difference vegetation index (NDVI) (NDVI Image Bank Africa, 1991) and population density were obtained. A number of published data sets were available for this purpose. (Hutchinson *et al.* 1995; African Data Sampler, 1995). The resultant array of variables consisted of: monthly rainfall, monthly average maximum temperature, monthly average minimum temperature, monthly NDVI and population density. In addition, the number of months with rainfall in excess of 60mm (regarded as suitable for malaria transmission) was computed for each location. Using GIS, the distance to the nearest water body was also calculated.

Fig 2.1. Map showing survey sites



All climatic variables were available as long term averages for each calendar month, but not by individual year. The individual monthly averages of the climatic variables are highly correlated within climatic seasons. The question arises over what period climatic variables should be sensibly averaged. The shorter the aggregation period the stronger the likelihood of a high degree of serial autocorrelation in the values. For the purpose of selecting climatic variables for explaining the variation in malaria prevalence it was decided to average monthly climatic data over climatic seasons in order to reflect the variation in weather. Temperature and rainfall were averaged over 3 months periods, with the first quarter starting in December to coincide with the beginning of the dry season. The vegetation index NDVI was aggregated over two six-month periods corresponding approximately to the dry season (December to May) and the wet season (June to November) respectively.

Methods and results

The first stage of this analysis involved ordinary logistic regression analysis to determine the relationship between malaria prevalence and ecological predictors of malaria. From this a first prediction map for the whole of Mali was produced. In the second stage we investigated spatial pattern in the residuals of the model and used residual spatial dependence in the data to improve prediction at local level.

1. Regression analysis

The relationship between malaria parasite prevalence and each individual potential explanatory variable was first investigated by inspection of scatter-plots and by single variable regression analysis. Since parasite prevalence data are binomial fractions, a logistic regression model for grouped (blocked) data was used as is standard practice for the analysis of such data (Hosmer and Lemshow, 1989). Predictions of prevalence made from the logistic model will always fall within the interval 0 to 1. Larger surveys are implicitly accorded more weight than the smaller ones. The glm command in the statistical package STATA (Stata Corp, 1997) was used for the analysis.

Each of the explanatory variables was adjusted for all of the others by performing multiple regression in the usual way. Non-linearity in the relationship between parasite prevalence and a predictor variable was explored by adding polynomial terms and then grouping the values of continuous variables into categorical ones. Variable selection for the multiple logistic regression model was carried out by a combination of automatic (stepwise) procedures, goodness of fit criteria and by using judgement in selecting variables that explain malaria prevalence in terms of vector, host and parasite dynamics of malaria. An additional criterion for selection of the final model was the degree of spatial correlation of the model residuals (see below).

The final multiple logistic regression model contained four significant explanatory variables for the prediction of malaria prevalence. These were distance to water (categorical), average NDVI during the wet season(June to November, also categorical), number of months with more than 60 mm rainfall, and average maximum temperature during the quarter March to May. The detailed results are discussed in the companion paper. Table 2.1 summarises these results.

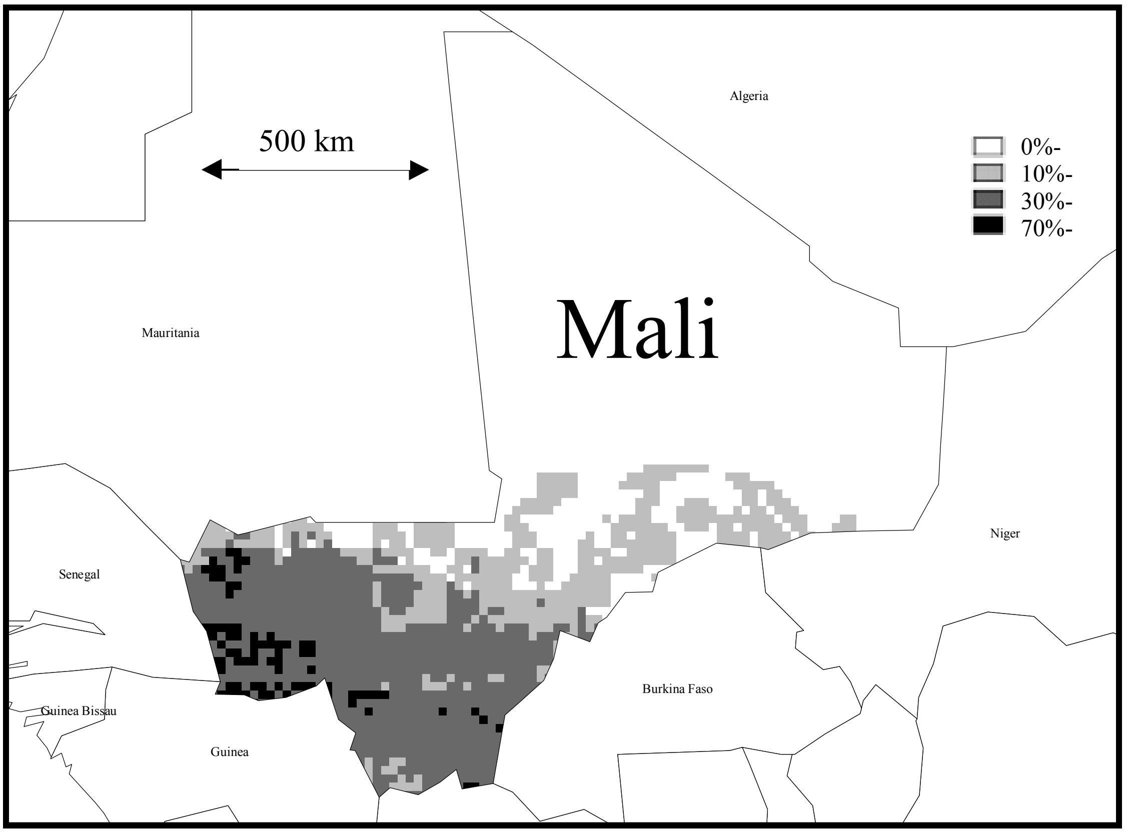
Table 2.1. Factors associated with malaria parasite prevalence. Adjusted odds ratios obtained by multiple logistic regression.

	Unadjusted		Adjusted	
Variable	**Odds Ratio**	**95% Confidence Interval**	**Odds Ratio**	**95% Confidence Interval**
Vegetation index(NDVI) in rainy season (relative to NDVI of 0.50 or less)				
0.50 > NDVI <=0.7	16.17	4.96 – 52.74	4.13	1.37 – 12.47
NDVI>0.7	36.30	11.00-119.74	4.90	1.29 – 18.55
Distance to water (relative to less than 4km)				
between 4 and 40 km	2.63	2.52 - 2.74	2.55	1.90 –3.423
more than 40km	0.19	0.17 – 0.23	0.70	0.24 – 2.11
Average maximum temperature, March to May				
Change per °C	0.75	0.63-0.88	1.40	1.14 – 1.72
Length of rainy season (months) change for each month of season length	1.62	1.59 – 1.64	1.76	1.33 – 2.34

The final model explains about 65% of the total variation in malaria if one takes the reduction in deviance as a measure of variation. It must be noted that the final model is ‘overdispersed’ i.e. the residual deviance is larger than would be expected for the number of degrees of freedom. This has been taken into account in the model by using a deviance based extra dispersion parameter , which results in inflating the standard errors of the model parameters by the square root of the dispersion factor (Littell *et al.* 1996). The inclusion criteria for the variables selected for the final model can therefore be regarded as conservative.

For each variable used in the model an image covering the whole of Mali was produced in the GIS package IDRISI (Clark Labs, 1998). In the case of categorical variables this entailed creating the equivalent boolean indicator variables as used in the statistical model. The prediction formula of the model was then used with the IDRISI image calculator to produce a prediction image. The predicted risks were then grouped into 4 categories: below 10%, from 10% to 30%, from 30% to 70% and above 70%. As an additional validation exercise, the predicted frequencies in these 4 categories were compared with those of the known values. Of the 101 survey results, 70 fall within their predicted group. The resulting map of malaria risk is shown in figure 2.2.

Fig. 2.2. Map of predicted malaria risk based on regression model only



2. Investigation of spatial pattern

For geographical data of the type of the malaria survey data, it is of interest to know whether the data display any spatial auto-correlation, i.e. do surveys that are near in space have values (of malaria prevalence) that are similar, in contrast to surveys that are far apart. Put another way, does nearness in space go together with nearness in value? This is important because spatially correlated data cannot be regarded as independent observations. If the analysis does not take account of the correlation structure of the data, the estimates obtained from modelling may be inaccurate.

The malaria prevalence data and the residuals of the regression model were analysed for the presence of spatial pattern. We used two separate methods to investigate spatial pattern: the D-statistic and the variogram.

The non-parametric D statistic (Walter, 1992) is a weighted average of rank differences in the values of observations, with the average taken over all pairs of points. If y_i refers to the rank of the value at any point *i*, then D is defined by

$$D = \frac{\sum\sum w_{ij}\left|y_i - y_j\right|}{\sum\sum w_{ij}}$$

Weights w_{ij} refer to pairs of points. Weights can be chosen in different ways, but should be large for points that are near in space and small or zero for points that are distant in space. In this analysis two approaches to assigning weights were used: a) all pairs of points that were within a particular distance of each other were assigned a weight of 1, all other points were assigned a weight of zero(binary neighbourhood weights); and b) the weight for each pair of points was assigned the inverse of the distance between them. If there is spatial autocorrelation, rank differences for nearby pairs of points will be small values, whilst the weights for these pairs of points will be large values. Distant pairs of points on the other hand would be expected to display large differences in rank, but these would be multiplied by low or zero values of weights. The overall effect is that D will be a smaller value if there is spatial pattern in the data, than if the ranks of points were randomly distributed i.e. near and far pairs of points showing no significant differences in rank difference.

A significance test was obtained by simulation. The simulation consists of randomly assigning ranks to the data points and then calculating D assuming the particular pattern of weights given by the spatial layout of the data. This process is repeated many times over, and the distribution of the simulated D is then compared to the actual value of D calculated from the observed data. This directly yields a p-value for significant evidence of spatial autocorrelation. For mutual binary weights an analytical test was used (Walter, 1994), which is computationally less demanding.

Since it is based on the ranks of the data rather than the actual values, the D-statistic is not dependent on normality of the data. In the malaria data (and generally) negative autocorrelation is not likely, since this would assume distant points to be more similar than near ones. Therefore, a one sided significance test was used, rejecting the null hypothesis of random spatial pattern if the value of D is sufficiently small.

The semi variogram (Oliver *et al.* 1992; Carrot and Valleron, 1992; Diggle *et al.* 1998) (often simply called the variogram) also measures spatial dependency, but there is no significance test associated with this measure. It is normally used to obtain a spatial model for kriging, but it also serves to examine spatial pattern. The semi variance γ(h) measures half the average squared difference between pairs of data values separated by the so-called lag distance, h.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i}^{N(h)} (y_i - y_j)^2$$

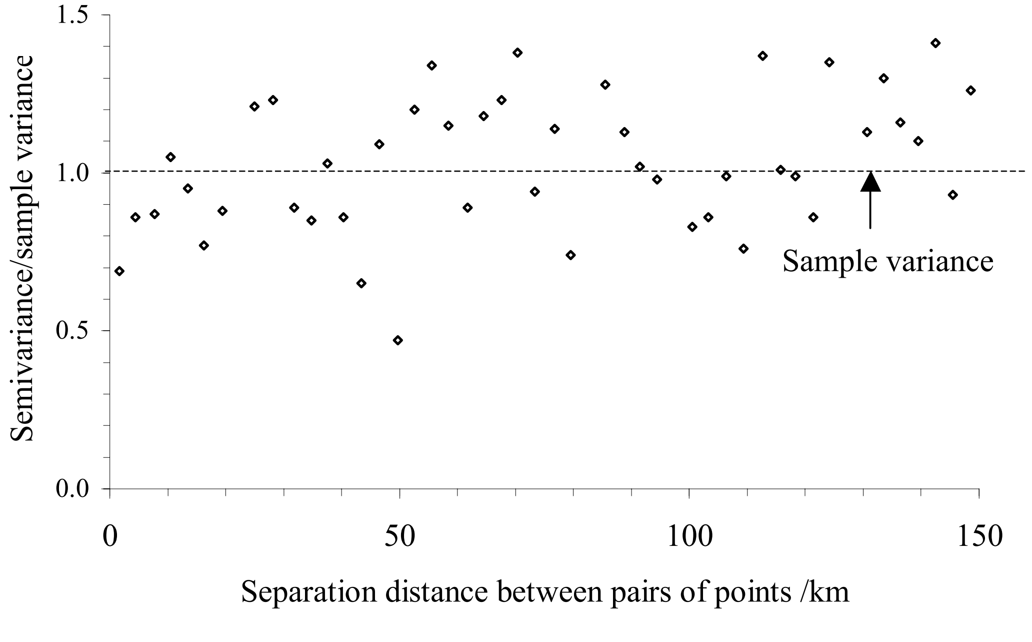
where N(h) is the number of pairs of sample points at a distance in the range h±h/2 from each other. Computations of γ(h) are repeated for 2h, 3h, 4h … etc. The semi-variogram is a plot of the semi-variance γ(h) against lag distance h. If the semi variance is markedly small for low values of h it is taken as an indication of spatial autocorrelation i.e. values at short distance from each other are more alike (less variable) than those at large distances.

Table 2.2 shows that the observed malaria prevalence for Mali is highly autocorrelated in space, as one would expect on account of its strong link with climatic factors. The model residuals still show evidence of spatial pattern, but some of this has been removed by the modelling process. This result holds whether spatial pattern is assessed using the D-statistic with inverse distance weights or binary neighbourhood weights. It can be seen from the p-value for binary weights, that the spatial pattern is more distinct over short distances. The semi-variogram of residuals (fig. 2.3) shows that there is some evidence of spatial correlation over short ranges of below 20km.

Table 2.2. Results of tests for autocorrelation by non-parametric D(p-values)

Type of weight for pairs of points	Autocorrelation of observed Malaria prevalence	Autocorrelation of model residuals
Binary neighbourhood weights, 50km	<0.0005	0.05
Binary neighbourhood weights, 15km	<0.0005	0.001
Inverse distance weights	<0.0005	0.006

Fig. 2.3. Variogram of model residuals (lag=8km)



3. Geo-statistical prediction (Kriging)

Prediction by kriging (Krige, 1966; Oliver *et al.* 1992; Carrot and Valleron, 1992; Diggle *et al.* 1998) is based on the assumption that covariance between points is entirely a function of distance between them as modeled by means of the variogram.

A further assumption is that the underlying mean of the quantity that is being predicted is constant (the assumption of stationarity).

Since the variogram describes the spatial dependence between the observed measurements as a function of the distance between them, it allows us to estimate the value of malaria prevalence at any point from the observed data. The value of prevalence, *Z*, at the coordinates(x_0, y_0) can be estimated from the *n* nearest sampling values $Z_{obs}(x_1, y_1)$, $Z_{obs}(x_2, y_2)$, …. $Z_{obs}(x_n, y_n)$ by the linear formula

$$\hat{Z}(x_0, y_0) = \sum_{i=1}^{n} a_i Z_{obs}(x_i, y_i)$$

The a_i are found by introducing a Lagrange multiplier λ and solving the system:

$$\sum_{i=1}^{n} a_i \gamma\,(h_{i,j}) + \lambda\ = \gamma\,(h_{i,0}), j = 1, \ldots.., n$$

under the constraint

$$\sum_{i=1}^{n} a_i = 1$$

where $h_{i,j}$ is the distance between two points located at (x_i,y_i) and (x_j,y_j), at which malaria prevalence has been measured, and $h_{j,0}$ is the distance between a measured point and the point (x_0,y_0) at which the prediction is to be made. γ(h) is the semi-variance as previously defined.

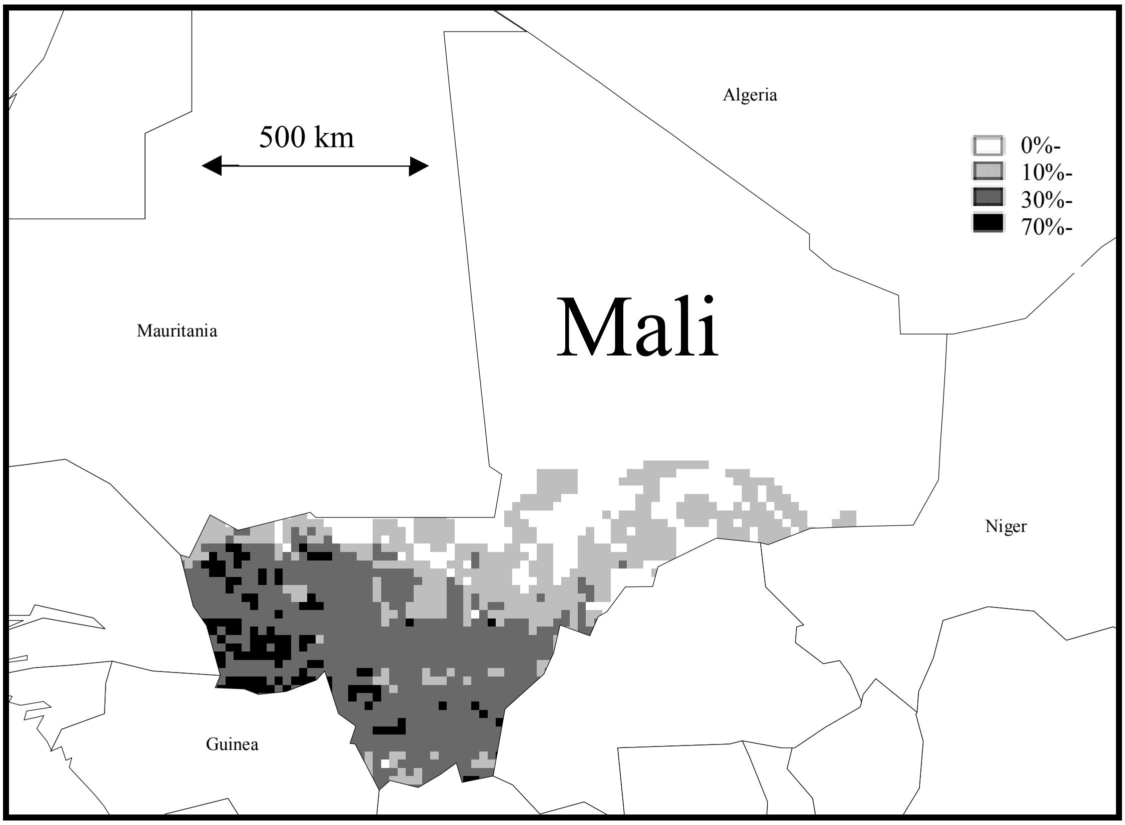
The extreme variation in the Mali malaria prevalence data invalidates the assumption that a common mean exists. There is clearly a need to take co-variates into account due to the strong association between malaria risk and climatic factors, and due to the wide variation of the latter across Mali. Residuals from the logit model should be free of covariate effects and the logit transformation will moderate any non-homogeneity in variance of the residuals.

Inspection of the variogram based on the residuals (fig 2.3) shows that there is spatial dependence (not taken into account by the model) over short distances up to about 15 or 20 km. A variogram of logit scale model residuals was constructed, confirming a short range spatial pattern up to distances of about 18km, although the relatively small number of pairs of points that are less than this distance apart makes the variogram less reliable in this region. This means that there is small area variation in malaria prevalence which cannot be modeled well by climatic factors presumably because these do not vary much over this short distance.

Kriging performed on residuals is equivalent to kriging a variable which has an underlying (stationary) mean of zero. To carry out this process residuals for all observed points were calculated on the logit (ln(p/1-p)) scale of the logistic model. Spatial dependence of these was modeled using the previously constructed variogram. An exponential model was fitted to the variogram using a sill and nugget of 0.7 and 0.4 respectively, and a range of 18 km. This geo-statistical model was then used in the kriging procedure of the package GEO-EAS (Geostatistical Environmental Assessment Software, 1991) to map predictions of residuals in an 18 km radius around each observation. These logit scale 'kriged' residual predictions were then added to the logit scale predicted values produced from the original logistic model. The resultant map predictions were transformed back to prevalences in the usual way (exp(Xβ+kriged residuals)/(1+exp(Xβ+kriged residuals))) to produce a new prediction map (fig.2.4). This map takes into account local spatial dependence and allows local deviation from the prediction of the logistic model.

To see how much improvement was achieved by local kriging, another map was produced showing the difference between the final map (fig. 2.4), and the original map produced by regression only (fig 2.2). This difference map is shown in fig 2.5. The new map results in an improvement of 5 additional surveys whose observed prevalence falls within the predicted prevalence bands of the map. (We would expect that this can be improved upon with a higher grid resolution). A weighted inter-rater kappa statistic (Altman, 1991) for agreement between observed and predicted map values for the surveys shows an improvement from 0.624 for the map based on regression only to 0.727 for the map based on the 2 stage procedure. This takes into account not only agreement/non agreement between observed and expected prevalence bands, but also the seriousness of discordance, if any.

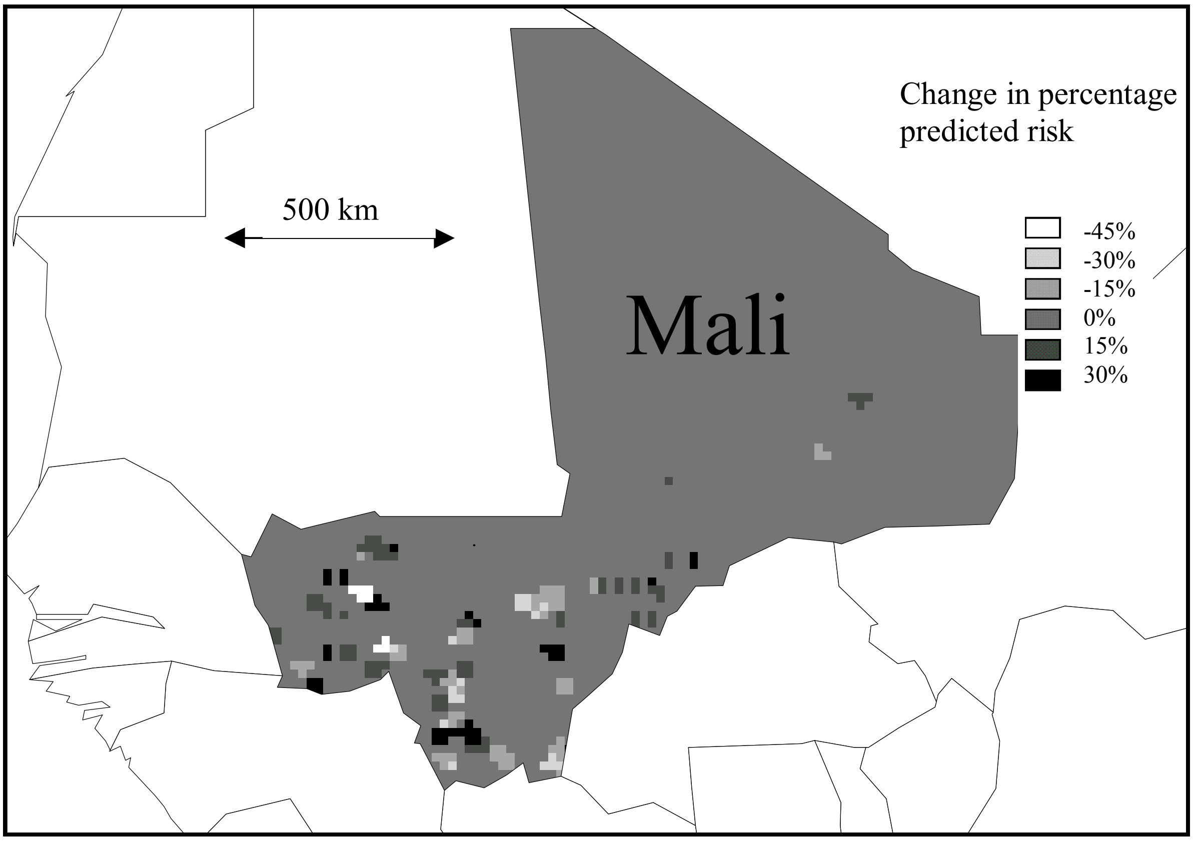
Fig. 2.4. Map of predicted malaria risk using regression model plus kriging



Discussion

The final malaria prediction map is in agreement with eco-geographical descriptive epidemiology of malaria in Mali (Doumbo *et al.* 1989). Kriging has significantly improved the prediction of malaria risk in parts of the map, particularly where the density of surveys is high, which coincides with areas of high risk. However, given that the data used for obtaining the model are not a random sample of the population or a spatially well distributed set of sampling points, one needs to be cautious in extrapolating the predicted risk to points outside the data set as has been done here.

Fig 2.5. Map showing difference in predicted malaria risk as a result of kriging



A concern with spatial data is the potential for spatial correlation in the observations, which could lead to incorrect estimates. Spatial clustering of disease is almost inevitable since human populations generally live in spatial clusters rather than random distribution of space. An infectious disease that is heavily associated with climatic variables is likely to be spatially clustered even if population distribution was not clustered. The model derived here explains some of the spatial pattern of malaria risk, but there is still significant spatial correlation, particularly over short distances of under 20km. (This result holds for differing ways of defining ‘nearness’ in the D-statistic and is confirmed by the variogram method.) The reduction in spatial structure in the residuals lends credence to the correctness of the model.

Overdispersion in the logistic model does indicate that there may be important covariates missing from the model. Some of these unknown predictors are likely to be spatially distributed, particularly at a local level.

Kriging with a non-stationary mean ('universal kriging') is a refinement of ordinary kriging in that it allows for co-variate adjustment by means of regression modelling (Diggle *et al*. 1998). This would be more appropriate in the case of malaria risk where we know that climatic factors are strong predictors. Since the mean prevalence is now a function of the co-variates, rather than a constant, the model assumptions would not be violated as in the case of ordinary kriging. Universal kriging offers the most comprehensive approach to the mapping of malaria risk: it uses the values of the co-variates (climate data) at the point at which the prediction has to be made, as well as the position of the point in relation to points at which observed values of malaria risk are available. Universal kriging applied to generalised linear models such as the logistic model, is currently not available and we have therefore not been able to apply it as such.

The two stage approach that we used offers an appealing alternative to universal kriging and it is somewhat similar in approach. The non-spatial model provides the covariate adjustment and prediction of mean risk in an area. It thereby allows for non-stationarity in the data by modelling the long range differentials in the malaria risk pattern. Kriging of the resulting residuals allows for local deviation from the predicted mean and for spatial dependence in points that are close together. In the MARA project it is unlikely that local predictors affecting malaria risk over and above what is predicted by climatic factors will ever be available. For this reason local variation from the more global area prediction has to be taken into account by spatial modelling.

Whilst the kriging process will give minimised unbiased prediction error (of residuals) on the logit scale, this cannot be guaranteed for the backtransformed predictions (Cressie, 1993). However, the kriged logit scale residuals are only a component (in most cases a small component) of the linear predictor which is backtransformed to produce the final prediction for the point on the map.

Prediction based on regression alone has a tendency to produce predicted values that are pulled towards the mean. For example, two observations in different parts of the country with very similar climatic data may differ in their observed malaria prevalence value. Regression modelling would predict for these two places a value close to the mean prevalence of the two points. This would result in large residuals. Kriging the residuals and adding the predicted residuals to the model predictions will produce predictions that are closer to the observed prevalences in each neighbourhood, particularly if the deviation from the model prediction is supported by other points in the neighbourhood.

As one might expect therefore, the range of final predictions from the two stage method is wider than that produced by the regression model alone, with predictions ranging from about 0% to 92% (compared to a range of 0% to 80% for the logistic model alone). As can be seen from the new prediction map (fig 2.4) and the difference map (fig 2.5), the changes brought about by this process are confined to areas around most of the survey locations. For the rest of the map the data are too sparse to be affected by this process i.e. most places are more than 18 km removed from the nearest survey.

A problem with this approach is that often there are insufficient data points to give us a good basis for estimating the local variability. In the case of malaria maps this problem is less serious in those areas where malaria prevalence is highest, simply because the frequency of surveys is greatest in these areas. The map is therefore likely to be at its most accurate where it matters most: in places where malaria prevalence is high.

It should be noted that universal kriging might have resulted in a different model to the one obtained here, since it attempts to simultaneously obtain good estimation of covariate effects and allow for residual spatial pattern. In this particular example, however, the residual spatial correlation was weak and therefore we would not expect that universal kriging would have produced a model that differs much from the present one. We are currently investigating an iterative approach that would be applicable in situations were the residual spatial pattern is substantial.

The specification of a nugget variance makes allowance for measurement error at a location. This avoids the prediction ‘honouring’ every observation, which would result in a very spiky map. Future development in this area should include a method of weighting the observations in such a way that large surveys draw the map prediction closer to their observed value than small surveys.

Additional further work in this area would be to develop ‘goodness of fit’ indicators for this two stage method. For example, how much of the overdispersion in the model has been taken up by local kriging? What proportion of variation in the data is ‘explained’ by kriging? It would also be important to produce combined prediction errors for the whole map, taking into account both components of the process of prediction.

In conclusion, our view is that the model produced here is a reasonable representation of malaria risk in Mali. The reduction of residual spatial pattern enhances our confidence in the fidelity of the model and residual spatial dependence has been modeled by kriging wherever the density of observed points allows for this. Kriging has been made possible by ‘leveling’ the map through the regression model, and applying the kriging process to the residuals. The final predictions make sense from the entomological perspective. However, a more systematic approach to this work in future would be a full mixed model with universal kriging to take account of spatial pattern.

Chapter 3

Use of generalised linear mixed models in the spatial analysis of small area malaria incidence rates in KwaZulu Natal, South Africa

Kleinschmidt I,[1] Sharp BL,[1] Clarke GPY,[2] Curtis B,[1] Fraser C.[1]

Abbreviations

Generalised Linear Mixed Model (GLMM)
Geographic Information System (GIS)

[1] Medical Research Council (South Africa), 771 Umbilo Road, Congella, Durban 4001, South Africa.

[2] Department of Statistics and Biometry, University of Natal, Pietermaritzburg, South Africa.

This paper has been published in *American Journal of Epidemiology* (2001); 153(12):1213-21.

Summary

Spatial statistical analysis of small area malaria incidence rates of the northern-most districts of KwaZulu Natal in South Africa was undertaken in order to identify factors that may explain very strong heterogeneity in the rates. A method for adjusting the results of the regression analysis for strong spatial correlation in the rates by making use of generalised linear mixed models and variogram methods is described. The results of the spatially adjusted multiple regression analysis show that malaria incidence is significantly positively associated with higher winter rainfall and higher average maximum temperature and significantly negatively associated with increasing distance from water bodies. The statistical model is used to produce a map of predicted malaria incidence in the area taking account of local variation from the model prediction where this is supported by the data. The predictor variables show that even small differences in climate can have very marked effects on malaria transmission intensities, even in areas that have been subject to malaria control for many years. These results have important implications for malaria control program activities in the area.

Keywords: malaria incidence, spatial analysis, statistical models, variogram

The districts of Ngwavuma and Ubombo in the northern part of the province of KwaZulu-Natal experience the highest malaria incidence rates in South Africa (Sharp and Le Sueur, 1996). Very large variation in malaria incidence rates within these districts has hitherto not been properly accounted for although it has been ascribed as much to human factors such as cross-border migration as to natural forces such as climate and environment.

The purpose of this study has been to undertake a spatial statistical analysis of malaria incidence in order to identify important predictor variables and to produce an incidence map of the area that illustrates the variation of malaria risk. A secondary but important aim of this analysis has been to advance methodology for the spatial analysis and modelling of malaria transmission data in the context of the MARA/ARMA project (MARA/ARMA, 1998) which amongst other objectives seeks to produce maps of malaria risk for the continent of Africa (Snow *et al.* 1998; Kleinschmidt *et al.* 2000).

In Africa the predominant species of the malaria causing-parasite is *Plasmodium falciparum*. When a person is bitten by an infected anopheles mosquito, the parasite, called sporozoite at this stage of its cycle, enters the human body via the saliva of the mosquito which is injected into the blood. The parasites multiply in the liver, and re-invade the blood via red blood cells as merozoites. The merozoites multiply sexually and some of them form into gametocytes. Uninfected anopheles mosquitoes become infected if they feed on a person with gametocytes in their blood. The gametocytes will undergo another phase of reproduction inside the insect called the "sporogony" cycle. At the end of this cycle the mosquito will become infective as a new generation of sporozoites are able to infect another human host.

Malaria as a disease is therefore closely bound to conditions which favour the survival of the anopheles mosquito and the life cycle of the parasite. These conditions are predominantly determined by climatic factors, by vegetation coverage and by the vector's access to water surfaces for breeding requirements (Molineaux, 1988; Gillies and De Meillon, 1968; Ghebreyesus, 1999). Human population movement from areas where malaria is endemic to areas where the disease has been at least partially eradicated can also contribute to malaria transmission (Martens and Hall, 2000).

Accurate knowledge of the distribution of malaria is an important tool in planning and evaluating malaria control (Snow *et al.* 1996). Explaining this distribution is important since it provides a rationale for interventions, and because it makes it possible to predict transmission intensity in places where it has not been measured. The area under consideration has been subject to insecticide house-spraying over several decades (Sharp and Le Sueur, 1996), resulting in reduced incidence rates compared to the era before the introduction of malaria control measures (Sharp *et al.* 1988). In recent years incidence rates have risen again steeply. The area is situated on the southern fringe of climatic suitability for endemic malaria distribution in Africa (Craig *et al.* 1999). Summer rainfall (six-monthly average = 68 mm per month) and temperatures (average maximum daily temperatures = 29.4°C) are generally suitable for malaria transmission. On the other hand, winter conditions are sub-optimal and could be limiting transmission (Molineaux, 1988) (average rainfall = 19 mm per month and average maximum daily temperature = 25.9°C). We sought to test the hypothesis that it is the spatial distribution of climatic conditions in winter that accounts for much of the variation in malaria transmission intensity in this region. To this end we undertook a spatial statistical analysis of small area malaria incidence rates in relation to rainfall and temperature. We included proximity to permanent water bodies and distance to the border with Mozambique in the analysis due to their potential confounding effects. The latter was used as a proxy for migration from Mozambique where routine malaria control had not been implemented. The water bodies included all permanent fresh water surfaces, except rivers.

In both Ngwavuma and Ubombo districts a small area malaria incidence reporting system has been in operation for a number of years as part of the provincial malaria control programme (Sharp *et al.* 1999). A component of the control strategy is to identify and treat all infected individuals. Passive and active case finding is therefore practised. The latter consists of screening measures by which teams go into the community to encourage individuals who may be suspected of having malaria, to be tested. Whilst this system may not achieve 100 percent coverage, it is thought to identify the vast majority of cases. Low levels of exposure to *Plasmodium falciparum* in the past have made it unlikely that the host population possess naturally acquired immunity, apart from recent immigrants from Mozambique. Individuals are therefore unlikely to possess clinical tolerance to parasite infection and to recover from it without treatment (Molineaux, 1988, pp 936-938).

A population census at homestead level was carried out in 1994 by the control programme, thus making it possible with the use of Geographic Information Systems (GIS) to derive population counts for the same small areas for which cases were being reported. It was therefore decided to base the analysis on the malaria season from July 1st 1994 to June 30th 1995 since this would have the most reliable denominator data available for the calculation of incidence rates.

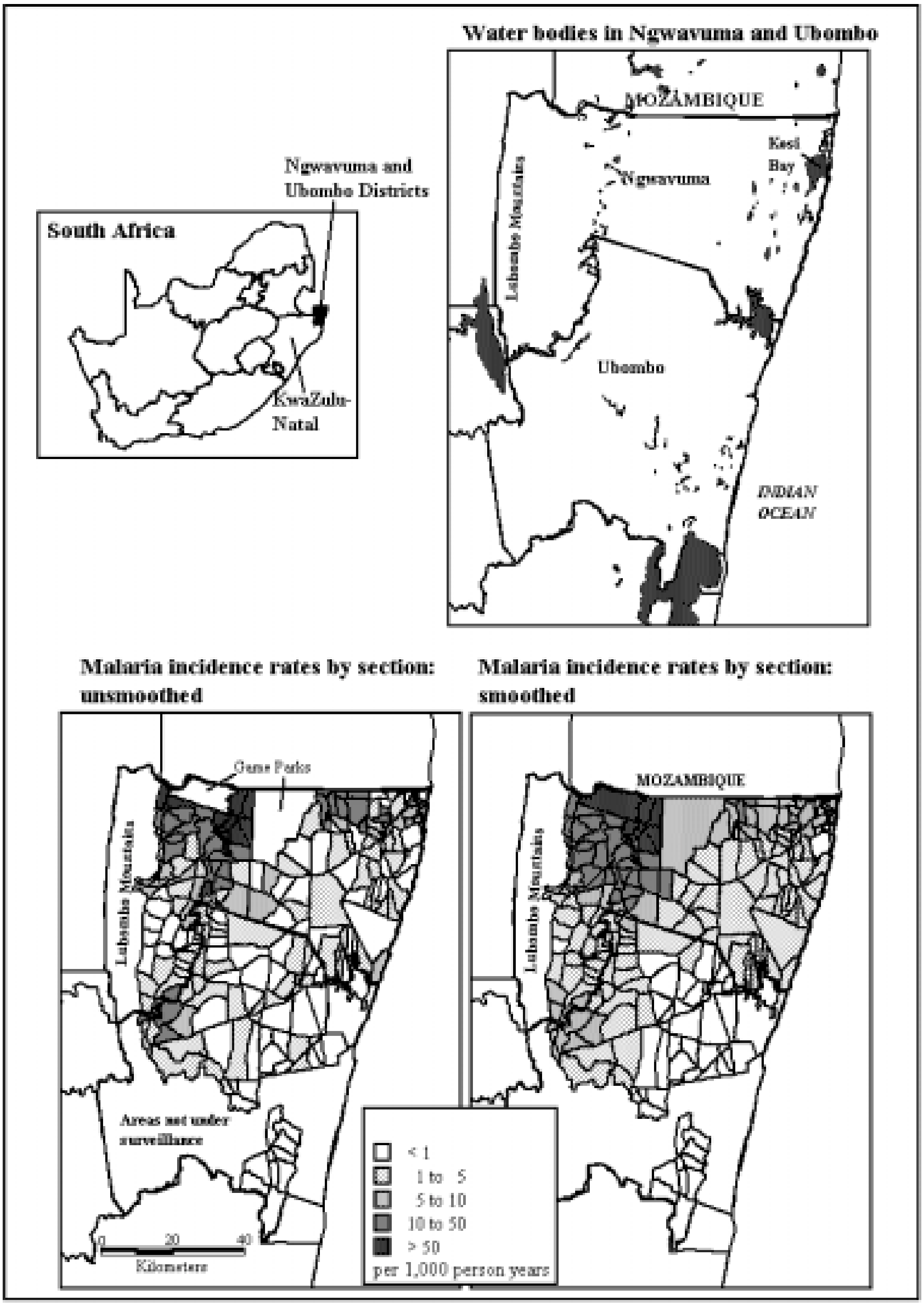
Data

Malaria cases (by passive and active detection) for the season mid-1994 to mid-1995 for the districts of Ngwavuma and Ubombo in northern Kwazulu Natal were extracted from the malaria control programme database and allocated to 220 magisterial subdivisions referred to as sections. Population totals for each section were obtained from the same source, based on a census carried out during the year 1994/95.

The total number of cases was 2,418 including 400 patients whose place of residence was not known and who were therefore excluded from all of the spatial analysis. Many of the latter are likely to be imported cases from Mozambique with no fixed residence in the area. The total population count for the study area was 241,397 persons, resulting in a crude overall incidence rate of 8.4 per 1000 person years.

Population per section ranged from 11 to 5,482 persons (median=858, mean = 1102, SD = 886). Area per section ranged from 1.1 km^2 to 263.2 km^2 (median=19.0, mean = 25.9, SD= 27.5).

Figure 3.1. Water bodies, Malaria Incidence and Smoothed Malaria Incidence for the population of the districts of Ngwavuma and Ubombo, KwaZulu Natal, July 1994 to June 1995 (Source: KwaZulu Natal Malaria Control Programme)



For each section a crude incidence rate for the year 1994/1995 was calculated. For the purpose of calculating rates it was assumed that the entire population of an area was exposed to malaria risk for the period that was studied i.e. that each individual contributed exactly one person year of exposure. Crude incidence rates per section ranged from 0 to 306 cases per 1000 person-years (mean = 11.2, SD = 32.0, interquartile range 0 to 8 per 1000). In all but 14 of the sections the rate was less than 50 per 1000 person years. In 66 sections there were no cases. The map in figure 3.1 shows the distribution of incidence rates by section.

Long term averages of climatic data (rainfall, average daily maximum temperature) by calendar month were obtained from climatic databases for Africa (Hutchinson *et al*. 1995). By means of GIS (Clark Labs, 1998) the average value of each variable was calculated for each of the sections by averaging the pixel values of the variable over the area of the section. The monthly averages were combined into six-monthly averages to coincide with the winter (April to September) and summer (October to March) seasons. Average daily maximum temperature in winter was highly correlated with average daily maximum temperature in summer (correlation coefficient =0.995). We therefore combined these two variables into a single average daily maximum temperature for the whole year. We also calculated the average monthly rainfall and average maximum daily temperature for the midwinter months of June and July combined.

The centroid of each section was derived, also by GIS. For each of these centroids distance to water bodies (lakes, reservoirs and dams – see figure 3.1) and distance to the Mozambican border were calculated.

Analysis and results

Smoothed incidence map

Relatively small population totals per section result in rates that are subject to considerable random error (Cuzick and Elliot, 1992). The map of smoothed rates in figure 3.1 was produced using the smoothing method proposed by Kafadar (Kafadar, 1996). The smoothing effect of the incidence rate of one area on another was limited to areas whose centroids were less than 15km apart.

Heterogeneity test

A heterogeneity test due to Potthoff-Whittinghill (Potthoff and Whittinghill, 1966) was applied to the observed rates, using Monte Carlo simulation to distribute cases randomly amongst areas, but in proportion to populations in each area. This showed that the observed distribution of cases is very unlikely to have come about due to chance alone ($p<0.0001$), given a null hypothesis of no differences in underlying risk between areas.

Test for spatial pattern

The non-parametric D-statistic (Walter, 1994) was used to test the observed incidence rates of the sections for significant spatial pattern. The statistic is calculated as a weighted average of rank differences in incidence rates for all pairs of sections, the weights favouring pairs of areas in close proximity to each other. This results in a low value of D if there is marked spatial correlation. We used binary mutual neighbourhood weights (i.e. a weight of 1 if two areas share a common boundary, 0 otherwise) which were obtained using a GIS program written for this purpose. The calculated value of D for the observed incidence rates was 34.3, whilst the expected value of D given the adjacency configuration of the areas is 194 (SD=5.67), indicating that there is strong evidence of spatial pattern in these rates ($p<0.0001$), as one might expect.

Association between malaria incidence and climatic and environmental factors

In order to determine climatic and environmental co-variates that are associated with observed incidence rates whilst allowing for the spatial correlation of the data, an iterative approach of variogram methods and generalised linear mixed models(GLMM) was used (Littell *et al*. 1996). We have chosen this approach since we would expect it to inherit the optimal properties (Best Linear Unbiased Prediction) of the mixed model.

Appendix 1 gives details of how the spatial correlation of the data can be derived from a variogram of model residuals and how this spatial correlation can be taken into account by the SAS (SAS System 1996) implementation of the GLMM. The overall strategy for modelling spatially correlated incidence data was as follows:

The counts of malaria cases for each section were represented by a GLMM with a Poisson distribution, a logarithmic link function, a correlated error structure as described in appendix 1, and with the population of the section as an offset. (The offset is a term added to the Poisson model on the logarithmic scale so that the count of cases is adjusted for population size.) Potential explanatory variables were average values for winter and summer seasons separately of monthly rainfall, annual average daily maximum temperature, distance to nearest water body and distance to the Mozambican border. Average monthly rainfall and average daily maximum temperature combined for the midwinter months June and July were also included as candidate explanatory variables. The specification of the correlated error structure in the GLMM due to spatial pattern in the data, was iteratively improved by examining the model residuals and re-specifying the covariance matrix, as detailed in appendix 1.

The final model contained three variables, namely annual average daily maximum temperature, average monthly rainfall during March through to September, and distance to water bodies, which were significantly associated with malaria incidence. Once these variables had been selected, we investigated for each of the three variables in turn whether any of a set of transformations of the variable would result in a significant reduction in deviance in the multiple regression Poisson model, following a procedure suggested by Royston et al (Royston *et al*.1999). The transformation producing the biggest reduction in residual deviance was chosen if this reduction in deviance was at least 3.84 compared to the untransformed variable. Transformations that were tried for each variable x were $1/x^2$, $1/x$, $1/x^{0.5}$, ln(x), $x^{0.5}$ and x^2, without simultaneous inclusion of the untransformed variable (x^1) in the model. The transformations are useful to represent relationships in which the log incidence rate increases more rapidly than a straight line at low values of x and more slowly at high values, or vice versa. Any resulting improvement in the fit of the model would enhance its utility for the purpose of prediction.

The variogram of residuals of the final model shows that there is residual spatial dependence after fitting the model (Figure 3.2). The specification of a covariance structure in the GLMM based on the spatial correlation of residuals ensures that the results are adjusted for this spatial correlation.

The incidence rate ratios of the final model are shown in table 3.1, together with the estimated effect these variables by themselves would have on incidence rates. Distance to water and winter rainfall were fitted as square root and logarithmic transformations respectively, whereas maximum temperature fitted best without any transformation. To facilitate the reporting and interpretation of transformed variables, we have calculated the model based incidence rate ratio for the transformed variable for two separate values of the variable, relative to a third (referent) value of the variable (Royston *et al.* 1999). The referent value chosen was the mean value of the untransformed variable minus one standard deviation, with the other two values being the mean itself and the mean plus one standard deviation. The two incidence rate ratios demonstrating the association between the transformed variable and malaria incidence are therefore one and two standard deviations removed from the referent value, and give some indication of the non-linear nature of the association.

Throughout this analysis there was evidence of considerable overdispersion of the Poisson model with a variance greater than the mean (dispersion factor in the final model = 11.5). This was accounted for by inflating the standard errors in the model by the square root of the overdispersion factor (Littell *et al.* 1996 pp 445). Overdispersion indicates other unmeasured sources of variation that could not be taken into account.

TABLE 3.1. CLIMATIC and topographic factors and their effect on malaria incidence in the districts of Ngwavuma and Ubombo. For the multiple regression model, incidence rate ratios (IRR) have been calculated from the model:
log(IRR) = 4.91*TMAX + 7.70*ln(RAI0409) – 0.46*sqrt(DSTWTR)
All results have been adjusted for spatial autocorrelation.

		Malaria incidence			
		Single variable analysis		Poisson multiple regression analysis	
Climatic/topographic variable	**Reference point**	**Incidence Rate Ratio**	**95% Confidence Interval**	**Incidence Rate Ratio**	**95% Confidence Interval**
Average daily maximum temperature[1] (TMAX), Change per °C		6.7 (p<0.0001)	3.5 - 13.0	135.6 (p<0.0001)	61.2 – 300.4
Average monthly rainfall, March to September (RAI0409), mm	13 (reference)	1		1	
	19	0.48	0.29-0.80	18.6	10.0 – 34.4
	25	0.29 (p = 0.005)	0.12 – 0.68	153.8 (p<0.0001)	53.1 – 445.5
Distance to water (DSTWTR), km	3 (reference)	1		1	
	8	0.43	0.32 – 0.59	0.60	0.50 – 0.73
	13	0.24 (p<0.0001)	0.14 – 0.41	0.43 (p<0.0001)	0.31 – 0.58

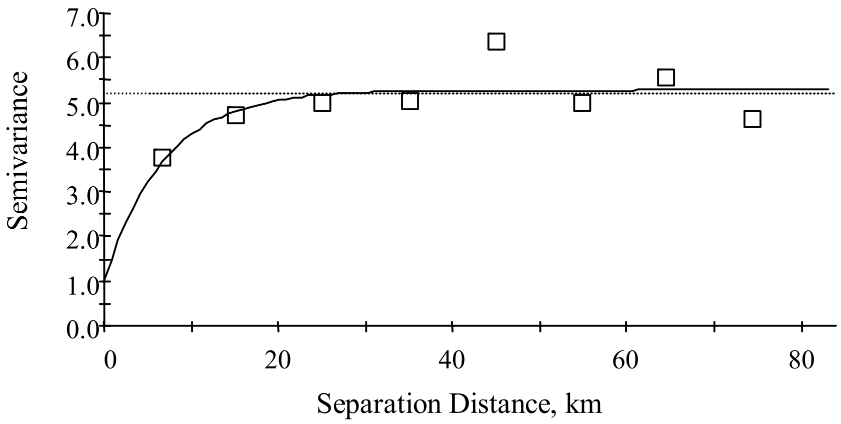
[1]Note: Range: 26.5 °C to 29 °C

Grouping observed and predicted incidence rates into categories of less than 1 per 1000, between 1 and 5 per 1000, between 5 and 10 per 1000, between 10 and 50 per 1000 and more than 50 per 1000, gives a weighted kappa statistic on agreement between observed and predicted categories of 83%.

Model based prediction map

Using the model derived above, a map of predicted malaria incidence rates was produced following a two-stage approach previously developed for the mapping of malaria prevalences (Chapter 2). Kriged residuals from the final model were added to the model predictions and exponentiated (exp(linear prediction+kriged residual)) to produce a map of incidence rates that takes into account local deviation from the model predictions where this is supported by the data (Figure 3.3).

Figure 3.2. Variogram of deviance residuals of final model of incidence rates for the population of the districts of Ngwavuma and Ubombo, KwaZulu Natal for the season July 1994 to June 1995



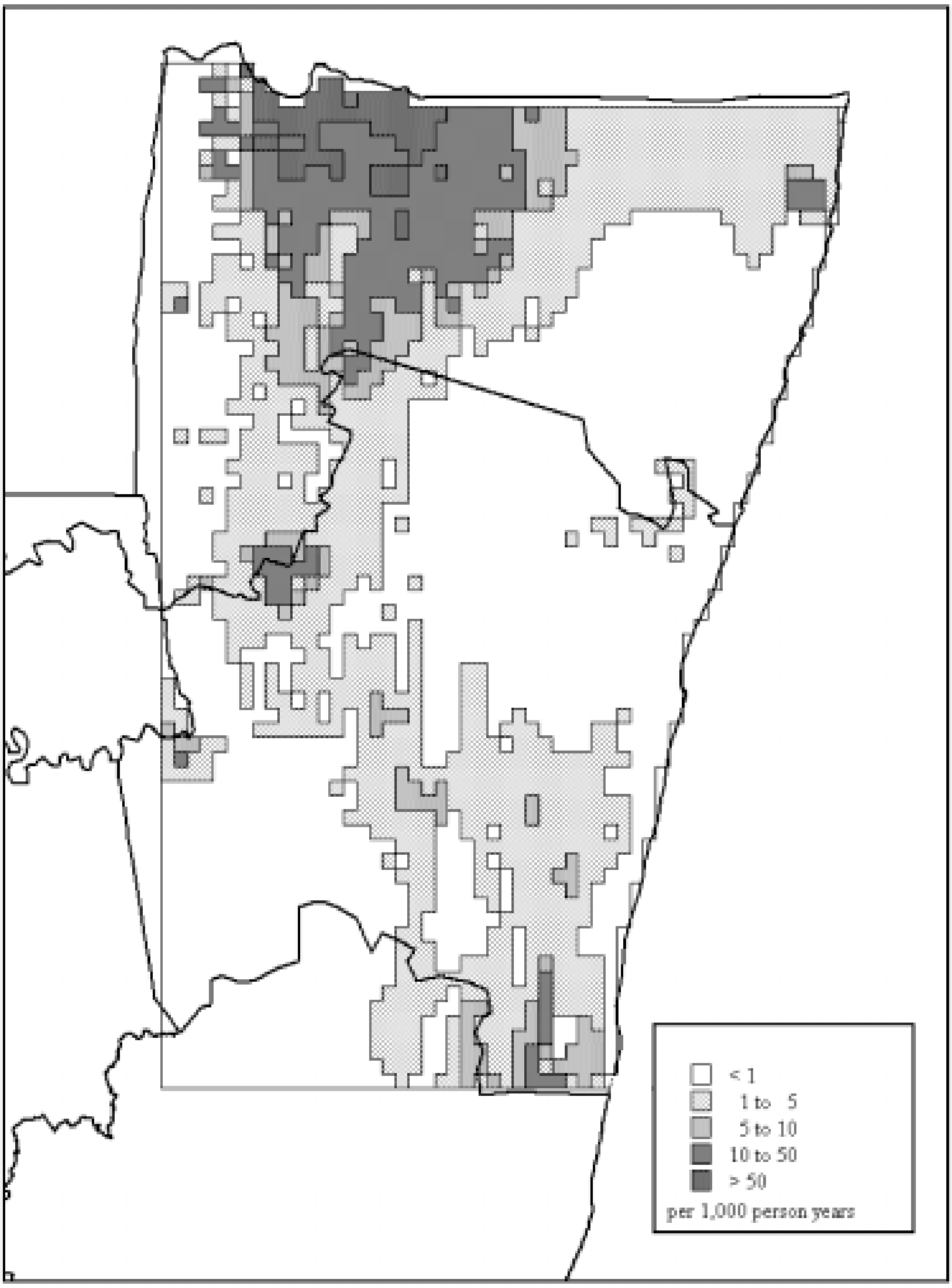
Discussion

Historical comparisons of malaria incidence with the period prior to the introduction of malaria control (Sharp and Le Sueur, 1996) and comparisons with neighbouring countries suggest that house spraying has resulted in a significant reduction in transmission intensity in KwaZulu Natal. Our analysis shows that even at this relatively low level of transmission, the overwhelming factors that are related to variation in malaria incidence are climatic factors that are also associated with malaria distribution in endemic areas (Sharp and Le Sueur, 1996; Molineaux, 1988). The prediction map (Figure 3.3) shows that, the combination of the three factors of proximity to water bodies, winter rainfall and maximum temperature are closely associated with high malaria risk in the north west of the study area and moderately high transmission intensity in the border areas along the north and west of the area. Although cross-border migration of infected individuals and the proximity of uncontrolled areas across the border may further add to transmission intensity in border areas, our model shows that natural factors go some way towards explaining the higher transmission levels found in these areas. Distance to the Mozambican border, the surrogate measure of cross-border migration, summer rainfall, average rainfall and average daily maximum temperature in June and July were not significant once other variables had been included in the model.

Our model underlines the importance of adjusting for confounding effects when investigating the association between malaria incidence and climatic factors. Our study area is marked by significantly higher winter rainfall along the coast than in the interior. The coastal areas on the other hand experience lower daily maximum temperatures than inland areas. Both these factors are closely associated with malaria incidence, but due to the negative correlation between winter rainfall and maximum temperature (correlation coefficient = -0.89), the former on its own is negatively associated with malaria incidence (table 3.1). The same negative confounding between rainfall and temperature is responsible for the large difference between the sizes of the adjusted and unadjusted rate ratio of maximum temperature. The multiple regression model shows that after correcting for maximum temperature, winter rainfall is positively associated with malaria incidence. Similarly the effect of temperature, after adjusting for the effect of rainfall and to a lesser extent distance to water, is much larger than the single predictor relationship between temperature and malaria incidence would suggest. The large magnitude in rate ratios should be seen in the context of very low incidence rates in much of the study area, with the areas of higher risk constituting incidence rates many tens of times higher. For this reason the incidence rate ratios cannot be extrapolated to situations of higher endemicity levels found in areas further north. Furthermore, the likely absence of immunity in the study population has the effect that any increase in transmission intensity is directly translated into higher incidence rates.

The very high collinearity between average daily maximum temperature in summer and in winter has meant that we were unable to test the hypothesis that variation in average daily maximum temperatures in winter is a factor in determining variation in malaria risk. In the study area there is virtually no variation in average daily maximum temperature in winter independent of average daily maximum temperature in summer.

Figure 3.3. Map showing model based prediction of malaria incidence for the population of the districts of Ngwavuma and Ubombo, KwaZulu Natal for the season July 1994 to June 1995



According to our results, average daily maximum temperature has a large effect on malaria incidence with a rate ratio of 135.6 per degree centigrade. This study area is at the southern limit of malaria distribution and an association with temperature, particularly winter temperature, would not be surprising due to its effect on both parasite and vector development (Craig *et al*. 1999). In the study area annual average daily maximum temperatures vary from 26.5°C to 29°C, with some places having much lower maxima over some of the winter period. The water temperature of small water bodies around which the vectors over-winter is likely to be well below the maximum daily air temperature. Low water temperatures have been shown to have a significant negative impact on mosquito abundance due to long larval duration (Le Sueur, 1991), whilst fairly high air temperatures for at least part of the day keep adult life spans reasonably short. At the same time any reduction in air temperature will lead to an increase in the incubation period in the vector (the time needed for production of sporozoites that infect man in the female mosquito) (Molineaux, 1988) thereby reducing the chances of adult vectors surviving long enough to become infective. This result is therefore in accordance with the Macdonald model which expresses the dependence of the basic reproduction rate of malaria in terms of the daily survival probability of the vector and the length of the incubation period (Bruce-Chwatt, 1980; Molineaux, 1988 p. 923).

Winter rainfall obviously affects winter vegetation and sustains smaller water bodies throughout winter in which vector populations can survive. The range in winter rainfall is quite large in the study area and some places are fairly dry, with averages below 15mm per month during the winter half of the year. This would suggest a lengthy period during which no breeding sites are available in these areas. The non-linear nature of the relationship between winter rainfall and malaria incidence suggests that some places are below a threshold which is needed for additional rain to be associated with a large increase in malaria incidence. The distribution of winter rainfall in this area is a plausible constraint on vector survival in winter, and hence on malaria transmission.

Under these circumstances mosquitoes may be forced to overwinter around permanent water bodies from which populations can spread out as more transient water bodies form after early summer rains. Permanent water bodies may constitute a last resort as breeding habitat for *Anopheles gambiae* due to the presence of predators (Le Sueur and Sharp, 1988). However these water bodies are often surrounded by smaller less permanent water bodies in their vicinity which constitute more suitable breeding sites. Whatever the dynamics of survival around these sites, our model shows that their proximity constitutes an additional risk factor for malaria. As a result of the non-linearity of the relationship the effect of unit increase in distance from water bodies attenuates with distance from these.

Spatial correlation in the residuals of the model up to 20km between pairs of sections (Figure 3.2) suggests that there are unmeasured spatially structured sources of variation not accounted for by the model. The range of this spatial pattern in residuals is in excess of the flight distances of mosquitoes (Gillies and Meillon, 1968 pp 213). Some of this unexplained variation could be due to deviation of climatic effects during the particular study year from the pattern of the long term averages which were used in the analysis. Other possible sources of spatially dependent variation in incidence rates may be the diligence with which insecticide spraying was carried out by teams in particular sections, the emergence of insecticide and drug resistance as well as differences in the host populations.

The two climatic factors of significance in our analysis are both factors that influence the survival of vector populations in winter. An important conclusion therefore is that control activities in this area can be made more effective by focussing on eradication of winter populations of vectors. Any areas with a combination of high average daily maximum temperature and high winter rainfall are at risk. The additional risk posed by water bodies should be considered when new irrigation schemes in the area are envisaged. The highly uneven distribution of malaria in Ngwavuma and Ubombo districts would justify an uneven application of control activities, with more effort being concentrated in the higher risk areas.

This study had a number of limitations which could have affected the results. 1) Only 83 percent of the cases could be allocated to their geographical area. We did not use the remaining 17 percent of cases in the spatial analysis assuming that they are either transients, or a random sample of all cases. 2) Incidence rates could not be adjusted for age and sex since demographic data were not available for the population counts. Previous studies have shown that the age distribution of cases in this area closely follows the age distribution of the population, so that adjustment for age would have had a negligible effect on the results (Sharp *et al.* 1988). 3) Large databases of reporting data are always prone to inaccuracies, omissions and duplicates (Kleinschmidt et al. 1994). As long as these errors are randomly distributed this would lead to an attenuation of any significant associations (MacMahon *et al.* 1990). 4) Potential bias caused by the system of active case finding cannot be ruled out, but we have no reason to believe that such under-reporting favours some areas over others. A recent health survey in this area has shown that health seeking behaviour in general does not appear to be affected by distance of residence from health facilities (Joyce Tsoka, Medical Research Council, Durban, personal communication, 1999).

In this study we have accounted for the spatial correlation in the data by iteratively improving the measurement of residual correlation and subsequently specifying the covariance matrix of the data in a generalised linear mixed model. Ignoring spatial correlation can result in explanatory variables apparently being associated with incidence, as a result of overstatement of the degrees of freedom in the data and consequent under-estimation of the sizes of standard errors. In our analysis the standard errors of regression coefficients were under-estimated by as much as 35 percent in the model ignoring spatial effects, compared to the model that adjusted for spatial effects. The advantage of using the GLMM approach is that it can be extended to incorporate additional data requiring further random effects to be specified. For example, we would like to extend this analysis in future to take into account several years of data for the same small areas, by specifying the year of each annual count as a random effect and allowing for temporal correlation. Such spatial-temporal analysis would show whether unusual weather conditions in a given winter predict unusual malaria incidence during the ensuing malaria season.

Acknowledgements

The authors would like to thank Dr. Tom Smith for his critical comments on an earlier draft of this paper and Carrin Martin for providing the digitised data on water bodies.

Chapter 4

An empirical malaria distribution map for West Africa

Immo Kleinschmidt[1], Judy Omumbo[2], Olivier Briët[3], Nick van de Giesen[4], Nafomon Sogoba[5], Nathan Kumasenu Mensah[6], Pieter Windmeijer[7], Mahaman Moussa[3], Thomas Teuscher[3]

This paper has been published in *Tropical Medicine and International Health* 2001; 6(10): 779-786.

[1] South African Medical Research Council, P.O. Box 17120 Congella, Durban 4013, South Africa.

[2] Kenya Medical Research Institute/Wellcome Trust Collaborative Programme, PO Box 43640, Nairobi, Kenya

[3] West Africa Rice Development Association (WARDA),01 BP 2551 Bouake 01, Cote D'Ivoire

[4] Center for Development Research, Bonn University, Walter-Flex-Str 3, 53113 Bonn, Germany.

[5] Malaria Research and Training Center DEAP/FMPOS, Universite du Mali, Bamako, Mali.

[6] Navrongo Health Research Centre, P.O.Box 114, UE/R Ghana.

[7] Alterra, PO Box 47, 6700 AA Wageningen, The Netherlands.

Summary

The objective of this study was to produce a malaria distribution map that would constitute a useful tool for development and health planners in West Africa.
The recently created continental database of malaria survey results (MARA/ARMA Collaboration 1998) provides the opportunity for producing empirical models and maps of malaria distribution at a regional and eventually at a continental level. This paper reports on the mapping of malaria distribution for sub-Saharan West Africa based on these data.

The strategy used in this study was to undertake a spatial statistical analysis of malaria parasite prevalence in relation to those potential bio-physical environmental factors involved in the distribution of malaria transmission intensity, which are readily available at any map location. The resulting model was then used to predict parasite prevalence for the whole of West Africa. We also produced estimates of the proportion of population of each country in the region exposed to various categories of risk to show the impact that malaria is having on individual countries.

The data used in this study represent a very large sample of children in West Africa. It constitutes a first attempt to produce a malaria risk map of the West African region, based entirely on malariometric data. We anticipate that it will provide useful additional guidance to control programme managers, and that it can be refined once sufficient additional data become available.

Introduction

Accurate knowledge of the distribution of malaria is an important tool in planning and evaluating malaria control (Snow *et al.*1996). A report to the recently held first sub-Saharan regional African summit meeting on malaria cites a "dire lack of extensive and comparable data about malaria," and calls, amongst other things, for more research on trends in incidence and prevalence, epidemic outbreaks and clinical epidemiology (Sachs 2000).

Global, continental and regional maps of malaria distribution in the past have been largely based on expert opinion (Molineaux 1988), and more recently on climatic suitability (Craig *et al.* 1999). Empirical maps based on malariometric data have hitherto been produced only at country or district level (Snow *et al.* 1998, Kleinschmidt *et al.* 2000, Thomson *et al.*1999). These have the advantage of approximate homogeneity of factors related to malaria control and health services, but they ignore the "wider picture" of effects outside the political boundaries of the country being studied. Since transmission intensity and the factors that determine it are rarely confined to these political boundaries, a country or district map is subject to inaccuracies due to spatial effects acting across such boundaries.

The recently created continental database of malaria survey results (MARA/ARMA Collaboration 1998) provides the opportunity for producing empirical models and maps of malaria distribution at a regional and eventually at a continental level. This paper reports on the mapping of malaria distribution for West Africa based on these data. With a total population of nearly 300 million people, sub-Saharan West Africa represents the region with the largest population exposed to high levels of malaria transmission intensity. More detailed knowledge of the distribution of malaria transmission intensity in this region can be used as a basis for more targeted malaria control and health service provision for a very large number of people.

The objective of the present study was to produce a malaria distribution map that would constitute a useful tool for development and health planners in West Africa. We also produced estimates of the proportion of population of each country in the region exposed to various categories of risk to show the impact that malaria is having on individual countries.

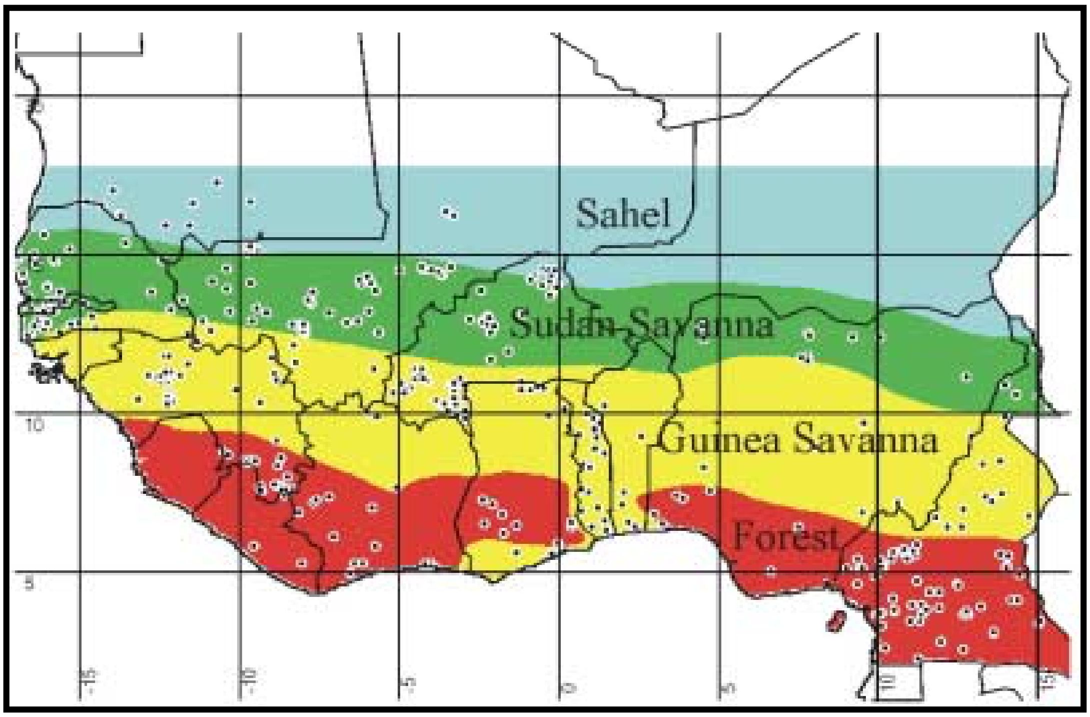
Methods and materials

Previous studies using the MARA database for the production of malaria distribution models have described methodological approaches that we have essentially followed in this study (Craig *et al.*1999; Snow *et al.*1998; Chapter 2, Chapter 3). In this paper we describe the methods and data used for this study, the results obtained and the implications for malaria control in West Africa. Further detail relating to the methods and the results are contained in a technical report (Appendix 2).

Data

The entomological inoculation rate (EIR) (the number of sporozoite positive bites per person per time unit) would have been the ideal malariometric measure to model for the purpose of mapping the distribution of transmission intensity (Snow *et al*. 1996). Since EIR is not widely available, we modelled parasite prevalence, which is far more commonly available and which is a reasonable proxy for EIR (Beier *et al*. 1999). Results from parasite prevalence surveys used for this analysis, were restricted to those of childhood populations of less than 10 years of age, in order to avoid the effects of population immunity in endemic areas moderating the survey results.

Figure 4.1. Locations of surveys, and agro-ecological zones for West Africa



The MARA / ARMA database of geographically referenced survey reports on malaria endemicity in sub-Saharan Africa has been described elsewhere (MARA/ARMA Collaboration 1998). For this study all data relating to community based surveys between latitudes 1° and 22° North and longitudes 17° West to 16° East, in which at least 50 children between 1 and 10 years of age were examined for the presence of *Plasmodium falciparum* in blood smears, were extracted from the database. In a few instances where no further age breakdowns were available, surveys on populations between 1 and 15 years were also included. Surveys conducted during known epidemics were excluded, as were those that may represent biased samples, such as those that were restricted to school attenders only. Data from island populations were also excluded. The survey dates covered several decades from about 1970 onwards, and surveys conducted more than once at the same location were combined (summing numerators and denominators). An implicit assumption therefore is that malaria endemicity has remained relatively stable over this period, so that the surveys taken at different time points can be conceptually regarded as a cross-section of surveys, taken at many locations. A total of 450 data points resulted from this process representing approximately one quarter of a million children surveyed for malaria parasites. The locations of these points are shown in figure 4.1.

Distribution of malaria is governed by a large number of factors relating to the parasite, the vector and the host (Molineaux 1988). Predominant among these are climatic and environmental factors, particularly those that effect habitat and breeding sites of the *anopheline* vectors such as temperature, precipitation, humidity, presence of water, vegetation and man to vector contact. The data used in this study for modelling and mapping malaria parasite prevalence were long-term averages of monthly rainfall, monthly averages of daily minimum and maximum temperature (Hutchinson *et al.* 1995), normalised difference vegetation index (NDVI Image Bank Africa 1991), drainage density (Windmeijer and Andriesse 1993), and estimated population density (Deichman 1996). Monthly climate and vegetation data were aggregated into quarterly averages, from December onwards (to approximately coincide with the drier and wetter seasons respectively).

Four agro-ecological zones (AEZ) were distinguished on the basis of the length of the growing period, i.e. the period that water is available for vegetative production on well drained soils. This is a function of precipitation, evaporation, and the amount of available water in the soil (FAO 1978). The definition of the zones is as follows: Equatorial Forest zone (> 270 days), Guinea Savanna zone (165 – 270 days), Sudan Savanna zone (90 –165 days) and the Sahel zone (< 90 days), shown in figure 4.1. Such zones are well established environmental entities with specific agricultural potential (FAO 1978).

Statistical modelling

For the purpose of this study, the data were divided into 3 groups corresponding to the agro-ecological zones described above, with Sahel and Sudan Savanna combined into one group. A statistical model was derived for each of these three zone specific groups. This approach was based on the assumption that the factors affecting malaria risk such as rainfall would be different in the four agro-ecological zones. Parasite prevalence values varied from 0 to 100%. Of the total number of individuals surveyed, 48.8% tested positive. A variogram (Krige 1966, Carrat and Valleron 1992) of prevalence values showed that spatial dependence of the survey results extended over a distance of about 160 kilometres.

Initial variable selection for each model was done by performing a stepwise procedure using a generalised linear model (GLM) with logit link function (Hosmer and Lemshow 1989, StataCorp. 1997) and with the parasite prevalence of a point being the response variable. The criterion for inclusion of a variable into the model was set to $p<0.01$.

In order to account for spatial correlation in the data we followed a previously documented iterative procedure (Chapter 3) for improving the specification of the co-variance structure of the data using a generalized linear mixed model (GLMM) (Littell *et al*. 1996; SAS 1996). Deviance residuals were calculated for each statistical model that was derived from the initial GLM. Semivariance (Carrat & Valleron, 1992) of the deviance residuals of all pairs of observations was calculated and a variogram constructed to determine if there was evidence of residual spatial correlation i.e. if the semivariance of pairs of residuals that are close together is markedly less than that of observations which are further apart. The parameters of the function that describes the relationship between semivariance and separation distance (the spatial model) is then used to specify the correlation structure of the data in the GLMM thereby taking account of any residual non-independence in the data. Allowing for spatial correlation may therefore lead to removal of some variables from the model due to the resultant inflation of the standard errors. Deviance residuals of the spatially adjusted model are calculated and a new variogram is constructed. This process is iterated until the variogram no longer changes indicating that a covariance structure corresponding to the model residuals is adequately specified (Chapter 2; Appendix 1).

In order to improve the fit (i.e. reduce residual deviance), each variable that survived the above procedure was transformed into 7 different fractional polynomials (Royston *et al.* 1999). The transformation producing the biggest reduction in residual deviance was chosen if this reduction in deviance exceeded 3.84, compared to the untransformed variable. Transformations that were tried for each variable x were $1/x^2$, $1/x$, $1/x^{0.5}$, $\ln(x)$, $x^{0.5}$, x^2 and x^3.

Once the zone specific models had been derived, these were used to produce map based on the predictor variables which are available as map images. The zone boundaries represent a somewhat arbitrary cut-off, with places near such a boundary sharing characteristics of the zones on both sides of the boundary. Predictions of parasite prevalence along a boundary between two zones were therefore based on a weighted mean of the predictions obtained from the models for the two adjoining zones, with the weights being a function of the distances from the boundary (see appendix 2, p.156). This interpolation of predictions along zone boundaries was carried out up to a distance of 160km from each zone boundary, since the previously constructed variogram showed that spatial effects were limited to approximately this distance.

To improve prediction in places where there is considerable divergence between model predictions and observations in a local neighbourhood we used a previously developed method (Kleinschmidt *et al.* 2000) based on kriging (Krige 1966) of the residuals of the final model predictions. A kriged map of deviance residuals is calculated, which is added to the predicted values on the logit scale before transforming the result back to proportions. The addition of kriged residuals will allow the map to deviate from the model and move closer to the observed values, if such deviation is supported by other observed values in the neighbourhood. This improves the final map in the sense that it does not deviate too severely from the observations, which is particularly important if the model does not adequately explain the observed variation in transmission risk.

Our method therefore involves a combination of modelling (predictions based on the values of climatic and environmental variables at each location) and kriging (interpolation of prevalence values at points between observed survey locations). This has the effect that the map predictions are primarily model driven in areas with a paucity of points, whereas in areas with an abundance of survey locations the map values will be primarily determined by the actual observed values at these points.

Predicted population at risk

We overlayed the final predicted prevalence map on a population density map (Deichman 1996), to calculate the population at risk for different endemicity categories for each country, excluding urban areas.

Results and discussion

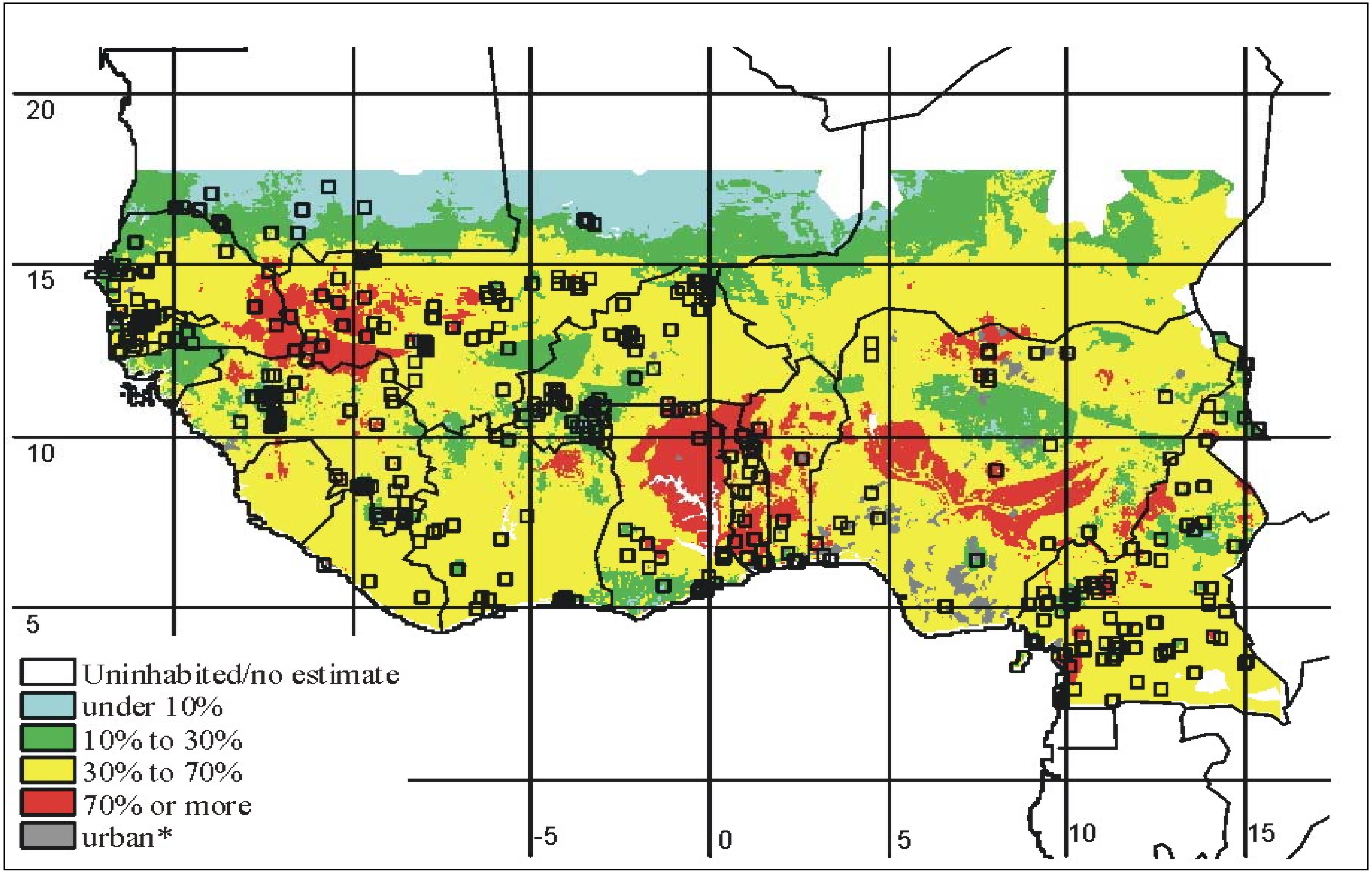
Significant explanatory variables for the model for the Sahel and Sudan Savanna zone were: average monthly rainfall from March to May, average minimum temperature from September to November and from December to February, average maximum temperature from March to May and from September to November, average vegetation index from March to May and drainage density. For the model for the Guinea Savanna zone the significant variables were average monthly rainfall from September to November, average vegetation index from December to February, and from March to May, average minimum temperature from December to February and from June to August, average maximum temperature from September to November, difference in maximum monthly and minimum monthly vegetation index, drainage density and population density. Finally, the model for the Forest zone contained average maximum temperature from September to November and from June to August, and average monthly rainfall from September to November. Since all the models are multiple variable models, each variable is corrected for all the other variables in the model. The relationship between these quantities and parasite prevalence is complex, and we give details of model coefficients and their plausibility in the technical report (Appendix 2).

Figure 4.2 shows the final map of predicted risk of malaria infection for children under 10 years during a location's main malaria season that was predicted from our models after processing the predictions in the way described above. The grouping of the map predictions into the four categories of risk shown in the map are the same as were used for a country level malaria map for Mali (Chapter 2).

Our data contained a handful of points (n=21) that could be regarded as urban on account of their 1995 population density being above 386 per sq km (US Bureau of Census 1995). Average parasite prevalence in these "urban" surveys was 45.1%, compared to a mean of 46.7% for non-urban surveys (two-sample t-test, p = 0.77). This result was not sensitive to the particular population density cut-off chosen for the definition of urban sites, and it was true in all three zones. It was only in the Guinea Savanna zone that there was a significantly higher prevalence for points with population densities below 1 per sq km after adjusting for other factors in the model. Despite this lack of evidence in the MARA database for lower parasite ratios in urban areas, we considered our data too unrepresentative of urban areas to make any predictions in such areas. Urban areas were therefore excluded from the prediction map, and from the population at risk calculations. It is quite likely that some surveys were taken in places which were rural outskirts of urban areas at the time of the surveys, but which are now urban. Whilst climatic factors might justifiably have been regarded as constant over the time that the surveys were conducted, this assumption is almost certainly not uniformly valid for population density, and this may be the reason for it not featuring more prominently as a significant explanatory variable.

Figure 4.2. Predicted prevalence of *P.falciparum* in children aged 2 to 10 years for West Africa

* Differing map resolutions have caused some digitisation error along the coast, causing some coastal urban areas not to show on the map



Comparing our final map predictions with the observed prevalence values of the 450 surveys, 77.6% (349/450) of the surveys were correctly classified, i.e. the predicted prevalence category agreed with the observed prevalence category (kappa=0.62, p<0.0001). Of the points where there was a disagreement between the observed and predicted prevalence categories, only 3 were misclassified by more than one category value.

Visual comparison of our map with previous "expert" opinion maps (Wernsdorfer and McGregor 1988, Haworth 1988) and with the suitability map by Craig *et al.* (1999), shows broad agreement. The map is also in agreement with a map of Mali, that was previously derived from MARA data (Chapter 2). Our map offers more differentiation in the category of "highly suitable" of the climatic suitability map, but less differentiation in the areas designated as "unstable malaria" in the suitability map. This is in part due to the fact that comparatively few malaria surveys are done in areas of unstable malaria, and this is reflected in the MARA database. Nevertheless, a visual comparison of our map with the suitability map shows many similar features, which is not surprising since climatic factors were involved in the production of both maps. We should caution that there were several countries in the regions which were either poorly covered by surveys, or not at all. We are optimistic that this situation will improve in future and this will allow a more accurate map to be produced. However, in the meantime our map predictions for these areas are entirely based on our models that were derived from data from neighbouring countries. This may still give reasonable predictions for smaller countries or those that are surrounded by countries with an abundance of data points, but it is bound to give inaccurate estimates for countries on the periphery of our map window, such as Niger. We excluded Niger from the calculation of populations at risk (table 4.1) for this reason. Most of Nigeria, and the central parts of Ghana also suffered from a sparse coverage of points, and hence the predictions in these regions are model dependent, rather than interpolation driven. A current shortcoming in our modelling methodology is the fact that we are unable to give an estimation error for the various parts of the map.

The proportion of population in each country exposed to each of the four risk categories varies considerably between countries in the region (table 4.1). For example, the population living in areas with less than 30% prevalence make up 17% of the population of the entire region, with high proportions of the population in this category living in Mauritania (50%), Guinea Bissau (30%), Mali (31%), and Senegal (23%). Some of these could be populations with low levels of immunity and it can reasonably be expected that exceptional rainfall will cause significant morbidity in all age groups. Often such areas are remote and interventions are hampered by poor health service infrastructures. On the other hand, populations in areas with predicted prevalences above 30% (categories 3 and 4 on the map) are more likely to have some measure of immunity with young children and pregnant women being the groups most vulnerable to morbidity and mortality due to malaria. According to our map, 58% of the population of West Africa (168 million people) fall into this category. In Côte d'Ivoire, Togo, Burkina Faso, Sierra Leone and Liberia 70% or more of the population is exposed to this level of transmission intensity.

Although the highest prevalence category, namely 70% to 100%, occupies a considerable area on the map, the proportion of population living in these areas is reasonably small in all countries except Togo. For the West African region as a whole about 16 million people are exposed to this high level of transmission intensity. Marsh and Snow (1999) suggested that vector-contact reducing measures such as insecticide treated materials (ITM) may change severe-disease patterns of malaria and consequently case fatality in high endemicity settings. The introduction of ITMs on a large scale should be accompanied by more intense monitoring efforts in such circumstances. Our prediction map helps to identify areas where such long term morbidity monitoring might need to accompany ITM deployment.

Ideally, we would like to have a map that clearly identifies two types of areas requiring two quite distinct types of intervention packages. These would be epidemic prone areas, and areas with stable malaria endemicity. In areas of unstable malaria transmission, surveillance efforts, the stocking of efficacious insecticides such as DDT for in-house spraying as well as appropriate and affordable diagnosis and treatment algorithms play a primary role. In holoendemic areas on the other hand, rapid diagnosis and treatment, intermittent treatment during pregnancy, behavioural aspects related to the large scale use of ITMs and innovative strategies to ensure the availability of high quality first line treatments at home might be considered high priority by country control programs. Our map cannot provide such clear division into endemic and epidemic areas, but it can be used to guide such decisions.

Table 4.1. Predicted percentage of population at risk by country and risk category (excluding urban populations)

	Percentage of total population in each risk category[1]			
Country[2]	**Predicted prevalence of less than 10%**	**Predicted prevalence of 10 to 30%**	**Predicted prevalence of 30 to 70%**	**Predicted prevalence above 70%**
Benin	0	5	43.4	12
Burkina Faso	0	17	76	0
Cameroon	1	16	58	2
Côte d'Ivoire	0	4	75	0
Gambia	0	8	44	0
Ghana	1	15	46	17
Guinea	1	12	57	3
Guinea Bissau	2	30	13	0
Liberia	0	1	81	2
Mali	3	28	66	1
Mauritania	20	30	6	1
Nigeria	0	8	48	8
Senegal	1	22	41	2
Sierra Leone	0	0	79	2
Togo	0	0	39	38
Entire region	2.4	14.8	52.7	5.4
Total population at risk	7,006,869	42,941,669	152,779,264	15,698,929

[1] Percentages do not sum to 100% since urban populations have been excluded and parts of some countries lie outside the map window.

[2] Excluding Niger

It is well known that malaria transmission intensity exihibits strong spatial heterogeneity even at a local level. It is therefore likely that the map may be at variance with local experience in some places. Where this occurs, it ought to motivate further investigation through well conducted local surveys.

A possible source of variation that is not determined by natural factors such as climate and drainage density may be differences in socio-economic development, which has played a part in malaria control and eradication elsewhere, probably coinciding with other factors (Bruce-Chwatt and de Zuleta 1980, Molineaux 1988, Packard 1984, Wernsdorfer and Wernsdorfer 1988). Socio-economic development could reduce malaria transmission in a variety of ways. For example, increases in household income of women and poverty reducing measures in general have the potential to reduce exposure to malaria and to improve health seeking behaviour and quality of treatment. However, socio-economic development in a high transmission tropical setting could equally increase malaria transmission due to changes such as forest clearing or the migration of people with little or no immunity into areas of high endemicity. We have been unable to model such factors in our analysis due to the fact that such data for the entire region are currently not available with adequate spatial resolution. It is highly likely that there are other unmeasured, perhaps more local factors that determine variation in parasite prevalence.

A further source of variation that has not been taken into account in this study is variation in prevalence by season and by age (Sissoko *et al*., *submitted*). The impact of these factors will differ according to the endemicity level of an area. It was our opinion that the differentiation that was available within the results of many surveys was inadequate to stratify the data by these factors.

A regional malaria risk map, such as the one produced in this study, will allow planners to assess the possible health impacts of measures aimed at improving food security through the promotion of large scale irrigation and wetland management projects. Elsewhere in Africa such developments have significantly increased malaria infection and morbidity in epidemic prone areas of unstable malaria (Ghebreyesus *et al*. 1999). However, the same agricultural production methods are unlikely to affect the malaria risk profile of rural populations living in areas characterized by high parasite prevalences (Dossou-Yovo *et al*. 1998, Faye *et al*. 1995).

Finally, the map will also help guide public health research managers in identifying appropriate study environments for intervention trials as well as assist with the identification of populations potentially benefiting from new interventions.

The data used in this study represent a very large albeit imperfectly sampled population of children in West Africa. This study is a first attempt to produce a malaria risk map of the West African region, based entirely on malariometric data. We anticipate that it will provide useful additional guidance to control programme managers, and that it can be refined once sufficient additional data become available.

Acknowledgements

This regional analysis of the MARA database was carried out with the aid of a grant (center file 00397) from the International Development Research Centre (IDRC), Ottawa, Canada. We gratefully acknowledge support for the MARA project by the Multilateral Initiative on Malaria (MIM), the IDRC and Roll Back Malaria. We are also thankful to the West Africa Rice Development Association (WARDA) for providing the necessary infrastructure to host the workshop. Particular thanks are due to Mylene Kamara (WARDA) for producing the population at risk calculations and to Nicole Teuwsen for assisting with GIS work.

Chapter 5

Rise in malaria incidence rates in South Africa: a small area spatial analysis of variation in time trends

Immo Kleinschmidt[1], Brian Sharp[1], Ivo Mueller[2], Penelope Vounatsou[3]

[1] South African Medical Research Council, Durban, South Africa.
[2] Tropical Health Program, University of Queensland, Brisbane, Australia.
[3] Swiss Tropical Institute, Basel, Switzerland.

This paper is in press at *American Journal of Epidemiology*

ABSTRACT

Spatial and temporal variations in small area malaria incidence rates for the period mid-1986 to mid-1999 for two districts in northern KwaZulu Natal, South Africa were investigated using Bayesian statistical models. Maps of spatially smoothed incidence rates at different time points and spatially smoothed time trend in incidence gave a visual impression of the highest increase in incidence occurring where incidence rates previously had been lowest. This was confirmed by conditional autoregressive models, which showed that there was a significant negative association between time trend and smoothed baseline incidence before the steady rise in caseloads began. Growth rates also appeared to be higher in the areas close to the Mozambican border. The main findings were that: (1) the spatial distribution of the rise in malaria incidence is uneven and strongly suggests a geographical expansion of high-risk malaria areas; (2) there is evidence of a stabilisation of incidence in areas which had the highest rates before the current escalation of rates began; (3) areas immediately adjoining the Mozambican border appear to have undergone larger increases in incidence, in contrast to the general pattern of low growth in the more northern, high baseline incidence areas, but this was not confirmed by modelling; (4) smoothing of small area maps of incidence and growth in incidence (trend) is important for the interpretation of the spatial distribution of disease incidence, and the spatial distribution of rapid changes in disease incidence.

Malaria cases in South Africa have risen steadily and steeply over the past few years. The total number of cases reported nationally during the first six months of 1999 was over 34,000, representing an increase of 80 percent compared to the same period of the previous year (Department of Health, 2000a). Isolated cases of local malaria transmission have recently been identified in the Durban municipal area, which is several hundred kilometers to the south of malarious areas, giving rise to considerable public concern(Pillay, 2000). A question that has arisen is whether the increase in case loads is associated with an expansion of South Africa's malaria transmission area or whether it is the same areas as before suffering increased transmission rates. The sharp increase in malaria incidence could have potentially severe consequences not only for public health but also for tourism and economic development.

The cause for the steep increase in malaria cases in South Africa has been linked to a variety of factors, namely the El Nino effect on weather patterns in Southern Africa, proximity to areas where no malaria control systems are in place and migration from such areas, the development of drug-resistance in the malaria parasite (Bredenkamp *et al.* 2000) and insecticide resistance in the malaria vector *Anopheles funestus* (Hargreaves *et al.* 2000), and the possible effects of HIV infection on a substantial proportion of the population (Department of Health, 2000b), Whitworth *et al.* 2000). Our analysis attempts to document the spatial changes in malaria transmission in two particular districts for which high resolution reporting systems are available. These are the neighbouring magisterial districts of Ngwavuma and Ubombo in Northern KwaZulu Natal which have hitherto had the highest malaria incidence rates in the country.

The objective of this study was to investigate whether there was geographical expansion of malaria transmission. By modelling the spatial variation of time trend in incidence rates we sought to establish whether the additional cases come predominantly from areas that have always had the highest transmission levels, or whether they originate from previously low transmission or malaria free sub-regions. We used Bayesian statistical methods to produce maps of smoothed incidence at different time points, and maps of spatially smoothed rate of change of incidence. By modelling the time trend we investigated whether there was an association between time trend and baseline malaria incidence.

Materials and methods

In both Ngwavuma and Ubombo districts a small area malaria incidence reporting system has been in operation since 1986 as part of the provincial malaria control programme(Sharp *et al.* 1999). This records all parasitologically confirmed cases, both passive as well as those found by active surveillance. The latter consists of screening measures by which teams go into the community to encourage individuals who may be suspected of having malaria, to be tested. Active case finding forms part of the control strategy of treating all infected individuals. Whilst such active case finding may not achieve 100 percent coverage, it is thought to identify the vast majority of cases. Since malaria incidence is generally low (average annual incidence rates have been as low as 2.3 cases per 1000 in the last 10 years), individuals living in this area generally have, until very recently, had infrequent or no exposure to malaria. The low levels of exposure to *Plasmodium falciparum* in the past have made it likely that the actively found cases are recent infections, rather than asymptomatic semi-immune cases. This assumption may no longer be valid in those areas which experience the highest incidences.

A population census at homestead level was carried out in 1994 by the malaria control programme, thus making it possible to use a Geographic Information System (GIS) to derive population counts for the same small areas for which cases were being reported. Unfortunately the boundaries for small area counts for both the 1991 and 1996 censuses are considered unreliable and do not coincide with those of the malaria reporting boundaries. We therefore used population totals based on the 1994 census, applying a constant and uniform growth rate of 2 percent per annum(Statistics South Africa, 1999).

Figure 5.2 shows that the overall malaria incidence rate for the study area over the 13 year period fluctuated strongly around approximately 10 cases per 1000 person years annually with neither an upward nor a downward trend during the years 1986/7 to

1994/5 (We have used mid-year to mid-year aggregations since these correspond roughly to a malaria season). After 1995 there has been a steep and consistent increase in malaria incidence. We have therefore modelled the data separately for the two time periods 1986/7 to 1994/5 and 1995/6 to 1998/9.

Modelling

Crude maps of disease incidence are often subject to considerable random error, particularly if either the disease is rare or the population per spatial unit is small, so that the rate may be influenced by a relatively small number of cases. This leads to maps in which attention is drawn to those areas whose rates are based on the least stable estimates (Cuzick and Elliott, 1992). Moreover, the estimation of the standard errors of explanatory variables will be biased if spatial correlations are not taken into account. These problems can be overcome by spatial smoothing of the rates, which is based on "borrowing strength" from neighbouring regions. In this study we have followed the approach that uses hierarchical fully Bayesian spatial modeling as described by Bernadinelli and Montomoli (1992). This approach models spatial variation via conditional autoregressive (CAR) priors (Clayton and Kaldor, 1987).

Let Y_{it} and P_{it} denote the observed counts of cases and population respectively and let $\eta_{it} \equiv \mathrm{E}(Y_{it})$ denote the mean count of cases for the i^{th} area in the t^{th} year. It is assumed that the Y_{it} are conditionally independent given the η_{it} and follow a Poisson distribution, i.e. Y_{it}~Poisson(η_{it}). The η_{it} are defined using customary linear models which may include covariate terms as well as random time and area effects.

The following model was used to estimate smoothed incidence rates for each area for the nine year period from 1986/7 to 1994/5:

$$\log(\eta_{it}) = \log(P_{it}) + \mu + \varphi_i + \omega_t \qquad \text{(model 1)}$$

where μ represents the mean incidence rate over all areas over all time periods, φ_i is a random effects term that allows for spatially structured variation in rates and ω_t is a random term representing between year variation, assumed independent and normally distributed.

Bayesian statistical inference is based on posterior distributions which combine information available from the data via the likelihood function and any prior knowledge about the model parameters by specifying appropriate distributions for these parameters. We incorporate our prior information about the structure of the map by assuming CAR models for the area random effects. According to the CAR model the area specific spatial effects φ_i are modeled (conditional on neighbouring random effects) as normally distributed with mean equal to the mean of the effects of its neighbours ($\overline{\varphi_i}$) and a variance that is inversely proportional to the number of neighbours n_i, i.e. $\varphi_i \mid \varphi_{-i} \sim \mathrm{N}(\overline{\varphi}_i, \sigma_\varphi^2 / n_i)$ where $\overline{\varphi}_i = \frac{1}{n_i} \sum_{j \in neighbours\, of\, i} \varphi_j$. The effect of this prior distribution is to shrink the incidence rates of areas to that of the local mean, where the local mean is the mean of all contiguous areas excluding the area *i* itself. The posterior distribution of the rate of an area is therefore a compromise between the prior, which is based on the rates of neighbouring areas, and the data for the area, thus stabilising the rate in areas where the data are sparse due to small populations.

Since no information is available for the remaining parameters we adopt standard conjugate priors, i.e. vague inverse gamma priors for the variances σ_ω^2 and σ_φ^2 and vague normal priors for all other parameters.

We used a second model to analyse the data for the last 4 years of the series, namely 1995/6 to 1998/9. To be able to determine the spatial variation of increases in malaria incidence the model included a spatially smoothed time trend instead of random time effects used in model 1.

Bayesian models for the analysis of space-time variation of disease rates have been considered by a number of authors (Heisterkamp *et al.* 2000; Knorr-Held and Besag, 1998; Waller *et al.* 1997, amongst others). Bernadinelli *et al* (1995) and Sun et al (2000) assumed the temporal variation of disease rate to be linear, which we judged to be a reasonable constraint in our data given a relatively short period of 4 years for the second time period. We used the following model to estimate spatially smoothed time trend for each area

$$\log(\eta_{it}) = \log(P_{it}) + \mu + \varphi_i + (\alpha + \delta_i)t \qquad \text{(model 2)}$$

where t represents the years from 1995/6, α represents an overall time trend for all areas, and the random term δ_i represents the smoothed local deviation in trend from the overall trend. The latter, which has been termed *differential trend* (Bernadinelli *et al.* 1995) is assigned a CAR Normal prior distribution, as described above, to allow for spatial smoothing of time trends, thereby facilitating the interpretation of patterns in time trend from a map.

A third model was used to investigate the association between time trend and baseline incidence i.e. average incidence before the period of steady increases in incidence. For this model the data for the entire time series (13 years) were used so that baseline incidence (first period) and its effect on differential trend during the second period could be estimated simultaneously from the same model. This has the advantage that uncertainty in the estimates of baseline incidence are incorporated into the estimates of parameters that express the association between baseline incidence and time trend. The following model was used:

$$\log(\eta_{it}) = \log(P_{it}) + z_{1t}[\mu_1 + \varphi_{1i}] + z_{2t}[\mu_2 + \varphi_{2i} + (\alpha + \delta_i)(t\text{-}9)]$$

Subscripts 1 and 2 refer to the first and second period of the data respectively so that μ_1 and μ_2 denote the overall mean log of incidence rates during the first and second periods respectively and φ_{1i} and φ_{2i} denote the area specific random effects for the first and second periods respectively. z_{1t} and z_{2t} are indicator variables to distinguish between the first and second time periods i.e. $z_{1t} = 1$ for $1 \le t \le 9$ and $z_{1t} = 0$ for $10 \le t \le 13$ whilst $z_{2t} = 1 - z_{1t}$. The φ_{1i} 's and φ_{2i} 's are assigned separate CAR prior distributions, as previously described. The number of years after the start of the second period is represented by t-9. The area specific differential trend is denoted by δ_i as before. Following Bernadinelli *et al* (1995), δ_i are assumed to be independent, conditional on the φ_{1i} i.e.

$$[\delta_i \mid \varphi_{1i}, \sigma_\delta^2] \sim \text{Normal}(\beta\varphi_{1i}, \sigma_\delta^2) \qquad \text{(model 3)}$$

The term β allows for baseline incidence and trends to be correlated in the prior. It is assigned a vague normal prior. Note that δ_i is modelled conditional on the first period area effect φ_{1i} to determine its correlation with incidence during the earlier period. The variance term σ_δ^2 represents the variance of the differential trend and is assigned a non-informative gamma distribution. Figure 5.1 is a graphical representation of model 3, using graphical conventions outlined by Bernadinelli and Montomolli (1992), and used in the software package WinBUGS (2000)

A further variant of model 3 was used to determine whether areas whose centroids are within 4km of the Mozambican border follow a different time trend. The investigation of this model was primarily motivated by inspection of the smoothed trend map (see below). The distance of 4km was chosen since it corresponds approximately to maximum distances of dispersal of the main vector (*Anopheles gambiae*) (Gillies and De Meillon, 1968) in the area. This was done by modeling the differential trend as

$$[\delta_i \mid \varphi_{1i}, x_i, \sigma_\delta^2] \sim \text{Normal}(\beta\varphi_{1i} + \gamma x_i, \sigma_\delta^2) \qquad \text{(model 4)}$$

where x_i represents a dummy variable denoting whether an area is within 4km of the Mozambican border, γ allows the differential trend in border areas to differ from that of other areas and all other terms have the same meaning as before.

Markov Chain Monte Carlo simulation was used to obtain estimates of the posterior and predictive quantities of interest. The models were implemented using Gibbs sampling in the software package WinBUGS. In order to properly monitor convergence a sampling scheme was designed using 3 independent chains and a 'burn-in' of 12000 iterations. After convergence a final sample of 5000 was collected to obtain summaries of posterior distributions of the parameters. Convergence was assessed using the method of Gelman and Rubin (1992).

In order to compare models 3 and 4 we calculated the expected predictive deviance (EPD) (Carlin and Louis, 1996) for each model. A brief description of the EPD is given in the appendix. A lower EPD is indicative of a better model. We also calculated the likelihood ratio statistic (LRS), which assesses model fit.

Figure 5.1. Graphical representation of model 3. Symbols as defined in text.

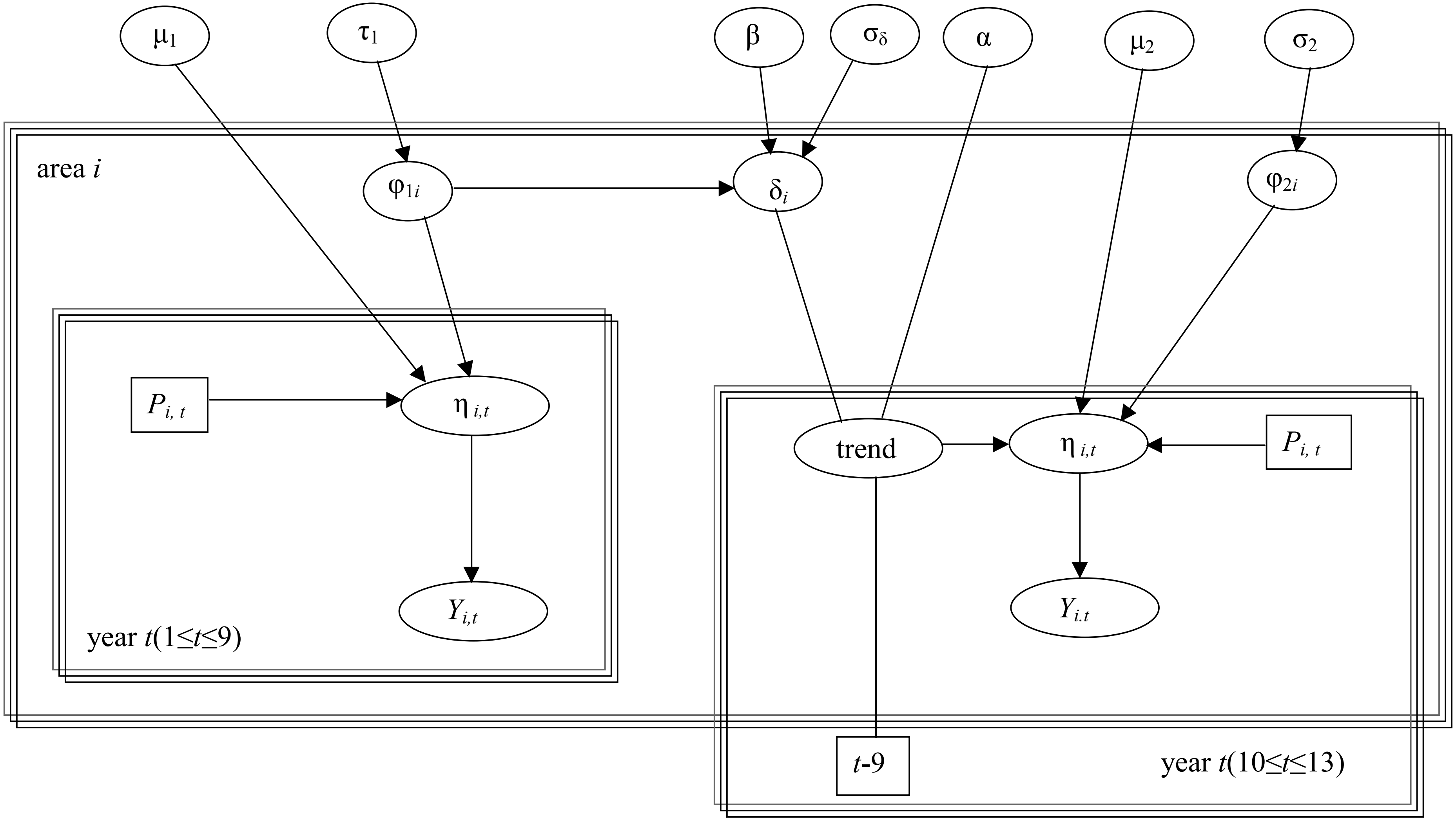


Figure 5.2. Annual malaria incidence rates for the population of Ngwavuma and Ubombo districts, and proportion of areas with incidence of less than 1 per 1,000 person-years by year from mid 1986 to mid 1999

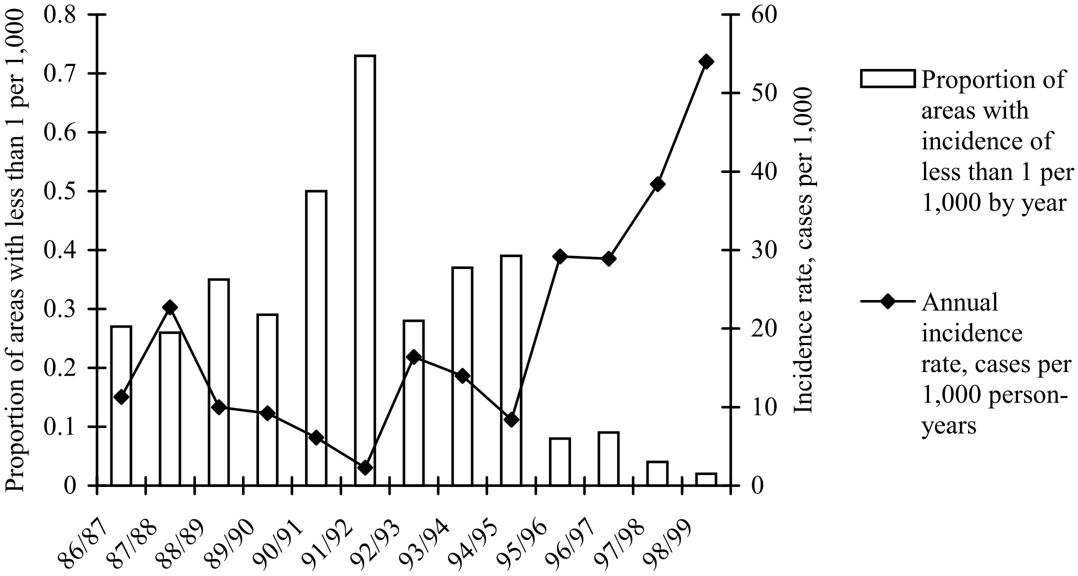


TABLE 5.1. Description of areas

Number of areas	220
Total population (1994)	239 000
Average size of area (SD), km^2	25.9 (27.6)
Average population per area in 1994 (SD)	1086 (884)
Average population density per area (SD), persons per km^2	65.1 (61.9)
Overall average incidence rate per area (SD), cases per 1000 person years	28 (91)
Lowest annual average incidence rate per area (SD), cases per 1000 person years	3.4 (16)
Highest annual average incidence rate per area (SD), cases per 1000 person years	78 (133)

Results

There was considerable variation in population size, area, population density and malaria incidence rates between the areas that formed the unit of analysis (also know as sections) and between years (table 5.1 and figure 5.2). The proportion of very low risk areas with incidence rates of less than 1 per 1000 reduced sharply over the period 1995/6 to 1998/9 (figure 5.2). The map of smoothed incidence rates for the period 1986/7 to 1994/5 as derived from model 1 (figure 5.3) clearly shows a trend of highest incidences in the north west of the region, with the lowest incidence in the south-east. There are also some high incidence "hot spots" in the north along the Mozambican border, and in the south-west along the Pongola river. The steep rise in incidence rates across the area is evident from the smoothed map of incidence for the final year of the time series (figure 5.4).

TABLE 5.2. Posterior medians and 95% credible intervals(C.I.) for model estimates of mean log incidence rate, log overall trend, effects of baseline incidence and being a border area, and standard deviation of log differential trend, for the population of Ngwavuma and Ubombo districts, 1995/6 – 1998/9.

Parameter	Description	Model 3		Model 4	
		Median	95% CI	Median	95% CI
μ_1	Mean log incidence rate during period 1 (log cases/person)	-5.42	-5.45, -5.38	-5.42	-5.45, -5.38
μ_2	Mean log incidence rate during period 2 (log cases/person)	-4.78	-4.82, -4.72	-4.78	-4.83, -4.73
α	Log overall trend during period 2 (log incidence rate ratio)	0.35	0.31, 0.41	0.34	0.29, 0.38
β	Effect of log baseline incidence on log differential trend (log incidence rate ratio/log incidence rate)	-0.11	-0.14, -0.08	-0.12	-0.14, -0.09
σ_δ	Standard deviation of log differential trend	0.086	0.069, 0.11	0.084	0.067, 0.11
γ	Effect of being a border area on log differential trend (change in log incidence rate ratio)			0.15	0.017, 0.28

Mean values of spatially smoothed local trend for each area were obtained from model 2. Smoothed local trends for areas, expressed as incidence rate ratios per annum, varied from 0.6 to 3.5, with a median value of 1.4, and an inter-quartile range of 1.2 to 1.7. Figure 5.5 shows how the smoothed time trends over the 4 year period from 1995/6 to 1998/9 are distributed across the study area. This map, when compared with the map of relatively stable "baseline" rates for the period 1986/7 to 1994/5 prior to the period of steady growth in annual cases, gives a visual impression of an inverse relationship between the gradient of incidence, and baseline incidence. The areas with the lowest baseline incidences appear to have been subjected to the steepest increases, and vice versa, with some of the high baseline incidence areas having either stabilised or undergone a small negative trend. Sections immediately bordering Mozambique appear to be an exception with moderately high time trends despite high baseline incidences. The impression of an inverse relationship between trend in incidence and baseline incidence is confirmed by figure 5.6, which is a scatter plot of the log of the trend against the log of average incidence during the first period.

According to model 3, baseline incidence is significantly negatively associated with trend (tables 5.2 and 5.3). According to both models 3 and 4, incidence rose by just over 40 percent per annum and this annual rate of increase in incidence rate is reduced by a factor of about 0.90 for each doubling in the baseline rate (incidence in the first period).

According to model 4, the time trend is 16 percent higher in the border areas than in other areas after adjusting for the effect of baseline incidence on time trend. However, model comparison (table 5.3) shows that there is no difference in EPD between model 3 and 4. The data therefore provide no evidence that the differential trend in border areas is different from what it is in other areas.

TABLE 5.3. Estimates of mean trend, effects of baseline incidence and proximity to Mozambican border on growth in incidence rates per annum for the population of Ngwavuma and Ubombo districts, 1995/6 – 1998/9, and model fit criteria

	Model 3		Model 4	
	Median	95% CI	Median	95% CI
Mean trend per annum(e^{α}): Incidence rate ratio	1.42	1.37, 1.50	1.41	1.33, 1.46
Differential trend*: Effect on annual trend of doubling of baseline incidence rate ($e^{\beta \ln 2}$)	0.93	0.91, 0.95	0.92	0.91, 0.94
Effect of being a border area (e^{γ})			1.16	1.02, 1.32
Expected predictive deviance (EPD)	21580		21572	
Likelihood ratio statistic (LRS)	18639		18625	

* Change in trend expressed as ratio of incidence rate ratios

Discussion

Our model of smoothed time trends in malaria incidence over the 4 years from 1995/6 to 1998/9 shows that there is considerable variation in average annual growth between areas (model 2, figure 5.5). The average annual increase for all areas was 52 percent. At the two extremes, over 10 percent of areas (n=26) experienced more than 100 percent annual increases, whilst just under 10% of areas (n=19) underwent a decline in incidence.

Our map of smoothed trends in incidence rates has enabled us to observe a spatial pattern in time trends in relation to baseline incidence. The map of unsmoothed crude trends for each area (not shown) makes it difficult to obtain an overall impression due to random noise in trends. For some areas it is impossible to calculate a stable crude trend value due to very low or zero incidences at the start of the four-year period. Smoothing of time trends is important to see underlying trends in a map, for the same reasons that smoothing of disease rates has become commonplace to stabilize estimates that are subject to high sampling variability (Wakefield *et al.* 2000).

We were able to confirm the visual impression from the smoothed trend map that increases in incidence rates have been steepest in areas with the lowest initial rates using a simple model incorporating initial incidence rates as a factor affecting time trend over the four years from 1995/6 to 1998/9. The border areas have been of special interest due to their proximity to Mozambique, where malaria vector control through house spraying is not practiced. The time trend in border areas appeared to be steeper than in other areas after adjusting for the association of trend with baseline incidence (model 4), but model comparison shows that we do not have evidence against the null hypothesis of border areas following the same pattern as other areas.

For the two districts in our study area it is therefore evident that there has been a steady geographical expansion of high-risk malaria transmission areas. It is likely that this expansion has progressed beyond the boundaries of the two districts which we have analysed, although this cannot be verified directly since the malaria reporting system in other districts does not permit the aggregation of cases in small geographic units as is possible for the districts of Ubombo and Ngwavuma.

The negative association of baseline incidence with time trend suggests that there is some degree of stabilisation of rates in high transmission areas. This could be due to vector and parasite related environmental factors. These would be the overall climatic suitability for malaria transmission in terms of temperature and rainfall of the region, which may impose constraints on further increases in transmission intensity, despite substantial cyclical variation (Craig *et al.*1999; Molineaux, 1988; Chapter3).

Alternatively, the stabilisation of incidence rates in high transmission areas could be due to population related factors in the form of a measure of immunity conferred by relatively high levels of exposure. It has been shown in settings of endemic malaria that a small number of infections with *P. falciparum* from birth can lead to an immune response that modulates disease outcome (Gupta *et al.* 1999). Our own data show that in high incidence areas of Ngwavuma and Ubombo, age-specific malaria incidence in adults is lower than it is in teenagers (Chapter 7). This observation would be consistent with the acquisition of clinical tolerance to infection by at least part of the population in high-incidence sub-regions of the study area, and it would explain the finding of lower increases in incidence in these areas.

Under circumstances of a steep overall increase in incidence, and plausible biological factors limiting the increase in incidence in some areas, it is unlikely that the observed inverse relationship between differential trend and baseline incidence is merely one of regression to the mean. The estimates of baseline incidence in model 3 are a fairly precise assessment of an area's underlying incidence during the first period, on account of being obtained from a long time series of data, and on account of being spatially smoothed estimates. Regression to the mean is unlikely if such stable estimates are used instead of individual unsmoothed observations relating to a particular year.

If immunity rather than leveling of transmission pressures is the cause for stabilisation of incidence rates in the high incidence areas, then it is possible that there has been an increase in transmission intensity in excess of that reflected by the incident cases. This would imply that the reported incidence rates are no longer a reliable indicator of transmission in these areas and the risk to non-immune visitors may have increased considerably more than the incidence rates. The consequences for tourism and economic development would consequently be more severe.

Our modelling approach represents a modest extension of the spatial temporal model developed by Bernadinelli *et al* (1995). We have applied this methodology to an infectious tropical disease with relatively high incidence rates and substantial spatial structure in the data (Chapter 3). We were able to model the data of the two time periods of the study comprehensively within the same overall model which enabled us to test the effects of incidence rates in the first period on time trends in the second period. Future work on this topic will investigate the potential association between possible model covariates, such as rainfall, during an earlier period, and malaria incidence in a latter period.

Figure 5.3. Smoothed mean malaria incidence rates by area estimated from model 1 for the population of Ngwavuma and Ubombo, mid 1986 to mid 1995

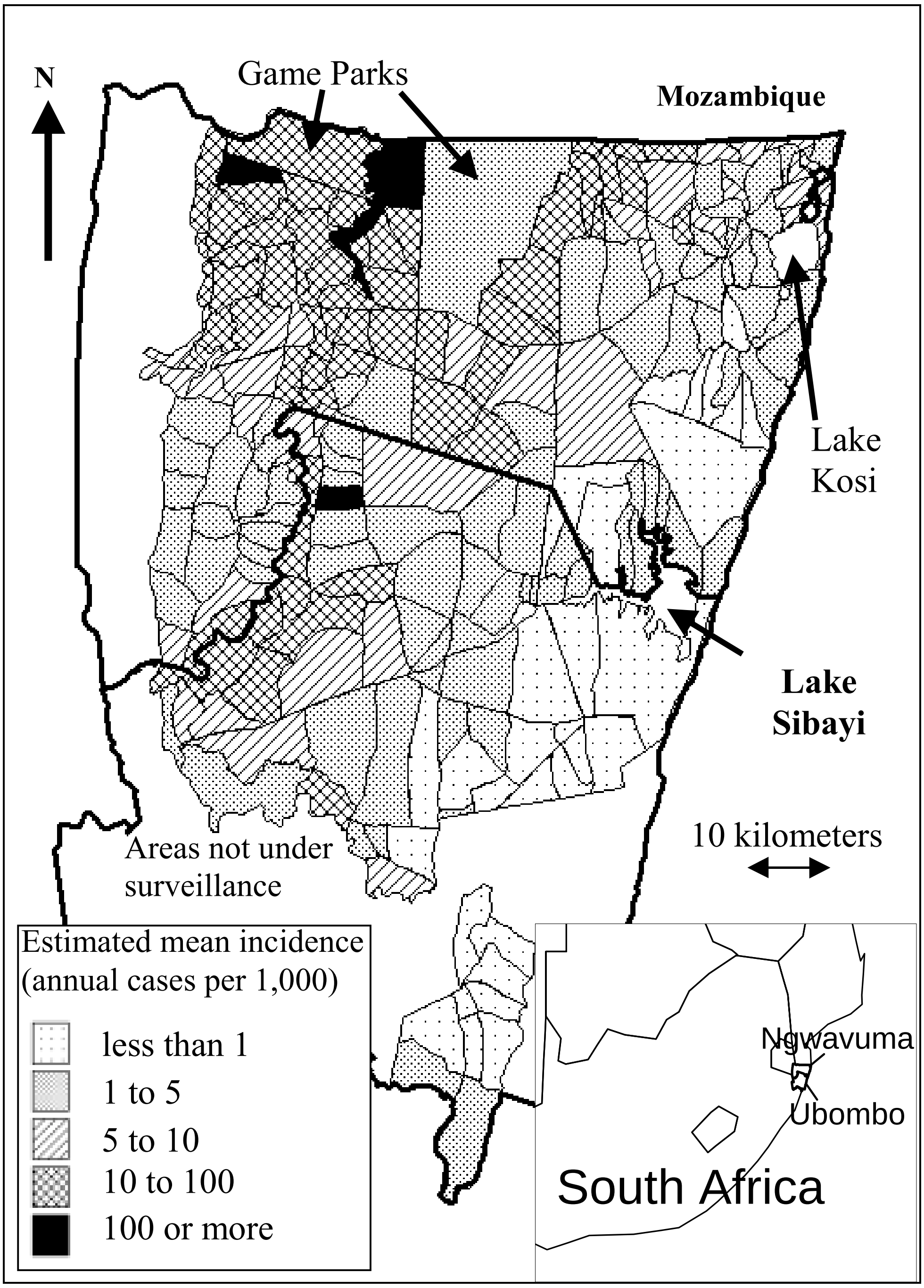


Figure 5.4. Estimated smoothed mean malaria incidence rates by area for the population of Ngwavuma and Ubombo, mid 1998 to mid 1999

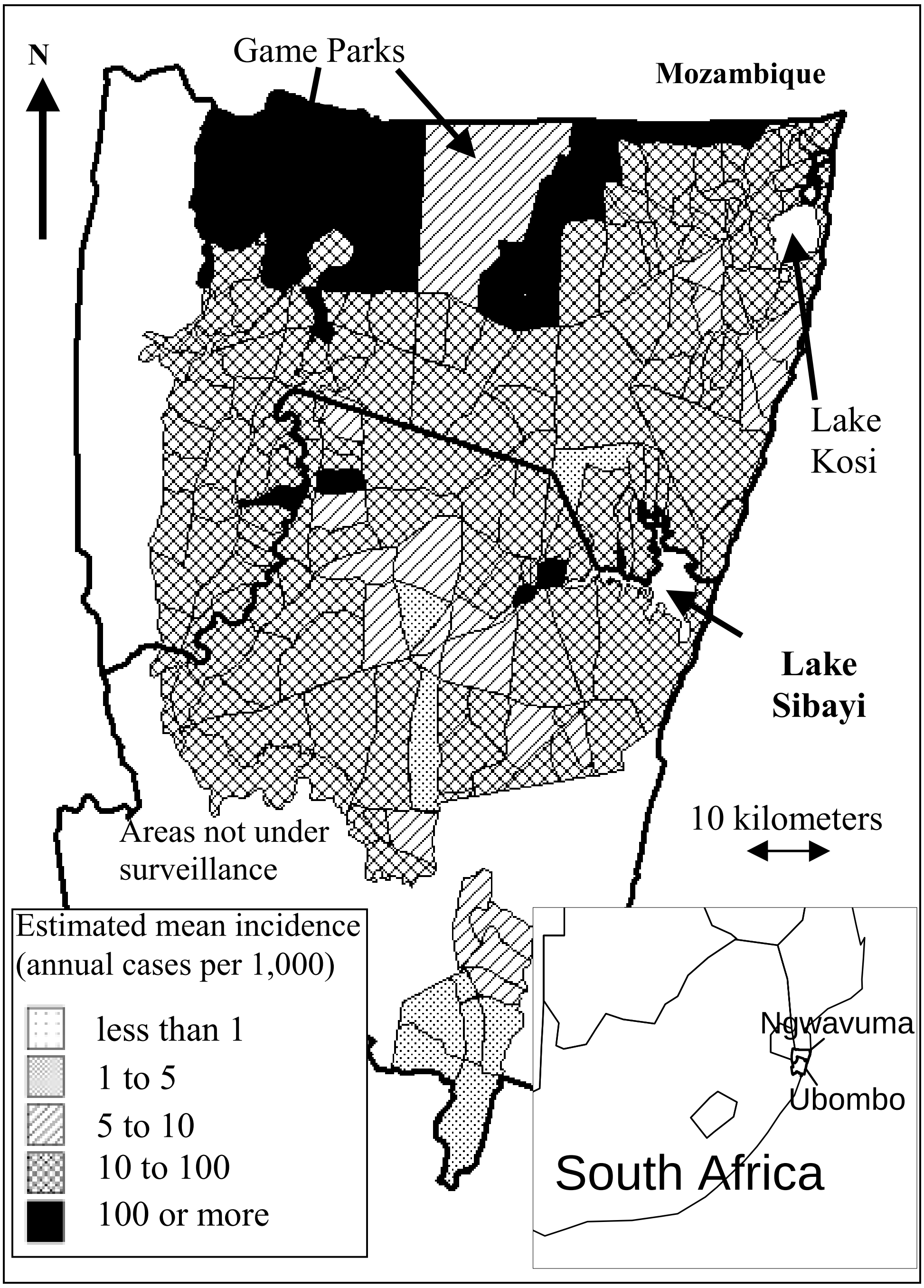
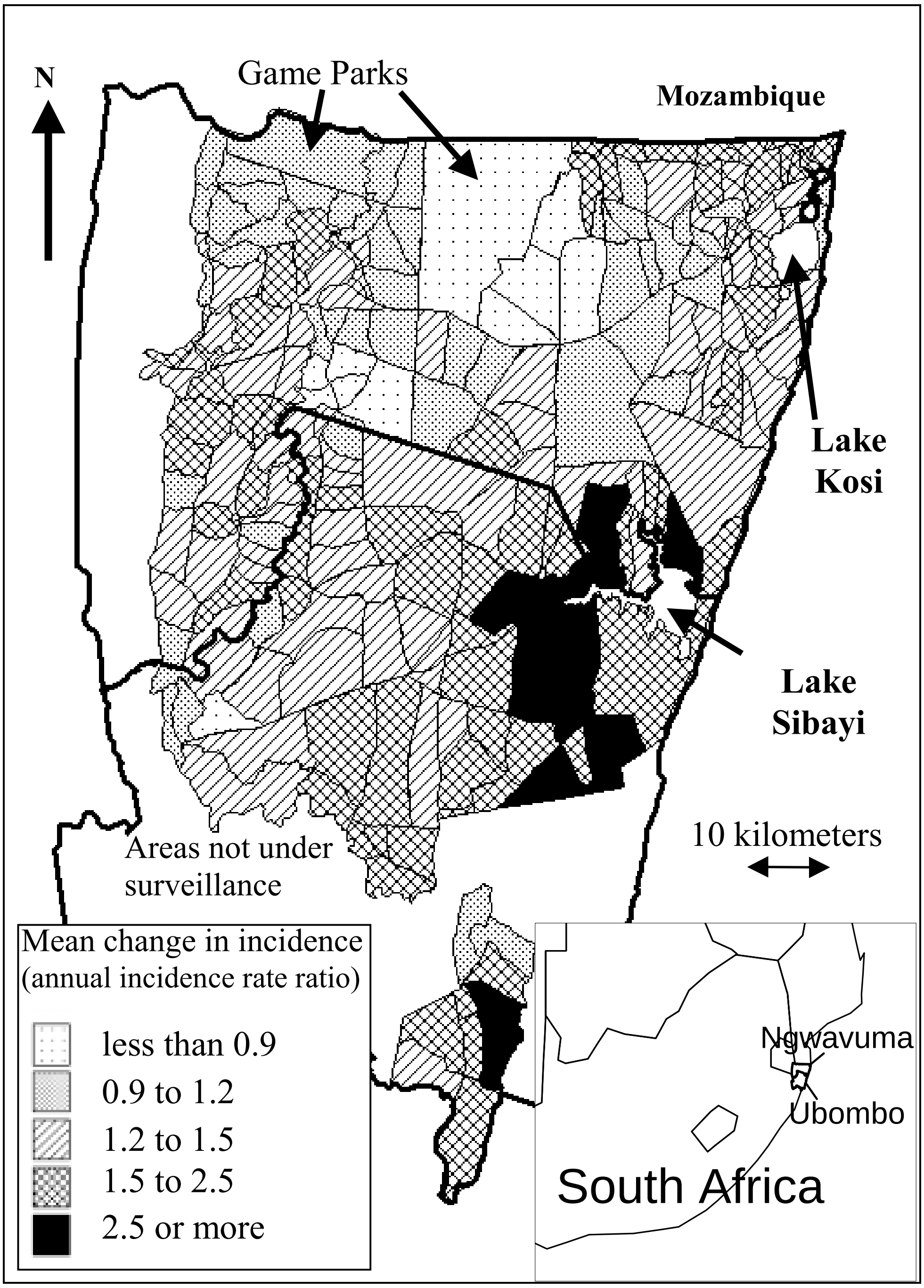


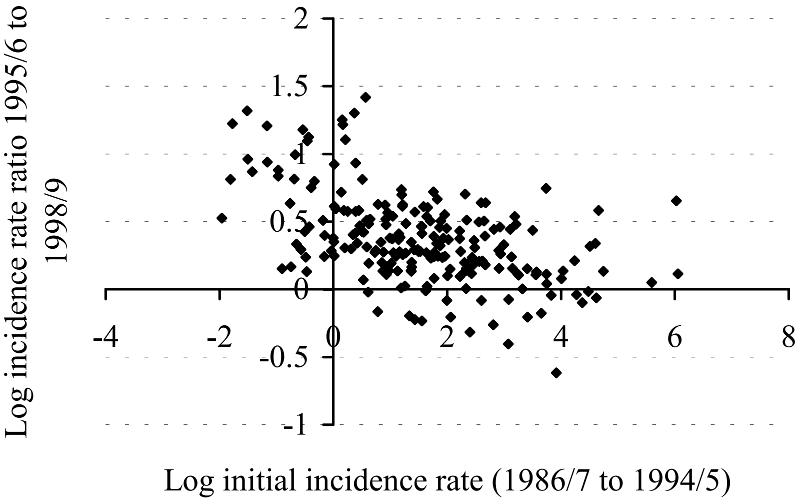
Figure 5.5. Smoothed trend in malaria incidence rates by area estimated from model 2 for the population of Ngwavuma and Ubombo, mid 1995 to mid 1999



The main limitation of this study is its dependence on data obtained from the provincial malaria control programme. We are concerned that under-reporting of cases may be worse in high incidence areas due to resources being over-stretched in these areas, which may have biased our trend analysis. Furthermore, we were unable to incorporate possible between-area population movements into our analysis.

In conclusion, the main finding of this retrospective analysis is that the spatial distribution of the rise in malaria incidence is uneven and strongly suggests an extension of high-risk malaria areas in South Africa. There is evidence of a stabilisation of incidence in areas with the highest rates before the current escalation of rates began. The impression that areas immediately adjoining the Mozambican border have undergone larger increases in incidence, in contrast to the general pattern of stabilizing of rates in the more northern, high baseline incidence areas was not confirmed by modelling. Smoothing of small area maps of incidence and growth in incidence (trend) is important for the interpretation of the spatial distribution of disease and rapid change in disease incidence. The analysis of time trend in relation to baseline incidence provides a useful means to describe geographical expansion of disease risk, provided that precise estimates of baseline incidence can be made.

Figure 5.6. Plot of log trend (1995/6 to 1998/9) in malaria incidence against log initial malaria incidence rate (1986/7 to 1994/5) of each area for the population of Ngwavuma and Ubombo



Acknowledgements

The authors would like to thank Dr Tom Smith and Dr Linda Haines for their valuable comments on earlier drafts of this paper. Partial funding of this work by the Swiss

National Science Foundation (grant SNF 3200-057165.99), and by the MARA/ARMA (MARA/ARMA, 1998) project is gratefully acknowledged.

Chapter 6

Space-time modelling of small area malaria incidence in relation to remote-sensed inter-annual climatic variation

Immo Kleinschmidt[1], Penelope Vounatsou[2], Brian Sharp[1], Bronwyn Curtis[1], Simon Hay[3], Jonathan Cox[4].

[1] South African Medical Research Council, Durban, South Africa.

[2] Swiss Tropical Institute, Basel, Switzerland.

[3] Trypanosomiasis and Land-use in Africa (TALA) Research Group, Dept of Zoology, University of Oxford, UK.

[4] London School of Hygiene and Tropical Medicine, UK.

This paper is being prepared for submission to *International Journal of Epidemiology*

Summary

Health services in areas of unstable malaria are often under severe strain as result of unexpectedly large numbers of malaria cases during epidemic outbreaks. Advance warning of such outbreaks can assist in putting in place adequate diagnostic and treatment facilities, as well as other counter measures. Remote sensors onboard earth orbiting satellites have been shown to have the potential for predicting unusual malaria seasons by monitoring changes in weather conditions. In this study remote sensed (RS) proxy measures for rainfall and temperature for small areas, namely cold cloud duration (CCD) and land surface temperature (LST) respectively, were modelled in relation to space-time variation in malaria incidence rates for small areas in a region of unstable malaria transmission in South Africa over a four year period. Autoregressive Bayesian space-time models using Markov Chain Monte Carlo (MCMC) methods were used to take account of and to estimate spatial and temporal correlation in the data. A number of simple models which included quarterly values of CCD and LST as covariates were investigated. There was strong correlation of incidence rates between years and between neighbouring areas for individual small areas. Although the posterior distributions of model coefficients for CCD and LST were significantly different from zero, comparison of EPD for models with and without these explanatory variables showed that they do not improve overall model fit. Our analysis therefore did not provide any real evidence of an association between seasonal RS data and malaria incidence. We recommend that this relationship should be further explored using longer time series, good quality disease and population data, and including data on appropriate confounding variables.

Introduction

In sub-Saharan Africa, malaria remains the major cause of morbidity and mortality, with an estimated 200 million clinical cases annually and approximately one million annual deaths. In areas of unstable malaria on the fringe of endemic malaria areas, and in highland areas of East Africa, malaria is characterized by strong seasonal and regional variation and by epidemics, with case fatality rates of about 1% in largely non-immune populations (Snow *et al.* 1999).

The disease is closely bound to conditions which favour the survival of the anopheles mosquito and the life cycle of the parasite. These conditions are predominantly determined by climatic factors which effect vegetation coverage and access to water surfaces for breeding requirements (Molineaux, 1988; Gillies and De Meillon, 1968; Ghebreyesus *et al.* 1999). Data obtained from sensors aboard earth orbiting satellites have been shown to possess powerful potential for mapping mosquito populations (Thomson *et al.* 1996) and to predict malaria seasons (Hay *et al*. 1998). This study investigates associations between remote sensed (RS) data and small area space-time variation in malaria incidence. It attempts to bring together the potential of on-line weather data at specific locations provided by RS technology, high resolution malaria incidence data, and the analytic framework provided by recent developments in spatial epidemiology (Elliott *et al*. 2000). The ultimate purpose of this investigation is to explore the potential for developing early warning forecasting models of unusually severe malaria seasons.

This study focuses on the neighbouring magisterial districts of Ngwavuma and Ubombo in northern KwaZulu Natal in South Africa. This area is traditional rural in character, and amongst the most underdeveloped in South Africa. According to the 1996 population census, 95% of households had no piped water, 97% of households had no electricity, 74% of households were built of traditional materials, and 52% of the adult population had no formal education. It has been argued that effective malaria control is at the same time both a pre-condition and a likely consequence of economic development of this area.

The two districts are unusual in the sense that a comprehensive small area malaria reporting system has been in operation for a long time (Sharp *et al.* 1999). Transmission is seasonal with strong inter-seasonal and spatial variation and with an exponential overall upward trend in recent years (Chapter 3; Chapter 5; Sharp and Le Sueur, 1996). Hospitals and clinics in the area have often struggled to cope with an influx of patients during periods of rapidly escalating case numbers (Dr. Harvey Williams. Comments made at KwaZulu Natal Malaria Conference, 19th September 2000, Richards Bay, South Africa). Anticipation of forthcoming high caseloads could be used to put counter measures in place by controlling mosquito vector populations, and by preparing health services for large numbers of cases in terms of diagnostic facilities, availability of drugs etc.

We have previously undertaken a small area spatial analysis of malaria incidence in this area (Chapter 3) which showed an association between malaria incidence and average rainfall and temperature of a locality, despite the malaria control measure that are in place. In this study we have extended this analysis to space-time modelling of malaria incidence in relation to the actual climatic conditions during a particular year by making use of small area, time specific climatic data that are available from satellite images.

Methods and materials

Population and case data

A component of the malaria control strategy in Ngwavuma and Ubombo districts is to identify and treat all infected individuals. Passive and active case finding is therefore practised. The latter consists of screening measures by which teams go into the community to encourage individuals who may be suspected of having malaria, to be tested. Whilst this system may not achieve 100 percent coverage, it is thought to identify the vast majority of cases. Low levels of exposure to *Plasmodium falciparum* in the past have made it unlikely that individuals possess clinical tolerance to parasite infection and recover from it without treatment (Molineaux, 1988, pp 936-938).

Although the malaria information system covers many years of case data, we were concerned to restrict our analysis to years for which reasonably good age- and sex specific population data are available at small area level, and to exclude years during which non-climatic factors are known to have had a major effect on malaria transmission in the area. We therefore confined this analysis to the years mid-1993 to mid 1997 which were reasonably close in time to the 1996 population census (the first of its kind in many years), and at the same time avoided the years after 1997, which have been characterized by very high increases in incidence, ascribed to the development of resistance to sulphadoxine-pyrimethamine (SP) in the parasite (Bredenkamp *et a*l. 2000), and resistance to synthetic pyrethroid insecticide in one of the vectors in the area (Hargreaves *et al*. 2001). Mid-year to mid-year totals were used rather than calendar years, to ensure that the cases in any one malaria season are grouped together.

A total of 20,754 malaria cases were extracted from the malaria information system for the period from 1st July 1993 to 31st July 1997 for the districts of Ngwavuma and Ubombo. Since homesteads in the area are geo-coded, cases could be allocated to census enumeration areas (EAs). The number of cases that had to be excluded since they could not be linked to enumeration areas was 4,414 (21%). EAs that straddle the outside boundary of the surveillance area, and cases belonging to them, were also excluded from the study. The remaining 15,166 cases were those that could be linked to corresponding census populations using geographic information systems (GIS). This resulted in a final set of 268 EAs after amalgamating eleven EAs that had zero population counts with a neighbouring EA.

Incidence rates by sex and by 5-year age group were calculated for the area as a whole over the entire time period of the study. These were then applied to the corresponding population strata of the EAs in order to calculate age- and sex adjusted expected numbers of cases for each EA, for each year.

The total population of the study area was 234,630 persons at the time of the 1996 census. Adjusting this downwards for years preceding the census to account for population growth (Statistics South Africa, 1999) leads to an overall crude mean incidence rate of 16.7 cases per 1000 person years for the study period. Population per area ranged from 6 to 2734 persons (median=855, mean = 875 , SD = 427). Crude incidence rates per area ranged from 0 to 473 cases per 1000 person-years (median = 3, interquartile range 0 to 356 per 1000).

Climate data

Climate data for the study area were derived using Advanced Very High Resolution Radiometer (AVHRR) and High Resolution Radiometer (HRR) imagery from National Oceanographic and Atmospheric Administration (NOAA) and European Meteorological Satellite Programme (EUMETSAT) satellites respectively.

1. HRR data for rainfall. In the African tropics, rain-bearing clouds are predominantly the result of strong convective currents and a predictable relationship has been shown to exist between cloud-top temperature and the probability of rainfall (Burt *et al.* 1995). Cloud-top temperatures can be obtained routinely using thermal infrared images from Meteosat's channel 2, which are collected every half hour. This information is most commonly presented as cold cloud duration (CCD), which is a measure of the amount of time a given image pixel is covered by cold cloud within the compositing period. Given that the relationship between cloud temperature and precipitation varies over space and between seasons, cold cloud thresholds should ideally be derived empirically for individual regions. In this study mean monthly CCD from FAO's African Real Time Environmental Monitoring Using Imaging Satellites (ARTEMIS) project, which covers the whole of Africa at a spatial resolution of 7.6 × 7.6 km, were used. The ARTEMIS data use seasonal thresholds (-50 °C in summer and –60 °C in winter) for areas between 0 and 27 °N (Snijders, 1991), and a single annual threshold of –40 °C for all other areas. Despite the simplicity of this approach, it has been shown that the relationship between CCD and measured rainfall remains relatively robust over the continent as a whole (Hay and Lennon, 1999) and for this reason CCD remains a plausible proxy for rainfall even in the absence of local calibration.

2. AVHRR data for temperature. AVHRR data were obtained from the Global Inventory Monitoring and Modelling Systems group, GIMMS, at Goddard Space

Flight Center. The data have a spatial resolution of 7.6 × 7.6 km and represent maximum value monthly composites. Land surface temperature (LST) was derived from surface brightness channels 4 and 5 using a split window algorithm which takes into account the attenuating effect of water vapour in the atmosphere (Price, 1984; Hay and Lennon, 1999). It should be stressed that while LST is commonly used as a proxy for ambient temperature, the relationship between the two variables is not always straightforward.

For each EA mean monthly values of CCD and LST were calculated from the pixel values that cover the individual areas. These monthly values were then averaged over three month periods leading to quarterly values of CCD and LST for each year, for each EA. Each unit of analysis therefore represents one EA for one particular year, with eight potential explanatory variables consisting of early winter, late winter, early summer and late summer quarters for CCD and LST respectively.

Quarterly averages of CCD had a mean value of 137 hours per month over all areas over the entire period (SD=117, range 0 to 465, median 88, inter-quartile range 54 to 211). Quarterly values of LST had a mean value of 33.5°C over all areas over the entire period (SD=6.1, range 18.1 to 45.3, median 33.0, inter-quartile range 19.2 to 38.4°C).

Modelling

Mapping and modelling the geographical distribution of disease incidence is complicated by the fact that the rates are often subject to considerable random error. This is particularly true if either the disease is rare or the population per spatial unit is small, so that the rate may be influenced by a relatively small number of cases. This leads to maps which show a misleading picture of the true underlying relative risks since attention is drawn to those areas whose rates are based on the least stable estimates (Cuzick and Elliott, 1992). Moreover, failing to account for the anticipated similarity of relative risks in nearby or adjacent regions will lead to bias in the estimation of co-variate effects and their standard errors. These problems can be overcome by spatial smoothing of the rates, which is based on “borrowing strength” from neighbouring regions. Hierarchical fully Bayesian spatial modelling using Markov Chain Monte Carlo (MCMC) as described by Bernadinelli and Montomoli (1992) is an approach to spatial modelling of disease rates that has found widespread application in recent years (Elliott *et al.* 2000). This approach models spatial correlation via conditional autoregressive (CAR) priors (Clayton and Kaldor, 1987). The Bayesian approach to spatial modelling has been extended to spatial-temporal models by several authors (Heisterkamp *et al.* 2000; Waller *et al.* 1997; Bernadinelli *et al.* 1995; Sun et al. 2000; Xia and Carlin, 1998). We have applied these methods to the problem of modelling the spatial and temporal distribution of malaria incidence in relation to space and time varying climatic factors.

Let Y_{it} and E_{it} denote the observed and expected counts of cases respectively for the i^{th} area in the t^{th} year. E_{it} are the expected counts under the null hypothesis of homogenous risk over space and time. If μ_{it} denotes the log of relative incidence then the expectation of $Y_{it} = E_{it}\exp(\mu_{it})$ denotes the mean count of cases. It is assumed that the Y_{it} are conditionally independent given the μ_{it} and E_{it} and follow a Poisson distribution, i.e. $Y_{it} \sim \text{Poisson}(E_{it}\exp(\mu_{it}))$. The μ_{it} are defined using appropriate linear models which may include covariate terms as well as random time and area effects.

Appendix 3 gives details of how the prior probability distributions of the area effects can be specified to model spatial auto-correlation, and how prior distributions for the year effects can be specified to model temporal auto-correlation.

We used an adaptation of the CAR model described by Sun *et al.* (2000). This enabled us to estimate an index of spatial dependence ρ_1 ($|\rho_1| <1$), which expressed the extent of spatial correlation between neighbouring areas. Between-year correlation was assessed by estimating a temporal correlation coefficient ρ_2 in a first order autoregressive (AR(1)) model which was used to specify the temporal term δ (see Appendix 3).

We fitted the model

$$\mu_{it} = \mu + \beta x_{it} + \varphi_{it(t=1)} + \delta_{it(t>1)}$$

where φ_{it} and δ_{it} are random effect terms that capture spatial structure (area effects) and temporal heterogeneity (time effects) respectively, μ is the overall log-relative risk for all areas and time periods, x_{it} represents one of the RS variables for area *i* and year *t*, and β represents the corresponding regression coefficient. In this model the area effects ($\varphi_{it(t=1)}$) are fitted for the first year only and the time effects ($\delta_{it(t>1)}$) are fitted for years after year 1 only. The area-specific time effects for year 2 are allowed to be correlated with the area effects of year 1 (see Appendix 3 for details). We chose this model because it explicitly models between year correlation in incidence rates for individual areas, as well as spatial correlation between neighbouring areas. We did experiment with other possible models, but they either resulted in a much poorer fit, or they failed to converge due to over-parameterisation.

We used the model without covariates to obtain posterior distributions of predicted values by sampling from Poisson($E_{it}\exp(\mu_{it})$). After converting the predictions to standardized incidence rates, we mapped these together with 2.5 and 97.5 percentiles so that the map estimates can be displayed together with the variation of prediction error in our map.

In order to compare the models, we calculated the expected predictive deviance (EPD). Smaller values of EPD are indicative of a better model. The EPD consists of two components, the likelihood ratio statistic (LRS), which assesses goodness of fit, and, a penalty term PEN which penalises for over- or under fitting (details are given in Appendix 3).

Results

The results (table 6.1) show that whilst all but one of the regression coefficients of the eight climatic covariates were significantly different from zero, the reduction in deviance of models including a co-variate was very modest compared to the model without co-variates. Most multiple covariate models (not shown) resulted in convergence failure, probably due to significant correlation between these covariates.

Furthermore, the results indicate very strong spatial correlation (ρ_1) between neighbouring EAs, and temporal correlation between years for individual EAs (ρ_2)

(table 6.1) The spatial correlation index remained virtually unchanged indicating that the spatial correlation in incidence is not explained by any of the climatic co-variates.

Table 6.1. Posterior medians (95% credible intervals) for spatial correlation index (ρ_1), temporal correlation coefficient (ρ_2), coefficients of climatic co-variates (β), and deviances to assess model fit, using the model $\mu_{it} = \mu + \beta x_{it} + \varphi_{it(t=1)} + \delta_{it(t>1)}$ where φ refers to effects with CAR priors, δ to effects with AR(1) priors and μ_{it} is the log relative incidence

	Median β (95% CI)	**Spatial correlation: Median ρ_1 (95% CI)**	**Temporal correlation: Median ρ_2 (95% CI)**	**EPD**	**LRS**	**PEN**
Model without co-variates		0.989 (0.986 – 0.990)	0.85(0.81 – 0.88)	1860	568	1292
Models with covariate: Climatic variable, x						
Temperature[1]						
early winter[2]	1.87(1.54 – 2.19)	0.989 (0.984 – 0.99)	0.83(0.80 – 0.87)	1852	561	1291
late winter	0.52(0.43 – 0.61)	0.989 (0.985 – 0.99)	0.82(0.79 – 0.85)	1843	549	1294
early summer	0.86(0.73 –0.99)	0.988 (0.982 – 0.989)	0.86(0.82 – 0.89)	1843	570	1273
late summer	-0.82(-1.1 - -0.56)	0.989 (0.986 – 0.99)	0.83(0.80 – 0.86)	1851	549	1302
Rainfall[3]						
early winter	0.18(0.16 – 0.21)	0.988 (0.983 – 0.99)	0.86(0.83 – 0.88)	1855	578	1277
late winter	0.017(-0.001-0.037)	0.989 (0.986 - 0.99)	0.85(0.81 – 0.88)	1857	558	1299
early summer	0.009(-0.002 – 0.021)	0.989 (0.985 – 0.99)	0.85(0.81 – 0.88)	1857	571	1286
late summer	0.10(0.08 – 0.12)	0.989 (0.984 - 0.99)	0.84(0.82 – 0.87)	1855	577	1278

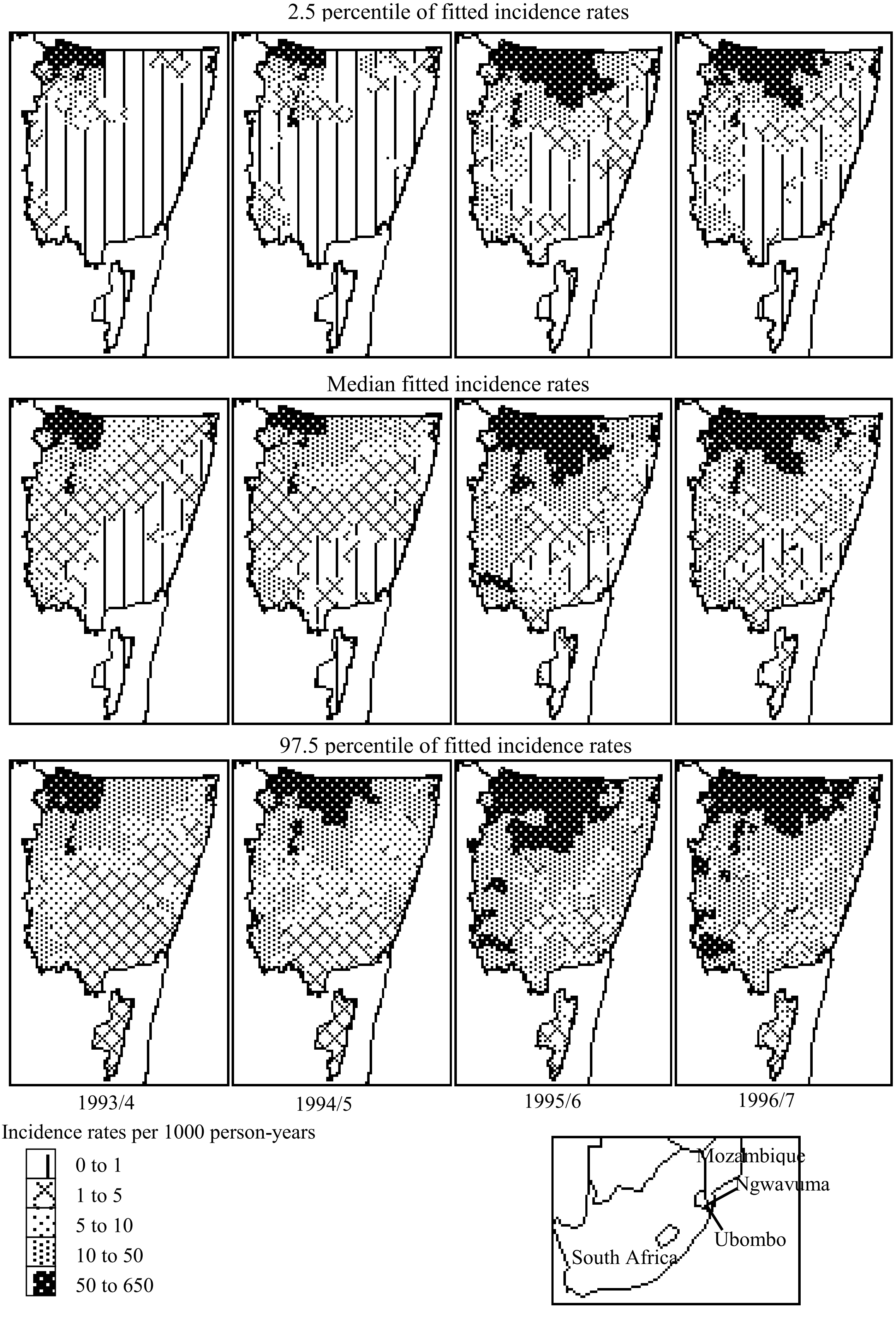
[1] Temperature = Land surface temperature

[2] Early winter = Refers to quarter from April to June of following season, where season extends from 1st July to 30th June of following year. Other quarters follow chronologically.

[3] Rainfall = rainfall proxy as given by cold cloud duration

Figure 6.1. Maps of median and 95% credible intervals of fitted incidence rates predicted from model

$$\mu_{it} = \mu + \varphi_{it(t=1)} + \delta_{it(t>1)}$$



2.5 percentile of fitted incidence rates

Median fitted incidence rates

97.5 percentile of fitted incidence rates

Our map of model predictions (fig. 6.1) shows the spatial and temporal distribution of incidence together with the corresponding credible interval. This shows that the estimate of incidence is less reliable in the southern low incidence areas of the region, than in the north and confirms the conclusion of very strong spatial correlation in rates. The maps of median estimates show that the highest incidence rates are concentrated in the north-west corner of the study area during the first year of the study. By years 3 and 4 the highest incidence category had extended eastwards mainly along the Mozambican border. This eastward extension is also evident in the 2.5 percentile map, which can be regarded as a minimum estimate of underlying incidence. The 97.5 percentile maps can be thought of as indicating the distribution of maximum estimates of transmission. These maps show that the highest incidence rates in this upper bound of estimates of underlying transmission are located in areas along the northern border as well as some areas in the south west of the region.

Discussion

The paucity of detailed epidemiological data on malaria in Africa has limited the scope of analysis of RS in relation to malaria transmission (Thomson et al. 1997). Nevertheless RS data have been successfully used to predict malaria seasons by using hospital admissions covering extensive catchment areas (Hay *et al.* 1998), and to establish correlations between RS variables and vector abundance at specific sites where entomological data were available (Thomson *et al.*1996). In this study we have sought to exploit geo-referenced reporting data for two entire districts to analyse variation of incidence in space and time in relation to variation in weather data in space and time. Furthermore, we have been able to use Bayesian modelling approaches to disease mapping, which have hitherto not been employed in this context.

In our results, the absence of any appreciable difference in model fit between models with co-variates, and the model without co-variates, leads us to conclude that our data do not provide any real evidence in favour of an association between the two RS weather variables LST and CCD, and seasonal malaria incidence. Nevertheless, the coefficients of six of the eight weather variables, when included in the model individually, were significantly different from zero. Apart from average temperature during the second summer quarter, the direction of these associations were all positive. Increased rainfall and temperature would generally be expected to have a positive effect on malaria incidence, particularly in malaria “fringe” areas, where these two climatic factors limit malaria transmission intensity (Molineaux 1988; chapter 3). The negative effect of late summer maximum LST on malaria incidence is plausible since the highest temperatures reduce mosquito longevity with the result that fewer of them survive the incubation period (sporogony) of the parasite (Craig *et al.* 1999). Alternatively, the negative coefficient of late summer temperature could be due to confounding with rainfall, i.e. the hottest places and years are also the driest.

Our conclusion of a negative finding regarding the association between malaria incidence and the two RS variables, is therefore somewhat tempered by the fact that the significant model coefficients presented in table 6.1 are all biologically plausible. A negative finding does, of course, present no evidence against the hypothesis of an association between variation in malaria incidence and temporal variation in weather.

In our analysis we explicitly modelled spatial and temporal correlation between neighbouring areas and successive years. This is plausible in terms of malaria, and is confirmed by the estimates of spatial and temporal correlation from our model. According to our results malaria incidence in an area is significantly correlated with incidence during the preceding year i.e. low incidence in one year in an area will have some protective effect on incidence in the following year and vice versa. This may be due to the time required for the environment to respond to changes in weather that lead to an increase or decrease in transmission intensity. For example, it may take more than a single season to replenish diminished water tables following a dry period, or conversely for ground water tables to deteriorate following an above average wet season. Such momentum in environmental factors will add to the momentum of vector populations to build up or decrease, and to the delay due to changes in the number of infectious individuals in the host population supplying a sufficient pool of gametozytes to increase or decrease disease transmission.

Our previous spatial analysis of malaria incidence in the same population (chapter 3) showed significant association between malaria incidence for a particular year, and average temperature and average winter rainfall of an area. This apparent contradiction with the absence of an association in this study poses the question of whether it is possible for malaria incidence to be associated with long term average climatic factors of individual areas, but not with temporal variation of these factors. Could average temperature and rainfall explain spatial heterogeneity in malaria incidence in the area, whilst temporal variation in incidence is driven by other factors? Long term averages in climate determine the reservoir of infection in the human population which probably accounts for much of the strong correlation of incidence between years shown in our results. Changes in weather from one year to the next can therefore only play a limited role in changing transmission in an area. Another possibility is of course that our data of four years simply lacked the power to demonstrate an association between weather and malaria, if it does exist, particularly in view of the multi-factorial causes of malaria transmission intensity. The strong spatial and temporal correlation in incidence rates implies that the number of degrees of freedom in the data set, and hence its power to demonstrate any associations, is considerably less than it would appear to be from the total number of observations.

In this study we have used the annual number of cases of an area as a response variable. This way we were deliberately excluding within-year variation of malaria incidence and RS variables from our analysis, because we were interested in determining the factors that are responsible for between year and between area variation. However, within year variation generally far exceeds between year variation in disease incidence and in weather and this variation is ignored in a year by year analysis. Analysis that uses annual area counts is further weakened by the fact that cases which occur early in the malaria season (early summer) precede the climatic events of the peak of the season (late summer). It may therefore be necessary to carry out an analysis that uses incidence data that are further disaggregated in time. Such analysis would also reveal whether the strong annual temporal correlation we observed is also present over shorter time periods of months or weeks.

In this analysis we have simply used three month averages of CCD and LST. However, it may be necessary in future studies to differentiate, for example, between rainfall in the form of a torrential downpour lasting only a few days on the one hand, and more prolonged gentle rain of the same overall quantity on the other.

A more intractable difficulty in using RS data for small area disease analysis is the considerable mismatch in resolution between satellite data, and small area disease data. The pixel size of 57.8 km^2 of the ARTEMIS and AVHRR data that were used in this study is of much lower resolution than what is usually thought of as small area resolution of disease and population data. The EAs that formed the areal unit of analysis in this study had a mean area of 21 km^2 (SD=27, median area= 13, inter-quartile range 5 to 25). As a consequence we had to allocate the same value of CCD and LST for a particular time period to all EAs within the same pixel. This difference in spatial resolution between satellite and disease data is likely to be even bigger in more densely populated areas.

There are also likely to be substantial errors of under-reporting and misallocation of both cases and populations in our small area disease data. These shortcomings and the ones mentioned above all contribute towards a dilution of any potential association between weather and malaria incidence. Some of these difficulties can only be overcome by prospective studies, for example by conducting consecutive malariological surveys that are located in a grid that coincides with the satellite images on the ground.

Our results show that malaria incidence in an area is significantly related to incidence during the previous season but show no conclusive evidence of an association between variation in malaria incidence and variation in climatic factors. We would therefore suggest that studies need to be undertaken that consist of longer time series, to further explore the potential of RS data to forecast malaria incidence and to validate such models against data of subsequent years. Other RS variables may need to be explored, and possibly other data reduction techniques need to be applied (Hay *et al.* 1998). Factors that restrict such modelling at present are the availability of good data, particularly age-sex stratified population data, and data on important factors that confound the relationship between climate and malaria transmission intensity, such as insecticide house spraying and drug and insecticide resistance. We therefore recommend that the collection of comprehensive data for such modelling should include these factors.

In this study we have shown that by plotting disease maps together with the estimation confidence interval it is possible to extend to map predictions the general principle that estimated quantities should be quoted together with a measure of uncertainty in the estimate. This is easily accomplished as part of the MCMC procedure and we recommend that this practice should be carried out wherever possible in disease mapping.

The strong temporal correlation in incidence rates implies that for most areas between year changes in incidence rates are only very minor. However, there are obviously exceptions to this, and these are the instances when health services become overwhelmed with the influx of new patients. To be able to predict these instances with a reasonable degree of probability would be the real value of an early warning system.

Acknowledgements

We would like to thank the following: Compton Tucker at NASA's Goddard Space Flight Centre for making available the AVHRR data; Fred Snijders of the FAO African Real Time Environmental Monitoring Using Imaging Satellites (ARTEMIS) project for permission to use Meteosat CCD data. We are also grateful to Dr Linda Haines and Dr Tom Smith for important comments on an earlier draft of this paper. Partial funding of this work by the Swiss National Science Foundation (grant SNF 3200-057165.99), and by the South African Medical Research Council is gratefully acknowledged.

Chapter 7

Patterns in age-specific malaria incidence in a population exposed to low levels of malaria transmission intensity

Immo Kleinschmidt and Brian Sharp

South African Medical Research Council, P.O.Box 17120, Congella, Durban 4013, South Africa.

This paper is in press at *Tropical Medicine and International Health*.

Summary

The population of the northern part of the province of KwaZulu Natal in South Africa has experienced low levels of malaria transmission intensity for many years. We investigated the widely held assumption that individuals in this population do not develop clinical tolerance to infection with *P.falciparum*. We calculated malaria incidence rates by five year age groups from a comprehensive small area malaria reporting system and from national census data for the period from mid 1990 to mid 1999. Incidence rates were plotted against age groups for each of the nine malaria seasons, and by quintile of crude incidence rate. These show that age specific incidence varied considerably in areas of high incidence and in years of high incidence. In these areas malaria incidence rose with age until the late teens, and either remained constant or decreased in young adults. This finding appears to be consistent with results from settings of much higher transmission intensities which show that clinical tolerance to infection with *P.falciparum* in adults may be acquired as a result of a small number of infective bites in early childhood and implies that even in this relatively low transmission area, there is an asymptomatic reservoir of infection in older people. The results also show that in high incidence sub-regions the lowest incidences are reported for children under five years of age, which may be the result of greater protection offered to this age group by malaria vector control through indoor house spraying.

The relationship between the pattern of age-specific malaria morbidity and malaria transmission intensity has been well documented (Molineaux, 1988; Snow *et al.* 1997; Snow and Marsh, 1998b). In areas of high transmission intensity this generally shows that the incidence of clinical attacks peaks in early childhood and then declines rapidly with increasing age due to the acquisition of clinical immunity in such populations. In areas of moderate transmission intensity the age of peak transmission occurs at a later age, whereas in populations exposed to very low levels of transmission or to epidemic malaria, the risk of infection remains constant across all ages. This has been shown to be the case for both mild as well as severe clinical malaria (Snow and Marsh, 1998b). The pattern of age specific malaria incidence can therefore serve as an indication of the presence of any naturally acquired immunity in a population.

Historically the population of northern KwaZulu Natal exhibited the features characteristic of an endemic malaria setting, but due to long term intensive vector control through house spraying with residual insecticide, it changed to that of epidemic malaria (Sharp *et al.* 1988). Data from the provincial malaria information system show that overall incidence rates in the two northernmost magisterial districts of Ngwavuma and Ubombo (figure 7.1) varied from less than 2 cases per 1000 pa to 25 cases per 1000 pa during the years from 1985 to 1995. Despite this low level of overall malaria transmission there have been pockets of much higher transmission within this area, even in years of low incidence (Chapter 3). In the second half of the 1990's there has been a sharp increase in incidence in almost all parts of this region (Sharp *et al.* 2000).

In this study we report on variation in the pattern of age-specific incidence over the years 1990 to 1999, and on variation in this pattern between low and high incidence sub-regions of the area.

Methods and materials

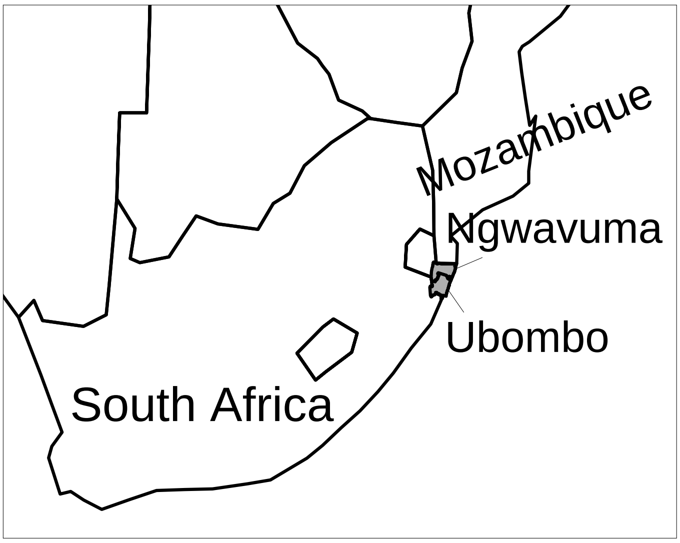
The malaria information system that has been in operation in Ngwavuma and Ubombo for many years has been described elsewhere (Sharp and le Sueur, 1996; Sharp *et al.* 1999). In essence, this records all parasitologically confirmed cases, both passive as well as those found by active surveillance. The latter consists of screening measures by which teams go into the community to encourage individuals to be tested. These may be individuals with non-specific symptoms such as fever, or simply people living near or in the same homesteads as recent confirmed cases. However, the distinction between active and passive cases is not always strictly adhered to, and some cases classified as active are in actual fact patients presenting for treatment. Low levels of exposure to *Plasmodium falciparum* in the past have made it unlikely that individuals possess clinical tolerance to parasite infection and to recover from it without treatment. It has therefore been assumed up to now that active and passive cases together represent the overall sum of all new infections, and that there are generally no asymptomatic cases, apart from recent migrants from Mozambique.

Every case is linked to their residential homestead. All homesteads have been located by global positioning systems (GPS) and their co-ordinates stored in a database. The exact residential location of each incident case is therefore known.

South Africa's first post-apartheid national census, which was carried out in 1996, gives population counts by age and sex for small areas known as enumerator areas (EAs). Boundaries of EAs are available in digitized form. By means of geographic information systems (GIS) software it was therefore possible to link all cases to EAs that are wholly contained within the malaria surveillance area of Ngwavuma and Ubombo. EAs that straddle the outside boundary of the surveillance area, and cases belonging to them, were excluded from the study. This resulted in 279 EAs whose populations could be linked to all cases arising from them. A uniform annual population growth rate of 2% was assumed to project population totals forward and backward in time from the census year (Statistics South Africa, 1999). The study was restricted to the time period from 1st July 1990 to 31st June 1999 in order to limit the effect of the simplifying assumption of uniform population growth, and zero population movement between EAs. Annual case totals by five year age group and by annual malaria season were extracted from the malaria information system for each EA. A malaria season was taken to be the period from the beginning of July to the following end of June. Cases and populations over 40 years of age were aggregated into one age-group.

For each of the nine malaria seasons, age specific incidence rates for the entire area were calculated by aggregating cases and populations over all EAs for individual years and age-groups. To see if the pattern of age-specific incidence rates varies between areas of low and high transmission intensities, the crude incidence rate of each enumeration area was used as a proxy for its transmission intensity. EAs were divided into quintiles of crude incidence rate, excluding EAs with zero incidence. This was done for incidence rates of EAs within individual years, and for all annual incidence rates of individual EAs combined over the study period. Incidence rates for five-year age-groups were then computed separately for each quintile of malaria incidence. The number of EAs per quintile varied from 13 for the year of lowest incidence, to 48 for the year of highest incidence.

Figure 7.1. Map showing location of study area.



Results

At the time of the 1996 census, the total population of the study area as delineated by the 279 EAs was 228,806 persons. For this area, a total of 37,303 cases were recorded during the nine years of the study period. Annual incidence rates for the whole area range from 5 per 1000 pa in 1990/91 to 43 per 1000 pa in 1998/99.

Figure 7.2 shows that there has been a gradual change in the pattern of age specific incidence rates for the entire area over the nine-year study period. Incidence rates rise quite sharply with age up to age 15, and then remain nearly constant or decline with age from the 1992/1993 season onwards. In the earlier years the variation in incidence with age is less pronounced but not absent.

Figure 7.3 shows that the pattern of rising incidence rates with age for children up to their mid- to late teens is mainly a phenomenon of areas with incidence rates above 14.4 per 1000 pa (quintile 4 and above). The marked decline in incidence in adulthood is confined to areas with incidence above 37.9 per 1000 pa (quintile 5). Plotting age-specific incidence rates by quintile of overall incidence rate for individual years shows a similar pattern for the entire time period, although it is less marked for low incidence years (data not shown). Figure 7.4 shows the location of EAs belonging to the highest quintile of incidence for the first and the last year of the time series.

Discussion

Since the data are obtained from a large database of all recorded cases, the observed variation of incidence by age is very unlikely to be due to chance. Our analysis shows that there is a highly uneven distribution of malaria incidence across age-groups in high incidence areas. This is evident since at least the early nineties, but the phenomenon was masked by the majority of areas which had very low incidence rates in the earlier years of the decade (figure 7.2). Once the age curves are shown separately for low and high transmission intensity areas, variation in incidence with age becomes apparent in the latter. Closer scrutiny of published data covering a much earlier time period also suggest that incidence amongst under fives was lower than for other age-groups (Sharp *et al.* 1988 p.104).

Figure 7.2. Age specific incidence by malaria season. Age is by midpoint of 5-year age-group, with over 40 combined into one group. Overall incidence rates in cases per 1000 pa (by season) were: 4.6 (1990/91); 1.7 (1991/92); 13.5 (1992/93); 12.5 (1993/94); 8.3(1994/95); 26.0(1995/96); 24.0(1996/97); 30.4 (1997/98) and 42.9(1998/99).

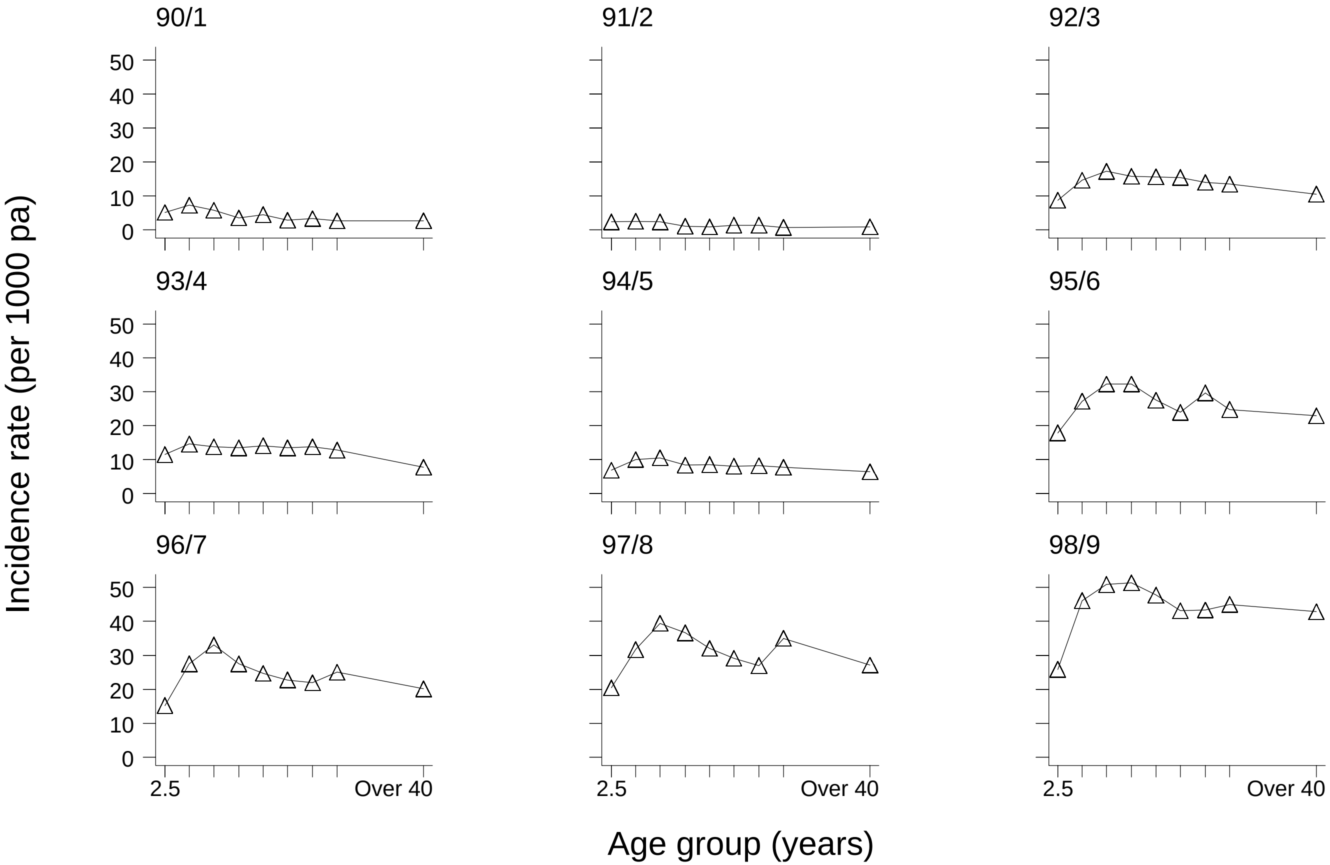
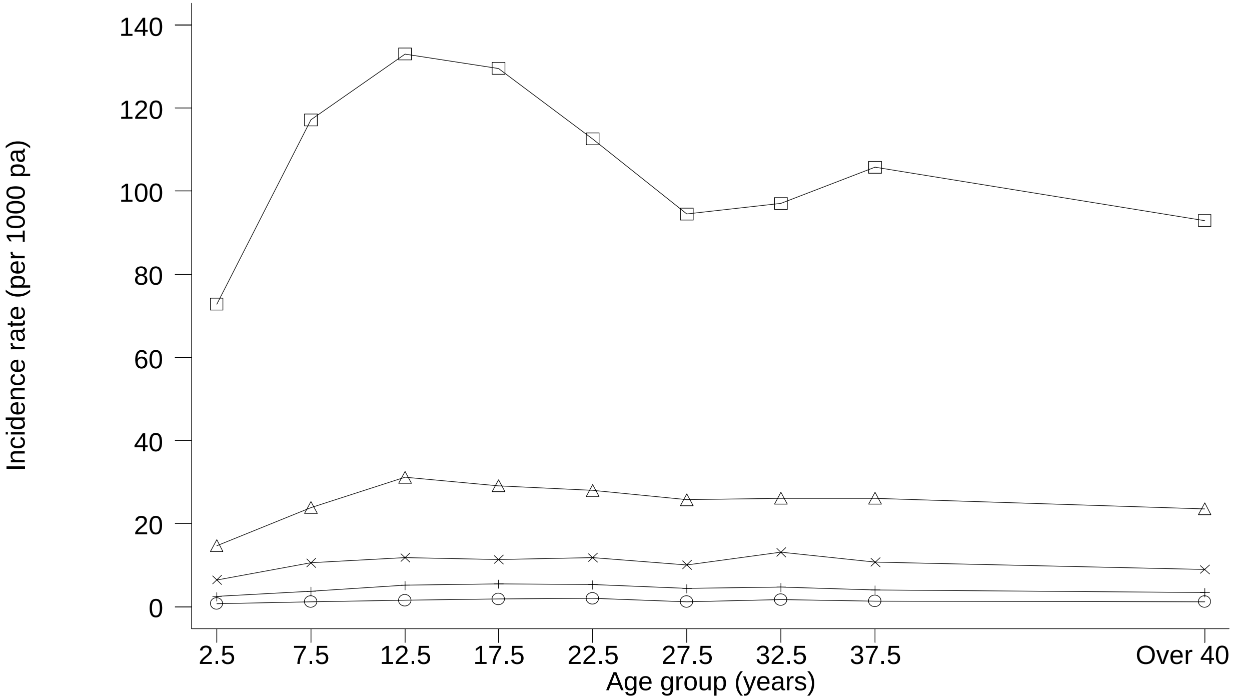


Figure 7.3. Age specific incidence by quintile of overall incidence of EA, all years combined. EAs with zero incidence excluded. Age is by midpoint of 5-year age-group, with over 40 combined into one group. Quintile 1 ()) corresponds to overall incidence rates below 2.2 per 1000 pa; quintile 2 (2) to incidence rates between 2.2 and 6.1 per 1000 pa; quintile 3 (3) to incidence rates between 6.1 and 14.4 per 1000 pa; quintile 4 (+) to incidence rates between 14.4 and 37.9 per 1000 pa; quintile 5 (э) to overall incidence rates above 37.8 per 1000 pa (median 77.8)



In endemic malaria areas the peaking of incidence at a much younger age is thought to be associated with the acquisition of clinical immunity due to intense challenge of the immune system early in life, when other protective mechanisms may still be operating (Snow *et al.* 1997). The incidence curves shown here are fundamentally different in that they show a peak in late teenage years, and a decline with age in adulthood. We do not have sufficiently accurate data to investigate single year age-specific incidence for early childhood, but it seems likely that any effects between single years would be swamped by the steep variation between five-year age groups that are evident from our data. The only possible explanation for the steadily rising incidence with age during childhood can be an increase in chance infections as children become older. All curves show that incidence is lowest in the under 5 age group, possibly because children at this age are indoors at night and benefit from the protection offered by indoor house spraying. As they get older, their sleeping patterns may be less regular and they may therefore become more at risk of infective bites by mosquitoes. However, it is unlikely that adults are at lower risk of exposure to P. *falciparum* than teenagers. It has been shown that a small number of infective bites in early childhood may be sufficient to yield clinical protection in adults (Gupta *et al.* 1999; Snow *et al.* 1998c). It is therefore plausible that some adults in high transmission areas are less susceptible to clinical attack due to the acquisition of some immunity early in life. This does not appear to be the case in low transmission intensity areas, where the incidence curve is more or less flat. A possible alternative explanation would be some behavioural factor, for example the possibility that adults spend more time outside the area to seek work on a temporary or commuting basis. There is however, no reason why this should be the case for high incidence areas in particular.

Although all cases, both active and passive, were parasitologically confirmed, it is possible, though unlikely that misdiagnosis, for example of paediatric fevers, may have lead to lower recorded incidence in young children. Misdiagnosed children would be unlikely to recover without anti-malarial therapy, unless they already possessed a measure of clinical tolerance, which is unlikely, given the rise in incidence with age in children. The existence of real differences in the rate of detection in different age-groups would therefore lend support to the hypothesis of immunity in some age-groups.

Our results appear to contradict those of a study by Baird *et al.* (1998) in Irian Jaya which showed that severe disease increased with age in a population of non-immune migrants from Java, after relocation to a hyperendemic area. However, the subjects in the Irian Jaya study were exposed to much higher levels of transmission intensity than the population in our study, whilst having had no previous exposure to *P.falciparum*. The decline in incidence with age in adults living in areas of moderate transmission intensity in KwaZulu Natal, in contrast to the adult migrants in Bairds study, can therefore plausibly be explained by the acquisition of immunity due to exposure to *P.falciparum* earlier in life.

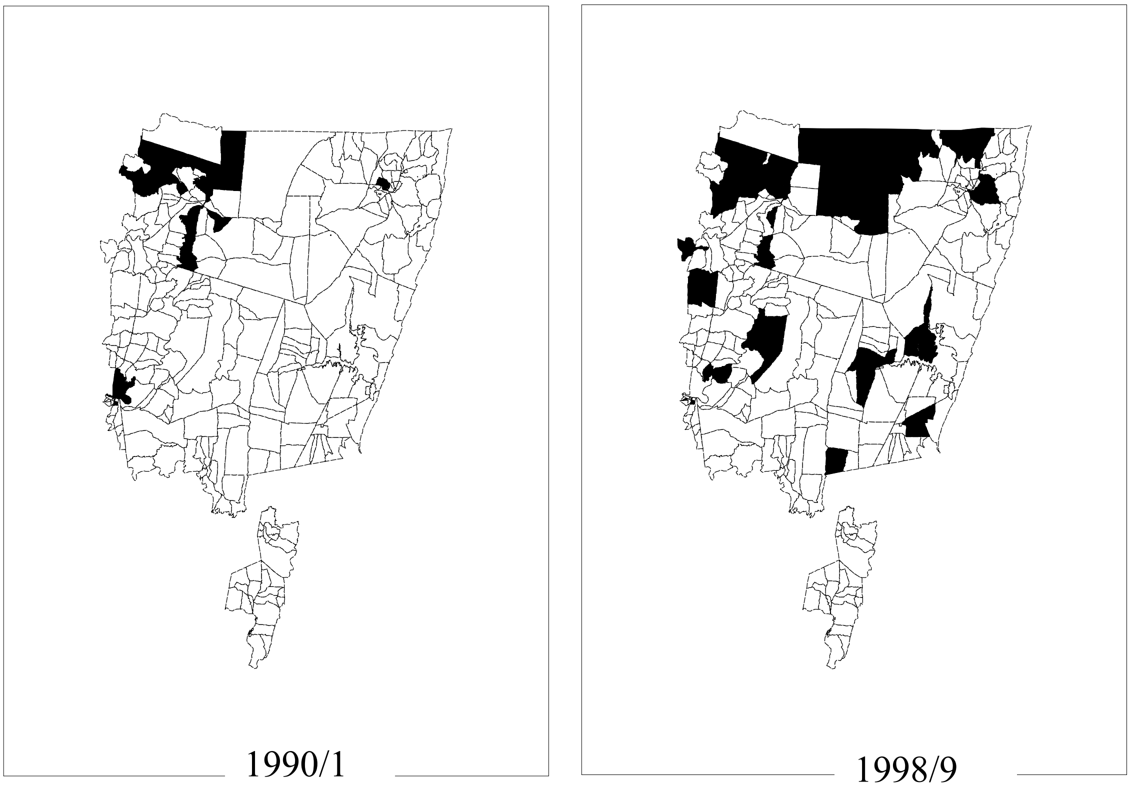
Figure 7.4 shows that EAs in the highest quintile of incidence were primarily located in the northernmost border areas, but this pattern has started changing with the geographic expansion of high malarious areas in recent years. The differentially greater protection against malaria afforded to children under five years of age in these high incidence areas is an important beneficial feature of the vector control measures being practiced. This is underlined by the fact that 71% (26510/37303) of all cases over the nine year period originate from EAs in the top quintile of incidence.

We sought to further investigate the hypothesis of a measure of clinical immunity in adults from high transmission areas by distinguishing between active and passive cases. Unfortunately the practice of recording cases as active only when they have not presented themselves for treatment does not appear to have been followed rigorously, and there are reports of mobile field stations recording cases as active, when in fact they were patients seeking treatment. What the data do show, however, is that the proportion of passive cases decreases sharply with increasing incidence rates, and also with time. If this trend is not artefactual it would support the concept of increasing clinical tolerance with increasing transmission intensity.

A limitation of our study is that it is based entirely on reporting data and it is not possible to derive independent confirmation of our finding from these data. It is, however, unlikely that the pattern we have observed would be maintained over such a long time-period if it was simply due to a reporting error. An additional shortcoming is that the measure of transmission intensity we are using is the overall incidence rate rather than a measurement of exposure derived from a separate source. If regular data on parasite ratios or entomological inoculation rates had been available, we could have investigated the variation in age-specific incidence in relation to transmission intensity derived from these. Nevertheless, what our investigation shows, is that where incidence is high, it is always highly unevenly distributed across age groups.

If it is true that a measure of clinical tolerance to infection with *P. falciparum* is developing in adults in high transmission areas, this does have implications for the current control strategy, since it challenges the assumption that there are no asymptomatic cases. The existence of asymptomatic infected individuals, who are not being treated, undermines the strategy of eliminating sources of transmission by ensuring parasitological cure of all infected individuals. It would also provide an explanation for our previous finding that the recent rise in malaria incidence in this area is highest in areas that previously experienced the lowest incidence rates and vice versa (Chapter 5).

Figure 7.4. Maps showing enumerator areas belonging to the highest quintile (shaded areas) of crude malaria incidence for the years 1990/1 and 1998/9 within the malaria surveillance area of the districts of Ngwavuma and Ubombo in KwaZulu Natal, South Africa.



Chapter 8

Discussion and conclusions

The history of modern epidemiology is often considered to start with the investigation of an outbreak of cholera in London in 1854 by the physician John Snow and his successful intervention to control it (Snow J, 1855). Disease mapping is therefore as old as epidemiology itself. John Snow's study depended on relating 3 essential data sets: a reliable and unbiased count of incident cases, good small area census data and detailed knowledge of the distribution of the exposure, namely the contaminated water. In addition it was necessary to have a geographical entity in which each of these data items could be linked. The same requirements still hold today for the successful execution of studies based on spatial analysis of disease data. All of the studies in this thesis have had to deal with bringing together malariological data, populations at risk, and exposure data, linked in geographical entities that become the unit of analysis. These studies have produced a more detailed picture of the distribution of malaria in parts of Africa. They have also given us a clearer idea of the factors that are associated with malaria incidence and parasite prevalence in children in these areas and how incidence in an area of unstable malaria has changed over time.

One disease, many scenarios

In this thesis spatial analysis and modelling of malaria distribution has covered a wide range of different scenarios, not only in terms of the disease but also in terms of scale and in terms of the data and methodological approaches. These differences were: 1) vastly different levels of spatial resolution: regional (or sub-continental) level in the case of West Africa, country level in the case of Mali, and sub-district level in the case of Kwa Zulu Natal. 2) Vastly differing levels of endemicity ranging from areas of all year transmission in parts of the forest and Guinea savanna zones of West Africa, to seasonal and epidemic prone areas in Kwa Zulu Natal, and in the northern parts of sub-Saharan West Africa. 3) Differences in the measure of transmission intensity: parasite prevalence derived from random surveys on the one hand, reported incidence on the other hand. 4) Differences in the spatial organisation of the data: points representing the population and area in which they are located on the one hand, aggregations within contiguous areas on the other hand. 5) Differences in modelling approaches used: classical geostatistical, generalized linear mixed model on the one hand, hierarchical fully Bayesian spatial models on the other. 6) Differences in the number of dimensions of the data: cross-sectional spatial analysis in some studies, space-time analysis in others.

Maps are outputs of interest in all the studies. In most cases they are the primary end-product that motivates the study. In other cases the models and co-variates that were used are more important, since they reveal the potential for forecasting. In four of the studies, climatic data were used. In five of the studies the issue of spatial dependence in the data had to be addressed. Spatial dependence was approximately inversely related to the scale of the study: weak in the regional and country level analyses, strong in the small area analyses. Spatial dependence can be a complication for which allowance has to be made in the analysis, or it can be an aid to mapping, as in the case of kriging which would not be possible without it, or map stabilisation (smoothing) which is based on spatial correlation. Much the same is true for temporal autocorrelation in the space-time analyses. All the studies were based on historical or observational data, which is a source of problems in the analysis and in the interpretation of the results.

Throughout these studies there has been a duality of purpose: to evaluate, apply and advance spatial modelling methodology to problems of geographical distribution of malaria on the hand, and to produce tools that will assist in the control of malaria on the other. This discussion will attempt to draw together the methodological insights gained from this experience and to summarise the conclusions for malaria control strategies.

Methodological insights gained

What should be done differently if this same exercise was attempted again, with the benefit of hindsight, and what should be done the same way again?

One of the major problems facing the researcher in observational studies is the problem of interpretation. The plethora of potential explanatory variable data make it tempting to try out what is available, resulting in many chance associations. The solution to this problem is to ensure that the study is based on a few solid hypotheses, rather than many vague ones, and to make the threshold for inclusion of an explanatory variable relatively high, to avoid spurious associations from dominating a model. Strategies for combining essentially similar co-variate data should be made explicit beforehand. Ideally, even a retrospective study should be carried out against an agreed protocol. Recent examples of mapping lymphatic filariases based on length of rainy season and maximum and minimum temperature (Lindsay and Thomas, 2000) show what can be achieved with good hypotheses and relatively few explanatory variables. The proven relationship between seasonality and malaria parasite ratios make such an approach feasible for malaria mapping as well (Tanser *et al.* 2000).

When this work was started the MARA database was in its infancy and there was a temptation to include all available data. It might have been better to be more selective, for example in excluding surveys that are atypical or that are of questionable quality, or that reflect urban malaria when the main focus of the study is on rural malaria. It would also be advisable to always exclude a random sample of points against which the validity of the resulting model can be tested.

The experience of the West Africa model showed that it is not feasible to attempt to derive a single model for vastly differing ecological and climatic areas, when dealing with distribution models at this scale. Even once a reasonable division of such a region into ecological zones has been achieved, there is the problem of producing a smooth map for the entire region. A satisfactory empirical approach to this was achieved in the case of West Africa (see appendix 2).

Quite different data problems are encountered when dealing with small area reporting data. Since such data are derived from one information system, rather than many surveys, one needs to be vigilant in looking for artefactual patterns in the data, as opposed to scrutinizing the validity of a single record.

Despite methodological and conceptual difficulties, small area studies of geographical distribution of disease have become widespread in developed countries due to the potential they offer of studying relationships between environmental exposures to pollutants or other hazards, and disease (Elliott et al. 1992; Cuzick and Elliott 1992; see Elliott *et al.* 2000 for further references). Developments in spatial statistical methodology (for example, Bernadinelli *et al.* 1995; Diggle *et al.* 1998; Cressie, 2000) and improvements in small area disease and census data have given the impetus for carrying out such studies, as much as heightened interest in questions of environmental health (see for example, Elliott *et al.* 1996; Heisterkamp *et al.* 2000). The paucity of appropriate population and disease reporting data in developing countries has limited such studies on diseases in poor countries in general and on tropical diseases in particular, despite some notable exceptions (for example Smith *et al.* 1995; Vounatsou *et al* 2000; Schellenberg *et al.* 1998; Mueller 2000). This thesis has in part been an attempt to transfer these developments in spatial epidemiology to a tropical disease setting.

The use of small area population and disease data for the study of the geographical distribution of disease is one such development that has been applied in chapters 6 and 7. The study on age distribution of malaria cases (chapter 7) would not have been possible without age-specific population data being available at small area level. Chapters 6 and 7 are probably the first studies that have used the South African enumeration area census data in this particular way. Even so, inaccuracies in such small area data may have contributed towards the negative result in chapter 6. However, the advantage of using census data is that expected numbers of cases based on the age and sex structure of populations can be computed, and that many socio-economic indicators are available for defined populations.

In these studies essentially two modelling approaches were used. For Mali, West Africa and the first KwaZulu Natal analysis generalized linear or generalized linear mixed models (GLMM) were used, with variograms of model residuals providing the means by which the specification of the co-variance structure of the data could be iteratively improved. Kriging of the residuals of the final model was carried out in each case in order to adjust model predictions where local deviation of observations from the model supported this. This adaptation of standard kriging has proved to be a useful tool whenever disease modelling resulted in unexplained variability which resulted in spatially correlated residuals. Details are given in appendix 1.

In the two space-time studies of the KwaZulu Natal small area incidence rates, Bayesian autoregressive spatial models were fitted. In this approach the posterior distribution

$$[\boldsymbol{\varphi}, \boldsymbol{\beta}, \lambda \mid \mathbf{E}, \mathbf{Y}] = [\mathbf{Y}|\mathbf{E}, \boldsymbol{\varphi}, \boldsymbol{\beta}]\ [\boldsymbol{\varphi}|\ \lambda]\ [\lambda]\ [\boldsymbol{\beta}]$$

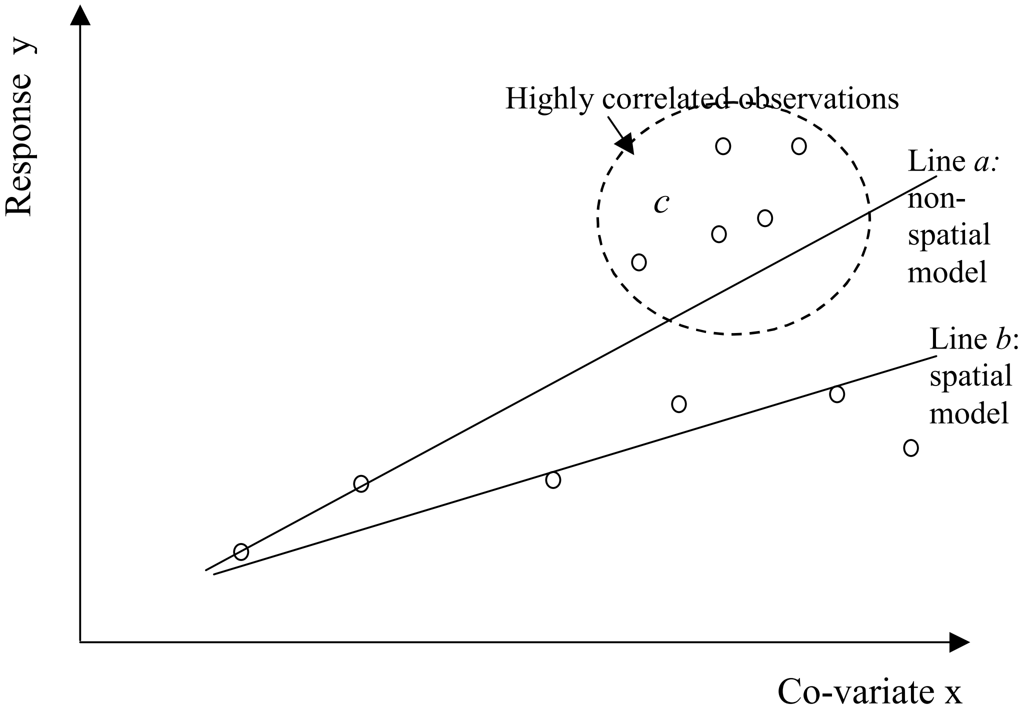
is computed, where **Y** = the vector of observed cases, **E** = vector of expected cases, **φ**= vector of area effects, **β** = vector of regression coefficients of area-specific explanatory variables, and λ represents geographical variability and controls the amount of variation in risk distribution in the map (Bernadinelli and Montomoli, 1992). Prior distributions are specified for the parameters **φ**, μ and λ, as described in chapters 5 and 6, and in appendix 3. Since it is not possible to obtain parameter estimates of the posterior distributions analytically, these have previously been estimated via the empirical Bayes approach (Clayton and Kaldor, 1987). The empirical Bayes approach leads to parameter estimates which are over-precise since the uncertainty associated with λ is not incorporated into the estimation of **φ** and μ. The advent of the implementation of Markov Chain Monte Carlo (MCMC) simulations in readily accessible software has made fully Bayesian approaches to estimating the parameters of the above distribution feasible. As Wakefield et al (2000) point out “All MCMC algorithms generate a sequence of dependent values which will eventually resemble a sample from the required posterior distribution”. However, high correlation between model parameters often lead to problems of non-convergence of MCMC simulations.

Most recent spatial statistical disease mapping developments (see Elliott *et al.* 2000 for examples) have been carried out in the context of fully Bayesian spatial models using MCMC simulation. One advantage of this method is that it produces probability distributions of all parameters, including area effects, so that “confidence interval” maps can easily be produced. Unfortunately this methodology has not been widely adapted to spatial data representing points rather than areas, which currently limits its application to the vast majority of spatial data in the field of vector borne diseases, where reporting systems for areal units are the rare exception. It remains an important future research priority to remedy this shortcoming.

No direct comparisons of the two approaches, GLMMs on the one hand, Bayesian autoregressive models on the other, were attempted by analysing the same data by the two different methods. However, chapters 4 and 6 both analyse malaria incidence in relation to climatic factors in KwaZulu Natal using the GLMM approach in the first instance, and the Bayesian autoregressive models approach in the second instance. The two studies had many differences: the first study dealt with a single year only, used long term averages of climate data for small areas and used population counts obtained from a census carried out by the malaria control programme; the second study dealt with several years of incidence, used remote sensed earth observation climatic proxies obtained from satellites for the particular time periods and small areas of interest, and used aggregates of population data obtained from the general population census of 1996. Despite these various differences in the modelling approach and in the two data sets for the same area, the two studies produced essentially similar associations between malaria incidence and climatic factors, even though model fit criteria led us to reject the model containing climatic co-variates in the MCMC analysis. A general conclusion from all the studies in this thesis is that both of the methods employed are feasible in a tropical disease setting, and can make a considerable contribution to more evidence based disease control.

In the analysis of spatial data it is often claimed that the point estimates of regression coefficients in a spatial model are the same as those that are obtained from the non-spatial model, and that it is merely the standard errors of the regression coefficients that are different (for example, Cressie 1993, p 14). This is only true in particular circumstances, and it should not be generalised to all spatial configurations. This can be illustrated by the graph in figure 8.1 which represents a scatterplot of a response, y, against a covariate x. Non-spatial regression would assume all the observations to be independent, and result in the regression line *a*. If however, the observations denoted by *c* are obtained from a spatial cluster and hence are more strongly correlated with each other than other pairs of observations, the regression line will be less steep in the spatial model (line *b*). This is because the degrees of freedom of the group of observations c is exaggerated in the non-spatial model, whereas the spatial model would accord them much less influence. This situation, where the degree of spatial correlation is uneven across the map, and is associated with the co-variate, was common in the data sets analysed in this thesis, and hence the regression coefficients of spatial and non-spatial models were rarely the same, in contrast to the widely held belief that they should not differ.

Fig. 8.1. Regression line for spatial and non-spatial model



The ambiguous findings of chapter 6 deserve further investigation of the statistical issues. On the one hand the coefficients of the co-variate effects were significantly different from zero and therefore suggest evidence of an association with malaria incidence. On the other hand the absence of any real difference in model fit criteria between models with and models without co-variates undermines the apparent evidence of an association. This conflict between the message we get from regression coefficients, and the one we get from model fit considerations should be investigated more closely to see how it can be resolved.

A more general point is that studies of the kind carried out in this thesis are highly interdisciplinary in nature. Quite detailed expertise is required from diverse fields such as malaria epidemiology, spatial statistics, geographic information systems, entomology, remote sensing and climatology. Such studies therefore should be collaborative in nature. Writing up such multi-disciplinary work poses substantial challenges, particularly for the coordinating author, since any write-up must be comprehensible to a general epidemiological audience without loss of rigour in the reporting of detail in any of the areas which made up the study. Often it is difficult to ensure that co-authors let alone their readership, understand the contributions of their colleagues from other fields of expertise. There is really no substitute for the collaborating researchers to familiarise themselves thoroughly with all the fields of expertise that come together in such studies. They may prefer to leave some of the decisions and execution of some of the work to those with experience and expertise in an area, but they should nevertheless be well acquainted with all the issues that are involved.

The potential contribution of this work to malaria control

The ultimate purpose of modelling malaria distribution must be to improve control of malaria as a disease. This has already been indicated in the discussion in various chapters. The following is therefore an attempt to draw conclusions for malaria control from the combination of studies.

Evidence based maps, derived from observed parasite ratios must be a more reliable indication of malaria risk than those based on expert opinion. These studies have produced a number of such maps, one of which (chapter 4) is already being used by the Roll Back Malaria programme in conjunction with country control managers. Often the absolute level of prevalence is not the decisive factor, but the broad category of prevalence, which can guide control efforts. Low endemicity areas will require an approach that assumes little or no immunity in the host population. In such circumstances an epidemic will affect all age-groups leading to large scale morbidity and mortality, often through complications such as cerebral malaria. Control efforts therefore need to be directed towards early detection and preparation for epidemic outbreaks. On the other hand high prevalence areas should be prepared to cope with severe disease presenting as severe anaemia in young children and in illness in pregnant women. Although the majority of the population may possess a certain measure of clinical tolerance, preventative measures need to concentrate on insecticide treated materials (ITMs) to offer protection to infants in particular, and to ensure the availability of effective drugs and rapid diagnostic facilities as part of the management of severe disease. More research effort needs to be directed to classify areas as endemic stable versus epidemic unstable on the basis of parasite prevalence values.

Although the four studies based on the Ubombo and Ngwavuma districts in KwaZulu Natal focus on a population of only one quarter of a million, this area represents populations living in areas of unstable malaria generally. The first of these studies showed that climatic differences still play an important part in malaria distribution in an area, even if intensive malaria control measures have reduced overall levels to well below their historical levels (Sharp *et al.* 2000). Of particular importance to malaria control is the fact that increased winter rainfall can create conditions of all year transmission and hence high annual caseloads. Proximity to water bodies also constitutes a risk factor particularly if transmission factors are otherwise sub-optimal, unless measures are taken to intensify malaria control in these areas. This has obvious consequences for the location of irrigation and water storage schemes, and may give rise to difficult choices in places where access to water is a critical development target.

The results of chapter 5 show that the increase in malaria incidence over recent years, for whatever reason, has been associated with a geographical spread of malarious areas. This shows that areas to the south of the high incidence area, where malaria had been all but eradicated for a long period, are vulnerable to resurgent malaria if neighbouring areas to the north experience a breakdown of control systems, for example, through failing drugs, or failing insecticides, or climate change, or a combination of these factors. It is also important for malaria control programs to sustain vector control through house spraying throughout the area, instead of focusing only on the historically high-incidence sub-regions. Such selective application of house-spraying may have given rise to the faster rise in incidence in previously low incidence areas.

The results of chapter 7 give an indication of the reasons for the leveling of malaria incidence in areas that experience the highest incidence rates, showing that populations in these areas are developing a measure of clinical tolerance to infection with *Plasmodium falciparum*. This finding supports the notion that exposure to *P.falciparum* early in life, even if only by a few infections, may yield some clinical protection in adults (Gupta *et al*. 1999). In Kwa Zulu Natal it also means that disease incidence can no longer be regarded as a reliable proxy for transmission intensity.

Although the conclusions of chapter 6 are somewhat tentative, they hold potentially the greatest benefit for malaria control in areas of unstable malaria through early warning forecasting. Remote sensing has previously been applied to mapping malaria vector distributions in Africa (Hay *et al*. 2000; Hay *et al.* 1998). This study, however, is a first attempt at trying to demonstrate a relationship between malaria incidence in a particular year and a particular locality to remote sensed climatic variables during and preceding that year. We were unable to confirm a relationship between malaria incidence of an area and a number of climatic factors, but as we argue in the discussion to chapter 6, the potential of remote sensing technology for timely information on the location of areas at high risk of epidemics warrants further exploration. Being ill-prepared for the diagnosis and treatment of large numbers of cases has been identified as a serious problem in these districts due to the severe fluctuation in case loads. Advance warning of periods of unusually high peaks in incidence therefore affords the opportunity for being better prepared for them. The groundwork done in this study has in part been the basis of a grant application to investigate the relationship between malaria and climate in this area prospectively. In particular, this project would collect malaria prevalence data prospectively from spatially dispersed surveys at numerous instances in time over a number of years, to provide the basis for a more reliable data set on which a space-time model could be based. Such a data set would be independent of small area population data derived from the census, which has been a weakness in the current data set, and it would include important covariate data such as intensity of insecticide spraying and entomology.

What the analysis in chapter 6 does show quite convincingly is the strong correlation between incidence in an area in a particular year, with incidence in the same area in the previous year. In other words, change in incidence in an area usually does not happen quickly from one season to the next, and last year's incidence is perhaps the best predictor of incidence for the current year, even if there is an overall time trend in incidence. However, in some years and in some places there is a more dramatic change, and the task of an early warning system would be to predict these instances with reasonable probability.

This thesis therefore combines a number of studies that have shown the applicability of methods that have previously been largely employed in developed country disease settings. As argued above, the results of these studies have refined knowledge of the distribution of malaria, of the factors that affect the distribution of malaria, and of the factors that have promoted or inhibited changes in malaria distribution in one area. The development of this work has laid the foundation for future more focused studies that aim to usefully exploit the relationship between malaria and climate.

Appendix 1

Iterative procedure for specifying spatial dependence in the generalised linear mixed model

(This appendix forms part of the paper which constitutes chapter 3)

The SAS procedure MIXED is an implementation of mixed model methodology for Gaussian response variables. For non-Gaussian responses, such as incidence rates, the macro GLIMMIX implements the *Generalised* Linear Mixed Model. GLIMMIX produces estimates via PROC MIXED and hence provides similar functionality. We used GLIMMIX to adjust for spatial dependence in the regression analysis

1. *The standard Generalised Linear Model*(GLM) (McCullagh and Nelder, 1989).
With the classical GLM, a vector of observations **y** is assumed to have uncorrelated elements. In our particular application the model assumes the y_i are Poisson distributed and if mean(**y**)=$\boldsymbol{\mu}$, then log($\boldsymbol{\mu}$)=$\mathbf{X}\boldsymbol{\beta}$, a linear function of the explanatory variables **X**, and var(**y**)= **V**, a diagonal matrix, with unequal elements since the variance of a Poisson variate depends on the mean.

2. *The Linear Mixed Model(LMM), allowing for a correlated error structure*
In the LMM, as implemented in procedure MIXED, it is possible for observations, **y**, which are assumed to be normally distributed, to have a spatial correlation structure. In particular, if d_{ij} denotes the distance between the points i and j where observation y_i and y_j were made,

$\mathbf{V}= \mathbf{I}\sigma_1^2 + \mathbf{F}\sigma^2$, where $F_{ij}=\exp(-d_{ij}/\rho)$

The unknown parameters in this model, namely σ_1^2, the nuggett, σ^2 , the sill, and ρ, the range, can be obtained from a variogram of the data, as described below. They are jointly referred to as $\boldsymbol{\theta} = \begin{bmatrix} \sigma_1^2 \\ \sigma^2 \\ \rho \end{bmatrix}$. This particular spatial model is known as the exponential model, and effectively postulates that the correlation increases as points occur closer together, whereas points at distances greater than the range from one another are uncorrelated. F_{ij} can be defined using other spatial models (Littell *et al.*, 1996 pp 305). In the PROC MIXED implementation only the variogram parameters $\boldsymbol{\theta}$, are specified for SAS to be able to calculate **V**. In the case of non-normally distributed data, $\boldsymbol{\theta}$ is specified in the macro call to enable GLIMMIX to estimate the appropriate correlation matrix for a spatially correlated Poisson model. However, since the Poisson model does not satisfy the requirement of constant variance, the estimation of θ is carried out on standardised residuals, as explained below.

3. *The variogram approach to the estimation of* $\boldsymbol{\theta}$

In classical kriging (Cressie, 1993), the following approach is adopted: If one can assume the mean value of *y* to not be changing from point to point (the characteristic of stationarity), a variogram is constructed. The pairs of observations are arranged so that all pairs a given distance h ± h/2 apart are pooled into one class, and the semi-variance $\gamma(h) = \frac{1}{2}\text{average}(y_i\text{-}y_j)^2$ calculated. The variogram is a plot of $\gamma(h)$ against distance for distances h, 2h, 3h, 4h … etc. (Littell *et al.* 1996 pp 307). Estimates of $\boldsymbol{\theta}$ are obtained graphically. We used the package GS-Windows (GS+) for this purpose.

In our application the variogram was not constructed from **y** because of the non-stationarity and the Poisson distribution (implying non-constant variance). Instead, we used the signed deviance residuals calculated from

$r = \text{sign}(y\text{-}\mu)\{2(y\log(y/\mu)\text{-}y\text{+}\mu)\}^{½}$

where y and μ are the observed and fitted values of the response variable respectively (McCullagh and Nelder, 1989 pp 37-40).

4. *Fitting the GLMM and iteration between the variogram and GLMM*

The mixed model estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (Littell *et al.* 1996 pp 229-251) can be obtained jointly in SAS via the procedure MIXED. However, in the context of the GLIMMIX implementation (i.e. non Gaussian models), interpretation of the output relating to the spatial parameters $\boldsymbol{\theta}$ is not straightforward, and convergence problems can arise. Our approach has been to fix $\boldsymbol{\theta}$, use GLIMMIX to estimate $\boldsymbol{\beta}$ and from the residuals, **r**, then use a variogram to revise the estimate of $\boldsymbol{\theta}$. In other words, accommodation of correlation between elements of **y** due to spatial effects is achieved by converting from the observed **y** to standardised residuals **r,** which have the property of constant variance and mean. The following steps were carried out.

(i) An initial estimate $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}$ was made assuming no spatial correlation i.e. with $\mathbf{V}$ being a diagonal matrix. The deviance residuals $\mathbf{r}_0$ were calculated using $\hat{\boldsymbol{\beta}}_0$ and a variogram constructed using GS-Windows. Initial estimates $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}$ were made as described above.

(ii) GLIMMIX was now used with $\boldsymbol{\theta}$ being fixed at the values $\hat{\boldsymbol{\theta}}_0$.

(iii) From GLIMMIX a new set of estimates $\hat{\boldsymbol{\beta}}_1$ are found and hence a new set of deviance residuals $\mathbf{r}_1$ are calculated. These are used in a new cycle to re-draw the variogram and hence derive fresh estimates $\hat{\boldsymbol{\theta}}_1$.

Steps (ii) and (iii) form an iterative cycle which continues until there is no further change in the estimates. Assuming that any spatial correlation is positive rather than negative, standard errors of $\hat{\boldsymbol{\beta}}_0$ may have been under-estimated rather than over-estimated. Adjusting for spatial correlation may therefore lead to removal of some variables from the model whose contribution was initially overstated. The criterion for dropping a variable from the model was $p>0.05$.

6. *Producing predicted values by kriging*

Figure 3.2 shows that the residuals of the final model remained spatially correlated. Map predictions could therefore be improved through kriging of the residuals. This was done by calculating the log(mean y) using $\hat{\boldsymbol{\beta}}$ and then adding to this (on the log scale), the kriged predictor for the deviance residual, r, at that point (Chapter 2). These kriged predictions are then transformed back to predicted incidence rates (Figure 3.3).

Appendix 2

ESHAW Technical Report

Immo Kleinschmidt[8], Judy Omumbo[9], Olivier Briët[10], Nick van de Giesen[11], Nafomon Sogoba[12], Nathan Kumasenu Mensah[13], Pieter Windmeijer[14], Mahaman Moussa[3], Nicole Teuwsen, Thomas Teuscher[3]

This appendix gives additional detail of the analysis and results of the MARA data for West Africa described in chapter 4. There is some overlap between this appendix and chapter 4 so that each can be read independently of the other.

This appendix is a report on the Ecosystem Health Analytical Workshop (ESHAW) held in Bouake, Cote d'Ivoire, November 1999. It is a description of the data used, the analytical methods employed and the results obtained in a project on mapping malaria risk in West Africa.

A malaria distribution model for West Africa

Introduction

The recently created continental database of malaria survey results(MARA/AMRA,1998) provides the opportunity for producing empirical models and maps of malaria distribution at a regional and eventually at a continental level. This appendix describes the methods used and the results obtained in the analysis, modelling and mapping of malaria distribution for West Africa. West Africa was chosen since a reasonably large set of survey results are available for this region on the MARA database, since it represents the region with the largest population exposed to high levels of malaria transmission intensity, and because at least two

[8] South African Medical Research Council, P.O. Box 17120 Congella, Durban 4013, South Africa.
[9] Kenya Medical Research Institute/Wellcome Trust Collaborative Programme, PO Box 43640, Nairobi, Kenya
[10] West Africa Rice Development Association (WARDA),01 BP 2551 Bouake 01, Cote D'Ivoire
[11] Center for Development Research, Bonn University, Walter-Flex-Str 3, 53113 Bonn, Germany.
[12] Malaria Research and Training Center DEAP/FMPOS, Universite du Mali, Bamako, Mali.
[13] Navrongo Health Research Centre, P.O.Box 114, UE/R Ghana.

country maps based on survey data are already available for West Africa (Thompson *et al*. 1999; chapter 2).

In the analysis and modelling of the geographical distribution of malaria the first question that arises is what measure of malaria intensity is to be investigated. If the distribution of transmission intensity is the main objective of such a study, then the ideal malariometric measure to be modelled would be the entomological innoculation rate (EIR), which measures the number of sporozoite positive bites per person per time unit (Snow *et al*. 1996). In practice this is not widely available. The most commonly available concept of intensity of malaria is that of parasite prevalence in a random sample of individuals, usually obtained by means of a local survey of the population. This gives rise to a geographically located binomial point response. The survey should be restricted to childhood populations of less than 10 years of age, in order to avoid the effects of population immunity in endemic areas moderating the survey results. However, using parasite prevalence as a proxy for transmission intensity has some distinct disadvantages. Most important amongst these is the variation of prevalence with age, and the fact that prevalence varies with dry and wet seasons in all age groups apart from the age group of peak prevalence (Molineaux and Gramiccia,1980).

The distribution of malaria is governed by a large number of factors relating to the parasite, the vector and the host. Many of these factors affect the interactions between parasite, vector and host in some way. For example, temperature effects the duration of sporogony and longevity of mosquitoes and hence determines the proportion of vectors surviving long enough to become infective. Similarly, vector populations depend on habitat and breeding sites which are largely determined by precipitation, humidity and presence of water, but also on man to vector contact. For efficient transmission of malaria the density of vectors in relation to man is also important, and hence human population density is a factor in transmission intensity (Molineaux, 1988). Due to the complexity of these relationships, it is unlikely that these factors are simply linearly related to measures of malaria intensity such as parasite prevalence.

[14] Alterra, PO Box 47, 6700 AA Wageningen, The Netherlands.

The strategy used in this study was to undertake a spatial statistical analysis of malaria parasite prevalence in relation to those potential factors involved in the intensity of malaria distribution which are readily available at any map location. The resulting regression model can then be used to predict parasite prevalence for the whole of West Africa. A geostatistical process is subsequently used to improve prediction in places where there is considerable divergence between model predictions and observations in a local neighbourhood (see chapter 2).

Data sources and data preparation

The MARA / ARMA collaboration database consists of over 8,000 reports on malaria endemicity in sub-Saharan Africa. Each report has been geographically referenced by means of longitude and latitude. The subset of this database that was used in this study contains discrete survey locations relating to community based surveys between latitudes 1° and 22° North and longitudes 17° West to 16° East, in which at least 50 children were examined for the presence of *Plasmodium falciparum* in blood smears. This represents over a quarter of a million children surveyed for malaria parasites. Surveys were screened to include only populations between 1 and 10 years of age, although in a few cases where no further age breakdowns were available, surveys on populations between 1 and 15 years were included. Surveys conducted during known epidemics were excluded, as were those that may represent biased samples, such as surveys that were restricted to school attendees only. The survey dates covered several decades from about 1970 onwards, and surveys conducted more than once at the same location were combined (summing numerators and denominators). An implicit assumption therefore is that malaria endemicity has remained relatively stable over this period, so that the surveys taken at different time points could be conceptually regarded as a cross-section of surveys, taken at many locations. A total of 450 data points resulted from this process. The locations of these points are shown in figure 4.1.

Environmental data were mostly derived from the Hutchinson Climate Dataset for Africa (Hutchinson *et al.* 1995). These comprised normalised difference vegetation index (NDVI), average maximum and minimum temperature and precipitation per calendar month, all averaged over a number of years. By means of Geographic Information Systems (GIS) values of each of these variables were assigned to each survey location.

Four agro-ecological zones were distinguished on the basis of the length of the growing period, i.e. the period that water is available for vegetative production on well drained soils. It is a function of precipitation, evaporation, and a fixed amount of available water in the soil (FAO, 1978). The zones are from south to north: the Equatorial Forest zone (> 270 days) Guinea Savanna zone (165 – 270 days), Sudan Savanna zone (90 –165 days) and the Sahel zone (< 90 days), shown in figure 4.1.

Population density data were derived for each survey location (Deichman, 1996).

A drainage density map was calculated based on agro-ecological zone and geology. This variable is a measure of the amount of surface water that is available in an area, and it is expressed as the total length of streams in an area per unit of area (Windmeijer and Andriesse, 1993). Drainage density values for each point were divided into categories of very low (<0.3 k/km^2), low (0.3 to <0.6 km/km^2), medium (0.6 to <1.2 km/km^2) and high (1.2 to <2.4 km/km^2).

Analysis, modelling and mapping: methods

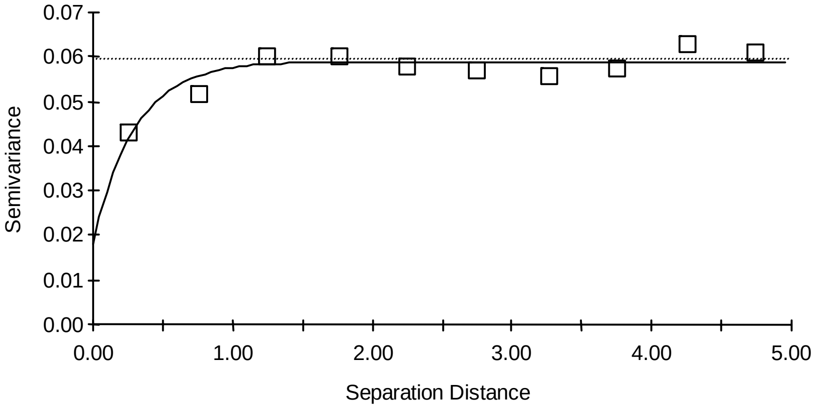
Parasite prevalence values varied from 0 to 100%, mean(SD) = 46%(24%). However, of the total number of individuals surveyed, 48.8% tested positive. Surveys varied in size from 50 to 10463 persons, mean(SD) = 539 (1197). Table A.2.1 shows how parasite prevalence varied between the 4 agro-ecological zones described previously.

Table A.2.1. Parasite prevalence by agro-ecological zone.

	Mean Prevalence(SD)	**Number of surveys**
Sahel	34(23)	34
Sudan Savanna	46.5(23)	128
Guinea Savanna	47(24)	159
Forest	50(26)	129

The variogram (Krige, 1966; Carrot and Valleron, 1992) in figure A.2.1 shows that the spatial dependence of the survey results extends over a distance of about 1.6 degrees.

Figure A.2.1. Variogram of parasite prevalence (all zones)



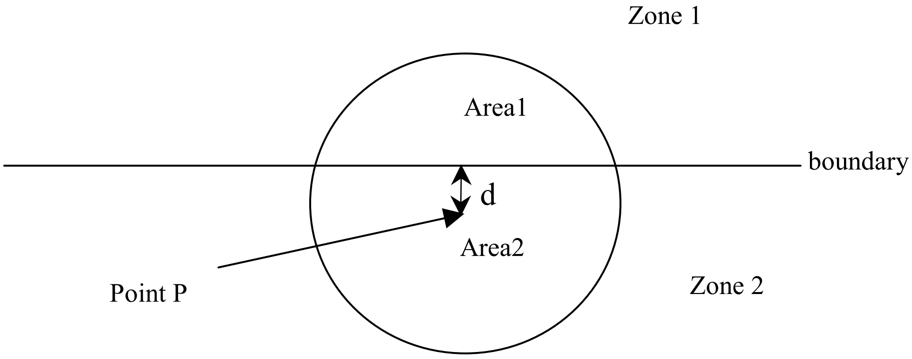
model1.txt, perev, lag=0.5(5.0)

Since climatic variables by calendar month are highly correlated with those of adjacent months, all rainfall, temperature and NDVI data were aggregated into quarterly (3-month) averages, from December onwards (to coincide with the drier and wetter seasons respectively).

The data were divided into 3 groups corresponding to the agro-ecological zones for West Africa, with Sahel and Sudan Savanna combined into one group. A statistical model was derived for each of these zone specific groups. This was done on the basis of the assumption that the factors affecting malaria risk would be different in the four agro-ecological zones (AEZ). (Sahel was combined with Sudan Savanna since there were only 34 observations in the former, and because these two zones are adjacent and have previously been analysed as one in a similar analysis for Mali.) An analysis by AEZ would therefore produce predictor variables that are appropriate for a particular zone, which would in turn produce models that fit the data better in terms of explained deviance.

The disadvantage of separate models for each zone is the somewhat arbitrary nature of the division between zones, with many points in the vicinity of a boundary between two zones having characteristics of both zones. If the prediction for such a point is based only on the model for the zone it is situated in, whilst a nearby (and therefore similar) point on the other side of the boundary is predicted entirely using the model for the zone it is situated in, this could lead to predictions that are very different from one another for two points that should be very similar. The result would be a sharp discontinuity along the zone boundary, which would be counterintuitive and likely not to reflect reality. (Fitting zone as a categorical variable to a combined model for all zones would create a similar problem). To avoid an artifact of this nature, map predictions in a buffer of 1.6 degrees (the extent of spatial dependence, from figure A.2.1) on each side of a zone boundary were based on a weighted average of the two predictions for the point, assuming it belongs to either the one or the other zone. The weights for the two predictions are dependent on the distance of the point from the boundary in the following way: For each point within 1.6 degrees of the boundary, a circle of radius 1.6 degrees is drawn, and the weights are proportional to the areas of the two parts of the circle lying on either side of the boundary (fig A.2.2).

Figure A.2.2. Prediction near zone boundaries



The prediction for point P (at the centre of the imagined circle) is therefore:

Z_p= (Area1*μ1 + Area2*μ2)/(Area1 + Area2) where $\mu 1$ and $\mu 2$ are the predictions at point P (i.e. using the covariates applicable to point P) obtained from the models for zone 1 and zone 2 respectively. The relative magnitude of the two areas, and hence of the weights, is a fuction of *d*, the distance of *P* from the boundary. The map prediction Z therefore provides a smooth transition between the two zones.

Initial variable selection for each model was done by performing a stepwise procedure using a generalised linear model (GLM) with logit link function (Hosmer and Lemshow, 1989; StataCorp, 1997) and with the parasite prevalence of a point being the response variable. The criterion for inclusion of a variable into the model was set to $p<0.01$.

In order to improve the fit (i.e. reduce residual deviance), each variable that survived the stepwise procedure was transformed into 7 different fractional polynomials. The transformation producing the biggest reduction in residual deviance was chosen if this reduction in deviance was significant by a χ^2 test on one degree of freedom. Transformations that were tried for each variable x were $1/x^2$, $1/x$, $1/x^{0.5}$, $\ln(x)$, $x^{0.5}$, x^2 and x^3. The transformations are useful to represent relationships in which parasite prevalence increases more rapidly than a straight line at low values of x and more slowly at high values, or vice versa. To facilitate the reporting and interpretation of transformed variables, we have calculated the model based odds ratio of parasite prevalence for the transformed variable for two separate values of the variable, relative to a third (referent) value of the variable. The referent value chosen was the mean value of the variable minus one standard deviation, with the other two values being the mean itself and the mean plus one standard deviation. The two odds ratios demonstrating the association between the transformed variable and parasite prevalence are therefore one and two standard deviations removed from the referent value, thereby giving some indication of the non-linear nature of the association (Royston *et al*. 1999).

Deviance residuals were calculated for each statistical model that was derived from the GLM. Semivariance of the deviance residuals of all pairs of observations was calculated and a variogram constructed to determine if there was evidence of residual spatial correlation i.e. if the semi-variance of pairs of residuals that are close together is markedly less than that of observations which are further apart. The parameters of the function that describes the relationship between semi-variance and separation distance is then used to specify the correlation structure of the data (R) in a generalised linear mixed model (GLMM), (Littell *et al.* 1996) thereby taking account of any residual non-independence in the data (see chapter 3 and appendix 1). Assuming that any spatial correlation is positive rather than negative, standard errors of the spatially naïve model may have been under-estimated rather than over-estimated. Allowing for spatial correlation may therefore lead to removal of some variables from the model due to the resultant inflation of the standard errors. Deviance residuals of the spatially adjusted model are calculated and a new variogram is constructed. If this variogram differs from the one that was used to specify the correlation structure of the data in the GLMM, then the model is fitted again using the improved spatial specification. This process is iterated until the variogram no longer changes indicating that a covariance structure corresponding to the model residuals is adequately specified (see details in chapter 3; appendix 1).

Once the models for the three areas had been derived, these were used to produce map predictions for the three zones based on the predictor variables which are available as map images. In the buffer regions of the boundaries between zones the interpolated predictions using the two adjoining models were calculated as described above. The resulting predicted map values at the survey points (observations) are then extracted and residuals calculated on the logit scale. Based on a variogram of these residuals, a kriged map of residuals is calculated, which is added to the predicted values on the logit scale before transforming the result back to proportions. The addition of kriged residuals will allow the map to deviate from the model and move closer to the observed values, if such deviation is supported by other observed values in the neighbourhood (details in chapter 2). This should improve the final map in the sense that it does not deviate too severely from the observations, which is particularly important if the model does not adequately explain the observed variation in transmission risk.

Results and model interpretation

Tables A.2.2, A.2.3 and A.2.4 show the odds ratios of the association between parasite prevalence and significant explanatory variables. In the interpretation of these results it is relevant to note that many of the climatic variables for this region broadly have a north to south gradient ranging from the arid Sahel zone in the north to the humid tropical forest zone in the south. Rainfall, NDVI and maximum temperature show a consistent gradient between the Sahel zone and the forest zone for all calendar months, the gradient being negative for maximum temperature throughout the year, and positive for rainfall and NDVI throughout the year. Minimum temperature on the other hand, has a positive gradient from the Sahel to the Forest zone for most of the dry months of the year, and a negative gradient for most of the wet months. This may explain at least partially its effect reversal for different months of the year.

Table A.2.5 shows the proportion of explained deviance that is achieved by each of the models. This shows that for the Sahel and Sudan Savanna and for the Guinea Savanna zone approximately 50% of variation is explained by the factors in the model, whereas only 17% of variation (as expressed by deviance) of the Forest zone model is explained by available predictor variables. The over-dispersion of parasite prevalence values has been taken into account in the analysis by inflating standard errors by the square root of the dispersion factor.

Inspection of variograms of deviance residuals of the models for the 3 zones shows that only the residuals of the model for the Guinea Savanna zone display a distinct spatial pattern (figure A.2.3), whereas the variograms for the other two zones (not shown) do not show a very distinct spatial pattern.

Predicted malaria risk was calculated for each pixel for the map of West Africa, based on the three separate models with interpolation between models applied in a buffer 1.6 degrees on each side of a boundary between zones. Figure A.2.4 shows that the residuals from these predictions show some spatial dependence. Kriging of these residuals could therefore be performed. Fig 4.2 is the result of adding kriged residuals to the map of model predictions. The map categories chosen are identical to those used in a previous study for producing a malaria distribution map for Mali (chapter 2).

Sahel and Sudan Savanna zones

Table A.2.2 shows the results that are obtained from the multiple regression model for this zone i.e. each effect is adjusted for all the others. This shows that for all places that have equal values for all the other variables in the model, any increase in rainfall in the second half of the dry season, is associated with a reduction in parasite prevalence. Vegetation index in the same quarter shows the same negative association. Whilst these two factors are somewhat counterintuitive in the direction of the association, this could be due to torrential floods flushing larvae out of pools that are used for breeding by vectors. Another possible explanation is overabundance of vegetation, particularly forest-like vegetation, which shuts out sunlight at ground level and around small water bodies which may have a negative impact on *A. gambiae* breeding potential (Ravoniharimelina *et al.* 1992; Imevbore, 1991). Higher average minimum temperatures in the wet season are associated with higher parasite prevalences, possibly due to the shortening of the duration of sporogony. Average maximum temperature in the second half of the dry season, other factors being equal, leads to a shortening of the duration of sporogony, thus facilitating increased transmission although very high temperatures would lead to smaller proportions of adult vectors surviving the maturation of the parasite. Low, as opposed to very low drainage density, provides more extensive and more stable breeding sites for vectors, whereas medium drainage density may reflect faster flowing streams that are less suitable as breeding sites, again other factors being the same. Average minimum temperature in the first half of the dry season is higher in the more southern, humid parts of this zone thus favouring malaria transmission. Average maximum temperatures in the second half of the wet season are much higher in the more arid north of the zone, thus making the negative association with parasite prevalence highly plausible.

Guinea Savanna zone

Table A.2.3 shows the results that are obtained from the multiple regression model for this zone. Increased rainfall in the second quarter of the wet season is likely to lead to increased vegetation density in the first quarter of the drier season. Both are positively associated with parasite prevalence, reflecting the superior vector breeding conditions that these two factors represent. Increased vegetation growth in the latter part of the drier season may represent those areas with very dense forest-like vegetation which inhibits vector proliferation on account of low levels of direct light, which may explain the negative association of this factor with transmission intensity. Areas with more humid climate generally experience higher minimum temperatures in the drier season on account of their greater cloud cover. A positive association of minimum temperatures in the early part of the drier season with parasite prevalence is therefore plausible. Average maximum temperature in the latter part of the more rainy season is higher in the drier parts of West Africa: the fact that this is nevertheless associated with parasite prevalence may be because other factors already represent the negative association with the more arid areas, so that this variable reflects an association between parasite prevalence and warmer places which are similar in terms of precipitation and humidity. As in the model for the Sahel and Sudan Savanna zone, parasite prevalence is significantly associated with areas that have a drainage density categorised as 'low' compared to those with a 'very low' drainage density reflecting the lack of vector breeding sites in areas of 'very low' drainage density. The two higher levels of drainage density, namely 'medium' and 'high' show no significant difference compared to areas of very low drainage density. This is probably due to very low numbers in some categories of drainage density with the point estimates indicating a raised risk compared to the 'very low' category.

Other factors in the model being equal, population densities below 1 per square km are associated with a raised risk of malaria parasite infection. This may be an indication of the importance of vector densities in relation to man, and it may also be a surrogate measure for low socio-economic development. There are no significant differences between categories of higher population densities which is in part due to small numbers of surveys in some of the categories. Nevertheless it is somewhat surprising to see such low population densities favouring malaria transmission.

There is a significant association between difference in maximum and minimum vegetation index and parasite prevalence. This variable has been added to model the 'pioneering' behaviour of the predominant vector in this region, *A. gambiae,* which is well equipped to take advantage of rapidly improving habitat and breeding sites. It is an indication that areas with all year round high vegetation density do not necessarily have the highest transmission intensities (Kuhlow and Zielke, 1976).

Table A.2.2. Factors associated with parasite prevalence in Sahel and Sudan Savanna zones. Odds ratios(OR) have been calculated from the model: log OR = - 0.081*rn0305 - 0.806*tmin0911 + 0.465*tmax0305 + 0.63*drdens2 - 0.62*drdens3 + 36615*vi0305^{-2} - 854*tmin1202^{-2} - 0.012*tmax0911^{2}

	Category	Reference point (or range)	Odds Ratio (model based)	
			Estimate	95% Confidence Interval
Average monthly rainfall March to May (rn0305), per mm			0.922	0.895 - 0.949
Average minimum temperature September to November (tmin0911), per °C			0.447	0.331 - 0.603
Average maximum temperature March to May (tmax0305), per °C			1.593	1.359 - 1.866
Drainage density km/km^2	Very low(referent)	<0.3	1.0	
(drdens2)	low	0.3 - <0.6	1.885	1.228 - 2.894
(drdens3)	medium	0.6 - <1.2	0.534	0.403 - 0.718
Transformed variables:				
Vegetation Index March to May (vi0305) , per unit on scale of 255	Low (referent)	100	1.0	
	Medium	110	0.53	0.40 - 0.70
	High	120	0.33	0.20 - 0.53
Average minimum temperature December to February (tmin1202) , °C	Low (referent)	14.3	1.0	
	Medium	15.6	1.95	1.42 – 2.67
	High	16.9	3.27	1.87 – 5.73
Average maximum temperature September to November (tmax0911), °C	Low (referent)	32.9	1.0	
	Medium	34.2	0.35	0.27 – 0.46
	High	35.5	0.12	0.07 – 0.20

Increases in average minimum temperature during the rainy season are associated with the more arid areas. The reduced risk of parasite infection associated with average minimum temperature from June to August is therefore plausible.

Table A.2.3. Factors associated with parasite prevalence in the Guinea Savanna zone. Odds ratios have been calculated from the model: log OR = 0.012*rn0911 + 0.031*vi1202 – 0.028*vi0305 + 0.467*tmin1202 + 0.584*tmax0911 + 2.665*drdens2 + 1.502*drdens3 + 1.898*drdens4 - 1.706*popdens2 - 1.521*popdens3 - 1.482*popdens4 - 1.532*popdens5 – 7202*$vidif^{-2}$ + 121*$tmin0608^{-0.5}$

	Category	Range or reference point	Odds Ratio (model based)	
			Estimate	95% Confidence Interval
Average monthly rainfall September to November (rn0911), per mm			1.012	1.005 - 1.020
Vegetation Index December to February (vi1202) , per unit on scale of 255			1.032	1.013 - 1.052
Vegetation Index March to May (vi0305) , per unit scale of 255			0.973	0.959 - 0.987
Average minimum temperature December to February (tmin1202), per °C			1.595	1.298 - 1.960
Average maximum temperature September to November (tmax0911), per °C			1.793	1.482 - 2.170
Drainage density km/km^2	Very low(referent)	<0.3	1	
(drdens2)	low	0.3 - <0.6	14.368	1.831 – 112.726
(drdens3)	medium	0.6 - <1.2	4.491	0.570 – 35.370
(drdens4)	high	1.2 - <2.4	6.673	0.741 – 60.049
Population density persons/km^2	Very low(referent)	<1	1	
(popdens2)		1 - <10	0.18	0.07 - 0.47
(popdens3)		10 - <50	0.22	0.11 - 0.44
(popdens4)		50 - <100	0.23	0.12 - 0.43
(popdens5)		>100	0.22	0.14 - 0.34
Transformed variables:				
Difference in maximum and minimum Vegetation Index (vidif) , scale of 255	Low (referent)	67	1	
	Medium	70	1.144	1.120 - 1.168
	High	83	1.749	1.603 - 1.909
Average minimum temperature June to August (tmin0608) , °C	Low (referent)	19.5	1	
	Medium	21	0.368	0.201 - 0.675
	High	22.5	0.150	0.047 - 0.475

Model for Forest zone

Table A.2.4 shows that only three factors were significantly associated with parasite prevalence in the tropical forest zone. Average maximum temperature between September to November is positively associated with malaria risk, presumably reflecting the shorter duration of sporogony having a favourable impact on transmission risk in a humid area where maximum temperatures do not reach a level at which adult vector survival becomes a limiting factor. This appears to be in contrast to the significant negative association between parasite prevalence and maximum temperature during June to August, for places of similar maximum temperature between September to November, and similar average rainfall between September to November. A possible explanation for this may be the fact that higher maximum temperatures are more prevalent in the relatively more arid areas of the region, particularly during the wetter part of the year.

The positive association with average rainfall between September and November is obviously due to the improved breeding sites that are provided by higher levels of precipitation, provided these are not of the type that flush out larvae from pools.

Table A.2.4. Factors associated with parasite prevalence in the Forest zone. Odds ratios have been calculated from the model: log OR = 1.092*tmax0911 - 0.0181*tmax0608^2 + 0.0000095*rn0911^2

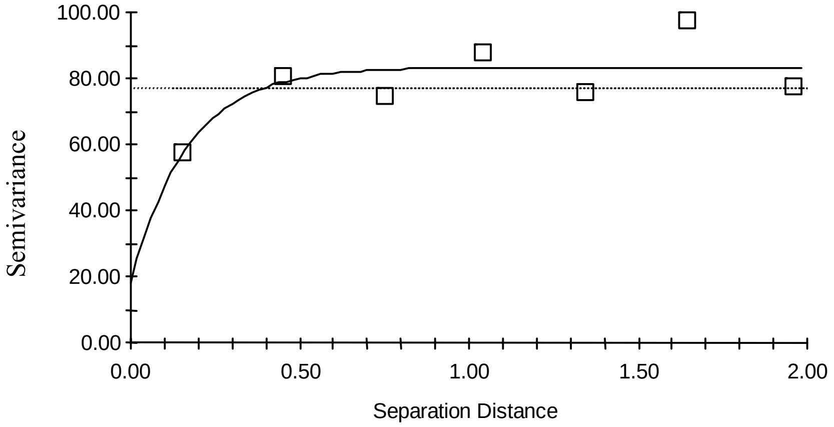
	Category	Range or reference point	Odds Ratio (model based)	
			Estimate	95% Confidence Interval
Average maximum temperature September to November (tmax0911), per °C			2.98	1.61 – 5.51
Transformed variables:				
Average maximum temperature June to August (tmax0608), °C	Low (referent)	26.5	1	
	Medium	28	0.227	0.0813 - 0.634
	High	29.5	0.0475	0.0056 - 0.392
Average monthly rainfall September to November (rn0911), per mm	Low (referent)	150	1	
	Medium	215	1.25	1.23 - 1.273
	High	280	1.70	1.63 - 1.77

Table A.2.5. Residual deviance

Zone	Total deviance*	Residual deviance	Percent "explained" by model
Sahel & Sudan Savanna	14096	7187	49
Guinea Savanna	27510	13516	51
Forest	10238	8512	17
Total			44%

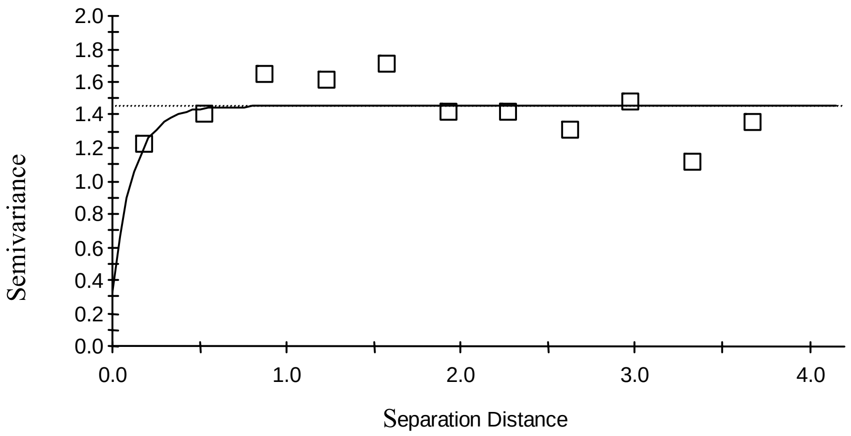
* = deviance of null model

Fig. A.2.3 Variogram of residuals of model for Guinea Savanna zone



aez3m1.txt, lag=0.3

Fig A.2.4. Variogram of residuals of model predictions



lgtres.txt, lgtres, lag=0.35(4.0)

Discussion

Note: This discussion deals only with technical aspects of this analysis. The implications of the map for malaria control in West Africa, are addressed in the discussion of chapter 4.

Our models for the three separate zones show that the relationship between malaria prevalence and climatic, environmental and population factors is complex. The difference in significant variables in the 3 models indicates that different factors are responsible for variation in malaria risk in the three zones.

In interpreting the plausibility of the models, it needs to be remembered that there is considerable confounding between explanatory variables, which are often highly correlated with each other. In some cases the other factors in the model can control for this confounding, in others there may be substantial residual confounding, or even no adjustment for confounding. This results in the interpretation being to some extent speculative. For example, sparsely populated areas are more common in very arid regions and hence would be associated with low transmission intensity since they are confounded with water-related factors, such as precipitation. However, once adjustment is made for precipitation and NDVI, it is quite plausible that for places that are equally wet (or dry), those with lower population densities have a higher transmission risk on account of the presence of suitable vector habitat etc.

A second interpretational problem is the fact that these models are essentially data driven, rather than hypothesis driven. This can lead to chance associations which have no biological or climatic explanation. We have sought to reduce this possibility by making it ‘harder’ for variables to enter the multiple regression model, firstly by setting the criterion for significance to p=0.01 instead of the customary 0.05, secondly by inflating standard errors by a factor equal to the square root of the over-dispersion, and thirdly by adjusting for spatial correlation in the data. We are therefore reasonably confident that the associations which we are reporting are strong associations, indicating some mechanism that is responsible for variation in transmission intensity, even if the interpretation is not always obvious.

Our analysis leaves a considerable amount of variation in malaria risk unaccounted for. A significant proportion of this variation is likely to be noise due to errors and non-uniform sampling methods employed in the surveys that yielded the observed parasite prevalences. This is to be expected in a large heterogeneous historical data set with no uniform standards in data collection having been applied. No amount of model fitting is going to overcome this shortcoming in the data.

Nevertheless these data represent a very large albeit imperfectly sampled population of children in West Africa, and it is highly likely that there are other unmeasured perhaps more local factors that determine variation in parasite prevalence.

The distribution of distances between pairs of surveys in our data is such that the lowest lag distances for which variograms can be constructed is about 0.25 kilometres without the number of pairs becoming too few to make a reasonable estimate of semi-variance. This leaves open the possibility of spatial dependence at a shorter range. Such spatial dependence may be due to small area socio-demographic or environmental factors but the spatial distribution of our survey data makes it difficult if not impossible to investigate these. Our analysis of residuals shows that except for the Guinea Savanna region unobserved factors determining malaria distribution are not spatially correlated unless the spatial correlation is at the micro scale referred to above. In the Forest zone most of the variation (83%) is not accounted for by the climatic factors that we have investigated. This may indicate widespread uniform climatic suitability for malaria transmission in this zone (Craig *et al*. 1999) with variation driven by other factors.

There are a number of additional limitations to this study that may have affected the results.

1. The combination of age-groups of survey subjects: Under conditions of the same transmission intensity, parasite prevalence will vary by age group in endemic areas. Since all age groups between 1 and 10 (in some cases 1 and 15) were combined this could have resulted in additional variation in prevalence if the mix of age groups was not constant. This problem could only have been avoided by rejecting all surveys that did not provide a more detailed breakdown by age. An additional problem related to the age of survey subjects is due to the fact that in areas of seasonal variation in transmission intensity, parasite prevalence will vary between dry and wet season for all ages except the age of peak prevalence (Molineaux and Gramiccia, 1980). This could only have been overcome by restricting prevalence values to those relating to the age of peak prevalence. Since this age varies depending on transmission intensity, such an approach would have required modelling first age specific prevalence and producing estimates of age specific prevalence so that at each map location a curve of prevalence versus age could be estimated. A parameter that expresses the shape of this curve and hence the transmission intensity could then be estimated at each location (This idea is due to Tom Smith, STI). Such a study would have entailed a modelling exercise of considerably more complexity than what has been attempted here.
2. Bias of sample surveys away from low prevalence values. There are generally fewer surveys in areas of low or zero transmission intensity and in urban areas than in areas of high transmission intensity, which causes a bias in the data towards higher prevalences. There are no immediate solutions to this problem, but it is an issue for the whole of the MARA database, and it is worth looking at ways of adequately compensating for this bias. This difficulty did not get directly addressed in previous country models using the MARA data but it was minimised by masking out areas where climatic suitability rules out malaria transmission.
3. Uneven sampling densities. As figure 4.1 shows sampling densities varied considerably between countries, and for the country with the highest population in

the region, namely Nigeria, only a small number of surveys were available. This has made map predictions less reliable in under-sampled sub-regions.

The very weak indication of a spatial structure of residuals in 2 of the 3 zones implies that the shortcomings of our model have no particular spatial distribution at distances above 25km. This makes it more difficult to identify areas and factors which contribute to the poor fit. By adding kriged residuals to the values predicted by the models (including the adjustments for boundary discontinuities), we have been able to adjust the map to take more account of the empirical observations and hence deviate from the model predictions in places were such deviation is supported by the data.

Despite the quoted limitations of this study, it does to our knowledge represent the first attempt to produce a malaria risk map of the West African region, based entirely on malariometric data. The map agrees broadly with expert opinion maps, (Wernsdorfer and McGregor, 1988) and represents a refinement of these. The complexity of the models does reflect the complexity of the factors that are responsible for malaria distribution. All the factors that are present in the three models have a plausible explanation, albeit a speculative one in some instances. In our view this first attempt at producing a regional model and map of malaria transmission can be built upon and improved in future through, amongst others, the suggestions we have put forward. Developments in Markov Chain Monte Carlo (MCMC) methods (Gelfand & Smith 1990; Diggle 1998) applied to the spatial analysis of point data may prove particularly useful for the type of modelling and mapping we have carried out. Although this will not overcome the problems that are inherent in the data, it will provide a comprehensive estimate of the prediction error associated with individual map locations.

Appendix 3

Some aspects of spatial disease modelling using hierarchical Bayes

Specification of priors in spatial and spatial-temporal models used in chapters 5 and 6

Bayesian statistical inference is based on posterior distributions which combine information available from the data via the likelihood function and any prior knowledge about the model parameters by specifying appropriate distributions for these parameters. We incorporate our prior information about the structure of the map by assuming conditional autoregressive (CAR) models for the area random effects. According to the CAR model the area specific spatial effects φ_i are modeled (conditional on their neighbours) as normally distributed with mean equal to the mean of the effects of its neighbours ($\overline{\varphi_i}$) and a variance that is inversely proportional to the number of neighbours n_i, i.e. $\varphi_i \mid \varphi_{-i} \sim \mathrm{N}(\overline{\varphi}_i, \sigma_\varphi^2/n_i)$ where $\overline{\varphi}_i = \frac{1}{n_i} \sum_{j \in neighbours\, of\, i} \varphi_j$. The effect of this prior distribution is to shrink the incidence rates of areas to that of the local mean, where the local mean is the mean of all contiguous areas excluding the area *i* itself. The posterior distribution of the rate of an area is therefore a compromise between the prior, which is based on the rates of neighbouring areas, and the data for the area, thus stabilising the rate in areas where the data are sparse due to small populations.

Sun *et al.* (2000) use a CAR prior that includes an index of spatial dependence, or shrinkage factor (Cressie 1993), ρ_1, so that

$$\overline{\varphi}_i = \frac{1}{n_j} \rho_1 \sum_{j \in neighbours\_of\_i} \varphi_j$$

If $|\rho_1| <1$ this leads to a proper prior. When $\rho_1 =0$, the φ_i 's are independent i.e. there is no spatial dependence between areas.

The temporal effects can be specified in a number of ways which will determine the prior distribution that is assumed for the term δ_t . The approach we used models time effects as first order auto-regressive AR(1) (Cryer, 1986). This allows for correlation between consecutive time periods which can be assessed via a temporal correlation coefficient ρ_2. The prior distributions for the time effects δ_t are specified as follows:

$\delta_1 \sim \text{Normal}(0, \sigma_\delta^2)$

where $\sigma_\delta = (\sigma_e)\surd(1- \rho_2^2)$ for year 1 (t=1), and

$\delta_t | \delta_{(t-1)} \sim \text{Normal}(\rho_2\delta_{t-1} , \sigma_e^2)$, for subsequent years (t>1).

Since no information is available for the remaining parameters we adopt standard conjugate priors, i.e. vague inverse gamma priors for the variances σ_e^2 and σ_φ^2 and vague normal priors for all other parameters. The prior for the correlation coefficient is specified $\rho_2 \sim$ uniform(-1,1).

The AR(1) effect described above can be applied to all areas uniformly for each consecutive year (δ_t), or we can specify a separate AR(1) effect for each area for each year (δ_{it}) thus allowing for space-time interaction. In our model we allow for one set of spatial effects for year 1 and then add AR(1) terms for each area for each consecutive year. The AR(1) term for year 2 is now conditional on the spatial term φ_i of year 1, i.e. $\delta_{it(t=2)} \sim \text{Normal}(\rho_2\varphi_i , \sigma_e^2)$, with the AR(1) effects for later years specified as described above. This allows for space-time interaction as well as spatial effects. The correlation coefficient ρ_2 will estimate the correlation of incidence rates in areas over time.

Markov Chain Monte Carlo simulation was used to obtain estimates of the posterior and predictive quantities of interest. The models were implemented using Gibbs sampling in the software package WinBUGS (2000). In order to properly monitor convergence a sampling scheme was designed using 3 independent chains. The number of iterations of ‘burn-in’ depended on convergence which was assessed using the method of Gelman and Rubin (1992). After convergence a final sample of 18000 was collected to obtain summaries of posterior distributions of the parameters.

Assessing model fit

We assessed model fit by comparing the expected predictive deviance (EPD) (Carlin and Louis, 1996) of different models. For the Poisson model this can be computed as follows: Let *l* denote all combinations of *i,t*, and let $y_{\mathrm{l,\ new}}$ refer to the posterior replicates of the observed data $y_{\mathrm{l,obs}}$. The $y_{\mathrm{l,new}}$ are obtained by sampling Poisson($\mathrm{E_l}\exp(\mu_{\mathrm{l}})$), with posterior samples of μ_{l} obtained after convergence of the MCMC run. The EPD is calculated as

$\mathrm{E}[d(\mathbf{y_{new}}, \mathbf{y_{obs}}) | \mathbf{y_{obs}}, \mathrm{M_i}]$, where

$$d(\mathbf{y_{new}}, \mathbf{y_{obs}}) = 2\Sigma_l\{ \mathrm{y_{l,obs}}\log(\mathrm{y_{l,obs}} / \mathrm{y_{l,new}}) - (\mathrm{y_{l,obs}} - \mathrm{y_{l,new}})\}$$

Carlin and Louis (Carlin and Louis, 1996, pp232-33) show that the EPD consists of two components, namely a likelihood ratio statistic (LRS) indicating goodness of fit, and a penalty term (PEN) which penalizes for under- or over fitting i.e. EPD=LRS+PEN. Smaller values of EPD are indicative of a better model. The two components can be calculated from

$$\mathrm{LRS} = d(\mathrm{E}[\mathrm{y_{new}}|\mathrm{y_{obs}}, \mathrm{M_i}], \mathrm{y_{obs}})$$

and

$$\mathrm{PEN} = 2\Sigma_l(\mathrm{y_{l,obs}})\{\log \mathrm{E}[\mathrm{y_{l,new}}|\mathrm{y_{obs}}] - \mathrm{E}[\log(\mathrm{y_{l,new}}) | \mathrm{y_{obs}}]\}$$

All three terms can be computed from MCMC samples.

References

1. Abdulla S, Schellenberg JA, Nathan R, Mukasa O, Marchant T, Smith T, Tanner M, Lengeler C (2001) Impact on malaria morbidity of a programme supplying insecticide treated nets in children aged under 2 years in Tanzania: community cross sectional study. BMJ; 322(7281): 249-50.
2. *Africa Data Sampler:* A Geo-Referenced Database for All African Countries. World Resources Institute. Washington DC, 1995.
3. Altman DG (1991) Practical Statistics for Medical Research. Chapman & Hall, London.
4. Armah GE, Adiamah JH, Binka FN, Adjuik M. (1997) The effect of permethrin impregnated bednets on the activity of malaria vectors in the Kassena-Nankana district of the Upper East Region of Ghana. In: Binka FN (1997) Impact and determinants of Permethrin impregnated bednets on child mortality in Northern Ghana. PhD dissertation, University of Basel.
5. Baird JK, Masbar S, Basri H, Tirtokusumo S, Subianto B, Hoffman SL. (1998) Age-dependent susceptibility to severe disease with primary exposure to Plasmodium falciparum. *J Infect Dis*; 178(2): 592-5.
6. Beck LR, Rodriguez MH, Dister SW, et al. (1994) Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *Am J Trop Med Hyg*. 51(3): 271-280.
7. Beier JC, Killen GF and Githure JI (1999) Short report: entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa. *Am J Trop Med Hyg*.; 61 (1), 109-13.
8. Bernadinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M (1995) Bayesian analysis of space-time variation in disease risk. *Statist. Med.*; 14: 2433-2443.
9. Bernadinelli L, Montomoli C (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11: 983-1007.
10. Binka FN (1997) Impact and determinants of Permethrin impregnated bednets on child mortality in Northern Ghana. PhD dissertation, University of Basel.

11. Bredenkamp BL, Sharp BL, Mthembu SD, Durrheim DN, Barnes KI (2000) Failure of sulphadoxine-pyrimethamine in treating Plasmodium falciparum malaria in KwaZulu-Natal province. *S Afr Med J.* (in press)
12. Bruce-Chwatt LJ, de Zuleta J (1980) The rise and fall of malaria in Europe. Oxford University Press, Oxford.
13. Bruce-Chwatt L J (1980) Essential Malariology. Heineman Medical Books Ltd, London, Chapter 1.
14. Burt, PJA, Colvin, J, Smith, SM (1995) Remote sensing of rainfall by satellite as an aid to Oedaleus senegalensis (Orthoptera, Acrididae) control in the Sahel. *Bulletin of Entomological Research*, 85: 455-462.
15. Carlin BP and Louis TA (1996) *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall. London.
16. Carrat F, Valleron AJ (1992) Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France. *American Journal of Epidemiology*; 135(11): 1293-1300.
17. Chandramohan D, Greenwood BM (1998) Is there an interaction between human immunodeficiency virus and Plasmodium falciparum. International Journal of Epidemiology; 27:296-301.
18. Clark Labs. Idrisi for Windows version 2.008 The Idrisi Project, Clark University, Worcester, MA. 1998.
19. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 1987 Sep; 43(3): 671-81.
20. Coetzee M, Craig MH, and le Sueur D (1998) Mapping the distribution of members of the Anopheles gambiae complex in Africa and adjacent islands. Parasitology Today 2000.
21. Craig MH, Sharp BL (1997) Comparative evaluation of four techniques for the diagnosis of Plasmodium falciparam infections. Trans R Soc Trop Med Hyg 91(3): 279-82.
22. Craig MH, Snow RW, le Sueur D (1999) A climate-based distribution model of malaria transmission in Africa. *Parasitology Today;* 5(3): 105-111.
23. Cryer JD (1986) *Time Series Analysis*. Duxbury Press. Boston.
24. Cressie NAC (1993) Statistics for spatial data. Revised Edition. John Wiley & Sons, Inc. New York.

25. Cressie N. (2000) Geostatistical methods for mapping environmental exposures. In: Elliott P, Wakefield JC, Best NG and Briggs DJ. (Eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford: 185-204.
26. Cuzick J, Elliott P (1992) Small-area studies: purpose and methods. In: Geographical and Environmental Epidemiology: Methods for Small Area Studies (Eds. Cuzick J, Elliott P). Oxford University Press, Oxford: 14-21.
27. Deichman U (1996) African population database. Digital database and documentation. Santa Barbara, CA, National centre for Geographical Information and Analysis.
28. Department of Health (2000a). Notifiable medical conditions. *Epidemiological Comments*. 1999; 1(3): 11-12.
29. Department of Health (2000b). Tenth national HIV survey in women attending antenatal clinics in public health services in South Africa, Oct/Nov 1999. Pretoria.
30. Diggle PJ, Tawn JA, Moyeed RA (1998) Model based geostatistics. J.R Statist. Soc. C; 47(3): 299-350.
31. Dossou-Yovo J, Doannio JMC, Diarrassouba S and Chauvancy G (1998) The impact of rice fields on the transmission of malaria in Bouake, Coted'Ivoire. Bulletin De La Societe De Pathologie Exotique; 91: 327-333.
32. Doumbo O, Ouattara N I, Koita O, Maharaux A, Toure YT, Traore S F, Quilici M (1989) Approche eco-geographique du paludisme en milieu urbain: ville de Bamako au Mali. Ecol. Hum; 8(3): 3-15.
33. Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J, Grundy C. (1996) Cancer incidence near municipal solid waste incinerators in Great Britain. *Br J Cancer*; 73(5): 702-710.
34. Elliott P, Wakefield JC, Best NG and Briggs DJ. (2000) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford.
35. Elliott P, Westlake AJ, Hills M, Kleinschmidt I, Rodrigues L, McGale P, Marshall K, Rose G. (1992) The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *Journal of Epidemiology and Community Health*; 46: 345-349.
36. *FAO.* Report on the agro-ecological zones project. Vol. 1: Methodology and results for Africa. World Soil Resources Report 48. Rome, Italy 1978.

37. Faye O, Fontenille D, Gaye O, Sy N, Molez JF, Konate L,Hebrard G, Herve JP, Trouillet J, Diallo S and Mouchet J (1995) Malaria and rice growing in the Senegal river delta (Senegal). Ann.Soc.Belg.Med.Trop; 75: 179-189.
38. Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*; 7: 457-472.
39. Gelfand AE and Smith AFM (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*; 85: 398-409.
40. Geostatistical Environmental Assessment Software. GEO-EAS 1.2.1. U.S. Environmental Protection Agency, Las Vegas, Nevada 1991.
41. Ghebreyesus TA, Haile M, Witten KH, Getachew A, Yohannes AM, Yohannes M, Teklehaimanot HD, Lindsay SW, and Byass P (1999) Incidence of malaria among children living near dams in northern Ethiopia: community based incidence survey. *BMJ*; 319: 663-666.
42. Gilles HM, Warrell DA (1993). Bruce-Chwatt's Essential Malariology. Edward Arnold. London.
43. Gillies MT and De Meillon B (1968) The Anophelinae of Africa South of the Sahara. The South African Institute for Medical Research. Johannesburg, 212-219.
44. GS+ Geostatistics for the Environmental Sciences, Version 3.10.2 Beta. Gamma Design Software, PO Box 201, Plainwell, MI 49080 USA.
45. Gupta S, Snow RW, Donnelly C, Marsh K, Newbold C (1999). Immunity to severe malaria is acquired after one or two exposures. Nature Medicine, 5: 340-343.
46. Hargreaves K, Koekemoer LL, Brooke BD, Hunt RH, Mthembu J, Coetzee M. (2001) Anopheles funestus is resistant to pyrethroid insecticides in South Africa. *Med Vet Entomol.* 2000 Jun;14(2):181-9.
47. Haworth J (1988) The global distribution of malaria and the present control effort. In: Wernsdorfer WH and McGregor I (eds) Malaria: Principles and practice of malariology. Vol. 2. Churchill Livingstone. London.
48. Hay SI, Omumbo JA, Craig MH, Snow RW (2000). Earth observation, geographic information syatems and *Plasmodium falciparum* malaria in sub-Saharan Africa. *Advances in Parasitology*; 47: 173-209.
49. Hay SI, Snow RW, Rogers DJ. (1998) Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Trans R Soc Trop Med Hyg.*;92(1):12-20.

50. Heisterkamp SH, Doornbos G, Nagelkerke NJD (2000) Assessing health impact of environmental pollution sources using space-time models. *Statist. Med.*; 19: 2569-2578.
51. Hosmer DW, Lemshow S. (1989) Applied Logistic Regression. John Wiley & Sons. New York.
52. Hutchinson MF, Nix HA, McMahon JP, Ord KD (1995) Climate data: a topographic and climate database (CD-Rom). Centre for Resource and Environmental Studies, The Australian National University, Canberra, ACT 0200, Australia.
53. Imevbore AM. (1991) Malaria and development in Africa: a cross sectoral approach. American Association for the Advancement of Science. Washington, DC; p105-12.
54. Kafadar K (1996) Smoothing geographical data, particularly rates of disease. Statistics in Medicine; 15: 2539-2560.
55. Kitron U, Pener H, Costin C, Orshan L, Greenberg Z, Shalom U (1992) Geographic information syatems in malaria surveillance: mosquito breeding and imported cases in Israel, Am J Trop Med Hyg. 1994, 50(5): 550-556.
56. Kleinschmidt I, Bagayako M, Clarke GPY, Craig M, Le Sueur D (2000) A spatial statistical approach to malaria mapping. *International Journal of Epidemiology*; 29(2): 355-361.
57. Kleinschmidt I et al. (1999) Ecosystem Health Analytical Workshop (ESHAW) Report: A malaria distribution model for West Africa: Technical report on outputs of workshop held in Bouake, Cote d' Ivoire, and subsequent analysis (http://www.mara.org.za/projects.htm).
58. Kleinschmidt I, Pattenden P, Walls P, Grundy C, Stevenson, S, Shaddick G, Elliott P (1994) A national health statistics database: data quality requirements for small area health studies. In Frederiksen P (ed.) *Proceedings of Eurocarto XII Conference on Geo-related databases*, Copenhagen, Denmark.
59. Kleinschmidt I, Sharp BL, Clarke GPY, Curtis B, Fraser C (2001) Use of generalised linear mixed models in the spatial analysis of small area malaria incidence rates in KwaZulu Natal, South Africa. *Am J Epid*; 153:1213-21.
60. Kleinschmidt I, Sharp B, Mueller I, Vounatsou P. Rise in malaria incidence rates in South Africa: a small area spatial analysis of variation in time trends. *Am J Epid* (in press).

61. Knorr-Held L, Besag J (1998) Modelling Risk from a disease in time and space. *Statist. Med.*; 17: 2045-2060.

62. Krige, DG (1966) Two dimensional weighted moving average trend surfaces for ore-evaluation. *J. S. Afr. Inst. Mining Metall*; 66: 13-38.

63. Kuhlow F, Zielke E (1976) Distribution and prevalence of Wuchereria bancrofti in various parts of Liberia. Tropenmedizin und Parasitologie. 27 (1): 93-100.

64. Le Sueur D (1991) The ecology, over-wintering and population dynamics of the pre-imaginal stages of the Anoptheles Gambiae Giles Complex (Diptera: Culicidae) in Northern Natal, South Africa. PhD thesis, University of Natal, Pietermaritzburg: 134-185.

65. Le Sueur D, Sharp BL (1988) The breeding requirements of three members of the Anopheles gambiae Giles complex in the endemic malaria area of Natal, South Africa. Bulletin of Entomological Research; 78: 549-560.

66. Lindsay SW and Thomas CJ (2000) Mapping and estimating the population at risk from lymphatic filariases in Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene. 94: 37-45.

67. Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) SAS® System for mixed models, Cary, NC: SAS Institute Inc., 423-460.

68. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J (1990) Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*;335(8692):765-74.

69. Marsh K, Snow RW (1999) Malaria transmission and morbidity. *Parassitologia*; 41(1-3): 241-6.

70. Martens P, Hall L (2000) Malaria on the move: Human population movement and malaria transmission. Emerg Infect Dis[serial online]; 6(2). Available from: URL: http://www.cdc.gov/ncidod/EID/eid.htm.

71. McCullagh P and Nelder JA (1989). Generalised Linear Models, 2nd Edition, New York: Chapman and Hall.

72. Mnzava AEP, Dlamini SS, Sharp BL, Mthembu DJ, Gumede K, Kleinschmidt I, Gouws E. Malaria control: bednets or spraying - trial in KwaZulu Natal, South Africa. *Trans. Roy. Soc. Trop. Med & Hyg.* 1999; 93: 1-2.

73. Molineaux L (1988) The epidemiology of human malaria as an explanation of its distribution, including some implications for its control. In: Wernsdorfer WH and

McGregor I (eds) Malaria: Principles and practice of malariology. Vol. 2. Churchill Livingstone. London.

74. Molineaux L, Gramiccia G (1980) The Garki project.Research on the epidemiology and control of malaria in the Sudan savanna of West Africa. World Health Organisation. Geneva.
75. Mueller I. (2000) Application and validation of new approaches in spatial analysis as tools in communicable disease control, health systems and environmental epidemiology in tropical countries. PhD thesis. University of Basel.
76. NDVI Image Bank Africa 1981-1991 (CD-ROM). Food and Agriculture Organisation (FAO) of the United Nations Remote Sensing Centre; Africa Real Time Environmental Monitoring Information System (ARTEMIS), NASA Goddard Space Flight Centre, Greenbelt, MD 20771, USA 1991.
77. Notifiable medical conditions, Department of Health, Republic of South Africa. *Epidemiological Comments*. 1999; 1(3): 11-12.
78. Notifiable medical conditions, Department of Health, Republic of South Africa. *Epidemiological Comments*. 2000; 2(1): 16.
79. Oliver MA, Muir KR, Webster R, Parkes SE, Cameron AH, Stevens MCG, Mann JR. (1992) A geostatistical approach to the analysis of pattern in rare disease. Journal of Public Health Medicine. 14(3): 280-289.
80. Packard RM (1984) Maize, cattle and mosquitoes: the political economy of malaria epidemics in colonial Swaziland. Journal of African History; 25: 189-212.
81. Pillay T. Malaria scare hits Durban. *Sunday Times* (South Africa), April 16, 2000: p5.
82. Potthoff RF, Whittinghill M (1966) Testing for homogeneity. II. The Poisson distribution. Biometrika; 53: 167-82.
83. Price, JC (1984) Land surface measurement for the split window channels of NOAA 7 advanced very high resolution radiometer. *Journal of Geophysical Research* 89: 7231-7237.
84. Ravoniharimelina B, Romi R, Sabatinelli G. (1992) Longitudinal study of the larval habitats of Anopheles gambiae s.l. in a canton of the province of Antanarivo (Central Highlands of Madagascar). Annales de Parasitologie Humaine et Comparee. 67(1): 26-30.
85. Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol;28(5):964-74

86. Sachs J (2000) Economic Analyses indicate the burden of malaria is great, in: The African Summit on Roll Back Malaria, Abuja; page 27-35, WHO/CDS/RBM/2000.17

87. SAS System version 6.12 (1996), SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

88. Schellenberg JA, Newell JN, Snow RW Mung'ala V, Marsh K, Smith PG, Hayes RJ (1998) An analyis of the geographical distribution of severe malaria in children in Kilifi District, Kenya. *International Journal of Epidemiology.* 27:323-329.

89. Sharp B, Craig M, Curtis B, Mnzava A, Maharaj R, Kleinschmidt I (2001) Malaria. *South African Health Review 2000*. Health Systems Trust, Durban. http://www.hst.org.za/sahr/2000/chapter18.htm.

90. Sharp B, Fraser C, Naidoo K, Dlungwane M. (1999) Computer assisted health information system for malaria control. MIM African Malaria Conference, Durban, South Africa.

91. Sharp BL, Le Sueur D (1996) Malaria in South Africa – the past, the present and selected implications for the future. *S Afr Med J*; 86: 83-89.

92. Sharp BL, Ngxongo S, Botha MJ, le Sueur D (1988) An analysis of 10 years of retrospective malaria data from the KwaZulu areas of Natal. *S Afr J Sci*; 84:102-106.

93. Sissoko MS, Dicko A, Briet OJT, Sissoko M, Sagara I, Keita HD, Sogoba M, Rogier C, Toure YT, Doumbo O. The impact of irrigated rice cultivation on the incidence of malaria in children in the district of Niono, Mali. (*Submitted*)

94. Smith T, Charlwood JD, Takken W, Tanner M, Spiegelhalter DJ (1995) Mapping the density of malaria vectors within a single village. *Acta Tropica, 59: 1-18.*

95. Snijders, FL (1991) Rainfall monitoring based on Meteosat data – a comparison of techniques applied to the western Sahel. *International Journal of Remote Sensing* 12: 1331-1347

96. Snow J (1855) On the mode of communication of cholera, 2nd edn. Churchill: London.

97. Snow RW, Bastos de Azevedo I, Lowe BS, Kabiru EW, Nevill CG, Mwankusye S, Kassiga G, Marsh K, Teuscher T (1994) Severe childhood malaria in two areas of markedly different falciparum transmission in east Africa. Acta Trop; 57(4): 289-300.

98. Snow RW, Craig MH, Deichmann U, le Sueur D (1999) A continental risk map for malaria mortality among African children. Parasitol Today. 15(3):99-104.

99. Snow RW, Marsh K, LeSeur D (1996) The need for maps of transmission intensity to guide malaria control in Africa. Parasitology Today; 12: 455-457.

100. Snow RW, Omumbo JA, Lowe B, Molyneux CS, Obiero JO, Palmer A, Weber MW, Pinder M, Nahlen B, Obonyo C, Newbold C, Gupta S, Marsh K (1997) Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet*, 349: 1650-4.

101. Snow RW, Craig M, Deichman U, Marsh K (1999) Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bulletin of the World Health Organisation*, 77(8): 624-640.

102. Snow RW, Gouws E, Omumbo JA, Rapuada B, Craig MH, Tanser FC, Le Seur D, Ouma J (1998a) Models to predict the intensity of Plasmodium falciparum transmission: applications to the burden of disease in Kenya. *Trans R Soc Trop Med Hyg.*;92(6):601-6.

103. Snow RW and Marsh K. (1998b) New insights into the epidemiology of malaria relevant for disease control. *British Medical Bulletin*; 54(2): 293-309.

104. Snow RW, Nahlen B, Palmer A, Donnelly CA, Gupta S, Marsh K (1998c). Risks of severe malaria among African Infants: direct evidence of clinical protection during early infancy. *Journal of Infectious Diseases*, 177: 819-822.

105. StataCorp. 1997-2000. Stata® Statistical Software: Release 5.0 to 7.0. College Station, TX: Stata Corporation.

106. Statistics South Africa (1999). Mid-year estimates. Statistical release P0302., Pretoria.

107. Sun D, Tsutakawa RK, Kim H, He Z (2000) Spatio-temporal interaction with disease mapping. *Statist. Med.*; 19: 2015-2035.

108. Tanser FC, Sharp B and Le Sueur D. Malaria seasonality and population exposure in Africa: a high resolution climatic model. Ch 7 In: Tanser FC. The application of geographical information systems to infectious diseases and health system in Africa. PhD thesis, University of Natal. November 2000.

109. Thomson MC, Connor SJ, 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. *Am J Trop Med Hyg*; 61(1): 2-8.

110. Thomson MC, Connor SJ, Milligan PJM, Flasse SP. (1996) The ecology of malaria – as seen from earth –observation satellites. *Annals of Tropical Medicine and Parasitology*; 90(3): 243-264.

111. Thomson MC, Connor SJ, Milligan P, Flasse SP (1997) Mapping malaria risk in Africa: What can satellite data contribute? *Parasitology Today*; 13: 313-318.

112. *Towards an atlas of malaria risk in Africa*: First technical report of the MARA/ARMA collaboration. MARA/AMRA, 771 Umbilo Road, Congella, Durban, South Africa. December 1998 (http://www.mara.org.za).

113. US Bureau of the Census (1995). Urban and rural definitions. Geography Division, U.S. Bureau of the Census, Washington, DC 20233 (http://www.census.gov/population/censusdata/urdef.txt)

114. Verhoeff FH, Brabin BJ, Hart CA, Chimsuku L, Kazembe P and Broadhead RL (1999) Increased prevalence of malaria in HIV-infected pregnant women and its implications for malaria control. *Tropical Medicine and International Health*; Vol.4 No.1 pp 5-12.

115. Vounatsou P, Smith T, Gelfand AE (2000) Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics*; 1(2): 177-189.

116. Wakefield JC, Best NG, Waller L (2000) Bayesian approaches to disease mapping. In: Spatial Epidemiology: Methods and Applications (Elliott P, Wakefield JC, Best NG and Briggs DJ, eds), 104-127. Oxford University Press, Oxford.

117. Waller LA, Carlin BP, Xia H, Gelfand AE (1997) Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*; 92(438): 607-617.

118. Walter SD (1992) The analysis of regional patterns in health data, Part 1. *Am. J. Epidemiol*. 136(6): 730-741.

119. Walter SD (1994) A simple test for spatial pattern in regional health data. *Statistics in Medicine*, 13: 1037-1044.

120. Wernsdorfer WH, McGregor I (1988) Malaria: Principles and Practice of Malariology. Churchill Livingstone, London.: 950-951.

121. Wernsdorfer G, Wernsdorfer WH (1988) Social and economic aspects of malaria and its control. In: Wernsdorfer WH and McGregor I (eds) *Malaria: Principles and practice of malariology*. Vol. 2. Churchill Livingstone. London.

122. White GB (1974) Anopheles gambiae complex and disease transmission in Africa. *Trans R Soc Trop Med Hyg*, 68: 278-301.

123. Whitworth J, Morgan D, Quigley M, Smith A, Mayanja B, Eotu H, Omoding N, Okongo M, Malamba S, Ojwiya A (2000) Effect of HIV-1 and increasing immunosuppression on malaria parasitaemia and clinical episodes in adults in rural Uganda: a cohort study. *Lancet*; 356: 1051-56.

124. WinBUGS (Bayesian inference Using Gibbs Sampling) (2000). Software Version 1.3. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK.

125. Windmeijer, PN and W Andriesse (Eds.) (1993) Inland valleys in West Africa. An Agro-ecological characterization of rice-growing environments. Publication 52. International Institute for Land Reclamation and Improvement, Wageningen, The Netherlands. p 160 .

126. World Bank. World Development Report 1993; Investing in Health. Oxford University Press. New York 1993. p. 329.

127. Xia H and Carlin BP (1998) Spatio-temporal models with errors in co-variates: mapping Ohio lung cancer mortality. *Statist. Med.* 17: 2025-2043.

Curriculum Vitae

Name

Immo Kleinschmidt

Address

Home: 66 Twelfth Avenue, Morningside, Durban 4001, South Africa.

Work: Medical Research Council, P.O.Box 17120, Congella 4013, Durban, South Africa.

Nationality

Dual South African/British.

Previous Higher Educational Qualifications

1) B.Sc.(Electrical Engineering) from University of Witwatersrand, Johannesburg, South Africa. (March 1971)

2) Post-Graduate Certificate in Education (with commendation) from Keele University, England. (July 1974)

3) M.Sc. in Computer Science (with distinction) from City University, London (March 1989).

4) M.Sc. in Medical Statistics (with distinction) from London School of Hygiene and Tropical Medicine (London University, November 1994).

5) Postgraduate Diploma in Medical Statistics from the London School of Hygiene and Tropical Medicine, February 1995.

Publications

Westlake AJ, Kleinschmidt I**.** The implementation of area and membership retrievals in point geography using SQL. In Statistical and Scientific Database Management, Proceedings of the 5th International Conference on Statistical and Scientific Database Management, Michaelawicz Z, ed. Springer Verlag, Lecture Notes in Computer Science 420. 1990

Elliott P, Hills M, Beresford J, Kleinschmidt I, Pattenden S, Jolley D, Rodrigues L, Westlake AJ, Rose G. Incidence of cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet*, 1992; 339: 854-858.

Elliott P, Kleinschmidt I, Westlake AJ. Use of routine data in studies of point sources of environmental pollution. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small area studies*. Oxford: Oxford University Press, 1992.

Elliott P, Westlake AJ, Hills M, Kleinschmidt I, Rodrigues L, McGale P, Marshall K, Rose G. The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *Journal of Epidemiology and Community Health*, 1992; 46: 345-349.

Kleinschmidt I. A database structure and analysis system for small area health statistics. In Westlake A.J. (ed.) *Geographical methods in small area health studies*. Proceedings of a workshop held at the London School of Hygiene and Tropical Medicine on 22 June 1990. London, Small Area Health Statistics Unit.

Kleinschmidt I. The Relational System: A Way of Looking at Data. In Westlake A.J. (ed.) *Relational Databases: Tutorial Papers on the Relational Model selected from an SGCSA Conference on Relational Databases in Survey Research*. Study Group on Computers in Survey Analysis, London, 1993.

Kleinschmidt I, Pattenden P, Walls P, Grundy C, Stevenson, S, Shaddick G, Elliott P. A national health statistics database: data quality requirements for small area health studies. In Frederiksen P (ed.) *Proceedings of Eurocarto XII Conference on Geo-related databases*, Copenhagen, Denmark, October 1994.

Sans S, Elliott P, Kleinschmidt I, Shaddick G, Pattenden S, Walls P, Grundy C, Dolk H. Cancer incidence and mortality near the Baglan Bay petrochemical works, South Wales. *Occupational and Environmental Medicine* 1995; 52: 217-224.

Kleinschmidt I, Hills, M, Elliott P. Smoking behaviour can be predicted by neighbourhood deprivation measures. *Journal of Epidemiology and Community Health* 1995;49:S72-S77.

Dolk H, Mertens B, Kleinschmidt I, Walls P, Elliott P. A standardised approach to the control of socio-economic confounding in small area studies of environment and health. *Journal of Epidemiology and Community Health* 1995;49:S9-S14.

Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J, Grundy C. Cancer incidence near municipal solid waste incinerators in Great Britain. *Br J Cancer* 1996; 73(5): 702-710.

Kleinschmidt I. Data requirements for epidemilogical research. *Newsletter of the Mine Medical Officers Association of South Africa.* January 1996.

Wilkinson P, Thakrar B, Shaddick G, Stevenson S, Pattenden S, Landon M, Grundy C, Elliott P, Kleinschmidt I, Walls P, Dolk H. Cancer Incidence and Mortality around the Pan Britannica Industries Pesticide Factory, Waltham Abbey. Report by the Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine, London 1996.

Elliott P, Kleinschmidt I. Angiosarcoma of the liver in Great Britain in proximity to vinyl chloride sites. *Occupational and Environmental Medicine* 1997; 54(1):14-18.

Jarvelin MR, Elliott P, Kleinschmidt I, Martuzzi M, Grundy C, Hartikainen AL, Rantakallio P. Ecological and individual predictors of birthweight among Northern Finland birth cohort for 1986. *Paediatric and Perinatal Epidemilogy* 1997; 11: 298-312.

Dolk H, Shaddick G, Walls P, Grundy C, Thakrar B, Kleinschmidt I, Elliott P. Cancer Incidence near Radio and Television Transmitters in Great Britain. *Am J Epidemiol* 1997;145(1):1-9.

Mqoqi N, Churchyard G, Kleinschmidt I, Williams B. Attendance versus compliance with tuberculosis treatment in an occupational setting: A pilot study. *S Afr Med J* 1997; 87(11):1517-1521.

Kleinschmidt I, Churchyard G. Variation in incidences of tuberculosis in subgroups of South African gold miners. *Occupational and Environmental Medicine* 1997; 54(9): 636-641.

Kleinschmidt I. Predictors of smoking in a cross-section of novice mineworkers. *Cent Afr J Med*. 1997 Nov; 43(11): 321-324.

Campbell C, Williams B, Mqoqi N, Kleinschmidt I. Occupational health, occupational illness: tuberculosis, silicosis and HIV on the South African mines. In: Banks and Parker (Eds) Occupational Lung Disease. Chapman and Hall. London 1998.

Sitas F, Pacella-Norman R, Peto R, Collins R, Bradshaw D, Kleinschmidt I, Kielkowski D, Saloojee Y, Yach D, Lopez A, Bah S. Why do we need a large study on tobacco-attributed mortality in South Africa? Editorial. *S Afr Med J* 1998; 88(8).

Colvin M, Mullick S, Kleinschmidt I. HIV Surveillance in South Africa. Letter. *S Afr Med J* 1998; 88: 1046.

Kleinschmidt I. South African Tuberculosis mortality data - showing the first sign of the AIDS epidemic? *S Afr Med J* March 1999: 269-273.

Churchyard GJ, Kleinschmidt I, Corbett E, Mulder D, DeCock K. Mycobacterial disease in South African gold miners in the era of HIV infection. *Int J Tuberc Lung Dis 1999; 3(9): 791-8.*

Beksinska M, Rees H, Kleinschmidt I, McIntyre J. The practice and prevalence of dry sex among men and women in South Africa: a risk factor for sexually transmitted diseases. *Sex Transm Inf* 1999; 75: 178-180.

Mnzava AEP, Dlamini SS, Sharp BL, Mthembu DJ, Gumede K, Kleinschmidt I, Gouws E. Malaria control: bednets or spraying - trial in KwaZulu Natal, South Africa. *Trans. Roy. Soc. Trop. Med & Hyg.* 1999; 93: 1-2.

Immo Kleinschmidt. Malaria among children living near dams - are the standard errors underestimated? *eBMJ* 25 October 1999. (Electronic Letter)

Churchyard G, Corbett EL, Kleinschmidt I, Mulder D, De Cock K, Williams B. Drug-resistant tuberculosis in South African gold miners: incidence and associated factors. *Int J Tuberc Lung Dis*. 2000 May;4(5):433-40.

Katharine F. Mallory, Gavin J. Churchyard, Immo Kleinschmidt, Kevin M De Cock, Elizabeth L Corbett. The impact of HIV infection on recurrent tuberculosis in South African gold miners. *Int J Tuberc Lung Dis*. 2000 May; 4(5):455-62.

Pettifor A, Rees H, Beksinska ME, Kleinschmidt I, McIntyre J. In-vitro assessment of the structural integrity of the female condom after multiple wash, dry and re-lubrication cycles. *Contraception.* 2000; 61: 271-276.

Churchyard GJ, Kleinschmidt I, Corbett EL, Murray J, Smith J, De Cock KM. Factors associated with increased case-fatality rate in HIV-infected and non-infected South African gold miners with pulmonary tuberculosis. *Int J Tuberc Lung Dis*. 2000 Aug;4(8):705-12.

Kleinschmidt I, Bagayako M, Clarke GPY, Craig M, Le Sueur D. A spatial statistical approach to malaria mapping. *International Journal of Epidemiology* 2000; 29(2): 355-361.

Desai DK, Adanlawo M, Naidoo DP, Moodley J, Kleinschmidt I. Mitral stenosis in pregnancy: a four-year experience at King Edward VIII Hospital, Durban, South Africa. *BJOG* 2000 Aug;107(8):953-8

Taylor M, Jinabhai CC, Couper I, Kleinschmidt I, Jogessar VB. The effect of different anthelminthic treatment regimens combined with iron supplementation on the nutritional status of schoolchildren in KwaZulu-Natal, South Africa: A randomized controlled trial. *Trans R Soc Trop Med Hyg*. 2001 Mar-Apr; 95(2):211-6.

Schlebusch L, Bosch DA, Polglase G, Kleinschmidt I, Pillay DJ, Cassimjee MH. A double blind placebo controlled, double centre study of the effects of an oral multivitamin-mineral combination on stress. *S Afr Med J* 2000; 90: 1216-1223.

Bagratee JS, Moodley J, Kleinschmidt I, Zawilski W. A Randomised Controlled Trial Of Antibiotic Prophylaxis In Elective Caesarean Delivery. *BJOG*. 2001 Feb;108(2):143-8.

Colvin M, Dawood S, Kleinschmidt I, Mullick S, Lalloo U. Prevalence of HIV and HIV-related diseases on the adult medical wards of a tertiary hospital in Durban, South Africa. *Int J STD AIDS*. 2001 Jun; 12(6):386-9.

Beksinska ME, Rees HV, Dickson-Tetteh KE, Mqoqi N, Kleinschmidt I, McIntyre JA. Structural integrity of the female condom after multiple uses, washing, drying and re-lubrication. *Contraception*. 2001 Jan;63(1):33-36.

Donnellan R, Kleinschmidt I, Chetty R. Cyclin E immunoexpression in breast ductal carcinoma: Pathologic correlations and prognostic implications. *Hum Pathol* 2001 Jan;32(1):89-94.

Sharp B, Craig M, Curtis B, Mnzava A, Maharaj R, Kleinschmidt I. Malaria. *South African Health Review 2000*. Health Systems Trust, Durban, March 2001. http://www.hst.org.za/sahr/2000/chapter18.htm.

Moodley M, Moodley J, Kleinschmidt I. Invasive cervical cancer and human immunodeficiency virus (HIV) Infection – a South African perspective. *Int J Gynecol Cancer*. 2001 May-Jun;11(3):194-7.

Kleinschmidt I, Sharp BL, Clarke GPY, Curtis B, Fraser C (2001) Use of generalised linear mixed models in the spatial analysis of small area malaria incidence rates in KwaZulu Natal, South Africa. *Am J Epid* 2001; 153:1213-21.

Kleinschmidt I, Sharp B, Mueller I, Vounatsou P. Rise in malaria incidence rates in South Africa: a small area spatial analysis of variation in time trends. *Am J Epid* (in press).

Kleinschmidt I, Omumbo J, Briët O, Van de Giesen N, Sogoba N, Mensah NK, Windmeijer P, Moussa M, Teuscher T. An empirical malaria distribution map for West Africa. *Tropical Medicine and International Health* 2001; 6(10): 779-786.

Kleinschmidt I & Sharp B. Patterns in age-specific malaria incidence in a population exposed to low levels of malaria transmission intensity. *Trop Med Int Health* (in press).

Present Employment (since February 1998)

Senior Specialist Scientist (Biostatistics) at Medical Research Council, Durban.

Honorary position (since May 2000)

Honorary lecturer in Department of Experimental and Clinical Pharmacology, Faculty of Medicine, University of Natal.

Previous Employment

November 1996 to January 1998

Director of household surveys and vital statistics, Statistics South Africa, Private Bag X44, Pretoria 0001.

December 1994 to November 1996

Senior Medical Statistician, Epidemiology Research Unit, Medical Bureau for Occupational Diseases, 144 De Korte Street, Braamfontein, Johannesburg.

October 1987 to December 1994

Senior Computer Scientist/Research Fellow in Environmental Epidemiology Unit, London School of Hygiene and Tropical Medicine.

Recent Consultancies

Temporary Technical advisor to Rapid Geographical Assessment of Lymphatic Filariases meeting, WHO workshop, Geneva.17 - 21 May 1999.

WHO Temporary Advisor and Facilitator at “Workshop on advanced spatial analysis of disease prevalence data”, Ouagadougou, Burkina Faso. 4-8th March 2001.

Professional Institutions

Member of the Royal Statistical Society.

Member of the South African Statistical Association.