Bayesian spatial models applied to malaria epidemiology

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Federica Giardina
aus Pescara, Italien

Basel, December 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. M. Tanner, P.D. Dr. P. Vounatsou, and Prof. Dr. A. Biggeri.

Basel, den 10 December 2013

Prof. Dr. Jörg Schibler
Dekan

Science is built up with facts, as a house is with stones.
But a collection of facts is no more a science than a heap of stones is a house.
(Henri Poincaré)

Summary

Malaria is a mosquito-borne infectious disease caused by parasitic protozoans of the genus *Plasmodium* and transmitted to humans via the bites of infected female *Anopheles* mosquitoes. Although progress has been seen in the last decade in the fight against the disease, malaria remains one of the major cause of morbidity and mortality in large areas of the developing world, especially sub-Saharan Africa. The main victims are children under five years of age. The observed reductions are going hand in hand with impressive increases in international funding for malaria prevention, control, and elimination, which have led to tremendous expansion in implementing national malaria control programs (NMCPs). Common interventions include indoor residual spraying (IRS), the use of insecticide treated nets (ITN) and environmental measures such as larval control. Specific targets have been set during the last decade. The Millennium Development Goal (MDG) 6 aims to halve malaria incidence by 2015 as compared to 1990 and to achieve universal ITN coverage and treatment with appropriate antimalarial drugs. In 2010, the Global Malaria Action Plan (GMAP), created by the Roll Back Malaria (RBM) Partnership, called for rapid scaling-up to achieve universal coverage with some form of vector control.

Transmission of malaria depends on the distribution and abundance of mosquitoes, which are sensitive to environmental and climatic conditions, such as temperature, rainfall, vegetation and land use. Geostatistical models can be used to estimate the environment-disease relation at fixed locations over a continuous study area, and predict the burden of malaria at places where data on transmission are not available. Data are correlated in space because common exposures of the disease influence malaria transmission similarly in neighboring areas. Geostatistical models take into account spatial correlation by introducing location-specific random effects. Bayesian model formulation is a natural and convenient choice for model fit via the implementation of Markov chain Monte Carlo (MCMC).

This thesis develops novel statistical methodology for (i) producing accurate disease

burden estimation (malaria parasitemia risk and number of infected) at high spatial resolution and (ii) assessing the coverage and effectiveness of vector control interventions. Produced maps and estimates make a significant contribution to the monitoring and evaluation of the progress toward the targets of disease reduction and intervention coverage scaling-up.

Contemporary information on malaria prevalence for this work was provided mostly by Malaria Indicator Surveys (MIS) and Demographic Health Surveys (DHS) with malaria modules. MIS are nationally representative surveys developed by RBM that collect parasitaemia data on children below the age of 5 years and are usually carried out during high malaria transmission seasons. Historical data were extracted from the Mapping Malaria Risk in Africa (MARA) database, that contains over 10,000 geographically positioned surveys from gray or published literature across all sub-Saharan Africa. Malaria confirmed cases data were gathered by the Health Management Information System (HMIS) in Zambia.

In *Chapter 2*, Bayesian geostatistical Zero-Inflated Binomial (ZIB) models were developed to produce spatially explicit parasitaemia risk estimates and number of infected children below the age of 5 years in Senegal. Geostatistical ZIB models were able to account for the large number of zero-prevalence survey locations (70%) in the Senegalese MIS 2008 dataset. Model validation confirmed that the ZIB model had a better predictive ability than the standard Binomial analogue. Bayesian variable selection methods were incorporated within a geostatistical framework in order to choose the best set of environmental and climatic covariates associated with the parasitaemia risk. Several ITN coverage indicators were calculated to assess the effectiveness of interventions.

Chapter 3 explores different modelling specifications of zero-altered models and suggests model formulations in a geostatistical setting. In particular, the work addresses the problem of selecting variables and assessing the need of incorporating a spatial structure in the modelling of the mixing probability and the non-degenerate distribution. Specific prior distributions for spatial process selection based on non-zero random effects variances are proposed and analyzed through a set of simulated data. The proposed approach was applied to obtain simultaneous estimation of suitability of malaria transmission and of conditional risk in Senegal. The renewed interest in malaria eradication suggests that more sparse data will be produced from parasitological as well as entomological surveys.

The impact of environmental predictors on malaria risk is commonly modeled as a linear effect, constant throughout the study area. However, more flexible functional forms, such

as piece-wise linear or splines may be required to capture non-linear relationships between the predictors and malaria risk. The area under study is often large, covered by different regions, (e.g. ecological zones) and the relationship between the disease and its risk factors may not be constant across the area. Furthermore, the spatial correlation is likely to vary not only as a function of distance but also of their geographic position. *Chapter 4* develops Bayesian spatial variable selection methods with spike-and-slab prior structure that allow the choice of different predictors and their functional forms in non-stationary geostatistical models for mapping disease survey data. Penalized spline effects are re-parameterized as mixed effects terms and their selection is based on non-zero random effects variance identification. Spatially varying weights are proposed to achieve smoothness across irregularly shaped regions. These methods are applied on the analysis of data from the Mali DHS to obtain spatially explicit estimates of the disease burden in the country.

Few studies have linked malaria survey data with Remote Sensing (RS)-derived land cover/use (LC) variables. *Chapter 5* assesses the effect of the spatial resolution of RS-derived environmental variables on malaria risk estimation in Mozambique. A proximity measure to define LC variables to be included as covariate in a geostatistical model for malaria risk is proposed and applied to the Mozambican DHS dataset in 2011. The model was validated using a LC layer at 5 m resolution produced by MALAREO, a Seventh Framework Programme (FP7) funded project which covered part of Mozambique during 2010-2012, and freely available Remote sensing sources. The predictive performance was compared.

When prevalence estimation relies on the compilation of historical data, surveys are commonly heterogeneous in season and sampled population (age groups). In *Chapter 6*, age and time heterogeneity between surveys is addressed by proposing a general formulation that couples spatial statistical models and mathematical transmission models allowing uncertainty incorporation. The proposed methodology is applied to obtain age/season-specific high resolution disease risk estimates in Zambia.

By 2013, six African countries had completed two rounds of MIS: Angola, Liberia, Mozambique, Rwanda, Senegal, and Tanzania. In *Chapter 7*, a spatio-temporal analysis was performed to estimate changes in malaria parasitemia risk across these countries. Additionally, the coverage and effectiveness of control measures (i.e., ITN and IRS) was quantified at national and subnational level in reducing malaria risk, after taking into account climatic factors. The analysis was performed with a Bayesian geostatistical model and spatially varying coefficients to study disease/interventions associations. Bayesian

variable selection procedures were developed to select the most relevant ITN measure in reducing malaria risk and spatial kriging over the study area was performed to produce intervention coverage maps. For the first time, smooth maps of probability of decrease in parasitemia were produced.

The methods described throughout this thesis may not be applied directly from field practitioners or NMCP personnel, since they require specialized knowledge. However, we are currently working on the implementation of the models with entirely free softwares and user-friendly interfaces to be distributed to the NMCPs and facilitate their work in monitoring and evaluating the progress in the fight of the disease.

Zusammenfassung

Malaria ist eine durch Mücken übertragene, ansteckende Krankheit, welche durch parasitäre Protozoen der Gattung *Plasmodium* ausgelöst und durch den Biss infizierter weiblicher *Anopheles* Mücken auf Menschen übertragen wird. Trotz des Fortschritts im Kampf gegen die Krankheit während der letzten Jahrzehnte bleibt Malaria eine der Hauptursachen für Morbidität und Mortalität in groen Teilen der Entwicklungsländer, insbesondere in Subsahara-Afrika. Kinder im Alter von unter fünf Jahren sind am meisten gefährdet. Beobachtete Rückgänge gehen einher mit einem beträchtlichen Anstieg der internationalen finanziellen Mittel für Malaria-Prävention, -Kontrolle und Eliminierung, welcher wiederum zu einer enormen Verbreitung der Implementierung nationaler Malaria-Kontrollprogrammen (NMCPs) fhrte. Interventionen umfassen üblicherweise das Versprühen von Insektiziden in Innenräumen (IRS), die Verwendung von mit Insektiziden behandelten Netzen (ITN) und Umweltmanahmen wie zum Beispiel der Kontrolle der Larven. Im letzten Jahrzehnt wurden spezielle Ziele festgelegt. Das Millenium-Entwicklungsziel (MDG) 6 strebt an, die Malaria-Inzidenz bis 2015 zu halbieren (im Vergleich zu 1990) und eine universelle Verbreitung von ITN und die Behandlung mit geeigneten Malariamedikamenten zu realisieren. Im Jahre 2010 rief der Global Malaria Action Plan (GMAP), ins Leben gerufen durch die Roll Back Malaria (RBM) Partnership, zu einer raschen Intensivierung aus, um eine universelle Abdeckung und einer Form der Vektorkontrolle zu ermöglichen.

Die Übertragung von Malaria hängt sowohl von der Verteilung als auch von der Häufigkeit der Mücken ab. Jene reagieren empfindlich hinsichtlich der Umwelt- und Klimabedingungen wie beispielsweise Temperatur, Niederschlag, Vegetation und Bodennutzung. Mittels geostatistischer Modelle kann die Beziehung zwischen der Umgebung und der Krankheit an fixen Lokalisierungen über ein stetiges Untersuchungsgebiet hinweg geschätzt werden. Des Weiteren ermöglicht diese Methode die Prognose der durch Malaria bedingten Last an Orten, für welche keine Daten bezüglich der Übertragung verfügbar sind. Die Daten sind räumlich korreliert, da benachbarte Gebiete denselben Risikofaktoren ausgesetzt sind und

somit der Einfluss auf die Übertragung von Malaria ähnlich ist. Geostatistische Modelle berücksichtigen räumliche Korrelation durch die Einführung von ortsabhängigen Random-Effekten. Die Modellformulierung nach Bayes ist eine natürliche und praktische Wahl hinsichtlich der Modellanpassung durch Implementierung von Markov chain Monte Carlo (MCMC).

Diese Arbeit entwickelt neuartige statistische Methodologien (i) zur genauen Schätzung der Krankheitslast (Malaria Parasitämie-Risiko und Anzahl der infizierten Personen) bei einer hohen räumlichen Auflösung und (ii) zur Abschätzung der Abdeckung und Effektivität von Interventionen zur Vektorkontrolle. Erstellte Karten und Schätzungen stellen einen wesentlichen Beitrag zur Überwachung und Evaluierung des Fortschritts zur Erreichung des Ziels, dem Rückgang der Krankheit und dem Ausbau der Abdeckung der Interventionen, dar.

Aktuelle Informationen bezüglich der Malaria-Prävalenz wurden zum gröten Teil von Malaria Indicator Surveys (MIS) und Demographic Health Surveys (DHS) mit Malaria Modulen für diese Arbeit zur Verfügung gestellt. MIS sind national repräsentative Umfragen, welche von RBM entwickelt wurden und in dessen Rahmen Parasitämie-Daten über Kinder unter 5 Jahren gesammelt werden. Die Studien werden üblicherweise während Perioden durchgeführt, in denen eine hohe Malariaübertragungsrate gegeben ist. Historische Daten wurden von der Mapping Malaria Risk in Africa (MARA) Datenbank extrahiert. Diese Quelle umfasst Daten aus grauer oder veröffentlichter Literatur aus über 10,000 geographisch positionierter Erhebungen quer durch Subsahara-Afrika. Daten über bestätigte Malariafälle wurden vom Health Management Information System (HMIS) in Sambia bezogen.

In *Kapitel 2* wurden Bayes'sche geostatistische zero-inflated binomiale (ZIB) Modelle entwickelt, um Schätzungen für das räumlich-explizite Parasitämie-Risiko zu erstellen und die Anzahl der infizierten Kinder unter 5 Jahren in Senegal zu bestimmen. Eine Vielzahl der Erhebungsstandorte des senegalesischen MIS 2008 Datensatzes wiesen eine Prävalenz gleich null auf (70%), welche durch die geostatistischen ZIB Modelle berücksichtigt werden konnte. Modellvalidierung bestätigte, dass das ZIB Model eine bessere Vorhersagefähigkeit aufwies als das Standard-binomiale Gegenstück. Bayes'sche Methoden zur Variablenauswahl wurden in einen geostatistischen Rahmen integriert, um die optimale Auswahl an Umwelt- und Klimakovariaten zu finden, welche mit dem Parasitämie-Risiko assoziiert sind. Verschiedene Indikatoren hinsichtlich der Verbreitung der ITN wurden berechnet, um die Effektivität der Interventionen zu ermitteln.

Kapitel 3 untersucht unterschiedliche Modellspezifizierungen von zero-altered Modellen und legt eine Modellformulierung innerhalb eines geostatistischen Rahmens nahe. Insbesondere wird das Problem der Variablenauswahl und Bewertung des Bedarfs der Berücksichtigung einer räumlichen Struktur innerhalb der Modellierung der Mixing-Wahrscheinlichkeit und der nicht-ausgearteten Verteilung behandelt. Es werden spezielle A-priori Verteilungen für die Abschätzung des räumlichen Prozesses, welche auf nicht-null Random-Effekt-Varianzen basieren, vorgestellt und anhand simulierter Daten analysiert. Der entwickelte Ansatz wurde angewandt, um eine simultane Schätzung der Angemessenheit der Malariaübertragung und dem bedingten Risiko in Senegal zu ermöglichen. Das wiederbelebte Interesse an der Auslöschung von Malaria suggeriert, dass weitere spärliche Datensätze durch sowohl parasitologische als auch entomologische Erhebungen erstellt werden.

Der Einfluss der Umweltprädiktoren bezüglich des Malariarisikos wird üblicherweise als linearer Effekt modelliert, welcher als konstant (über das Studiengebiet hinweg) angenommen wird. Jedoch bedarf es möglicherweise flexibleren funktionalen Formen wie zum Beispiel stückweis linear oder Splines, um die nicht-lineare Beziehung zwischen Prädiktoren und Malariarisiko zu erfassen. Das Untersuchungsgebiet ist häufig gro und wird durch unterschiedliche Regionen abgedeckt (zum Beispiel ökologische Zonen), wodurch die Relation zwischen der Krankheit und dessen Risikofaktoren nicht konstant über das Gebiet hinweg ist. Des Weiteren ist es wahrscheinlich, dass die räumliche Korrelation nicht nur als eine Funktion des Abstands sondern auch hinsichtlich der geographischen Lage variiert. In *Kapitel 4* wurden Bayes'sche Methoden räumlicher Variablenauswahl entwickelt, die eine spike-and-slab A-priori Struktur definieren. Hierdurch wird die Auswahl verschiedener Prädiktoren und deren funktionaler Form in nicht-stationären geostatischen Modellen ermöglicht, welche ihre Anwendung in der Kartierung von Erhebungsdaten über Krankheiten finden. Penalisierte Spline Effekte wurden als gemischte Effektterme umparametrisiert und deren Auswahl basiert auf nicht-null Random-Effekt Varianzidentifizierung. Räumlich variierende Gewichte werden vorgestellt, um *smoothness* über unregelmäig geformte Regionen zu ermöglichen. Diese Methoden wurden auf die Daten des Mali DHS angewandt, um räumlich explizite Schätzungen der Krankheitslast in dem Land zu erhalten.

Es existieren wenige Studien, die Erhebungsdaten über Malaria mit Variablen bezüglich der Bodenfläche/-nutzung (LC), welche durch Fernerkundung (Remote Sensing (RS)) erstellt wurden, verknüpfen. *Kapitel 5* ermittelt den Effekt der räumlichen Auflösung der

RS-erstellten Umweltvariablen bezüglich der Schätzung des Malariarisikos in Mosambik. Es wird ein Näherungsma präsentiert, welches definiert, ob eine LC-Variable als Kovariate in ein geostatistisches Model eingefügt wird. Diese Methodik wird im Anschluss auf den Mali DHS Datensatz aus dem Jahre 2011 angewandt. Modellvalidierung wurde mittels frei verfügbaren RS-Quellen und LC-Oberflächen mit einer 5 m Auflösung durchgeführt. Jene Oberflächen wurden von MALAREO, einem Seventh Framework Programme (FP7) finanzierten Projekt, welches Teile von Mosambik im Jahre 2010-2012 umfasst, erstellt. Die Vorhersagefähigkeit wurde verglichen.

Sofern die Schätzung der Prävalenz auf Zusammenstellung historischer Daten basiert, sind die Erhebungen für gewöhnlich heterogen bezüglich der Periode und der ausgewählten Bevölkerung (Altersgruppen). *Kapitel 6* beschäftigt sich mit der alters- und zeitlich bedingten Heterogenität zwischen Erhebungen, indem eine allgemeine Formulierung erläutert wird, welche räumliche statistische Modelle und mathematische Übertragungsmodelle unter Berücksichtigung von Unsicherheit verbindet. Die beschriebene Methodik wird angewandt, um hoch aufgelöste Schätzungen des alters-/periodenspezifischen Krankheitsrisikos in Sambia zu ermitteln.

Bis zum Jahre 2013 hatten sechs afrikanische Länder zwei MIS-Durchgänge vervollständigt: Angola, Liberia, Mosambik, Ruanda, Senegal und Tansania. In *Kapitel 7* wurde eine raum-zeitliche Analyse zur Schätzung von Veränderungen im Parasitämie-Risiko (Malaria) in den genannten Ländern durchgefhrt. Zusätzlich wurde die Abdeckung und Effektivität der Kontrollmanahmen (zum Beispiel ITN und IRS) sowohl auf nationaler als auch auf subnationaler Ebene hinsichtlich der Verringerung des Malariarisikos quantifiziert. Hierzu wurden klimatische Faktoren berücksichtigt. Die Analyse wurde anhand eines Bayes'schen geostatistischen Models durchgeführt, welches räumlich variierende Koeffizienten beinhaltete, um die Assoziationen zwischen Krankheit und Interventionen zu untersuchen. Bayes'sche Verfahren zur Variablenauswahl wurden entwickelt, um den relevantesten Indikator in Bezug auf die ITN Abdeckung (hinsichtlich der Reduzierung des Malariarisikos) zu selektieren. Darüber hinaus wurde die räumliche Kriging-Methode über das gesamte Untersuchungsgebiet angewandt, um Karten der Interventionsverbreitung zu erstellen. Zum ersten Mal wurden smooth Karten über die Wahrscheinlichkeit des Rückgangs der Parasitämie erstellt.

Die in dieser Thesis beschriebenen Methoden können möglicherweise nicht direkt von Feldarbeitern oder NMCP Personal angewandt werden, da Fachwissen erforderlich ist. Momentan arbeiten wir an der Implementierung der Modelle innerhalb einer kostenlosen

Software und benutzerfreundlichen Oberfläche, welche unter den NMCPs verbreitet wird und somit die Arbeit im Bereich der Überwachung und Evaluierung des Fortschritts im Kampf gegen die Krankheit untersttzt und erleichtert wird.

Acknowledgements

I had the great opportunity to carry out this PhD thesis at the Swiss Tropical and Public Health Institute in the frame of the Swiss South Africa Joint Reasearch Programme (SSAJRP) under the supervision of PD Dr. Penelope Vounatsou. This work has been made possible thanks to the support of a number of people whom I would like to acknowledge.

My sincerest thanks are addressed to my supervisor Penelope for believing in my capacities and accepting me on this PhD programme. Throughout these years of collaboration I could profit enormously from her knowledge and experience, which she was always willing to share. I greatly valued her patience and flexibility, as well as her generosity and encouragement to attend international conferences: the Spatial Statistics conference in Ohio and MIM meeting in South Africa, among others.

I am also grateful to Prof. Jürg Utzinger for all his constructive inputs and valuable comments to one of the chapters of my PhD thesis. Special thanks to Prof. Marcel Tanner for making the Swiss Tropical and Public Health Institute such a vibrant work place and a great environment to carry out my research.

I would like to acknowledge Prof. Annibale Biggeri, who kindly agreed to act as co-referee for this thesis, and Dr. Doleres Catelan, who accepted to be the expert member of the committee.

I received help and support from the Senegalese collaborators at Cheikh Anta Diop University, in Dakar. Their contribution, particularly from Dr. Lassana Konate and Prof. Ousmane Faye, is highly appreciated. Taking part of the MALAREO project has been a great experience and I would like to thank all members of the team, especially Jonas Franke and Ides Bauwens.

I am very grateful to Laura and Dominic for their support and friendship since my first days in Basel, to Nakul for his encouragement and to Sandra for always being there. I am thankful to Frédérique and Verena, with whom I started this journey, for the good time

spent together in and outside of work, Alex for sharing summer fruits and always showing his Italian language skills, Eveline for her contagious laughter and Sara for her positiveness and encouragement. I thank (and miss!) the MTIMBA group Amek-Susan-Simon with whom I had the pleasure to share the office.

The present PhD thesis was fully funded by the Swiss South Africa Joint Reasearch Programme Grant No IZLSZ3 122926. I would like to aknowledge the financial support received from the Studienstiftung of the University of Basel for printing this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Malaria disease and burden

Malaria is a preventable and treatable mosquito-borne disease, whose main victims are children under five years of age in Africa. According to the latest World Health Organization (WHO) estimates, there were about 219 million cases of malaria in 2010 and an estimated 660 000 deaths. Africa is the most affected continent: about 90% of all malaria deaths occur there. The six highest burden countries in the WHO African region (in order of estimated number of cases) are: Nigeria, Democratic Republic of the Congo, United Republic of Tanzania, Uganda, Mozambique and Côte d'Ivoire. These six countries account for an estimated 103 million (47%) of malaria cases. In South East Asia, the second most affected region in the world, India has the highest malaria burden (with an estimated 24 million cases per year), followed by Indonesia and Myanmar (WHO, 2012).

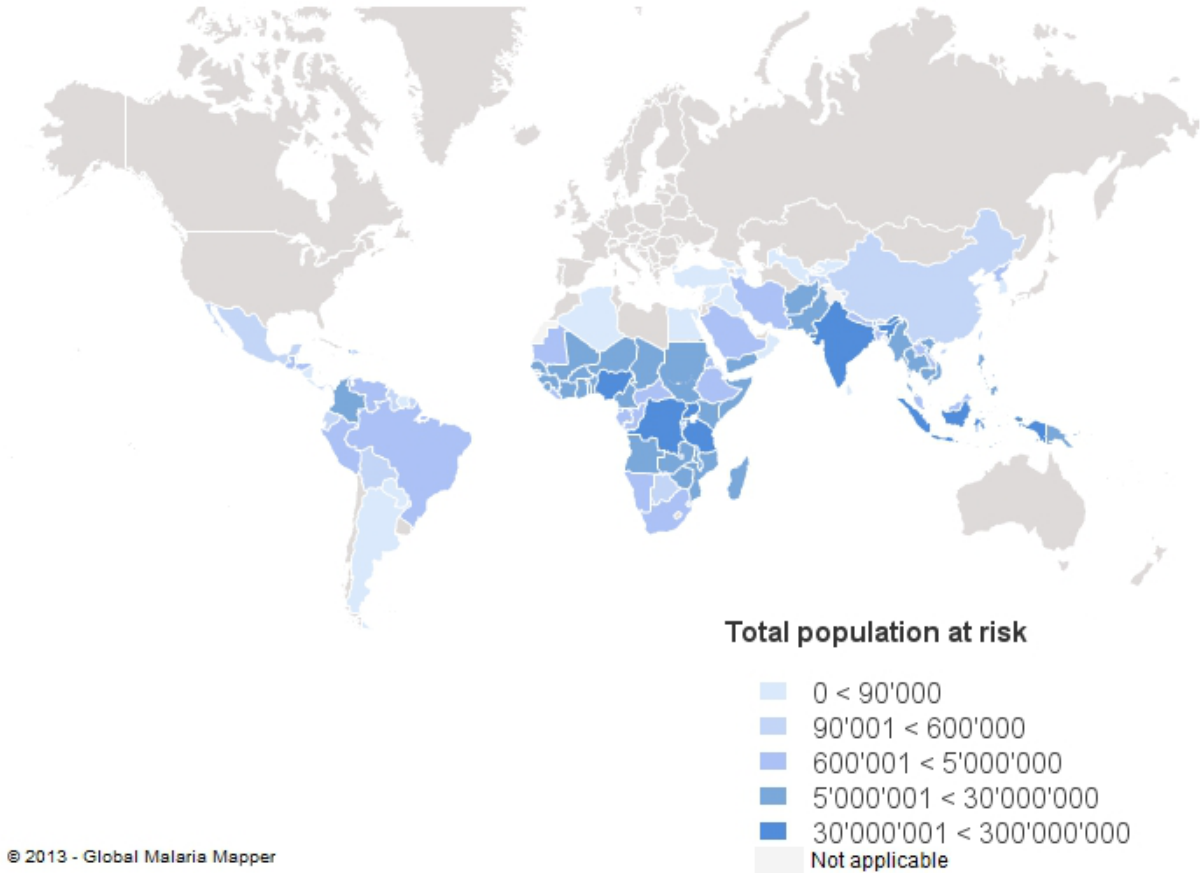


Figure 1.1: Global malaria distribution map: total population at risk. Map created with the Global Malaria Mapper. `http://www.worldmalariareport.org/`.

1.1.1 Malaria life cycle

The natural ecology of malaria involves malaria parasites, protozoa of the genus *Plasmodium* (*P. falciparum, P. vivax, P. ovale, and P. malariae*) and two types of hosts: humans and *Anopheles* mosquitoes. During a blood meal, a malaria-infected female *Anopheles* mosquito inoculates sporozoites into the human host. There, the parasites grow and multiply first in the liver cells and then in the red cells of the blood. In the blood, successive broods of parasites grow inside the red cells and destroy them, releasing daughter parasites (merozoites) that continue the cycle by invading other red cells. The blood stage parasites are those that cause the symptoms of malaria. When certain forms of blood stage parasites (gametocytes) are picked up by a female *Anopheles* mosquito during a blood meal, they start another, different cycle of growth and multiplication in the mosquito. After 10-18 days, the parasites are found (as sporozoites) in the mosquito's salivary glands. When the *Anopheles* mosquito takes a blood meal on another human, the sporozoites are injected with the mosquito's saliva and start another human infection when they parasitize the liver cells. Thus, the mosquito carries the disease from one human to another (acting as vector).

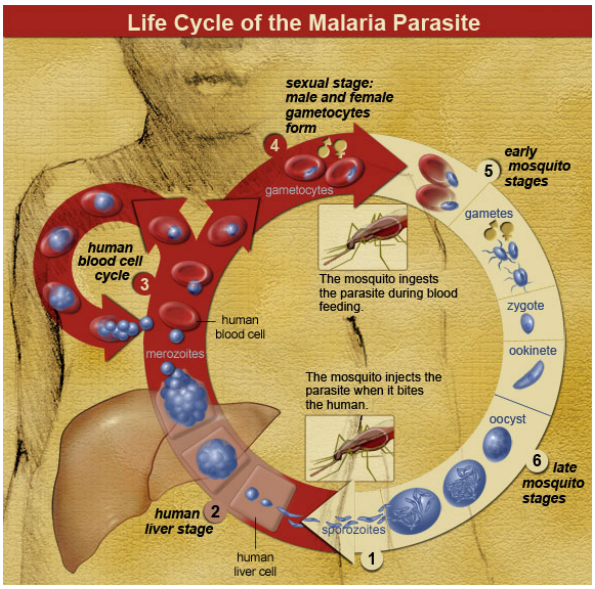


Figure 1.2: Life cycle of the malaria parasite. Source: `http://www.niaid.nih.gov`.

Differently from the human host, the mosquito vector does not suffer from the presence of the parasites. There are 430 *Anopheles* species, of which around 70 are malaria vectors, but only 40 of these are thought to be of major public health importance (Bruce-Chwatt et al., 1980). Among these, the *An. gambiae* complex and *An. funesus* are the primary malaria vectors in Africa. *An. gambiae s.s.* and *An. arabiensis* are the most widely distributed species of the *An. gambiae* complex in sub-Saharan Africa. Although these sibling species are morphologically indistinguishable, they exhibit different behavioral attributes. *An. gambiae s.s.* is predominant in humid areas, prefers feeding on humans (anthropophilic) and rests mainly indoors. On the other hand, *An. arabiensis* is more tolerant in the drier savanna regions, it often feeds on animals (zoophilic) and rests outdoors. Both species

breed in temporary habitats such as pools, puddles, rice fields. *An. funesus* prefers permanent water bodies with vegetation such as swamps and marshes, feeds both indoors and outdoors, mainly on humans and rests indoors.

1.1.2 Malariometric indices

In 1950s WHO suggested using the spleen rates (percent of children with enlarged spleen) as a proxy of malaria endemicity. Based on both the parasite and spleen rates, malaria endemicity has been classified as hypoendemic, mesoendemic, hyperendemic and holoendemic.

The parasite rate (i.e., prevalence) of human infections within a community is the most commonly used measure of malaria endemicity. Information on malaria prevalence is collected through community-based surveys by computing the percentage of individuals with malaria parasites, determined by Rapid Diagnstic test (RDT) or by analyzing thick or thin blood films on microscope slides. Malaria prevalence of the same population may vary in time, depending on the seasonality and stability of the disease.

Malaria incidence is a direct measure of the amount of malaria transmission because it represents the number of new malaria cases diagnosed during a given time interval in relation to the unit of population in which they occur. In some settings in sub-Saharan Africa is not possible to perform laboratory confirmation of malaria diagnoses, therefore incidence of fever is used as a proxy for incidence of malaria. However, the introduction of RDTs in health facilities as well as the ongoing commitment to strengthen the Health Management Information Systems (HMIS), has led to an improvement in the data quality and allowed availability of more reliable incidence data (Cibulskis et al., 2011).

The force of infection, i.e., the rate at which susceptible individuals become infected by malaria parasite, has long been proposed as an alternative measure of transmission, and different approaches to measuring it have been proposed, e.g. malaria parasite conversion rates in infants. With the wide acceptance of molecular approaches to malaria epidemiology, more precise measures can now be generated by genotyping individual parasite infections because natural superinfections can be monitored (Mueller et al., 2012).

Entomological inoculation rate (EIR) is the most used measure for assessing malaria transmission intensity. It represents the number of infective mosquito bites an individual is likely to be exposed to over a defined period of time, usually one year. EIR is expressed as the product of the anopheline mosquito density, the average number of mosquitoes biting each person in one day and the proportion of infective mosquitoes (sporozoite rate). The

product of the first two measures is known as human biting rate and is assessed using techniques like human bite catch, pyrethrum spray collection and light trap catch. The sporozoite rate is determined by dissection and examination of mosquito salivary glands or by the enzyme-linked immunosorbent assay (ELISA), a technique with high sensitivity and species specificity. Measurements of EIRs during longitudinal studies provides information on seasonal variations in transmission.

A measure of malaria transmission potential is the basic reproductive number R_0, defined as the average number of cases that a parasitemia infected individual is able to generate in an uninfected population.

1.1.3 Environmental determinants of malaria transmission

Malaria transmission is strongly influenced by climatic conditions which determine the abundance and seasonal dynamics of the *Anopheles* vector.

The amount and duration of malaria transmission is influenced by the ability of parasite and mosquito vector to co-exist long enough to enable transmission to occur. The distribution and abundance of the parasite and mosquitoes population are sensitive to environmental factors like temperature, rainfall, humidity, presence of water and vegetation.

Rainfall is one of the major factors influencing malaria transmission. It provides breeding sites for mosquitoes to lay their eggs, increasing the vector population and it increases humidity, improving mosquitoes survival rate. When humidity is below 60% the longevity of mosquitoes is drastically reduced. Mosquitoes are usually found in areas with annual average rainfall between 1100 mm and 7400 mm. However, excessive rain can have the opposite effect, by impeding the development of mosquito eggs or larvae, by flushing out many larvae and pupae out of the pools or by decreasing the temperatures, which can stop malaria transmission in areas at high altitudes.

Temperature plays an important role in the distribution of malaria transmission by influencing both the parasite and the vector. In particular, it has an effect on the survival of the parasite in the *Anopheles* mosquito. Optimum conditions for the extrinsic development of malaria parasite are between 25°C and 30°C, but as the temperature decreases, the number of days necessary to complete the extrinsic phase increases. At temperatures below 16°C the sporogonic cycle stops. For the vector, temperature affects the development rate of mosquito larvae and the survival rate of adult mosquitoes. Mosquitoes generally develop faster and feed earlier in their life cycle and at a higher frequency in warmer conditions. Development from egg to adult may occur in 7 days at 31°C, but takes about 20 days

at 20°C. In particular, water temperature influences larval development rates whereas air temperature determines adult longevity as well as the rate of parasite development within the adult mosquito (Garske et al., 2013).

Vegetation, as a result of rainfall and temperature, and the amount of green vegetation are important factors in determining mosquito abundance by providing feeding provisions and protection from climatic condition but also by affecting the presence or absence of the human hosts and therefore the availability of blood meals.

Land cover and land uses changes may influence the main determinants of the abundance and longevity of mosquitoes (Patz et al., 2005). Land cover concerns the physical material observed at the earth surface (natural factors such as forests, water bodies and bare rock) and land use is about anthropogenic elements (such as agriculture, irrigation, deforestation, urbanization and movements of populations) (Stefani et al., 2013).

1.1.4 Social determinants of malaria transmission

The relationship between poverty and malaria has long been recognized but its paths are multiple and complex. Recent studies suggest that causality works both ways, trapping communities in reinforcing cycles of poverty and disease (Sachs and Malaney, 2002). While malaria hits the poorest, those least able to afford preventative measures and medical treatment, simultaneously it affects the health and economic growth of nations and individuals. It has been estimated (WHO, 2009) that malaria is costing Africa about 12 billion a year in economic output, including direct and indirect costs as well as public expenditures. Malaria accounts for 3.3% of all disability-adjusted life years (DALYs) (Murray et al., 2013).

Maternal education plays an important role in malaria parasitemia in children (Siri and Lutz, 2012) and it is, in turn, intimately connected with economic conditions.

It has been reccomended (Tusting et al., 2013) to consider social and economical development as the main malaria control strategy. At the same time, malaria control should be seen as a poverty reduction strategy.

1.1.5 Control interventions and targets

The past decade has seen decreases in malaria in sub Saharan Africa. These reductions are going hand in hand with impressive increases in international funding for malaria prevention, control, and elimination, which have led to tremendous expansion in implementing national malaria control programs (NMCPs) (Alonso and Tanner, 2013).

Control measures are directed at each component involved in the malaria transmission cycle: the human host, the parasite and the mosquito vector.

Figure 1.3: MDGs. Source: `http://www.un.org/millenniumgoals`.

Vector control remains, in general, the most effective tool to prevent and control malaria transmission. The principal objective of vector control is to reduce malaria morbidity and mortality by reducing the levels of transmission. Common measures include indoor residual spraying (IRS), the use of insecticide treated nets (ITN) and environmental measures such as, in some specific settings, larval control. Applications of these techniques, alone or in combination, reduce human-mosquito contact, vector abundance and vector infectivity.

Other actions taken by the NMPCs concern the confirming malaria diagnosis through microscopy or RDTs for every suspected case; timely treatment with artemisinin-based combination therapies (ACTs) (O'Meara et al., 2010) and chemoprevention for the most vulnerable populations (pregnant women and infants). In particular, intermittent preventive treatment (IPT) with sulfadoxine-pyrimethamine for pregnant women living in high transmission areas and intermittent preventive treatment with sulfadoxine-pyrimethamine for infants living in high-transmission areas of Africa, alongside routine vaccinations. In 2012, WHO recommended Seasonal Malaria Chemoprevention (SMC) as an additional malaria prevention strategy for areas of the Sahel sub-Region of Africa. The strategy involves the administration of monthly courses of amodiaquine plus sulfadoxine-pyrimethamine to all children under 5 years of age during the high transmission season.

The Millennium Development Goal (MDG) 6 (UN, 2012a) aims to halve malaria incidence by 2015 as compared to 1990 and to achieve universal ITN coverage and treatment with appropriate antimalarial drugs. However, reducing malaria burden contributes significantly to the attainment of the MDG 4 target of reducing under-five mortality by two-thirds by 2015 and also to MDGs related to poverty reduction, education, and maternal health.

Renewed interest in malaria elimination and eradication has led to the definition of new targets in the last decade. In 2008, the Global Malaria Action Plan (GMAP), created by the Roll Back Malaria (RBM) Partnership (Roll Back Malaria, 2008), advocated reducing malaria cases by 75% (from 2000 levels) and malaria deaths to near zero, by 2015. Since 2007,

WHO has recommended universal coverage with ITNs (preferably long lasting insecticide-treated nets (LLINs)), rather than a pre-determined number of nets per household or exclusively targeting household members at high risk, i.e., pregnant women and children under five years of age (WHO, 2009). In 2010, the GMAP (Roll Back Malaria, 2008) called for rapid scaling-up to achieve universal coverage with some form of vector control.

Control interventions together with better case management and education have delivered positive results. Eleven African countries have reported a decrease of at least 50% in malaria cases between 2000 and 2009. By 2009, the annual number of malaria deaths had fallen by 20% in comparison with the beginning of the millennium. In 2010, Morocco and Turkmenistan were certified by WHO as having eliminated malaria. However, such strong pressure on vector and parasite populations has led to the selection and spread of resistant strains of mosquitoes and malaria parasites, respectively. Mosquito resistance to at least one of all four classes of insecticide available for malaria control has been identified in 64 countries around the world. Antimalarial drug resistance is a major concern for the global effort to control malaria: *P. falciparum* resistance to artemisinins has been detected in four countries in South East Asia and will probably spread globally. ACTs remain highly effective in almost all settings, as long as the partner drug in the combination is locally effective.

Figure 1.4: World War II poster. Source: U.S. National Archives and Records Administration.

1.2 Spatial epidemiology of malaria

All phases of control and monitoring activities starting form the planning, over to the implementation and coordination phase, and even evaluation of interventions, especially in resource-constrained settings, require an accurate geographic estimation of the disease risk. The study of the spatial epidemiology of malaria has made advances over the past years

and provided useful information for control programs. In this section, we review the data sources, tools and statistical modelling techniques for spatio-temporal analysis of malaria transmission indices.

1.2.1 Malaria data sources

The first database collecting information on malaria prevalence across sub-Saharan Africa compiled by the "Mapping Malaria Risk in Africa" (MARA/ARMA) project (Craig et al., 1999). The MARA database (`http://www.mara-database.org/`) contains malaria prevalence data collected over $10,000$ geographically positioned surveys from gray or published literature across the whole continent. The project was initiated in 1996 to provide comprehensive, empirical an accurate atlas of malaria risk for sub-Saharan Africa. However, maps from historical data may not reflect the current malaria situation at a given location, which could be influenced by control measures. On the other hand, historical data are useful for looking at temporal changes of the malaria situation. The malaria atlas project (MAP) (Guerra et al., 2007) is currently assembling historical as well as contemporary malaria data with corresponding geo-references. The major drawback of these type of data is their heterogeneity in season and age since they are collected during different seasons and include distinct (or sometimes not specified) age groups of the population. In addition, the data are sparse in time and space.

In order to provide standard and reliable information as well as to coordinate global efforts to fight malaria, RBM developed the Malaria Indicator Surveys (MIS). MIS are nationally representative surveys that collect both national and regional (or provincial) data from a representative sample of respondents. Surveys include measurement of malaria parasites and anemia among household members most at risk: children under five years and pregnant women; since 2000 they collect information about ITN ownership and use, IRS of insecticides, prompt and effective treatment of fever in young children, and the prevention of malaria in pregnant women. MIS are usually carried out during high malaria transmission seasons. These nationally representative household surveys represent the most precise benchmark of progress toward internationally agreed upon targets. Geo-reference is available at cluster (group of households of variable size) level. Sometimes MIS are conducted within Demographic and Health Surveys (DHS).

Malaria incidence data can be geo-referenced (collected at health facilities or residences), or aggregated over areal units (e.g., health districts). The quality and use of these data depend on the country surveillance system and presence of RDTs to confirm

suspected cases. Ongoing efforts to strengthen the HMIS have contributed to the increased availability and reliability of such data. Countries that have limited access to confirmation of cases, would report clinical malaria incidence, e.g. patients who are suspected to have malaria based on clinical signs and symptoms. In area with low malaria risk, like many countries in South-East Asia, such information may be not too biased since it is unlikely for people in the community to tolerate the parasite without being sick.

Entomological data provide direct measures of malaria transmission via estimates of EIR, sporozoite rates and other vector-related parameters. However, the data collection methods are not standardized, therefore the estimated transmission parameters could differ widely, depending on the techniques used. In regions with low malaria transmission the number of mosquitoes (infected mosquitoes) is very low, so the sampling error will be large. Because continuous collections of mosquitoes over a long period of time is difficult, the entomological data are usually derived from short/medium-term studies over small areas.

The most comprehensive database on entomological data was compiled by the Malaria Transmission Intensity and Mortality burden across Africa (MTIMBA) project, initiated in 2002 by the international network for the Demographic Evaluation of populations and Their Health in Developing countries (INDEPTH). The MTIMBA project provided reliable and standard entomological information that contributed to a better understanding of the links between malaria transmission intensity and mortality. MTIMBA was a multi-centre study that involved 18-malaria endemic sites in Africa collecting comprehensive disaggregated data at household or individual level on all-cause mortality and malaria transmission intensity (Kasasa et al., 2013; Amek et al., 2012).

1.2.2 Geographical information systems and remote sensing data

The study of malaria spatial epidemiology has benefited from the significant progress in the development of Geographic Information System (GIS), computerized systems capable of collecting, storing, handling, analyzing and displaying all forms of geographically referenced information, usually achieved by Global Positioning System (GPS).

Advances in earth observation (EO) via remote sensing (RS) technologies, have led to the development of high spatial resolution products. The growing availability of RS data, some of them accessible free of charge via the Internet, played a crucial role in determining the environmental predictors of malaria transmission (Ceccato et al., 2005).

RS data and spatial statistics have been used for mapping malariometric indices as

presence and persistence of vectors' breeding sites, larval densities, EIR as well as malaria prevalence, morbidity and mortality in the human (Machault et al., 2011).

The readily available up-to-date information on environmental variables pertinent to malaria transmission over large and remote regions makes RS a useful source of information for identification of pockets of transmission and epidemic early warning systems (EWS). RS can assist malaria control and elimination programs, through the development of spatial decision support systems enabling accurate and timely resource allocation (Clements et al., 2013).

1.2.3 Statistical models for spatial data

Exploratory spatial analysis such as variogram estimation (Cressie, 1993), Moran's I (Moran, 1950) and Gray's C (Geary, 1954) statistics, depending on the type of data, can be performed using cartographic representations in GIS.

Statistical models enable the identification of significant predictors of malaria transmission building an outcome-predictor relationship and can provide estimates of disease risk at unobserved locations. Locations in close proximity are characterized by similar infection risks due to shared spatial exposures. Unobserved spatially distributed exposures introduce spatial correlation to the data. Standard statistical modelling approaches are not appropriate for analysing spatially clustered data since they assume independence between locations. Accounting for spatial dependence can lead to better inference and predictions, and more accurate estimates of the variability of estimates. Spatial models introduce random effect parameters at each observed location or region to take into account potential spatial correlation.

When outcome data are geo-referenced, the more appropriate analysis is performed via geostatistical models (Diggle et al., 1998). An underlying spatial Gaussian process (GP) is assumed whose spatial covariance is commonly modelled as a function of distance between locations. Geostatistical models are often based on two common assumptions which are second order stationarity and isotropy. Second order stationarity implies that the mean of the process is constant and the covariance function depends on the spatial vector distance between two locations. When the covariance function only depends on the Euclidean distance between two locations, the process is called isotropic. The spatial covariance of a stationary and isotropic spatial process could be modeled using parametric functions of Euclidean distances. The Matérn covariance is the most commonly used family of parametric covariance functions. After fitting to data, geostatistical models are used for

spatial prediction (kriging) at unobserved locations.

When malaria data are aggregated over areal units, they usually consist of counts or rates. The focus of the analysis is to identify spatial patterns or trends and to assess association between malaria data and environmental factors that vary gradually over geographical regions. A Gaussian Markov random field (GMRF) is specified through full conditional distributions based on the Markov property in space. A GMRF introduces spatial associations in the model through the specification of neighborhoods based on the arrangement of the regions in a graphical representation. For example, two areas can be considered neighbors if they are within some specified distance of one another or they share a common boundary. The weights assigned to each neighborhood are determined in several ways. Common weight functions are binary functions with value 1 if two sites are neighbors and 0 otherwise, and scaled weights which are standardized by the row sum. The most popular choice is represented by conditional autoregressive (CAR) models (Besag et al., 1991).

Links and approximations between GP and GMRF have been showed in the work by Rue and Tjelmeland (2002).

1.2.4 Inference and software

Spatial models can be specified in a Bayesian framework by simply extending the concept of hierarchical structure, allowing to account for similarities based on the neighbourhood or on the distance, for area-level or point-reference data, respectively. Bayesian hierarchical models have become powerful methods in modeling spatial data due to development of simulation techniques like Markov chain Monte Carlo (MCMC) (Gelfand and Smith, 1990). These methods are employed to derive empirical approximation of the posterior distribution of parameters. Well-known methods include: Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), the Gibbs sampler algorithm (Gelfand and Smith, 1990) and reversible Jump MCMC (Green, 1995). MCMC methods, have become widespread for Bayesian computation thanks to the wide popularity of the BUGS software (Lunn et al., 2009) in its different releases of WinBUGS, OpenBUGS and JAGS Plummer (2003).

However, particularly in these cases, the main challenge in Bayesian statistics resides in the computational aspects. While extremely flexible and able to deal with virtually any type of data and model MCMC methods involve computationally and time-intensive simulations to obtain the posterior distribution for the parameters. Consequently, the complexity of the model and the database dimension often remain fundamental issues.

The Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)) approach has been recently developed as a computationally efficient alternative to MCMC. INLA is designed for latent Gaussian models, a very wide and flexible class of models ranging from (generalized) linear mixed to spatial and spatio-temporal models. For this reason, INLA can be successfully used in a great variety of applications also thanks to the availability of an R package named R-INLA (Martino and Rue, 2010).

Furthermore, INLA can be combined with the Stochastic Partial Differential Equation (SPDE) approach proposed by Lindgren et al. (2011) in order to implement spatial and spatio-temporal models for point-reference data.

1.2.5 Spatial modelling of malariometric indices

Lysenko and Semashko (1968) produced the first global malaria endemicity map combining data from historical documents and maps of several malariometric indices with expert opinion and simple climatic/geographical iso-lines. Malaria endemicity maps at national level were produced in the 50s (De Meillon, 1951; Nelson, 1959) although they made limited use of empirical evidence and they did not capture the spatial and temporal heterogeneity of malaria transmission.

The first in using a spatial statistical approach were the works by Kleinschmidt et al. (2000) and Kleinschmidt et al. (2001): malaria risk was mapped in Mali and West Africa, respectively, by fitting a standard regression model and applying classical kriging on the model residuals. A Bayesian statistical approach for the spatial epidemiology of malaria was used for the first time by Diggle et al. (2002) who fitted a geostatistical model on malaria survey data from The Gambia but did not provide a risk map. Gemperli et al. (2006a) and Gemperli et al. (2006b) developed Bayesian geostatistical models for mapping malaria risk in West Africa and Mali, respectively, using historical survey data extracted from the MARA database. They made use of the Garki transmission model to adjust for heterogeneous age groups. Sogoba et al. (2007) fitted Bayesian geostatistical models to identify the environmental determinants of the relative frequencies of *An. gambiae s.s.* and *An. arabiensis* mosquitoes species and to produce smooth maps of their spatial distribution in Mali. Gosoniu et al. (2006) and Gosoniu et al. (2009) developed Bayesian non-stationary models for malaria mapping in Mali and West Africa, respectively, using historical data extracted from the MARA database.

A global *Plasmodium falciparum* endemicity map depicting malaria levels in 2007 was produced by Hay et al. (2009) and the situation in 2010 was shown in Gething et al. (2011).

A global *Plasmodium vivax* endemicity map was presented for the first time in the work by Gething et al. (2012).Contemporary risk maps were produced on national levels using MIS and DHS with malaria module data in Zambia (Riedel et al., 2010), Angola (Gosoniu et al., 2010), Tanzania (Gosoniu et al., 2012) and Senegal (Giardina et al., 2012).

A Bayesian approach was adopted also to map malaria vector densities in a single village in Tanzania (Smith et al., 1995), malaria incidence rates in KwaZulu-Natal (Kleinschmidt and Sharp, 2001), South Africa, and to study malaria seasonality in Zimbabwe (Mabaso et al., 2005). Bayesian geostatistical models were developed to analyse data collected within the MTIMBA project: Amek et al. (2012) used zero-inflated models for the analysis of sparse malaria sporozoite rate data, (Kasasa et al., 2013) studied malaria transmission patterns in Navrongo and Rumisha (2013) modelled the seasonal and spatial variation of malaria transmission in relation to mortality.

1.2.6 Challenges in methodology

Several methodological issues arise from the spatial modelling of malaria prevalence data. Some of them are i) model formulation, variable selection and model choice for zero-inflated distributions, ii) analysis of non-stationary non-Gaussian geostatistical data, iii) modeling the non-linear effect of environmental/climatic factors on malaria risk, iv) dealing with misaligned data and v) spatially varying coefficients, vi) modeling prevalence from survey with heterogeneity factors, e.g. age and seasonality.

Sparse geostatistical data are likely to arise from parasitological surveys as well as entomological studies. The renewed interests in malaria elimination intensified malaria control activities and has led to a drastic decrease in the number of cases in some areas. This is mainly due to vector control strategies such as ITN and IRS. The factors leading to the onset/end of transmission in a specific area may differ from the ones causing an increase or decrease in malaria risk. Accurate spatially explicit estimates of transmission suitability as well as conditional number of infected represent an essential tool in the efforts towards elimination. Thus, each explanatory variable can have an effect on either or both (i) the probability of observing an (extra-) zero and (ii) the magnitude of the outcome. However, most studies employing zero-inflated models do not assess model specification and include either all or a specific subset of the potential explanatory variables in both equations. Furthermore, spatial dependence is commonly introduced via Gaussian processes but it is often ignored in the selection of explanatory variables, which can influence model formulation. Literature on variable selection methods for both zero-inflated and geostatistical

models is limited. To our knowledge, only Jochmann (2009) proposed a Stochastic Search Variable selection (SSVS) approach in zero-inflated count data and Scheel et al. (2013) defined Bayesian variable selection techniques in spatial Poisson hurdle models for areal data.

Most applications of geostatistical models assume that the spatial correlation is a function of the distance and independent of locations, that is, the spatial process is stationary. This hypothesis is not appropriate when malaria data are analyzed since local characteristics influence the spatial structure differently at various locations. A review of methods used for constructing non-stationary spatial processes can be found in Sampson (2010). These methods range ranging from spatial deformation models (Sampson and Guttorp, 1992) to spatial processes decomposition in terms of empirical orthogonal functions (Nychka et al., 2002) and process convolution models (Higdon, 1998). Smoothing and kernel-based methods (Fuentes, 2001) model non-stationarity as spatially weighted combinations of stationary spatial covariance functions. This approach was applied by Banerjee et al. (2004) to model house prices in California and by Gosoniu et al. (2009) to produce a smooth malaria risk map in West Africa. In the latter, the relation between climate factors and malaria risk was modelled separately in each ecological zone by penalized B-splines.

The relation between malaria transmission and climate is complex and often non-linear. However, in most applications, the impact of the predictors is modelled as a linear effect, constant throughout the study area. However, alternative functional forms, such as piecewise linear or splines may be more suitable to capture the relationships between the covariates and the response. Furthermore, the study area is often large and may contain different ecological zones which may influence the effect of the predictors on the disease outcome. In large areas the underlying spatial structure that models the geographical dependence among neighboring locations, may vary leading to non-stationarity. Therefore, a flexible model specification is required to enable choosing different predictors as well as different functional forms in each zone, while modelling a non-stationary spatial process.

Repeated MISs do not include always the same clusters. In the anaylsis of temporal trends the spatial misalignement between the different surveys has to be taken into account. Furthermore, when assessing the effectiveness of interventions, different endemicity levels can confound the relationship between parasitaemia and intervention coverages. Therefore, spatially varying coefficients modelling the impact of interventions may be required.

Malaria is seasonal and age dependent, therefore it is important when modeling survey data to account for seasonality and adjust for age. This task becomes challenging when

analyzing historical field survey data because they were collected in different seasons and at non-standardized and overlapping age groups of the population. Previous work on modeling heterogeneity in geo-referenced surveys for malaria mapping (Gemperli et al., 2006b; Gosoniu, 2008; Hay et al., 2009) were based on a 2-step procedure to (i) obtain age-correction factors and (ii) separately fit age-adjusted prevalence estimates in a geostatistical model, ignoring adjustment uncertainty. Moreover, the heterogeneity due to the different survey periods was not considered.

1.3 Objectives of the thesis

The overarching goals of this thesis are to develop methods for producing accurate disease burden estimation (malaria parasitemia risk and number of infected) at high spatial resolution and assessing the effectiveness of vector control interventions. The specific methodological objectives are:

i. Comparison of different model formulations and development of variable/random effect selection for zero-inflated data (Chapter 2 and Chapter 3);
ii. Modeling non-stationary non-gaussian geostatistical data (Chapter 4);
iii. Development of variable selection methods in non-stationary models that allow the choice between different functional forms (Chapter 4);
iv. Modelling malaria risk using RS-derived environmental covariates at very high resolution (Chapter 5);
v. Modelling age and season heterogeneity in the estimation of prevalence from historical survey data (Chapter 6);
vi. Development of geostatistical models for misaligned spatial data and spatially varying effects (Chapter 7).

The above mentioned methodological development were applied on MIS and DHS data, survey data extracted from the MARA database and confirmed incidence data provided by the HMIS to:

I. Produce smooth maps of malaria risk (Chapter 2) and suitability index (Chapter 3) in Senegal (MIS 2008) and select environmental predictors (Chapter 2 and Chapter 3) and intervention coverage indicators (Chapter 2);
II. Produce smooth maps of malaria risk in Mali (DHS 2010) (Chapter 4);

III. Select environmental predictors of malaria transmission and identifying their functional form in the three ecological zones in the analysis of the Malian DHS 2010 (Chapter 4);

IV. Produce spatially explicit estimates of risk and number of infected in Mozambique (DHS 2011) assessing the impact of very high resolution data (Chapter 5);

V. Produce age and seasonality adjusted malaria risk maps from heterogeneous malaria survey data (MARA and MIS 2006) in Zambia (Chapter 6);

VI. Estimate spatial and temporal trends, ITN and IRS coverage and effectiveness of interventions in six countries (two rounds of MISs in Angola, Liberia, Mozambique, Rwanda, Senegal and Tanzania) (Chapter 7).

Chapter 2

Estimating the burden of malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 data

Giardina F.[1,2], Gosoniu L.[1,2], Konate L.[3], Diouf M.B.[4], Perry R.[5], Gaye O.[6], Faye O.[3],Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Faculté des Sciences et Techniques, UCAD Dakar, Sénégal
[4] National Malaria Control Programme, Dakar, Sénégal
[5] Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America
[6] Faculté de Médecine, Pharmacie et Odontologie, UCAD Dakar, Sénégal

This paper has been published in *PLoS ONE* 7(3): e32625. (doi:10.1371/journal.pone.0032625).

Abstract

The Research Center for Human Development in Dakar (CRDH) with the technical assistance of ICF Macro and the National Malaria Control Programme (NMCP) conducted in 2008/2009 the Senegal Malaria Indicator Survey (SMIS), the first nationally representative household survey collecting parasitological data and malaria-related indicators. In this paper, we present spatially explicit parasitaemia risk estimates and number of infected children below 5 years. Geostatistical Zero-Inflated Binomial models (ZIB) were developed to take into account the large number of zero-prevalence survey locations (70%) in the data. Bayesian variable selection methods were incorporated within a geostatistical framework in order to choose the best set of environmental and climatic covariates associated with the parasitaemia risk. Model validation confirmed that the ZIB model had a better predictive ability than the standard Binomial analogue. Markov chain Monte Carlo (MCMC) methods were used for inference. Several insecticide treated nets (ITN) coverage indicators were calculated to assess the effectiveness of interventions. After adjusting for climatic and socio-economic factors, the presence of at least one ITN per every two household members and living in urban areas reduced the odds of parasitaemia by 86% and 81% respectively. Posterior estimates of the ORs related to the wealth index show a decreasing trend with the quintiles. Infection odds appear to be increasing with age. The population-adjusted prevalence ranges from 0.12% in Thillé-Boubacar to 13.1% in Dabo. Tambacounda has the highest population-adjusted predicted prevalence (8.08%) whereas the region with the highest estimated number of infected children under the age of 5 years is Kolda (13940). The contemporary map and estimates of malaria burden identify the priority areas for future control interventions and provide baseline information for monitoring and evaluation. Zero-Inflated formulations are more appropriate in modeling sparse geostatistical survey data, expected to arise more frequently as malaria research is focused on elimination.

2.1 Introduction

More than two hundred million cases of malaria were estimated worldwide in 2008 and the majority (85%) was in African countries. Malaria accounted for 850 thousand deaths in the same year, 89% of which occurred in Africa. Over 85% of deaths were in children under five years of age (WHO, 2009). Senegal is one of the 45 countries in Africa where malaria is endemic and represents the leading cause of morbidity and hospital mortality (WHO, 2008). The main parasite transmitted by anopheline mosquitoes is *Plasmodium falciparum* and transmission occurs seasonally in the entire country, from June to November. Rapid diagnostic tests (RDTs) have been provided free of charge since 2007. Two years later, almost 86% of suspected malarial fever cases were screened with an RDT (Thiam et al., 2011). Malaria incidence in children under five decreased from 400 000 suspected cases in 2006 to 30 000 confirmed cases in 2009 (Global Partnership to Roll Back Malaria, 2010). Routine surveillance provides some evidence that the number of malaria inpatient cases and deaths during the same period are decreasing. However, these estimates must be interpreted with caution since they are affected by poor reporting, introduction of RDTs as well as changes in case definition (WHO, 2009). Furthermore, the lack of nationally representative surveys makes these estimates unreliable. Almost all malaria surveys in Senegal were carried out in five parts of the country: Dakar and its suburbs, specific areas around the Senegal River, Fatick region and Niakhar province. Few studies have been conducted in the rest of the country, particularly in the regions of Tambacounda and Casamance.

The Senegal Malaria Indicator Survey (SMIS) is the second nationally representative household survey focusing on malaria-related indicators and the first that collected parasitological data. The survey was supported by the National Malaria Control Program (NMCP) and carried out between November 2008 and January 2009 by the Research Center for Human Development in Dakar (CRDH) with the technical assistance of ICF Macro and funding from the President's Malaria Initiative (PMI). Malaria control interventions have been implemented in the country recently. The SMIS collected information on interventions such as ownership and use of insecticide treated nets (ITNs) or long lasting impregnated nets (LLINs) as well as intermittent preventive treatment for pregnant women (IPTp). ITN coverage, measured by ownership of at least one mosquito net per household, reached 82% in 2010 (Global Partnership to Roll Back Malaria, 2010). In 2006, Artemisinin-based combination therapies (ACTs) were introduced and they were made freely available in 2010. However, indoor residual spraying (IRS) has not been implemented as a routine

intervention in the country. During the SMIS only three districts had introduced IRS as a mean of malaria control and therefore no related information was collected in the survey. The number of districts using IRS increased to six in 2010.

A national contemporary map of malaria distribution is an essential tool in order to prioritize control interventions in areas with higher burden and to achieve a better resource allocation and health management. Several maps presenting the distribution of malaria risk in Senegal have been generated over the last few years as part of mapping efforts covering larger areas. A West Africa malaria risk map (Gemperli et al., 2006a) was obtained using Bayesian geostatistical models on entomological inoculation rate estimates produced by applying the Garki transmission model (Dietz et al., 1974) on historical survey data from the MARA database (MARA/ARMA, 1998). An updated malaria risk map for West Africa was estimated using geostatistical models on MARA survey data considering a different effect of environmental factors on malaria depending on the ecological zones (Gosoniu et al., 2009). A Senegal malaria risk map was also embedded in a worldwide map based on historical survey data and geostatistical models (Hay et al., 2009). All these efforts made use of old and heterogeneous survey data, collected over different seasons, diagnostic tools and overlapping age groups across locations.

Common exposures such as environmental or climatic conditions as well as socio-economic status influence the transmission of malaria similarly in neighboring regions introducing spatial correlation. Geostatistical models including location-specific random effects were employed to model spatial correlation as a function of the distance between sampled locations. The data consisted of a large number of locations with zero prevalence; therefore the commonly used Binomial distribution may underestimate the zero-prevalence probability. Zero-Inflated Binomial (ZIB) models provide a flexible way to address this problem (Hall, 2000). ZIB models for prevalence data have not been applied before in the context of geostatistical modelling of infectious disease data. To our knowledge, the only application is in the modeling of sparse malaria entomological data (Amek et al., 2011). Zero-Inflated Poisson/Negative Binomial models have been formulated for geostatistical count data (i.e. mapping isopod nest burrows (Agarwal et al., 2002) and child HIV/TB mortality (Musenge et al., 2011)), however applications are rather limited.

In this paper, we provide spatially explicit burden estimates of malaria in Senegal using the SMIS data and Bayesian geostatistical Zero-Inflated Binomial models based on variable selection methods for spatial data.

2.2 Materials and Methods

2.2.1 Country Profile

Senegal is located in Western Africa, facing the North Atlantic Ocean between Guinea-Bissau and Mauritania. Its borders run south of the Casamance River and along the Senegal River respectively. The Gambia penetrates more than 320 km into the country, from the Atlantic coast to the centre along the Gambia River which bisects Senegal's territory. Northern Senegal is characterized by a Sahelian ecological zone with semiarid grasslands and acacia savannas. Malaria is unstable hypoendemic and immunity is acquired later in life. A Sudano-Sahelian zone in the centre of the country is dominated by a flat wooded savanna with very few prominent topographical features. Malaria is endemic in this area and immunity is acquired around the age of ten. The southern part of Senegal is occupied by a Sudano-Guinean ecological zone, with annual rainfall exceeding 800 mm. Malaria is hyperendemic and immunity is acquired in the first five years of life. The urban malaria burden is concentrated in the cities of Dakar, Rufisque, Kaolack and Saint-Louis where the anopheles vector density is very low. The high transmission season in Senegal occurs mainly between July and October. However, in the Senegal River delta area, there are two annual peaks of the disease caused by river flooding: one in the rainy and the other in the dry season.

2.2.2 Ethical statement

Participation in the survey was voluntary and written informed consent was obtained in the local language before questionnaire administration and blood collection for parasitaemia and anemia testing. Individuals were told about the general purpose of the survey, possible risks and benefits of the survey and those presenting malaria parasites and/or anaemia were treated. The survey protocol was submitted to and approved by the Ethical Review Committee at the National Malaria Control Program and the Institutional Review Board (IRB) of Macro International.

2.2.3 Malaria Data (SMIS 2008-2009)

A nationally representative random sample of 320 clusters and 9600 households was selected through a stratified two-stage sampling procedure. The clusters were the census units (CU) used by the National Agency for Statistics and Demography (ANSD) in the census carried out in 2002 (Recensement Genéral de la Population et de l'Habitat, RGPH-2002). However, in the three regions of Kaolack, Kolda and Saint-Louis, the health districts served

as sampling clusters. At the first sampling stage, 320 clusters were drawn with probability proportional to the number of households in each cluster. The sampling procedure was stratified by the area type (urban/rural) of the clusters: 67.5% of the selected ones were in rural areas and 32.5% in urban areas. At the second sampling stage 30 households were selected randomly from each cluster. Rural areas are slightly overrepresented due to over-sampling in the three regions of Kaolack, Kolda and Saint-Louis. Geographical information is available at cluster level. As part of the final sampling, one every third village was randomly selected and every child between 6 and 59 months of age was tested for parasitaemia. Two tests were performed, RDT and blood smear test (Centre de Recherche pour le Développement Humain (Sénégal) et ICF Macro, 2009). This study is based on the results of microscopic examination since thick blood smear test is considered as the gold standard (Wongsrichanalai et al., 2007).

2.2.4 Malaria predictors

Three sets of malaria predictors were considered in the study, namely environmental and climatic proxies, socio-economic factors and malaria intervention measures. The environmental and climatic variables were extracted from remote sensing sources. Decadal rainfall data were downloaded via the Africa Data Dissemination Service (ADDS). Weekly day/night land surface temperature (LST) and biweekly normalized difference vegetation index (NDVI) data were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS). Permanent rivers and lakes were extracted from Health Mapper. The shortest Euclidean distance between the centroid of each pixel and the closest water body was calculated in ArcGIS version 9.1 (ESRI; Redlands, CA, USA). Altitude data were obtained from an interpolated digital elevation model (DEM) developed by the U.S. Geological Survey - Earth Resources Observation and Science (USGS EROS) Data Center. The geographical distributions of the environmental factors are displayed in Figure 2.1. Data on the rural extents in Senegal are provided by the Global Rural-Urban Mapping Project (GRUMP). According to the UN definition for Senegal, agglomerations with more than 10 000 inhabitants were considered as urban (UN, 2006). The above data were available at 1km^2 spatial resolution, with the exception of rainfall which has a resolution of 8km^2.

Socioeconomic disparities were measured by a wealth index, included in the SMIS data and calculated by a weighted sum of household assets. The weights were estimated through principal components analysis (Rutstein and Johnson, 2011). ITN related information in the SMIS was used to calculate the following ITN coverage indicators (Thwing et al., 2011;

Kilian et al., 2010): i) a binary variable reporting whether the child has a bed net for sleeping; ii) the proportion of children under the age of 5 years reported to have slept under an ITN the night before the survey visit; iii) the total number of nets per household (irrespective of the number of household members); iv) a binary indicator representing the availability of at least one ITN per every two household members and v) at least one ITN per every two children under the age of 5 years in the household. Human population data estimates for the year 2010 were obtained from the Gridded Population of the World version 3 (GPWv3) database at $1km^2$ spatial resolution. These data were used to convert spatially explicit parasitaemia risk estimates into number of infected children under the age of 5 years. The total number of children under 5 years of age was obtained from the International Data Base of the U.S. Census Bureau, Population Division for the year 2010.

2.2.5 Bayesian geostatistical modeling

Let Y_i and N_i be the number of infected with malaria parasites and the number of screened children under the age of 5 years at location s_i (i.e. cluster centroid) respectively. Y_i is typically assumed to arise from a Binomial distribution, $Y_i \sim Bin(N_i, p_i)$ where p_i indicates the probability of parasitaemia at s_i. However, in the presence of excessive number of zeros, a Binomial model may be inadequate to estimate the zero-prevalence probability and to identify relevant covariates related to the outcome. To take into account the sparsity of the data, a Zero-Inflated Binomial (ZIB) model $Y_i \sim ZIB(N_i, p_i, \theta_i)$ was fitted and compared to the standard Binomial analogue. A ZIB model assumes two sources of zeros: $\theta_i\%$ (mixing probabilities) of the zeros are structural, not random and the remaining $(1-\theta_i)\%$ arise with a frequency defined by a Binomial distribution, see equation (2.1)

$$Y_i \mid p_i, \theta_i \sim \begin{cases} 0 & \text{with probability } \theta_i \\ Bin(N_i, p_i) & \text{with probability } (1-\theta_i) \end{cases} \tag{2.1}$$

In the above formulation, p_i does not have a direct interpretation of parasitaemia risk since it is influenced by the proportion of structural zeros.

The relation between p_i and the vector of k associated predictors $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ik})^T$ observed at location s_i is modeled via the equation $logit(p_i) = \mathbf{X}_i\boldsymbol{\beta} + \omega_i + \phi_i$, where $\boldsymbol{\beta}_i = (\beta_1, \beta_2, ..., \beta_k)$ is the regression coefficient vector, ω_i and ϕ_i are location-dependent random effects. Spatial dependence is introduced by assuming that the random effects $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_n)$ are distributed according to a MVN distribution with

mean 0 and covariance matrix $\mathbf{\Sigma}$ where each element σ_{ij} is defined by an exponential parametric function of the distance d_{ij} between two locations s_j and s_i, i.e. $\sigma_{ij} = \sigma^2_\phi exp(-\rho d_{ij})$. The parameter σ^2_ϕ represents the spatial variation and ρ is the parameter controlling the rate of correlation decay with increasing distance. In the case of exponential correlation function, $3/\rho$ can be used to calculate the distance above which spatial correlation is negligible, known as range. Any remaining non-spatial variation is estimated by the random effects ω_i, assumed independent and normally distributed with mean 0 and variance σ^2_ω.

Bayesian variable selection approaches were employed using the above geostatistical models to choose the best set of predictors. In particular, three variable selection methods, namely Gibbs variable selection (GVS) by Dellaportas et al. (2002), Stochastic Search Variable Selection (SSVS) by George and McCulloch (1993) and the variable selection sampler of Kuo and Mallick (1998) (KM) were compared. The best set of covariates was indicated by the model with the highest posterior probability. Details of the geostatistical variable selection methods are given in the Appendix.

The model includes over 330 parameters. To enable model fit and prediction a Bayesian formulation and MCMC estimation was adopted. To complete model specification, prior distributions were assigned to the parameters. An inverse-gamma prior was assumed for the variance and a gamma distribution for the spatial decay parameter ρ. The priors for the regression coefficients were non-informative Gaussian distributions with mean 0 and variance 100. Covariates were standardized in order to acquire better correlation properties and reduce MCMC computational time (Gelfand et al., 1995).

Bayesian kriging was employed to predict the parasitaemia risk at unsampled locations and produce a parasitaemia risk map at high spatial resolution (Diggle et al., 1998). A regular grid of 4 km^2 resolution covering the whole country was created, resulting in around 60 000 pixels. Predictions were based on a geostatistical model using only environmental/climatic factors since data on malaria interventions or socio-economic status are not available at high resolution scale for the whole country. Therefore, a two stage geostatistical variable selection approach was applied. In the first stage, only climatic predictors were included to identify the best prediction model. In the second stage, geostatistical variable selection was carried out to select among the five ITN coverage indicators defined above. The models were adjusted for age, wealth index and the climatic predictors determined during the first stage.

The predictive model was validated on a test subset of the data. In particular, a randomly selected sample of 269 locations (85% of the data locations) was used as a training

set for model fit. The predictive performance of the model was assessed by calculating the proportion of observed prevalence data at the remaining 15% of (test) locations, correctly estimated within Highest Posterior Density Intervals (HPDI) of probability coverage ranging from 50 to 100% (Gosoniu et al., 2006). The above validation procedure was also used to compare the ZIB model with its Binomial analogue. The number of malaria infected children under five years of age was estimated at pixel level by multiplying the geostatistical model-based risk estimates with the total number of children under the age of 5 years provided by the International Data Base of the U.S. Census Bureau, Population Division for the year 2010. The previous values were added to calculate the total infected children under the age of 5 years at district level. Subsequently, dividing by the number of children under the age of 5 years living in the district, population-adjusted estimates of parasitaemia risk were obtained.

Fortran 95 (Compaq Visual Fortran Professional 6.6.0) and standard numerical variables (NAG, The Numerical Algorithms Group Ltd.) were used to implement the MCMC code. OpenBUGS (Lunn et al., 2009) was also employed in the model fit.

2.3 Results

A total of 4138 children between 6 and 59 months of age from 320 clusters were tested for parasitaemia with both RDT and blood smear test. The overall observed malaria prevalence was 6.74%. The number of children under the age of 5 years tested with both Rapid Diagnostic Test and blood smear test was 3960. Almost 12.05% of the children under the age of 5 years tested with RDT were found positives. The percentage of children under the age of 5 years that were positives to both tests was 5.44%. Due to the observed discordance between the diagnostic tools, the standard microscopy test was considered in the analysis (Wongsrichanalai et al., 2007).

A large number of survey locations (around 70%) had zero prevalence. No children under the age of 5 years were tested in two clusters of Saint-Louis region and one cluster in Kaolack, thus reducing the actual number of GPS coordinates to 317. Figure 2.2 shows that the lowest malaria prevalence in the country was recorded in Saint-Louis (0%), followed by the regions of Dakar (1.72%) and Louga (1.43%).

Posterior model probabilities obtained from MCMC runs of 100 000 iterations using the GVS are presented in Table 2.1. Similar results were obtained with the other two variable selection methods, SSVS and KM. As shown in the table, the set of covariates that defined the Binomial as well as the ZIB geostatistical models with the highest posterior

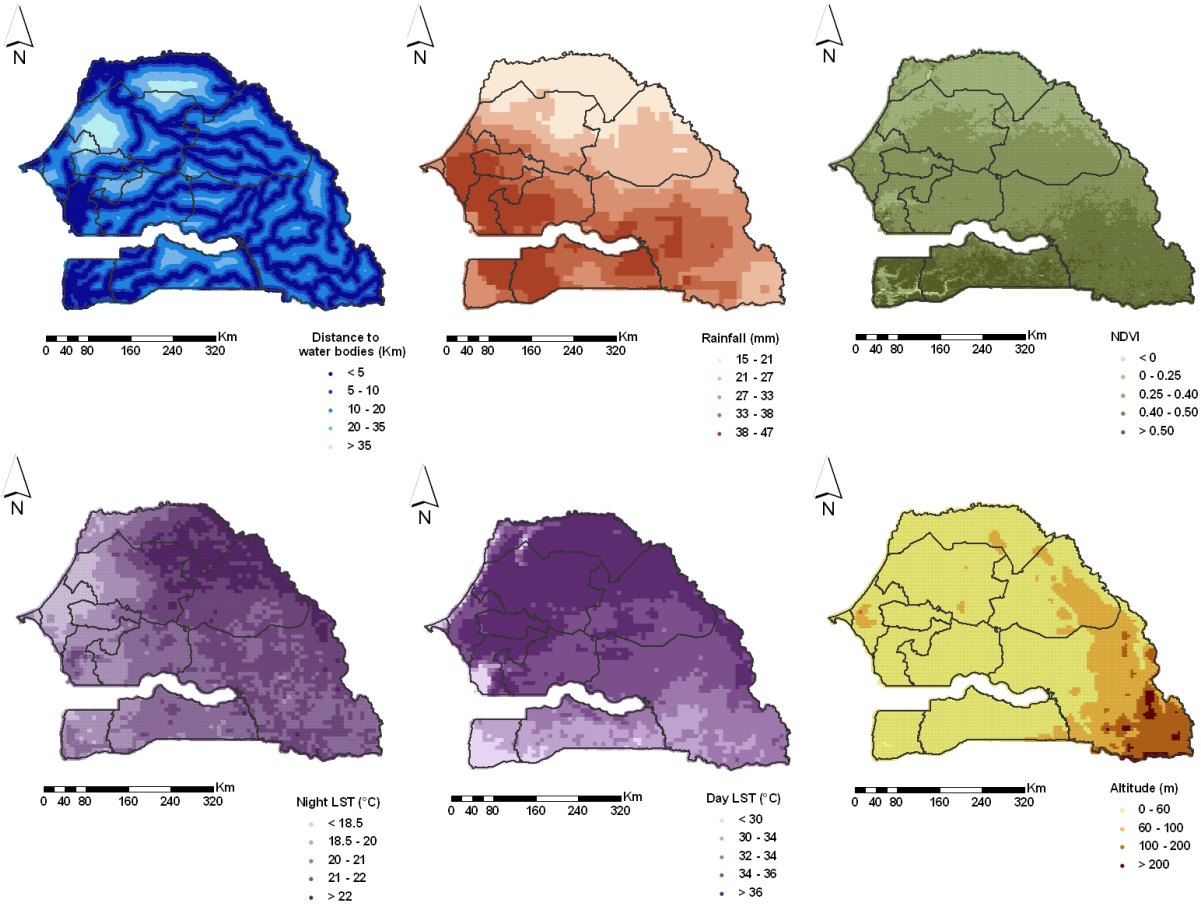


Figure 2.1: Environmental and climatic factors. Distance to water bodies, Rainfall, NDVI (Normalized Differenced Vegetation Index), Night and Day LST (Land Surface Temperature) and altitude at 4 km^2 resolution in Senegal. Regional boundaries are overlaid.

Table 2.1: Posterior model probabilities obtained using Gibbs Variable Selection (First stage). The shaded line indicates the selected model used to predict the malaria risk.

Model	Environmental variables	Binomial	ZIB
1.	Night LST, NDVI	2.46%	2.52%
2.	Night LST, NDVI, area type	72.21%	74.28%
3.	Night LST, Rainfall, NDVI, area type	12.13%	13.23%
4.	others	13.2%	9.97%

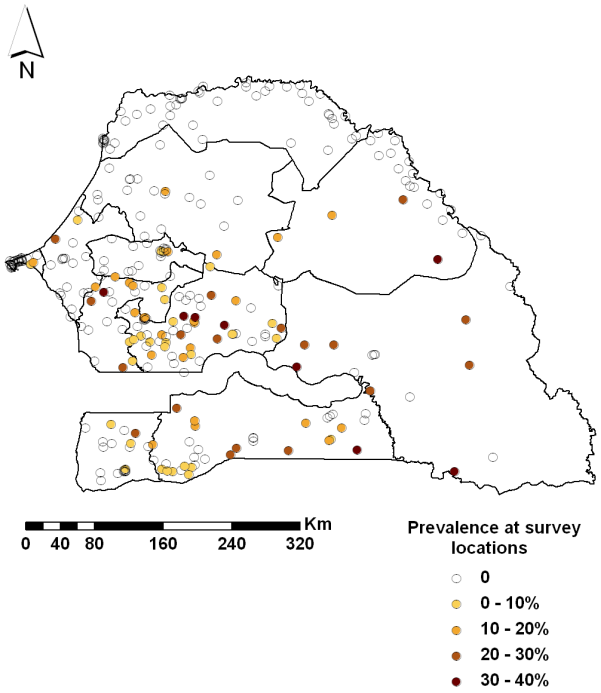


Figure 2.2: Prevalence at survey locations. Prevalence reported in the 317 locations of the SMIS 2008. Regional boundaries are overlaid.

probabilities consisted of night LST, NDVI and area type (urban/rural). The predictive performance of the selected models is shown in Figure 2.3. The proportion of test locations falling into the 50-95% HPDIs was constantly higher under the ZIB model. Furthermore, the latter model estimated narrower HPDIs. Based on the above results, the ZIB was adopted to predict the parasitaemia risk at high spatial resolution and to assess the effects of interventions on the infection risk.

Geostatistical ZIB model parameter estimates are given in Table 2.3. Model I includes only climatic covariates. The posterior estimate of the OR indicates a positive association between NDVI, night LST and parasitaemia, however the corresponding 95% credible intervals include one. Living in urban areas reduces the parasitaemia odds by 81% (95% BCI: 55%-93%). Raw data summaries estimate a parasitaemia prevalence of 1.3% in urban compared to the 8.47% in rural areas. The range parameter suggests that spatial correlation is present up to a distance of 2.4° which is equivalent to 265km (1°=111.12km). The spatial variance ($\sigma^2_\phi = 2.49$) was around 5 times higher than the non-spatial one ($\sigma^2_\omega = 0.52$) indicating high geographical variation. Model based predictions, obtained through Bayesian kriging over a grid of around 60 000 pixels of 2km x 2km spatial resolution are depicted in Figure 2.4. The plotted values correspond to the medians of the pixel-specific

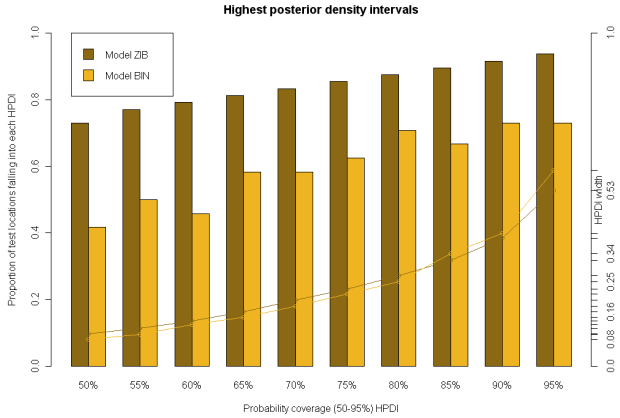


Figure 2.3: Model comparison and validation. Percentage of test locations with malaria prevalence falling in the highest posterior density intervals (HPDI) predicted from Binomial and Zero-Inflated Binomial models (bars). Lines indicate the corresponding HPDI length.

posterior predictive distributions. Low values of parasitaemia prevalence are concentrated in the northern Senegal, particularly in the region of Saint-Louis, Louga and Matam. Malaria risk increases in some areas of central Senegal and reaches the highest values in the southern Kolda and eastern Tambacounda where the predicted risk was 10.66% and 9.45%, respectively. Another high-risk area is located in the centre of Kaolak region with an estimated prevalence of 5.6%.

The predicted number of malaria infected children under the age of 5 years is displayed in Figure 2.5 and the estimates of population-adjusted prevalence obtained at the smallest administrative level (*arrondissement*) are summarized in Table 2.4. Kriging enabled the estimation of parasitaemia prevalence in areas where no survey locations were selected by the sampling procedure. For instance, the population-adjusted prevalence is 0.61% in the *arrondissement* of Barkedji, Louga region and 9.54% in Keniaba, Tambacounda region. The total number of infected children under the age of 5 years in the country below the age of five was estimated to be around 48 thousand. The map of the estimated number of children under the age of 5 years infected with malaria and the predicted parasitaemia prevalence show very different patterns, because of the population density, higher in the urban regions of Dakar and Saint-Louis.

Geostatistical variable selection among the five different ITN coverage indicators (Table 2.2) showed that having at least one available ITN per every two household members was most related with the parasitaemia risk after adjusting for climatic/environmental factors, age and wealth index. The posterior probability of the model was around 34% indicating

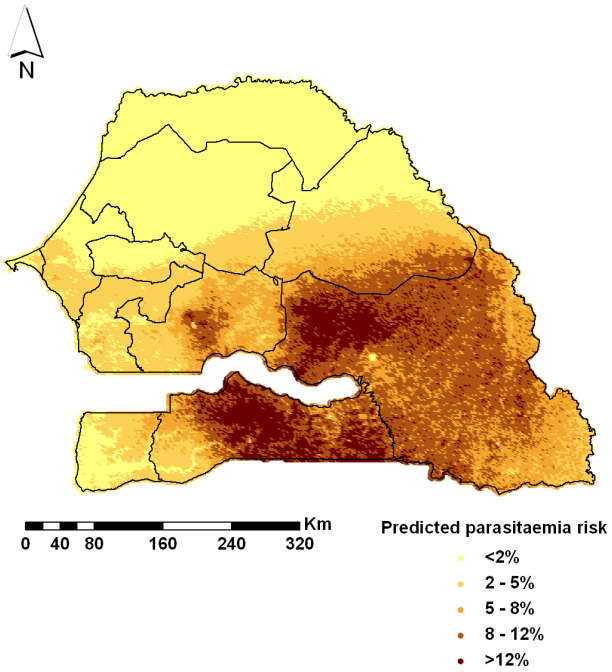


Figure 2.4: Predicted parasitaemia risk map. Predicted parasitaemia risk in children less than 5 years of age at 4 km^2 resolution in Senegal. Regional boundaries are overlaid.

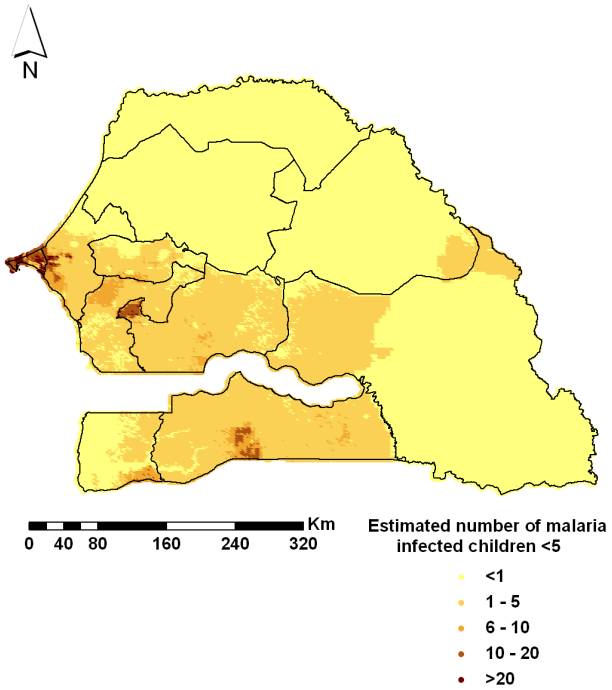


Figure 2.5: Estimated number of malaria infected children <5 years. The smooth map depicts the estimated number of malaria infected children less than 5 years of age at 4 km^2 resolution in Senegal. Regional boundaries are overlaid.

Table 2.2: Posterior model probabilities obtained using Gibbs Variable Selection (Second stage). The shaded line indicates the selected ITN coverage indicator.

Model	ITN coverage indicators	Posterior Probabilities
1.	None	25.20%
2.	Ownership of 1 ITN per 2 household members	34.00%
3.	Child has ITN for sleeping, ownership of 1 ITN per every 2 household members, n. of ITNs per household	7.80%
4.	others	33.0%

that the model was chosen 34% of the times among the $2^5 = 32$ possible models including all combinations of the five coverage indicators. Estimates of the posterior distribution of the parameters are given in Table 2.3 (Model II). Living in a household with at least one ITN per every two members was found to have a protective effect on parasitaemia, reducing the odds by 86% (95% BCI: 30%-97%). This result was also seen in the raw data summaries as shown by the second column of Table 2.3. The observed parasitaemia risk in the two categories, i.e. "less than one ITN" and "at least one ITN per every two members" was 6.84% and 1.41% respectively. Posterior estimates of the ORs related to the wealth index show a decreasing trend with the quintiles. The second quintile (very poor) had an OR of 0.77 (95% BCI: 0.57-1.03) whereas the last one (least poor) was 0.09, (95% BCI: 0.01-0.26). A similar pattern was presented in the prevalence calculated from the raw data. The highest (13.75%) and lowest (0.65%) infection risk were observed in the most and least poor group, respectively. Infection odds appear to be increasing with age. For instance, the OR is 1.2 (95% BCI: 0.70-2.43) in children 1-2 years old and 2.77, (95% BCI: 1.44-5.21) in children 4-5 years old. Observed parasitaemia prevalence was the lowest in infants (3%) reaching 8.11% in children 4-5 years old.

Table 2.3: Association of parasitaemia risk with environmental/climatic factors, socio-economic status and malaria interventions resulting from raw data summaries and geostatistical Zero-Inflated Binomial models.

Variable	Raw Data	Geostatistical model I[a]		Geostatistical model II[b]	
	Prevalence	OR	95% BCI[c]	OR	95% BCI[c]
Night LST		1.16	(0.66, 1.86)	0.83	(0.53, 1.26)
NDVI		1.48	(0.88, 2.48)	0.91	(0.61, 1.83)
Area type					
Rural	8.47%	1		1	
Urban	1.30%	0.19	(0.07, 0.45)	0.43	(0.16, 1.06)
Wealth Index[d]					
Most poor	13.75%			1	
Very poor	6.51%			0.77	(0.57, 1.03)
Poor	1.51%			0.22	(0.08, 0.51)
Less poor	0.96%			0.12	(0.05, 0.41)
Least poor	0.65%			0.09	(0.01, 0.26)
Age					
0-1	3%			1	
1-2	4.54%			1.20	(0.70, 2.43)
2-3	8.07%			2.93	(1.62, 5.33)
3-4	7.95%			2.96	(1.66, 5.74)
4-5	8.11%			2.77	(1.44, 5.21)
ITNs[e]					
< 1	6.84%			1	
≥ 1	1.41%			0.14	(0.03, 0.7)
Spatial param.		**Post. Median**	**95% CI[c]**	**Post. Median**	**95% CI[c]**
σ^2_ϕ		2.49	(1.07, 6.41)	3.04	(2.22, 4.02)
σ^2_ω		0.52	(0.25, 1.03)	0.35	(0.15, 0.73)
range[f]		2.40	(1.11, 2.98)	1.689	(0.003, 2.93)
Mix. prob.					
θ		0.29	(0.19, 0.39)	0.35	(0.24, 0.46)

[a]Model I includes only environmental/climatic factors. [b]Model II includes ITN coverage, children's age and wealth index. [c]Bayesian Credible intervals. [d]Household wealth index [e]Number of available ITNs per every two household members. [f]The range parameter (degrees), defined as $3/\rho$ indicates the distance above which the spatial correlation becomes negligible.

Table 2.4: Estimates of infected children less than 5 years old at district (Arrondissement) level. Data based on the old administrative division (Decret n° 2002-166).

Region	Department	Arrondissement	OP[a]	EIC[b]	PEP[c]
Dakar	Dakar	Parcelles Assainies	0%	90	0.05%
Dakar	Guédiawaye	Guédiawaye	0%	28	0.04%
Dakar	Pikine	Niayes	0%	89	0.05%
Dakar	Rufisque	Diamnadio	12.5%	68	0.29%
Dakar	Rufisque	Rufisque-Bargny	4.35%	18	0.07%
Diourbel	Bambey	Baba-Garage	0%	44	0.22%
Diourbel	Bambey	Lambaye	0%	59	0.29%
Diourbel	Diourbel	Ndindy	0%	129	0.35%
Diourbel	Diourbel	Ndoulo	0%	96	0.23%
Diourbel	Mbacké	Taif	15.38%	109	0.49%
Diourbel	Mbacké	Ndame	1.52%	89	0.11%
Fatick	Fatick	Diakhao	14.29%	105	0.74%
Fatick	Fatick	Niakhar	8.11%	159	0.69%
Fatick	Fatick	Tattaguine	7.55%	210	0.75%
Fatick	Foundiougne	Djilor	7.14%	32	0.49%
Fatick	Foundiougne	Colobane	6.98%	169	0.55%
Fatick	Gossas	Ouadiour	5.17%	135	0.75%
Kaolack	Kaffrine	Maka Yop	6.91%	560	1.38%
Kaolack	Kaffrine	Malem Hoddar	11.87%	435	1.54%
Kaolack	Kaolack	Sibassor	15.79%	302	1.27%
Kaolack	Kaolack	Ndiédieng	5.1%	113	0.75%
Kaolack	Kaolack	Koumbal	4%	221	0.35%
Kaolack	Nioro du Rip	Paoscoto	3.77%	316	0.87%
Kaolack	Kaffrine	Birkelane	8.2%	164	0.88%
Kolda	Kolda	Dabo	39.18%	803	3.3%
Kolda	Kolda	Médina Yoro Foula	27.17%	810	2.15%
Kolda	Sédhiou	Bounkiling	19.23%	408	1.71%
Kolda	Sédhiou	Diendé	3.49%	320	0.9%
Kolda	Sédhiou	Djibabouya	5.56%	115	1.31%
Kolda	Vélingara	Bonconto	5.45%	661	2.25%
Kolda	Vélingara	Kounkané	9.76%	368	2.62%
Louga	Kébémer	Ndande	0%	23	0.08%
Louga	Linguère	Barkedji	–	16	0.15%
Louga	Linguère	Dodji	4.92%	87	0.38%
Louga	Linguère	Yang Yang	0%	17	0.13%
Louga	Louga	Keur Momar Sarr	2.27%	18	0.08%
Louga	Louga	Sakal	0%	12	0.03%
Matam	Matam	Agnam-Civol	0%	94	0.39%
Matam	Matam	Ogo	3.09%	261	1.02%
Matam	Ranérou	Ranérou	4.65%	366	1.07%
Matam	Kanel	Sinthiou Bamambé	25%	133	1.27%
Saint-Louis	Dagana	Ross-Béthio	0%	30	0.07%
Saint-Louis	Podor	Gamadji Sarré	0%	18	0.08%
Saint-Louis	Podor	Thillé-Boubacar	0%	20	0.03%
Saint-Louis	Dagana	Mbane	0%	12	0.07%
Tambacounda	Bakel	Moudéry	27.08%	609	1.87%
Tambacounda	Bakel	Kéniaba	–	96	2.39%
Tambacounda	Bakel	Kidira	0%	51	1.78%
Tambacounda	Kédougou	Bandafassi	–	61	2.06%
Tambacounda	Kédougou	Salémata	–	43	1.72%
Tambacounda	Kédougou	Saraya	37.04%	52	1.59%
Tambacounda	Tambacounda	Koumpentoum	36.84%	866	2.62%
Tambacounda	Tambacounda	Koussanar	17.31%	200	1.03%
Tambacounda	Tambacounda	Makacoulibantang	0%	471	2.54%
Tambacounda	Tambacounda	Missirah	6.67%	106	1.43%
Thiès	Mbour	Ndiaganiao	–	56	0.62%
Thiès	Mbour	Sèsséne	5.08%	135	0.66%
Thiès	Mbour	Sindia	0%	171	0.46%
Thiès	Thiès	Keur Moussa	0%	328	0.33%
Thiès	Thiès	Thiénaba	–	30	0.28%
Thiès	Tivaouane	Méouane	1.59%	90	0.27%
Thiès	Tivaouane	Médina Dakhar	0%	39	0.25%
Thiès	Tivaouane	Niakhène	0%	72	0.23%
Thiès	Tivaouane	Pambal	27.78%	34	0.43%
Ziguinchor	Bignona	Sindian	6.25%	102	0.88%
Ziguinchor	Bignona	Tendouck	0%	52	0.55%
Ziguinchor	Bignona	Tenghory	3.33%	49	0.36%
Ziguinchor	Oussouye	Loudia-Ouoloff	0%	8	0.29%
Ziguinchor	Ziguinchor	Niaguis	–	6	0.34%
Ziguinchor	Ziguinchor	Niassia	1.25%	87	0.2%
Ziguinchor	Bignona	Diouloulou	6.25%	30	0.43%

[a]Observed Prevalence. [b]Estimated number of Infected children under 5 years of age.[c]Population-adjusted estimated prevalence.

2.4 Discussion

This study estimated the number of infected children under the age of 5 years at different geographical scales in Senegal and produced the first parasitaemia risk map in the country using contemporary data collected under the nationally representative malaria survey of 2008/2009. Geostatistical Zero-Inflated Binomial models were developed and Bayesian variable selection methods for spatially correlated data were employed to build a predictive model and assess the effectiveness of the ITN intervention adjusting for climatic and socio-economic confounders.

A large number of zeros was observed when modeling the number of infected children under the age of 5 years, probably due to the fact that the survey was carried out at the beginning of the dry season, when transmission starts to decrease. To address the issue of sparsity a ZIB model was derived. Model validation revealed that the ZIB model had higher predictive ability than the Binomial analogue suggesting that, when a large number of zeros occurs in the data, a ZIB model should be considered. Since malaria research is focused on elimination and eradication of the disease, it is expected that forthcoming surveys will include a large number of locations with zero prevalence and the ZIB models would provide a suitable alternative to the standard Binomial ones for geostatistical modeling.

Geostatistical variable selection is an important topic in malaria mapping. The predictive ability of a model depends on the covariates included in the multivariate regression setting. Modeling approaches in malaria mapping treat selection of predictors separately than the geostatistical model fit. Variable selection is often based on regression models that ignore spatial correlation, leading to wrong estimates of covariates effects and their significance. Geostatistical variable selection not only identifies the best set of predictors but builds parsimonious models with the best predictive ability (Gosoniu and Vounatsou, 2011). In addition, it can be used to avoid overfitting due to the inclusion of unnecessary predictors or random effects. In this work, we have employed three Bayesian variable selection methods within a geostatistical model formulation. The climatic model with the highest posterior probability selected by the three methods included the following combination of covariates: night LST, NDVI and area type. Altitude in Senegal presents very little variation throughout the country therefore it was not considered as a potential predictor of malaria transmission in the variable selection procedure.

As mentioned above, maps showing the distribution of malaria risk in Senegal can be found in Gemperli et al. (2006a); Gosoniu et al. (2009) and Hay et al. (2009) as part of efforts in mapping malaria risk at regional and continental level using historical data.

Nevertheless, compilations of historical data obtained from surveys, heterogeneous in the age groups involved and the seasons considered, require methods for standardizing risk estimates into a common scale for mapping purposes. Different statistical methods have been employed; the work by Gemperli and colleagues (Gemperli et al., 2006a), for instance, made use of the Garki transmission model to take into account the heterogeneity in the surveys. The model developed by Pull and Grab (1974) was instead employed by the MAP project (Hay et al., 2009), standardizing age-groups to produce a world map of Plasmodium falciparum malaria endemicity. The parasitaemia risk map presented in this paper, has been estimated from a contemporary survey and shows similar patterns to the one obtained from previous efforts (Gemperli et al., 2006a), especially in the Southern and Eastern part of Senegal, at the border with Mali where the risk is higher. However, Gemperli et al. (2006a) predicted a lower risk in the Central part of the country and higher in the urban areas of Dakar and Saint-Louis, as well as throughout the Sahelian region. In terms of absolute values, those results are uniformly higher than the current ones, due to the fact that the SMIS was carried out at the beginning of the low transmission season. The predicted pattern of malaria produced by the more recent work by Gosoniu et al. (2009) is more consistent with the map we generated, however the absolute values are still far from our estimates. The map of Senegal from the MAP project (Hay et al., 2009) does not show any relevant variations or geographical differences in the intensity of malaria risk throughout the country. For logistic reasons the survey took place at the start of the dry season, thus projections from our model are likely to underestimate the burden during the highest transmission season.

Furthermore, the differences between observed and population adjusted risk estimates are mainly due to low prevalence observed in highly populated areas. The urban area of Dakar, for example, is the most populated one, and the majority of surveys were carried out in that area although the parasitaemia risk is very low.

Geostatistical variable selection enabled the assessment of the effect on parastaemia risk of different ITN coverage indicators after taking into account climatic factors and socio-economic disparities. Recent work by Thwing et al. (2011) and Kilian et al. (2010) proposed a number of ITN coverage measures related to the ownership and use of nets at individual or household level. Five indicators have been assessed in the study and only one suggested a reduction in malaria risk with increasing coverage. This may explain the lack of relation between ITN coverage and malaria risk in similar analyses of MIS data. The Senegal data revealed that the presence of at least one ITN per every two household members reduced

the odds of parasitaemia by 86%. In a recent analysis in Tanzania (Gosoniu et al., 2012), ownership of at least one ITN was the only indicator assessed, showing no protective effect. On the other hand, the analysis of Zambia MIS 2005 (Riedel et al., 2010) measured ITN coverage by the ownership of at least one bednet per household and found a preventive effect on malaria risk. Gosoniu et al. (2010) reported a reduction in risk for areas having at least 0.2 ITNs per person, a measure similar to the one presented in this paper. Different indicators of ITN coverage were considered in a spatial analysis of the Liberia MIS data (Gosoniu and Vounatsou, 2011), however none of them was associated with a reduction in the infection risk. The model does not include some known risk factors for malaria such as maternal education, proximity to health services as this information was not readily available from the MIS data. It is however interesting to collect this data and include them in future MIS analyses aiming to assess ITN effects on parasitaemia.

This study found that the malaria risk in children less than five years old increases with age. Infants had the lowest risk. The risk rises especially after the age of two and levels off in older children. Similar results were observed in other low endemic settings.

All the results presented in the paper are based on the estimation of parasitaemia prevalence using the blood smear test. Malaria prevalence estimated using the RTDs was almost twice as high as the one based on the microscopy results. This confirms earlier findings suggesting that RDTs might present a large number of false positives when used in field conditions probably due to high temperatures during storage and transport as well as poor training on RDTs use.

In the model formulation, a linear relation between the parasitaemia odds and the environmental covariates was assumed. Geostatistical variable selection could be used to determine the best functional form that describes the above relation. Furthermore, a stationary geostatistical model was fitted assuming that spatial correlation depends only on the distance between locations irrespective of the locations themselves. This assumption may not be true when there are unobserved factors, such as health system performance, that vary across the country. The relation between climatic predictors and malaria may differ as well among the ecological zones.

Future control interventions can be planned and implemented by decision-makers according to the priority of the areas. A better resource allocation and health management can be achieved by monitoring the impact of prevention and control activities. The produced map and estimates generated in this study can be considered as baseline for comparisons with future national surveys to evaluate the effectiveness and progress of on-going

intervention programmes as well as the changes of the parasitaemia risk over space and time.

2.5 Appendix

Bayesian Geostatistical variable selection methods

Given a vector of potential regressors $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k)^T$, we aim at selecting the "best" subset $\mathbf{X}^* = (\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_q^*)^T$ $q < k$ to model the standard Binomial component p_i of the Zero-Inflated model. To this purpose, the geostatistical model was modified to let the MCMC scheme choose among the 2^k models: an auxiliary indicator variable g_i was introduced, where $g_i = 1$ indicates presence and $g_i = 0$ indicates absence of covariate j in the model. The prior that was used for the indicator g_j is $g_j \sim Bernoulli(1/2)$, i.e. the probability of inclusion in the model for each variable is 0.5.

The three different formulations of Bayesian Variable selection strategies in Geostatistical models implemented in the work are described and compared below.

Gibbs Variable Selection

The method relies on a linear predictor defined by the equation

$$logit(p_i) = \sum_{j=1}^{k} g_j \mathbf{X}_{ij} \beta_j + \omega_i + \phi_i \tag{2.2}$$

where g_j is the indicator defined in the previous paragraph. A mixture of independent normal distributions $\beta_j \sim N(0, \sigma_j^2) + (1 - g_j)N(\overline{\mu}_j, \tau_j^2)$, $j = 1, 2, ..., k$ was used as a prior for the coefficients where σ_j^2 is the prior variance when the $j\text{-}th$ term is included in the model and $\overline{\mu}_j$ and τ_j^2 re the mean and the variance respectively used when the $j\text{-}th$ term is not included in the model (pseudoprior).

Kuo & Mallick

The most straightforward method for variable selection has been proposed by Kuo and Mallick (1998). The method assumes that the indicators and the covariates effects are a priori independent, i.e. $f(\beta_j, g_j) = f(\beta_j)f(g_j)$, $j = 1, 2, ..., k$. It is easy to implement and requires only the specification of the prior distribution for the regression coefficients, usually assumed to be non informative Gaussian . The relation between the predictors and the outcome is given by equation (2.2), as for the Gibbs Variable Selection method.

Stochastic Search Variable Selection

The SSVS method is slightly different since the parameter vector retains its full dimension k of all potential covariates under all models. It assumes a prior distribution for the coefficients composed by a mixture of Normal distribution $\beta_j \sim N(0, \sigma_j^2) + (1 - g_j)N(0, l_j^{-2}\tau_j^2)$, $j = 1, 2, ..., k$ where l_j is specified in order to ensure that the coefficient β_j is close to 0 when $g_j = 0$, i.e. the *j-th* variable is not included in the model. In particular, the linear predictor is given by equation $logit(p_i) = \sum_{j=1}^{k} \mathbf{X}_{ij}\beta_j + \omega_i + \phi_i$.

For a comprehensive review of these methods, see O'Hara and Sillanpää (2009) while a simplified version for practitioners could be found in Gosoniu and Vounatsou (2011).

Chapter 3

Bayesian variable selection for geostatistical zero-inflated Binomial models in malaria epidemiology

Giardina F.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland

This manuscript has been submitted for publication.

Abstract

This paper explores different modelling specifications of zero-altered models and suggests model formulations in a geostatistical setting. In particular, the work addresses the problem of selecting variables and assessing the need of incorporating a spatial structure to be included in the modelling of the mixing probability and the non-degenerate distribution. Specific prior distributions for spatial process selection based on non-zero random effects variances are proposed and analyzed. The methods are illustrated on simulated and real data. The application uses data from the national malaria survey in Senegal which reported zero prevalence at over 70% of sampled locations. The median probability models are compared in terms of their predictive ability. The proposed approach allows the simultaneous estimation of suitability of malaria transmission and of conditional risk and it can be applied to other environmentally-driven diseases.

3.1 Introduction

Survey data with excess zeros arise frequently in many disciplines. A natural approach to model such data is to put an additional point mass at zero; the resulting zero-modified distributions include zero-inflated and hurdle models. Zero-inflated models (Lambert, 1992) are defined as two-component mixtures of a point mass at zero with a standard distribution allowing two types of zeros: "structural" that arise from the point mass at zero and "chance" zeros modelled by the standard distribution. Hurdle models combine a point mass at zero with a truncated distribution for the non zero values (Mullahy, 1986) treating zeros and non-zeros separately.

Increasingly, data collected in surveys are geo-referenced. This additional information allows the incorporation of spatial dependence in regression models and the study of relevant geographical patterns. Agarwal et al. (2002) introduced spatial random effects to model areal count data with excess zeros in a zero-inflated Poisson model. Rathbun and Fei (2006) proposed a zero-inflated Poisson model in which the excess zeros are generated by a spatial probit model. Fernandes et al. (2009) modeled zero-inflated spatio-temporal processes for both continuous and discrete responses. Finley et al. (2011) formulated a geostatistical hurdle model for continuous responses applied to forest variables and Recta et al. (2012) illustrated a hurdle geostatistical model for count data. Following a similar approach, Neelon et al. (2013) developed a hurdle Poisson model for areal count data. Amek et al. (2011), Musenge et al. (2011), Giardina et al. (2012) and Kasasa et al. (2013) showed applications of geostatistical zero-inflated Binomial and Poisson models to the epidemiology of HIV/AIDS, Tuberculosis and Malaria.

The probability of zeros (or excess zeros) is commonly modeled via an appropriate link function (e.g. logit or probit). A (generalized) linear model with a suitable link function is used to model the rest of the data. Thus, each explanatory variable can have an effect on either or both (i) the probability of observing an (extra-) zero and (ii) the magnitude of the outcome. However, most studies employing zero-inflated models do not assess model specification and include either all or a specific subset of the potential explanatory variables in both equations. Furthermore, spatial dependence is commonly introduced via Gaussian processes but it is often ignored in the selection of explanatory variables, which can influence model formulation.

Literature on variable selection methods for both zero-inflated and geostatistical models is limited. To our knowledge, only Jochmann (2009) proposed a Stochastic Search Variable selection (SSVS) approach in zero-inflated count data and Scheel et al. (2013) defined

Bayesian variable selection techniques in spatial Poisson hurdle models for areal data. Bayesian variable selection methods for geostatistical data with application to malaria and neglected tropical diseases are presented in Giardina et al. (2012), Chammartin et al. (2013a,b,c) and Giardina et al. (2013b).

In general, zero-inflated models for unbounded count data have been widely studied while less has been done on zero-inflated Binomial models.

The data that motivated this work are frequently observed in the field of malaria epidemiology: sparse geostatistical data are likely to arise from parasitological surveys as well as entomological studies. The renewed interests in malaria elimination intensified malaria control activities and has led to a drastic decrease in the number of cases in some areas. This is mainly due to vector control strategies such as Insecticide Treated Nets and Indoor Residual Spraying. Unfortunately, accurate measures of intervention coverages and use are not available (and sometimes not reliable). However, they are likely to be spatially structured due to the distribution process or socio-economic factors. Their impact on disease reduction or elimination varies also in space, due to the so-called "community effect", although difficult to quantify. Furthermore, the factors leading to the onset/end of transmission in a specific area may differ from the ones causing an increase or decrease in malaria risk. Identifying risk factors for a disease provides guidance for policy making and prevention programming. Accurate spatially explicit estimates of transmission suitability as well as conditional number of infected represent an essential tool in the efforts towards elimination.

In this paper, we explore different specifications of zero-altered models for geostatistical data and propose Bayesian variable selection methods to allow the choice of both fixed and random effects in modelling the probability of (extra-) zeros as well as the rest of the data. Model performance was assessed by evaluating the predictive ability of the median probability model (Barbieri and Berger, 2004). The proposed methodology is illustrated through simulated datasets and applied to the analysis of the National Malaria survey in Senegal of 2008 which showed no parasitaemia in over 70% of the observed locations (Giardina et al., 2012).

3.2 Methods

3.2.1 Geostatistical Zero-Inflated Binomial Model

Geo-referenced prevalence survey data $Y(s)$ are commonly modelled via a Binomial distribution with parameter $N(s)$ and $\theta(s)$ where $s = 1, \ldots, n$ indexes locations $s \in S \subset R^2$, $N(s)$ indicates the total number of observations and $\theta(s)$ the probability of success at location s . However, in some cases the observed data results in a larger number of zeros than expected under the Binomial distributional assumptions. Two alternative specifications of discrete mixture models have been proposed for zero-altered data: zero-inflated and hurdle models. The zero-inflated distribution is defined as a two-component mixture model combining a standard non degenerate distribution with a point mass at zero, i.e.:

$$Y(s)|p(s), \theta(s), N(s) \sim \begin{cases} 0 & \text{with probability } p(s) \\ \pi(y(s)|\theta(s), N(s)) & \text{with probability } 1 - p(s) \end{cases}$$

The likelihood can be written as:

$$f(y(s)|p(s), \theta(s), N(s)) = \begin{cases} p(s) + (1 - p(s))\pi(0|\theta(s), N(s)), & y(s) = 0 \\ (1 - p(s))\pi(k|\theta(s), N(s)), & y(s) = k, k = 1, \ldots, N(s) \end{cases} \tag{3.1}$$

where $\pi(\cdot)$ is the Binomial distribution and $p(\cdot)$, $0 \leq p(\cdot) \leq 1$ is the mixing probability. In practice, zero-inflated models allow two sources of zeros: one is the implicit Bernoulli trial associated to the parameter $p(s)$ and the other is through the Binomial distribution.

The hurdle Binomial model is a two-component mixture model consisting of a point mass at zero and a truncated Binomial for the nonzero observations:

$$f(y(s)|p^*(s), \theta(s)) = \begin{cases} p^*(s), & y(s) = 0 \\ (1 - p^*(s))\pi^*(k|\theta(s)), & y(s) = k, k = 1, \ldots, N(s) \end{cases} \tag{3.2}$$

where π^* denotes the truncated Binomial distribution (i.e. $P(y(s) = k|k \geq 0, \theta(s)) = \frac{\binom{N(s)}{k}\theta(s)^k(1-\theta(s))^{N(s)-k}}{1-(1-\theta(s))^{N(s)}}$). Hurdle models can always be re-written as zero-inflated models replacing $p^*(s) = p(s) + (1 - p(s))(1 - \theta(s))^{N(s)}$ in Equations (3.1) and (3.2).

In geostatistical models is common to introduce covariates and spatial random effects through appropriate link functions:

$$\begin{aligned} f(p(s)) &= \mathbf{Z}'(s)\boldsymbol{\alpha} + \omega_1(s) \\ g(\theta(s)) &= \mathbf{X}'(s)\boldsymbol{\beta} + \omega_2(s) \end{aligned} \tag{3.3}$$

where f and g are commonly taken as the *logit* link function, $\boldsymbol{X} = (X_i, \ldots, X_m)$ and $\boldsymbol{Z} = (Z_i, \ldots, Z_n)$ are collection of predictors linked to the outcome throught the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_m)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_n)$, respectively. In equation (3.3), $\Omega_1 = \{\omega_1(s)\}_{s \in S}$ is a stationary Gaussian spatial processes with mean 0 and variance-covariance matrix $\boldsymbol{\Sigma}_1 = \sigma_1^2 \mathbf{R}(||s - s'||; \rho_1, \nu)$ and $\mathbf{R}(\cdot)$ is a valid correlation function of the Euclidean distance $||s - s'||$ between sites s and s', smoothing parameter ν, and $\rho_1 > 0$ that controls the rate of correlation decay between observations as distance increases. The Matérn family describes most of the correlation functions used in geostatistical models:

$$\mathbf{R}(||s - s'||; \rho, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} (\rho||s - s'||)^{\nu} K_{\nu}(\rho||s - s'||) \tag{3.4}$$

where K_ν is the modified Bessel function.

The distribution of $\Omega_2 = \{\omega_2(s)\}_{s \in S}$ is defined conditionally on Ω_1, i.e., $\Omega_2|\Omega_1 \sim N(k\Omega_1, \boldsymbol{\Sigma}_2)$ where $\boldsymbol{\Sigma}_2$ can be written as $\sigma_2^2 \mathbf{R}(||s - s'||; \rho_2, \nu)$ with a different decay parameter ρ_2. The conditional speficification of spatial processes distributions as described above, corresponds to the definition of a linear model of coregionalization, as shown by Schmidt and Gelfand (2003). In fact, the bivariate process $\Omega = (\Omega_1, \Omega_2)$ can be written as $\Omega(s) = \boldsymbol{\Psi} v(s)$ where $\boldsymbol{\Psi} = \begin{pmatrix} \sigma_1 & 0 \\ k\sigma_1 & \sigma_2 \end{pmatrix}$ and $v_1(s)$ and $v_2(s)$ are mean 0 spatial processes with variance 1 and correlation functions $\mathbf{R}(||s - s'||; \rho_1, \nu)$ and $\mathbf{R}(||s - s'||; \rho_2, \nu)$, respectively. The resulting covariance structure for the bivariate process Ω is $\boldsymbol{\Sigma}_\Omega = \sum_{j=1}^{2} \mathbf{R}(||s - s'||; \rho_j, \nu) \otimes \boldsymbol{T}_j$ where $\boldsymbol{T}_j = \boldsymbol{\psi}_j \boldsymbol{\psi}_j'$ and $\boldsymbol{\psi}_j$ represents the j-th column vector of matrix $\boldsymbol{\Psi}$. Since $\boldsymbol{T} = \boldsymbol{\Psi}\boldsymbol{\Psi}'$ is a covariance matrix, a natural approach would be to assign Wishart prior to it. However, the equivalence of the conditional specification allows to work with the parametrization (k, σ_1, σ_2) with a Normal prior on k. Computational advantages as well as limitations of the conditional approach can be found in Gelfand et al. (2004b). Modelling a bivariate spatial process induces correlation between the Bernoulli and (truncated) Binomial components of the model. As shown by Neelon et al. (2013), addressing this source of correlation can improve inferences on model parameters.

A common approach in fitting the above models is to consider $\boldsymbol{X} = \boldsymbol{Z}$, i.e. to include the same set of covariates for both $\boldsymbol{p} = \{p(s)\}\ \forall s \in S$ and $\boldsymbol{\theta} = \{\theta(s)\}$ that may cause problems due to over-fitting of the model. Another approach is to assume a constant $p(\cdot)$ across the space, i.e. $p(s) = const\ \forall s \in S$ and perform a regression on $\boldsymbol{\theta}$ only. However, each explanatory variable can have an effect on either or both (i) the probability of observing a zero (or extra-zero) and (ii) the magnitude of the outcome. Therefore, we propose a Bayesian variable selection method that allows each regressor to be either included in or excluded from each of the two equations.

We assume a Normal mixture of inverse gamma distributions as priors for the SSVS scheme. We build the auxiliary variables γ_i and δ_i to indicate presence or absence of the covariate X_i in the first and second equation of (3.3) respectively and assign a Normal prior to the coefficients β_i and α_i, that is

$$\beta_i|\gamma_i, \tau^2 \sim N(0, \gamma_i\tau^2) \text{ and } \alpha_i|\delta_i, \tau^2 \sim N(0, \delta_i\tau^2) \tag{3.5}$$

where $\tau^2|a, b \sim \Gamma^{-1}(a_\tau, b_\tau)$. In particular, the indicators γ_i and δ_i are specified as follows:

$$\gamma_i|q \sim qI_1(\gamma_i) + (1-q)I_{v_0}(\gamma_i) \text{ and } \delta_i|q \sim qI_1(\delta_i) + (1-q)I_{v_0}(\delta_i)$$

where v_0 is some very small positive constant and the prior probability of inclusion for variable X_i is $q \sim Beta(a_q, b_q)$. This prior specification defines a continuous bimodal distribution on the hypervariance of β_i and α_i with a spike at v_0, that shrinks the coefficients that are not relevant for the model, and a right continuous tail (slab) to identify non-zero parameters. In particular, if $\gamma_i = 1$ the covariate effect β_i is estimated by assuming a Normal prior distribution with mean 0 and variance τ^2, otherwise $\gamma_i = v_0$ and β_i is shrunk towards 0, therefore the predictor X_i is not included in the second term of the regression.

The two spatial processes represent additional sources of heterogeneity in the data. We allow the model to choose the inclusion/exclusion of the geostatistical random effects by testing the associated variance components. In the proposed variable selection strategy, setting $\sigma_i = 0$ is equivalent to dropping the i-th spatial random effect from the model. Following the approach of Wagner and Duller (2012) we can apply data augmentation, and write $\boldsymbol{\Omega}_i = \pm\sqrt{\sigma_i^2}N(0, \mathbf{R}_i)$ and treat the standard deviation of the spatial process as a covariate effect assigning σ_i a spike and slab prior as in (3.5). The predictive distribution can be used to obtain the outcome Y at a set of unobserved locations $\mathbf{s}_0$ utilising posterior samples of the parameters β, α, Σ_Y (through $\sigma_1^2, \sigma_2^2, \rho_1, \rho_2$,k), e.g., in the case of the hurdle

model:

$$\int p(y(s_0)|\omega(s_0), x(s_0), z(s_0), \alpha, \beta)p(\omega(s_0)|\sigma_1^2, \rho_1, \sigma_2^2, \rho_2)p(\alpha, \beta, \sigma_1^2, \sigma_2^2, \rho_1, \rho_2|Y, T) \\ \mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}\sigma_1^2\mathrm{d}\sigma_2^2\mathrm{d}\rho_1\mathrm{d}\rho_2 \quad (3.6)$$

$$p(Y(\mathbf{s}_0) = k|\theta(\mathbf{s}_0), p(\mathbf{s}_0)) = \begin{cases} 0 & \text{if } T(\mathbf{s}_0) = 0 \\ \frac{\binom{N(\mathbf{s}_0)}{k}\theta(\mathbf{s}_0)^k(1-\theta(\mathbf{s}_0))^{(N(\mathbf{s}_0)-k)}}{1-(1-\theta(\mathbf{s}_0))^{N(\mathbf{s}_0)}} & \text{if } T(\mathbf{s}_0) = 1 \end{cases}$$

$$T(\mathbf{s}_0) \sim Bernoulli(p(\mathbf{s}_0))$$

$$\begin{aligned} p(\mathbf{s}_0) &= f^{-1}(\mathbf{Z}'(\mathbf{s}_0)\boldsymbol{\alpha} + \omega_1(\mathbf{s}_0)) \\ \theta(\mathbf{s}_0) &= g^{-1}(\mathbf{X}'(\mathbf{s}_0)\boldsymbol{\beta} + \omega_2(\mathbf{s}_0)) \end{aligned}$$

where $\omega_1(\mathbf{s}_0)$ and $\omega_2(\mathbf{s}_0)$ are the two components of the spatial bivariate process $\Omega(\cdot)$ evaluated in $\mathbf{s}_0$ that has distribution $p(\Omega(\mathbf{s}_0)) \sim N(C(\mathbf{s}_0, \mathbf{s})\Sigma_\Omega^{-1}\Omega, C(\mathbf{s}_0, \mathbf{s}_0) - C(\mathbf{s}_0, \mathbf{s})\Sigma_\Omega^{-1}$ $C(\mathbf{s}_0, \mathbf{s})')$ where $C(\mathbf{s}_0, \mathbf{s})$ and and $C(\mathbf{s}_0, \mathbf{s}_0)$ are the covariance matrices that consider the distance between observed and new locations and within new locations respectively.

3.3 Results

3.3.1 Simulation study

To evaluate the performance of the methods, we consider a series of synthetic datasets. We simulated a spatial zero-inflated Binomial process at 10000 pixels on a regular square grid $[0, 5]$x$[0, 5]$ to reproduce the large study areas that are common in many epidemiological applications. The process was generated via a zero-inflated likelihood (Equations (3.1) and (3.3)) with number of trials $N = 50$. With these assumptions, the probability of a "random" zero is very low, and p^* approaches p, (see Figure 3.3 in the Appendix for details).

The covariance functions of the underlying Gaussian processes were chosen to be exponential (parameter $\nu = 1/2$ in Equation (3.4)) with different values of variances and decay parameters as shown in Table 3.1. In the entire simulation study the two spatial processes were simulated independently, i.e. $k = 0$. For each model 20 different datasets were simulated under the two different scenarios (high or moderate zero inflation probability).

Table 3.1: Models and parameter values used to simulate the zero-inflated data. For each model 20 datasets were generated.

Model		Variance	Decay parameter	Frequency of zeros (high/moderate)
1.	Binomial	$\sigma_1^2 = 3$	$\rho_1 = 2$	0.74/0.37
	Zero-inflation	$\sigma_2^2 = 1$	$\rho_2 = 2$	
2.	Binomial	$\sigma_1^2 = 1$	$\rho_1 = 2$	0.72/0.42
	Zero-inflation	$\sigma_2^2 = 3$	$\rho_2 = 2$	
3.	Binomial	$\sigma_1^2 = 1$	$\rho_1 = 1.5$	0.73/0.36
	Zero-inflation	$\sigma_2^2 = 1$	$\rho_2 = 2.5$	
4.	Binomial	$\sigma_1^2 = 1$	$\rho_1 = 1.5$	0.71/0.41
	Zero-inflation	$\sigma_2^2 = 3$	$\rho_2 = 2.5$	

A set of 5 potential regressors were generated as independent standard Normal variates. Only three of them, namely X_1, X_2 and X_3 contributed in generating the Binomial regression part with coefficients β_1, β_2, and β_3, respectively. The zero inflation probability was simulated to depend on covariates X_3 and X_4 with coefficients β_3 and β_4, respectively. Since the outcome variable is spatially structured, in many applications it is likely that observed and/or unobserved predictors present spatial structure as well. For this reason, one of the covariates (X_3) was simulated from the same Gaussian process that generated the positive counts. In both regression components an intercept was included and fixed at $\beta_0 = 0.5$ and $\alpha_0 = -0.3$, respectively. Variable X_5 did not have an influence on any of the regression terms but it was included as a potential regressor in the variable selection procedure. The coefficients were chosen adequately to produce scenarios with moderate or high zero-inflation. Figure 3.1 shows a realization from Model I as defined in Table 3.1.

A subset of 200 data were randomly sampled after stratifying for the proportion of zeros/non-zeros and analyzed using zero-inflated and Hurdle models with variable selection. An independent Gamma prior distribution was assigned to each decay parameter ρ_i centered on a value corresponding to a spatial range of the half of the maximum inter-location distance. The hyperparemeters used for the variable selection priors were $v_0 = 0.00025$, $a_\tau = 5$ and $b_\tau = 25$, $a_q = b_q = 1$.

The models were run in JAGS (Plummer, 2003) for 100000 iterations with a burn-in of 1000. Convergence was monitored via examining trace plots as well as the Geweke diagnostic implemented in the package CODA (Plummer et al., 2006). In general, hurdle models took longer to reach convergence. Results of the variable selection are shown in Tables 3.2 and 3.3 in the case of "moderate" zero-inflation together with the true coefficients values.

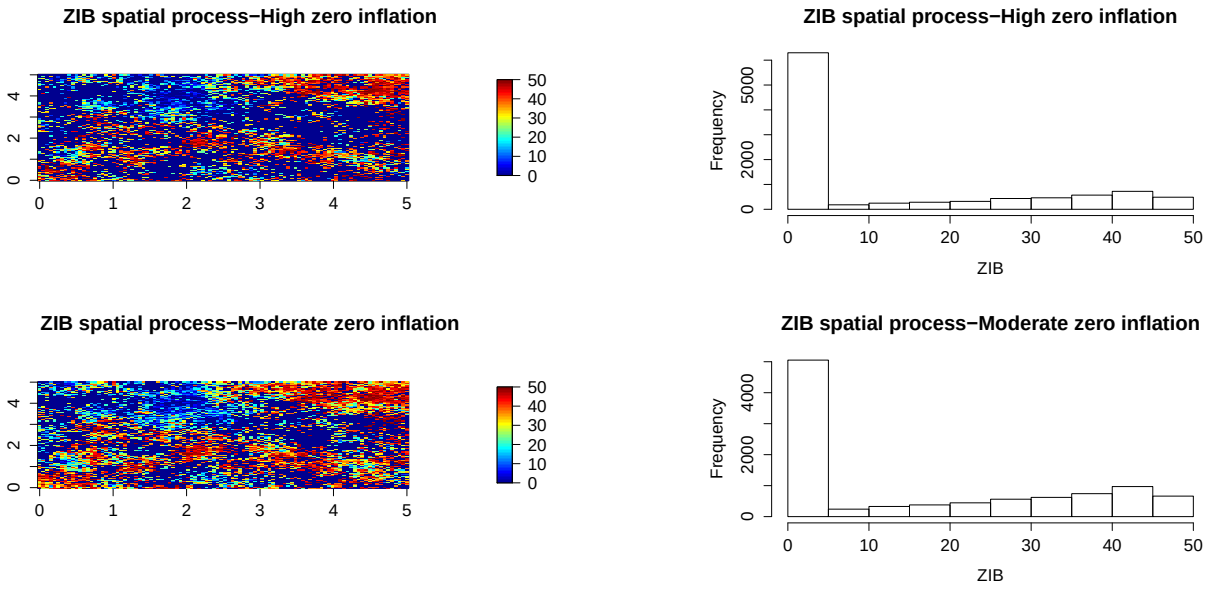


Figure 3.1: Two realizations of the spatial zero-inflated Binomial process (Model 1).

Results concerning variable selection for high zero-inflation are shown in the Appendix. Tables 3.2 and 3.3 report the average probability of selection over the 20 datasets for each of the predictor in the Binomial as well as in the zeros part under the zero-inflated and Hurdle models respectively.

Predictors in the Binomial part were in general better identified by the Hurdle model while in the zero (or extra-zero) part the probability of inclusion of the true covariates was only slightly higher. The zero-inflated model performed weakly in identifying covariates in common between the Binomial process and the extra-zeros (see X_3 in Table 3.2 and 3.3). The presence of a spatial residual structure was almost always well identified by the Hurdle model, while sometimes missed by the zero-inflated one. A big variance in the generated Gaussian spatial process modeling the residuals ($\sigma^2 = 3$) frequently resulted in lower selection probabilities for the predictors to be included in the mean structure. Zero-inflated models performed slightly better in the presence of lower percentage of zeros. For small effects (i.e. $\alpha_4 = 0.2$, Tables in the Appendix) the inclusion probability was on average smaller.

Although the zero-inflated Binomial model was not always able to identify the true model generating the data, we were interested in quantifying the effect of model "misspecification" on the ability of predicting counts at new locations. In particular, we randomly selected 50 locations (test set) from the original process generating the data and we

Table 3.2: Posterior inclusion probabilities of predictors and spatial processes estimated from the zero-inflated Binomial model. Estimates are averaged over 20 datasets

Variable (Bimomial)	Model 1 mean (sd)	Model 2 mean (sd)	Model 3 mean (sd)	Model 4 mean (sd)
$X_1 \quad (\beta_1 = 1.2)$	0.62(0.29)	0.72(0.23)	0.95(0.15)	0.73(0.21)
$X_2 \quad (\beta_2 = -2)$	0.58(0.23)	0.61(0.15)	0.63(0.15)	0.62(0.15)
$X_3 \quad (\beta_3 = 0.8)$	0.49(0.21)	0.48(0.22)	0.66(0.15)	0.46(0.17)
X_4	0.03(0.07)	0.12(0.09)	0.22(0.10)	0.15(0.09)
X_5	0.78(0.19)	0.55(0.21)	0.32(0.31)	0.59(0.23)
Ω_1	1.00(0.00)	0.88(0.21)	0.95(0.18)	0.89(0.20)
Variable (Zeros)				
X_1	0.12(0.18)	0.10(0.11)	0.09(0.05)	0.11(0.12)
X_2	0.22(0.21)	0.18(0.15)	0.22(0.20)	0.19(0.15)
$X_3 \quad (\alpha_3 = -0.8)$	0.41 (0.35)	0.55(0.32)	0.75(0.24)	0.51(0.21)
$X_4 \quad (\alpha_4 = 1.5)$	0.56 (0.21)	0.61(0.23)	0.93(0.11)	0.59(0.28)
X_5	0.19(0.12)	0.15(0.11)	0.17(0.09)	0.15(0.11)
Ω_2	0.32(0.29)	0.52(0.27)	0.41(0.33)	0.53(0.34)

Table 3.3: Posterior inclusion probabilities of predictors and spatial processes estimated from the Hurdle Binomial model. Estimates are averaged over 20 datasets.

Variable (Trunc.Binomial)	Model 1 mean (sd)	Model 2 mean (sd)	Model 3 mean (sd)	Model 4 mean (sd)
$X_1 \quad (\beta_1 = 1.2)$	0.98 (0.10)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
$X_2 \quad (\beta_2 = -2)$	0.92(0.12)	0.96(0.13)	0.92(0.10)	0.96(0.11)
$X_3 \quad (\beta_3 = 0.8)$	0.84(0.15)	0.89(0.15)	0.90(0.09)	0.90(0.10)
X_4	0.10(0.12)	0.07(0.04)	0.05(0.03)	0.07(0.03)
X_5	0.11(0.10)	0.12(0.10)	0.18(0.12)	0.13(0.11)
Ω_1	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Variable (Zeros)				
X_1	0.12(0.09)	0.10(0.10)	0.07(0.05)	0.09(0.04)
X_2	0.11(0.10)	0.12(0.09)	0.12(0.08)	0.11(0.09)
$X_3 \quad (\alpha_3 = -0.8)$	0.55(0.21)	0.61(0.19)	0.79(0.18)	0.62(0.21)
$X_4 \quad (\alpha_4 = 0.2)$	0.51(0.21)	0.54(0.32)	0.63(0.33)	0.54(0.31)
X_5	0.12(0.10)	0.13(0.08)	0.09(0.07)	0.15(0.12)
Ω_2	0.98(0.09)	1.00(0.00)	0.99(0.09)	1.00(0.00)

simulated from the predictive distribution of the median probability model. The latter is defined as the model that includes variables with estimated posterior selection probabilities higher than 0.5. The median probability model has been shown to have the best predictive ability by Barbieri and Berger (2004). Results of model predictive accuracy are expressed in terms of expected log predictive density (Gelman et al., 2013) and shown in Table 3.4

Table 3.4: Predictive ability: Hurdle vs zero-inflated Binomial model.

Model	Predictive log-score			
	Model 1 mean (sd)	Model 2 mean (sd)	Model 3 mean (sd)	Model 4 mean (sd)
Hurdle model	-225.1(34.00)	-227.1(35.23)	-226.1(34.15)	-228.2(35.68)
Zero-inflated model	-305.1(51.24)	-295.3(52.96)	-299.3(59.24)	-305.2(52.26)

which reports the mean and standard deviation over the 20 selected median probability models. The hurdle models performed better on average in all the scenarios.

3.3.2 Application to Senegal National Malaria Survey 2008

The methods described above were applied to the spatial analysis of the National Malaria Survey carried out in Senegal in 2008 which included a total of 9600 randomly selected households over 320 locations. Geographical information is available at cluster (set of households) level. Children between 6 and 59 months of age were tested for malaria. A large number of survey locations (70%) reported zero-prevalence.

Malaria is known to be an environmentally driven disease because the life cycle of the main vector (*Anopheles* mosquitoes species) is highly dependent on factors like the amount of precipitation, the distance to the water bodies among others.

The environmental/climatic variables used as potential explanatory variables were extracted from remote sensing sources. Dekadal rainfall data were downloaded from the Africa Data Dissemination Service; weekly day/night land surface temperature (LST) and biweekly normalized difference vegetation index (NDVI) data were obtained from Moderate Resolution Imaging Spectroradiometer. Permanent rivers and lakes were extracted from ArcGIS layers and the shortest Euclidean distance between the centroid of each pixel and the closest water body was calculated in ArcGIS version 10.0. Altitude data were obtained from an interpolated digital elevation model developed by the U.S. Geological Survey - Earth Resources Observation and Science Data Center. Data on the rural extents in Senegal are obtained by Afripop (Linard et al., 2012) which provides estimated population surfaces across Africa. According to the United Nations definition for Senegal, agglomerations with more than 10000 inhabitants were considered as urban. The above data were available at 1 km spatial resolution, with the exception of rainfall which has a resolution of 8 km.

The Hurdle model with the proposed variable selection was applied on this dataset to identify predictors that influenced the suitability of transmission (probability of zeros) and

the number of infected in each location in the area. The priors employed are the same specified in the description of the simulation study.

The model was run in JAGS for 100000 iterations with a burn-in of 1000 and a thinning of 10. Table 3.5 shows the posterior inclusion probability for each predictor.

Table 3.5: Posterior inclusion probabilities of predictors and spatial processes estimated by the Bayesian variable selection method using the Hurdle Binomial model for the analysis of malaria prevalence data in Senegal.

Variable (Trunc.Binomial)	$P(\gamma_i = 1)$
Rainfall	0.58
Vegetation Index	0.82
Night Temperature	0.52
Day Temperature	0.12
Area type (Urban=1)	0.23
Elevation	0.08
Distance to water bodies	0.12
Ω_1	0.92
Variable (Zeros)	$P(\delta_i = 1)$
Rainfall	0.88
Vegetation Index	0.21
Night temperature	0.25
Day temperature	0.65
Area type (Urban=1)	0.92
Elevation	0.42
Distance to water bodies	0.35
Ω_2	0.84

Environmental covariates show different selection probabilities in the two parts of the model. The cumulative precipitation during the year of the survey, the average NDVI and the average night LST were selected with a probability higher than 0.5 in the positive part of the model and therefore included in the model used for fitting the data. Precipitation, area type and day LST were chosen with a probability higher than 0.5 as predictors for the presence/absence of malaria (suitability index). In this application, the median probability model coincided with the model showing the highest posterior probability. Risk factor estimates as well as posterior estimates of the spatial parameters are presented in Table 3.6. Precipitation and NDVI were positively associated with the positive part of the model (number of malaria infected children); area type, day temperature and precipitation were important predictors of the suitability for transmission. In fact, economic development and urban type activities make cities and big agglomerations a less favourable environment for

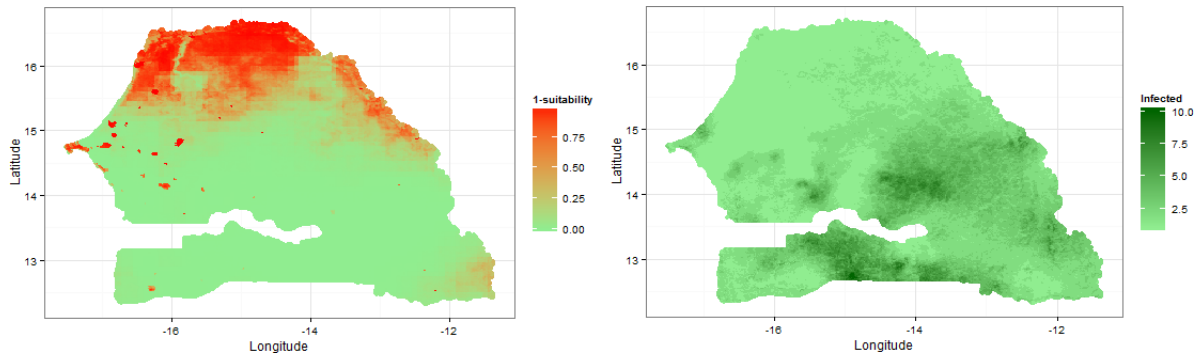


Figure 3.2: Predicted number of infected and suitability index in Senegal. Results obtained with the median probability model obtained via the Bayesian variable selection procedure applied on the geostatistical Binomial Hurdle model formulation.

vectors reproduction and survival as compared to rural areas. Very high temperatures and excessive rainfall can destroy potential breeding sites and impede transmission. The estimated correlation between the spatial processes is negative, suggesting that the variability in the intensity of the disease decreases with increasing spatial variability in the suitability for transmission. It is worth noting that the decay parameter ρ_2 and the spatial variance σ_2 are obtained using the conditional specification. The same model was used to predict suitability as well as the number of infected children at unobserved locations throughout the study area at a spatial resolution of 2 km (Figure 3.2). A common denominator of 30 children was chosen for the prediction of the malaria infected children.

Table 3.6: Binomial geostatistical hurdle model with the highest posterior probability: estimated effect of the selected environmental variables and spatial parameters estimates.

Model Component	Parameter	Median	95%CI
Trunc. Binomial	const.	-2.95	(-3.32,-2.45)
	Rainfall	0.15	(0.07,0.25)
	Vegetation Index	0.47	(0.32,0.81)
	Night Temperature	0.65	(-0.11,0.85)
Zeros	const.	-3.21	(-3.85,-2.55)
	Rainfall	-2.32	(-2.60,-2.08)
	Day Temperature	1.32	(0.84,1.88)
	Area type (Urban=1)	2.72	(1.54,3.11)
Spatial parameters	ρ_1	2.21	(1.42,3.65)
	ρ_2	3.41	(2.10,4.52)
	σ_1^2	1.97	(1.45,2.66)
	σ_2^2	0.74	(0.52,1.20)
	$corr(\Omega_1, \Omega_2)$	-0.49	(-0.63,-0.19)

3.4 Concluding remarks

We have assessed model specification of zero-inflated geostatistical Binomial data and proposed model formulation via variable selection methods. It has been shown (Hoeting et al., 2006b) that for geostatistical models it is essential to evaluate and assess spatial structure together with covariates effects. Furthermore, the inclusion of all potential covariates in the zeros and non-zeros regression parts may lead to over-parameterization. Therefore, we have proposed a variable selection approach that allows the selection of relevant predictors jointly with spatial structures in both the Bernoulli part of the model and the positive part. Through a large set of simulated examples, we found that the hurdle model showed higher ability to select the true model used to generate the data, probably because of poor identifiability of the spatial zero-inflated models. Furthermore, this parameterization allows for simpler interpretation of covariate effects especially with the choice of the logit link function (i.e., the exponentiated coefficients are odds ratios).

Our simulation results showed that model mis-specification arisen from variable selection reduces the predictive ability.

In the area of malaria epidemiology, the probability of a zero can be interpreted as disease transmission suitability, and the conditional mean of the counts represents the mean number of cases given that suitable conditions for transmission exist. Our application showed that different factors affected the number of infected children and the suitability of transmission for the disease. Spatial structure was present in both parts of the model. The hurdle model allowed the estimation of both the spatial processes associated with the Binomial (malaria risk) and the distribution of zeros (transmission suitability). The maps depicting the risk of being infected and the transmission suitability are useful tools for identifying priority areas for disease control. The focus on malaria elimination as well as on other parasitic disease suggests that more and more datasets with high percentage of zeros will be generated in the future.

3.5 Appendix

The zero-inflated Binomial model and the Hurdle model are linked by the relation: $p^* = p + (1-p)(1-\theta)^N$ where p^* represents the total probability of zeros, p is the mixing probability (zero-inflation), θ is the Binomial parameter and N is the number of trials. With the purpose of exploring the contribution of the Binomial part in the generation of zeros we have considered 3 different scenarios and plotted the relationship between (i) p and N for varying values of $\theta = 0.1, 0.2, 0.3, 0.4, 0.5$ when the total probability of zeros is fixed to $0.4, 0.5, 0.7, 0.8$ (ii) p and θ for varying values of $N = 5, 10, 20, 50$ and the same total probability of zeros and (iii) p and p^* for varying values of $N = 5, 10, 20, 50$ when θ is fixed to $0.1, 0.2, 0.3, 0.4$. We can observe how for increasing values of θ or N a higher number of zeros is generated by the degenerate distribution and therefore the total probability of zeros p^* can be safely approximated by the mixing probability p. Therefore, in our simulation study we choose $N = 50$.

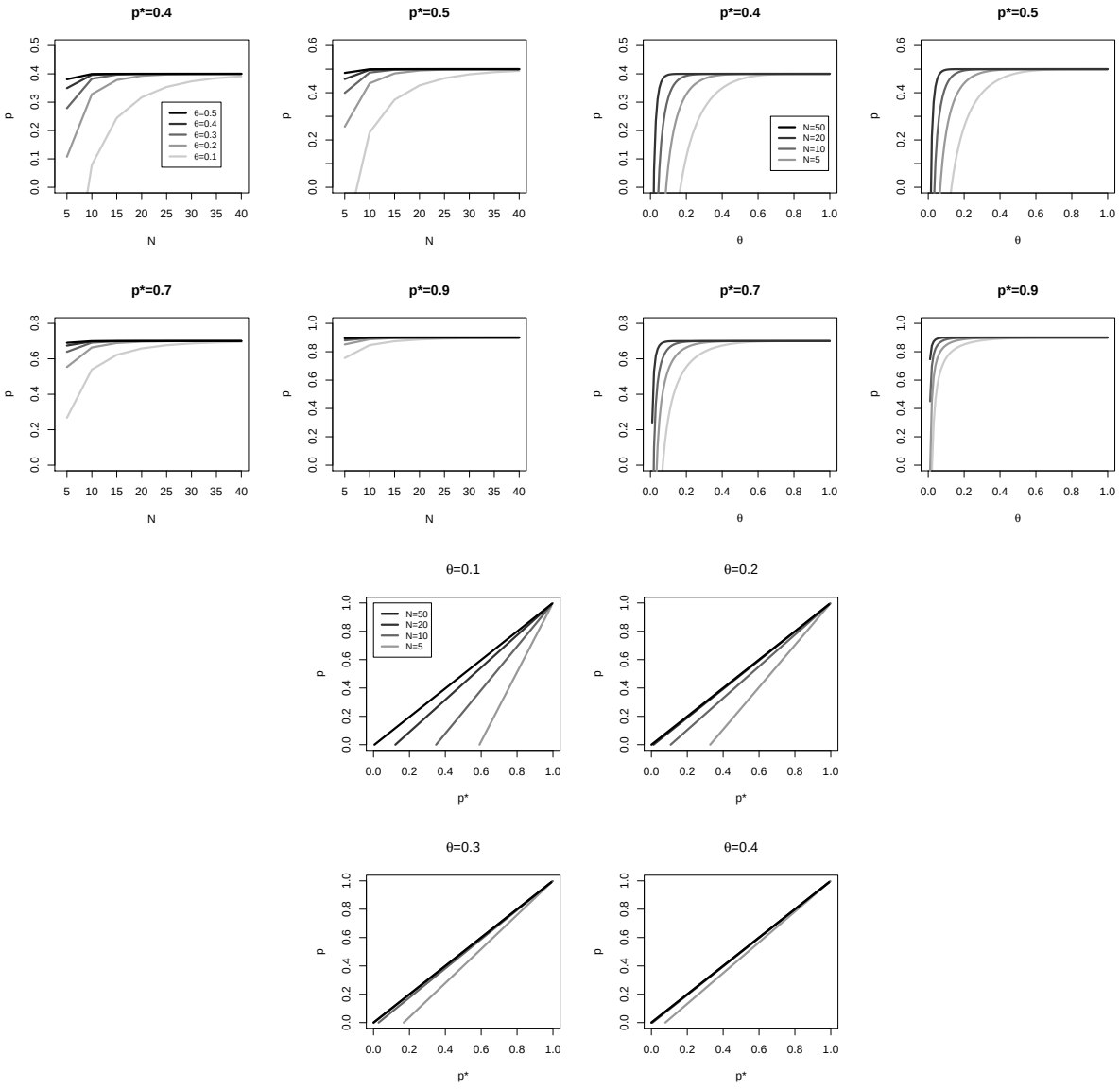


Figure 3.3: Relationship between Binomial distribution number of trials (N), Binomial distribution parameter (θ), mixing probability (p) and total probability of zeros (p*).

Chapter 4

Bayesian variable selection in semiparametric and non-stationary geostatistical models: an application to mapping malaria risk in Mali

Giardina F.[1,2], Sogoba N.[3], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Malaria Research and Training Center, Faculté de Médicine, de Pharmacie et d'Odonto-Stomatologie, Université du Mali, Bamako, Mali

This paper has been published as a book chapter in the *Handbook of spatial epidemiology*

Abstract

Geostatistical models applied in epidemiology aim to identify the main determinants of a disease and predict disease outcome measures (e.g. risk, incidence, mortality) at unobserved locations. The impact of the predictors is commonly modelled as a linear effect, constant throughout the study area. However, more flexible functional forms may be required to capture non-linear relationships between the covariates and the response. When the area of interest is large and covered by different regions, (e.g. ecological zones) the relationship between the disease and its risk factors may not be constant across the area. Furthermore, the spatial correlation is likely to vary not only as a function of distance but also of geographic position. In this work, we develop Bayesian spatial variable selection methods with spike-and-slab prior structure that allow the choice of different predictors and their functional forms in non-stationary geostatistical models for mapping disease survey data. Non linear functional forms are expressed as piece-wise constant or smooth terms (splines). Penalized spline effects are re-parameterized as mixed effects terms and their selection is based on non-zero random effects variance identification. Spatially varying weights are proposed to achieve smoothness across irregularly shaped regions. Markov chain Monte Carlo (MCMC) simulation is used for estimation and inference. Multiplicative parameter expansion methods are employed to allow mixing of the chains in selecting batches of coefficients that model non-linearity. The methods are illustrated by analysing recent national malaria survey data from Mali to obtain spatially explicit estimates of the disease burden in the country.

4.1 Introduction

Geostatistical models are used to analyze data collected at a discrete set of locations (geo-referenced data) within a continuous domain (Cressie, 1993). They have been widely applied to problems ranging from geology and ecology to epidemiology and public health (Gelfand et al., 2004a). Applications in epidemiology are mainly concerned with relating disease data to a set of predictors (i.e. environmental or climatic variables) with the aim of determining the main risk factors and predicting disease outcome measures (e.g. risk, incidence, mortality) at unobserved locations (Lawson, 2013). The Bayesian formulation of linear and generalized linear geostatistical models has been introduced by Diggle et al. (1998).

Bayesian geostatistical models have been widely used in mapping parasitic diseases such as malaria risk (Gemperli and Vounatsou, 2006; Gosoniu et al., 2006; Hay and Snow, 2006; Giardina et al., 2012; Noor et al., 2014), schistosomiasis risk (Raso et al., 2006b; Clements et al., 2008; Wang et al., 2008), filarial worm risk infection (Diggle et al., 2007; Crainiceanu et al., 2008), hookworm infection (Raso et al., 2006a; Chammartin et al., 2013b), and helminths co-infections (Pullan et al., 2008; Schur et al., 2011). Disease maps can be used to identify possible clusters, to define and monitor epidemics or provide baseline risk estimates at high spatial resolution. They represent an essential tool to guide disease control programs in planning targeted interventions and in evaluating their effectiveness. Recent developments in satellite-based remote sensing (RS) for environmental monitoring and geographical information system (GIS) have further boosted research in this area (Bauwens et al., 2011).

In most epidemiological applications, the impact of the predictors is modelled as a linear effect, constant throughout the study area. However, more flexible functional forms, are often more suitable to capture the relationships between the covariates and the response. Larges study areas can often be partitioned in different ecological zones which influence the effect of the predictors on the disease outcome. In large areas the underlying spatial structure that models the geographical dependence among neighboring locations, may vary also according to the geographic position. Therefore, a flexible model specification is required to enable choosing different predictors as well as different functional forms in each zone, while modelling a non-stationary spatial process.

Bayesian statistical methods for choosing an appropriate subset of covariates among

many potential predictors have received increasing attention in recent years. A comprehensive review of the most commonly used methods can be found in O'Hara and Sillanpää (2009). Chen and Dunson (2003) and Kinney and Dunson (2007) studied both fixed and random effects selection in linear and logistic models. Tüchler (2008) and Wagner and Duller (2012) proposed an approach that links Bayesian variable selection methods to random effects' variance selection by a re-parametrization of the random effects. Less work has been done on functional form selection methods including non-linear effects: only recently Scheipl et al. (2012) proposed a stochastic search based approach employing a modified spike and slab mixture prior for the coefficients and Bové et al. (2012) developed an extension of the classical Zellner's g-prior (hyper-g priors) to identify the presence of a variable and its spline transformation in generalized additive models. Curtis et al. (2014) provides a review of variable selection methods for additive models.

The literature on variable selection for spatial data is limited. Typically, spatial correlation is ignored in the selection of explanatory variables, influencing model selection as well as parameter estimation (Hoeting et al., 2006a). In the work by Smith and Fahrmeir (2007) an Ising prior is used to allow dependence among variable inclusion probabilities at neighboring locations for linear regression models defined on a regular lattice. A similar approach is adopted by Scheel et al. (2013) studying the effect of climate change on the insurance industry at local geographic scale (municipalities). Reich et al. (2010) proposed a stochastic search approach to select covariates with constant or spatially varying effects (Gelfand et al., 2003).

A review of methods used for constructing non-stationary spatial processes can be found in Sampson (2010). These methods range ranging from spatial deformation models (Sampson and Guttorp, 1992) to spatial processes decomposition in terms of empirical orthogonal functions (Nychka et al., 2002) and process convolution models (Higdon, 1998). Smoothing and kernel-based methods (Fuentes, 2001) model non-stationarity as spatially weighted combinations of stationary spatial covariance functions. This approach was applied by Banerjee et al. (2004) to model house prices in California and by Gosoniu et al. (2009) in malaria risk mapping in West Africa. In the latter, the relation between climate factors and malaria risk was modelled separately in each ecological zone by penalized B-splines. In this paper, we extend the work by Gosoniu et al. (2009) by developing Bayesian non-stationary geostatistical models that choose among different functional forms allowing variations across partitions of the area of interest. Furthermore, spatially varying weights are proposed to take into account irregularly shaped partitions of the study area. The

application that motivated the work comes from the area of malaria epidemiology. Over the last few years, national malaria surveys have been carried out routinely in several countries in Sub-Saharan Africa with the aim of monitoring and evaluating progress in disease control.

The paper is structured as follows: Section 4.2 describes the problem and the data used, Section 4.3 introduces variable selection methods for functional forms in non-stationary geostatistical models. Section 4.4 presents the results of the proposed methodology applied to a national malaria prevalence survey in Mali. Validation results compare predictions of the model determined by the developed methods to those obtained with a full B-spline model (Gosoniu et al., 2009) and with the same model with stationary covariance matrix. Section 4.5 provides concluding remarks and suggests further lines of research and areas of application.

4.2 Background

Malaria transmission is strongly influenced by climatic conditions which determine the abundance and seasonal dynamics of the *Anopheles* mosquito vector. The amount and duration of malaria transmission is influenced by the ability of parasite and mosquito vector to co-exist long enough to enable transmission to occur. The distribution and abundance of the parasite and mosquitoes population are sensitive to environmental factors like temperature, rainfall, humidity, presence of water and vegetation. Environmental factors affect the biological cycle of both vector and parasite allowing or interrupting the different development stages and therefore favoring or inhibiting transmission. Usually, *Anopheles* do not fly more than 2km but in certain circumstances they can fly up to 5km. The distance mosquitoes fly is determined largely by the environment: if suitable hosts and breeding places are nearby, mosquitoes do not to disperse far, but if one or more are more distant, greater dispersal may be necessary (Schlagenhauf-Lawlor, 2008).

Mali is divided into five ecological zones based on Food and Agriculture Organization (FAO) methodology (FAO, 2000): the Sahara desert, the South Saharan zone, the Sahelian zone, the central delta of the Niger river and the west Sudannian region, as shown in Figure 4.1. The northern part of Mali is occupied by the Sahara desert which is a hyper-arid zone with scarce water and precipitation; the first Sub-Saharan zone presents steppe and woodlands and it is as well an arid and desertic zone. These two regions do not represent a favorable environment for the malaria vector. The Sahelian zone is mainly characterized by acacia savanna; it is arid with rainfall between 250 and 550 mm. The central region of the

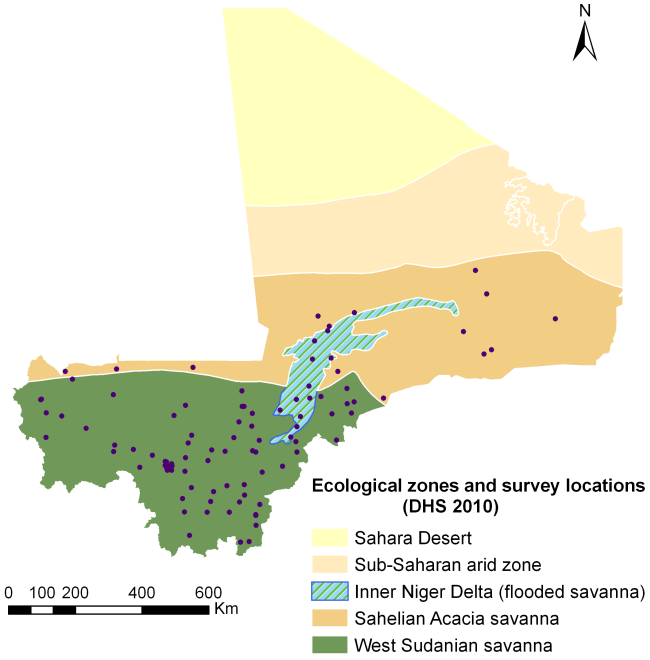


Figure 4.1: Ecological zones in Mali. FAO. Global Forest Resources Assessment 2000. `www.fao.org/forestry/fra/2000/report/en/`.

Niger delta presents similar characteristics in terms of rainfall but it is mainly constituted by flooded savanna. The Sudan zone, in the South-West of the country, is a semi-arid to sub-humid region with abundant rainfall (between 550 and 1100 mm).

4.2.1 National Malaria survey in Mali

A Demographic and Health Survey (DHS) was carried out in Mali between August and October 2010 by the National Malaria Control Program in collaboration with Macro International and the Malaria Research and Training Center in Bamako. The information collected in the survey consists of geo-referenced data with parasitaemia measurements (malaria test positivity) among 1788 children below the age of five years (Ministère de la Santé, Programme National de Lutte contre le Paludisme, INFO-STAT, ICF Macro, 2010).

4.2.2 Environmental predictors

RS and GIS have emerged as methods for exploring environmental factors potentially associated with malaria outcomes. With the purpose of deriving explanatory variables for our application, we have collated environmental and climatic data provided by satellite images. Vegetation measures such as Normalized Difference Vegetation Index (NDVI)

and Enhanced Vegetation Index (EVI), as well as temperatures proxies (night/day land surface temperature) were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) at 1km spatial resolution for the year 2010. Dekadal rainfall data were extracted at 8km resolution via Africa Data Dissemination Service (ADDS) and aggregated over a year previous to the survey time. Water bodies were identified using the world water bodies layer provided by the ArcGIS website. The shortest Euclidean distance between the locations and the water bodies was calculated in ArcGIS version 10.0 (ESRI, 2011). Altitude data were obtained from an interpolated digital elevation model by the U.S. Geological Survey - Earth Resources Observation and Science Data Center at a spatial resolution of 1km. Information on area type (rural/urban) was provided by the Global Rural-Urban Mapping Project (GRUMP) website and population density data by the Afripop project (Tatem et al., 2007).

All environmental and climatic data have been associated to observed locations with the shortest Euclidean distance from the layers.

4.3 Models

Let $N(s)$ be the number of individuals screened for parasitaemia at location s, $s = 1, \ldots, m$, $Y(s)$ be the number of those tested positives, and $\mathbf{x}(s) = (x_1(s), x_2(s), \ldots, x_p(s))^T$ be the vector of p potential predictors observed at location s. We assume that $Y(s)$ arises from a Binomial distribution:

$$Y(s)|\pi(s), N(s) \sim Binomial(\pi(s), N(s)) \qquad \forall s = 1, \ldots, m \text{ sites}$$

and the probability $\pi(s)$ of being infected at location s is modelled through an additive logistic regression,

$$log\left(\frac{\pi(s)}{1-\pi(s)}\right) = \mu(s) + \omega(s) \tag{4.1}$$

where $\mu(\cdot)$ represents the mean structure and $\omega(\cdot)$ models the spatial correlation through Gaussian processes. The mean structure takes the general form:

$$\mu(s) = \beta_0 + \sum_{i=1}^{p}\sum_{j=1}^{J}\sum_{k=1}^{K} f_{ijk}(x_i(s), \beta_{ijk})$$

where $f_{ijk}(\cdot)$ indicates each one of the J possible functional forms that relate the observed variable X_i to the disease risk $\pi(s)$ in ecological zone k via the coefficients β_{ijk} and β_0 is a

common intercept term.

We model non-stationarity in the spatial process through a mixture of stationary spatial processes smoothing at the borders between the zones through the definition of distance-dependent weights, as in Gosoniu et al. (2009). A stationary spatial process $\boldsymbol{\phi}_k$ is defined as $\boldsymbol{\phi}_k \sim N(0, \boldsymbol{\Sigma}_k)\ \forall k = 1, \ldots, K$ ecological zone where $(\Sigma_k)_{ss'} = \sigma_k^2 corr(||s - s'||; \rho_k, \nu)$ and *corr* is a parametric function of the Euclidean distance $||s - s'||$ between sites s and s'. The Matern family describes most of the correlation function used in geostatistical models:

$$corr(||s - s'||; \rho_k, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\rho_k||s - s'||)^\nu K_\nu(\rho_k||s - s'||)$$

where K_ν is a modified Bessel function with smoothing parameter ν, while $\rho_k > 0$ controls the rate of correlation decay between observations as distance increases. The choice $\nu = 1/2$ leads to the commonly used exponential correlation function, i.e. $corr(||s - s'||) = \exp(-\rho_k||s - s'||)$.

A non-stationary spatial process $\boldsymbol{\omega}$ is generated as a weighted sum of the above defined spatial processes as follows: $\boldsymbol{\omega} \sim N(0, \sum_{k=1}^{K} A_k \Sigma_k A_k)$ where A_k is a diagonal matrix with $(A_k)_{ss} = a_{sk}$. The weights a_{sk} are chosen as decreasing functions of the Euclidean distance between location s and "knots" of the subregion k. The "knots" are selected over a grid covering the entire region in order to take into account the irregularly shaped subregions. Further details on the choice of the weights are given in Section 4.3.2.

4.3.1 Mean structure selection

We describe a Bayesian variable selection procedure to choose an appropriate subset of potential covariates for malaria risk and determine whether a linear, piecewise constant or a smoother functional form is required to model the effect of the respective covariates, allowing them to vary across ecological zones. For each variable X_i in ecological zone k, we consider the following four scenarios: (i) there is *no relationship* between X_i and the infection probability π; there is a relationship that can be described by (ii) *linear*, (iii) *piecewise constant* or (iv) *smooth* functions.

The variable selection approach is defined by the following hierarchy:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)^T$$

$$\mathbf{p}_k|\boldsymbol{\alpha} = (p_{1k}, \ldots, p_{4k})^T \sim Dir(4, \boldsymbol{\alpha}) \quad \forall k = 1, \ldots, K$$

where $\mathbf{p}_k$ follows a Dirichlet distribution of concentration hyperparameters $\boldsymbol{\alpha}$. Each element $p_{jk}, j = 1, \ldots, 4$ corresponds to the selection probability of the different functional forms (1=linear, 2=piecewise constant, 3=smooth, 4=none) in region k. For each predictor i in region k a categorical variable c_{ik} can be defined to indicate the different functional forms, with probability mass function $p(c_{ik}|\boldsymbol{p}_k) = \prod_{j=1}^{4} p_{jk}^{\delta_j(c_{ik})}$where $\delta_j(\cdot)$ denotes the Dirac delta function evaluated at j. We build the auxiliary variables γ_{ijk} to indicate presence or absence of the j functional form of covariate X_i in region k

$$\gamma_{ijk} = \delta_j(c_{ik}) + \epsilon_0(1 - \delta_j(c_{ik})) \; \forall j = 1, \ldots, 3$$

and we assign a Normal prior to the coefficients β_{ijk}, that is

$$\beta_{ijk}|\gamma_{ijk}, \tau_{ij}^2 \sim N(0, \gamma_{ijk}\tau_{ij}^2)$$

$$\tau_{ij}^2|a, b \sim IG(a, b)$$

where ϵ_0 is some very small positive constant and the variance τ_{ij}^2 is sampled from an Inverse Gamma (IG) with shape parameter a and scale b.

This prior specification defines a continuous bimodal distribution on the hypervariance of β_{ijk} with a spike at ϵ_0, that shrinks the coefficients that are not relevant for the model, and a right continuous tail (slab) to identify non-zero parameters. In particular, if $\gamma_{ijk} = 1$ the covariate effect β_{ijk} is estimated by assuming a Normal prior distribution with mean 0 and variance τ_{ij}^2, otherwise $\gamma_{ijk} = \epsilon_0$ and β_{ijk} is shrunk towards 0, therefore the predictor X_i is not included in the model for region k in functional form j. The Dirichlet prior on the selection probability p allows flexibility in estimating model sizes by introducing another level of hierarchy in the model specification. If $\gamma_{i1k} = 1$, the relationship between the predictor X_i and the disease risk π is *linear* in region k and $f_k(x_i) = \beta_{i1k}x_i$. If $\gamma_{i2k} = 1$, the relationship between the predictor x_i and the disease risk is *piecewise constant* where x_i has been categorized into Q quantiles and $f_k(x_i) = \sum_{q=1}^{Q} \beta_{iqk}^{'} x_{iq}^{'}$. Spike and slab priors perform poorly in identifying non-linear forms of variables which include groups of coefficients (Scheipl et al., 2012). In particular, switching status (i.e. inclusion/exclusion of the coefficient's vector) becomes very unlikely, resulting in very poor mixing of the indicator variables. Parameter expansion (Gelman et al., 2008) offers a method to improve mixing in the MCMC while selecting simultaneously batch of coefficients. More specifically, we

define $\boldsymbol{\beta}'_{ik} = \beta_{i2k}\boldsymbol{\eta}_{ik}$ where

$$\eta_{iqk}|m_{iqk} \sim N(m_{iqk}, 1) \text{ and } m_{iqk} \sim 1/2N(-1, 1) + 1/2N(1, 1) \quad \forall q = 1, \ldots, Q \text{ quantiles.}$$

The two parameters η_{iqk} and β_{iqk} are not identifiable but inference can be obtained about their product β'_{ik}. If $\gamma_{i3k} = 1$, the relationship between the predictor X_i and the disease risk π includes non-linear terms in region k, expressed in the form of a penalized B-spline, i.e. $f_k(x_i) = bx_i + \sum_{l=1}^{L} u_{ilk} z_l(x_i)$ where $z_l, \forall l = 1, \ldots, L$ is an appropriate spline basis for covariate x_i, i.e. radial cubic basis function.

Following Ruppert et al. (2003) a quadratic penalty is placed on $\mathbf{u}$, that translates into the constrain: $\boldsymbol{u}_{ik}^T \boldsymbol{u}_{ik} \leq \lambda$ where λ is the smoothing parameter. The above functional form can be written in a mixed models representation (Zhao et al., 2006) as follows:

$$f_k(x_i) = \beta_{i3k} x_i + \mathbf{Z}_{x_i} \boldsymbol{u}_{ik}$$

where

$$\mathbf{Z}_{x_i} = \left[|x_i - \kappa_l|^3 \right] \left[|\kappa_l - \kappa_{l'}|^3 \right]^{-1/2}$$

and $\boldsymbol{u}_{ik} \sim N(0, \sigma^2_{u_{ik}} \mathbf{I})$. The knots κ_l are defined as the sample quintiles specific to each covariate X_i. To ensure identifiability of the model we do not include a constant term in the spline representation. We apply random effect selection methods to choose whether a smooth term has to be included in the models. We follow the approach suggested by Wagner and Duller (2012) reparametrizing the variance component and perform variable selection on the standard deviation treating it as a covariate effect. In particular, we re-write the random effects associated to the spline terms $\boldsymbol{u}_{ik}$ as $\boldsymbol{u}_{ik} = \pm\sigma_{u_{ik}} \boldsymbol{\theta}_{ik}$ where $\boldsymbol{\theta}_{ik} \sim N(0, \mathbf{I})$ and assign $\sigma_{u_{ik}}$ the same spike and slab prior as for the parameter β_{i3k}. The sign of both $\sigma_{u_{ik}}$ and $\boldsymbol{\theta}_{ik}$ is not identifiable but the product $\pm\sigma_{u_{ik}} \boldsymbol{\theta}_{ik}$ as well as the associated indicator γ_{i3k} can be estimated. In fact, as in the case of batches of coefficients, for the selection of the *piece-wise constant* functional form, this redundant parametrization has computational advantages in the MCMC implementation.

The procedure described above can be adopted only for continuous predictors. Categorical predictors, such as area type which is dummy variable, were modeled using piecewise constant functional forms.

It is not realistic to assume independence across the predictors selected in each ecological zone. To take into account spatial dependence in the mean structure and to obtain smooth prediction maps at the zone borders we introduce spatially varying weights $\psi_k(s)$ in the regression coefficients and define $\beta^*_{ij}(s) = \psi_k(s)\beta_{ijk}$. Details on the specification of the weights can be found in subsection 4.3.2. Equation (4.1) takes now the form $logit(\pi(s)) = \sum_{i=1}^{P}\sum_{j=1}^{J}\sum_{k=1}^{K} f_{ik}(x_i(s), \beta^*_{ij}(s)) + \omega(s)$. In most applications, model fit and prediction is performed using the model with the highest posterior probability. However, Barbieri and Berger (2004) shows that for normal linear models, the one with the best predictive ability, i.e. that minimize the squared error loss, is the so called median probability model. The latter is defined as the model consisting of those variables which have overall posterior probability greater than or equal to 1/2 of being included in a model. The median probability model may differ from the highest probability model.

4.3.2 Spatially varying weights

Spatially varying weights a_{sk} and $\psi_k(s)$ have been introduced to model the variance structure of the non-stationary Gaussian spatial process and the mean structure respectively. While a_{sk} smooths the values of the spatial process at the border of the zones, $\psi_k(s)$ takes into account that the risk in neighboring points across the borders of the zones should be affected similarly by covariates although the zones may have different predictors. Therefore $\psi_k(s)$ smooths the mean structure at the borders. For convenience, we chose $a_{sk} = \psi_k(s)$. We define $d_k(s)$ as the Euclidean distance between a given location s and the closest of the knots belonging to ecological zone k. The knots are equally spaced points over a grid covering the study area. We obtain weights that are decreasing function of the shared area between two circles of radius r, the first one centered in s and the other one in the point at distance $d_k(s)$. Following the definition of spherical correlation function in two dimension, (circular correlation function) we construct the spatially varying weights $\psi_k(s)$ as follows:

$$\psi_k(s) = \begin{cases} \frac{2}{\pi}\left(\arccos d_k(s)/r - (d_k(s)/r)\sqrt{1-(d_k(s)/r)^2}\right) & \text{if } d_k(s) < r \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the weights allow covariate effects of neighboring zones to be considered for all locations within a radius r from the border. Furthermore, each zone has a separate spatial process and the weights allow a mixing of neighboring process only for locations close to the borders (in proximity defined by the radius r).

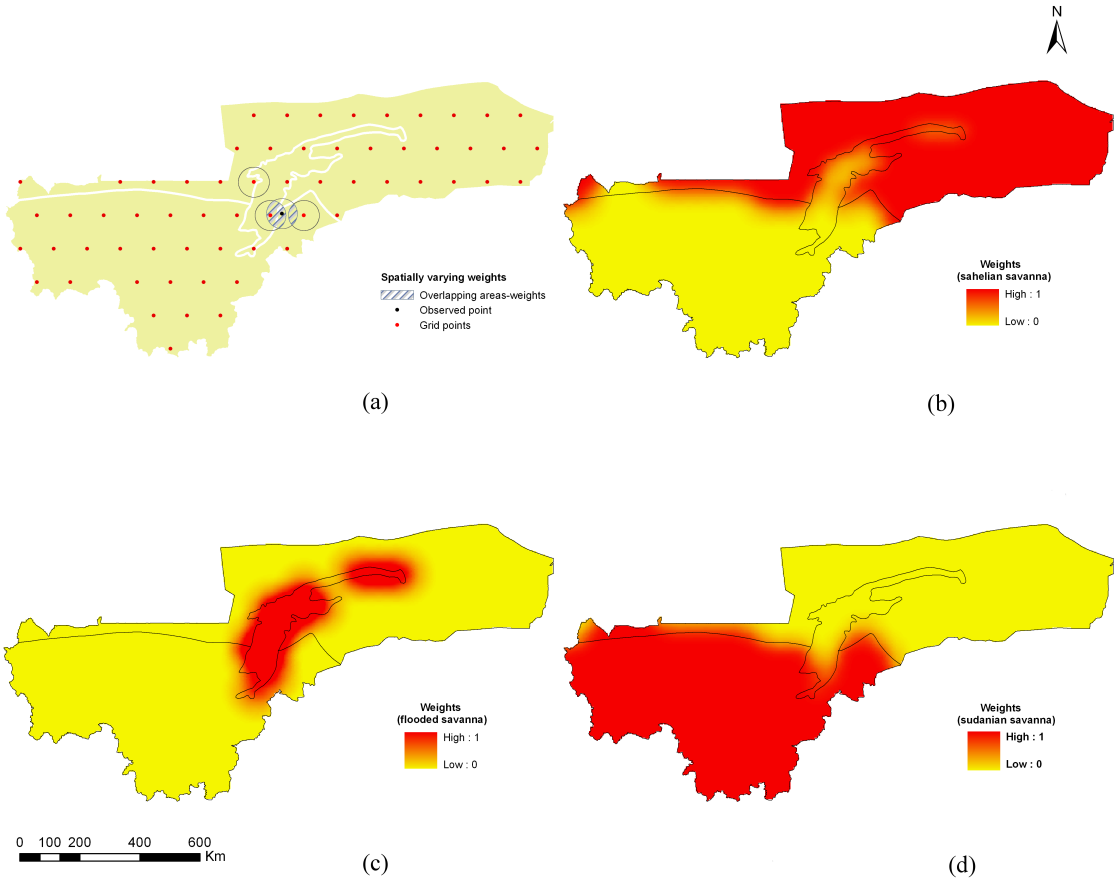


Figure 4.2: Spatially varying weights for an observed location and the three closest knots in each zone (a). Spatially varying weights for each prediction location: Sahelian zone (b), Flooded zone (c), Sudannian zone (d).

The weights were normalized (divided by their length) so that $\sum_{k=1}^{K} \psi_k(s)^2 = 1$. However, the choice of the grid spacing g and the radius r might influence posterior inference. Therefore, the main analysis has been carried out defining g and fixing $r = g$. Nevertheless, a sensitivity analysis has been conducted to assess the robustness of the results under different values of the radius and keeping the spacing of the grid knots constant.

4.4 Results

The model presented in Section 4.3 was applied on the national malaria survey data from Mali to identify the most important climatic predictors by ecological zone and perform spatial risk prediction over the study area at 2 km resolution. Three different malaria endemic ecological zones and eight predictors with three functional forms were considered in the analysis. All continuous covariates have been centered and standardized to obtain a better mixing of the Markov chains arising from simulations. The spatially varying weights defined for a specific observed point and for the whole study area are shown in Figure 4.2.

Posterior analysis was performed by MCMC using samples collected over 100.000 iterations after a burn-in of 10.000. Convergence was monitored by examining trace plots and auto-correlation plots for several representative parameters. Results of the variable selection procedure are given in Table 4.1 and Figure 4.3. Table 4.1 shows the models selected with the highest posterior probabilities (only the first three are listed). Figure 4.3 shows the posterior inclusion probability of the environmental variables for each zone and functional form, i.e. the overall posterior probability that each variable is in the model.

The model selected with the highest posterior probability (Model 1 in Table 4.1) coincided with the median probability model 4.1 and it was used for posterior inference on risk factors and spatial structure as well as for predictions. Model 1 includes the variable rainfall in linear form in the Sahelian zone, the day temperature in linear form and the area type in the flooded zone of the Niger delta, the NDVI as smooth term and the area type in the Sudannian zone. The functional forms of the selected predictors are shown in Figure 4.4. Living in rural areas is associated to a reduction in the odds of being infected with malaria by 23%, 95% BCI:$(19\% - 41\%)$ in the flooded zone and by 52%, 95% BCI:$(41\% - 63\%)$ in the Sudannian zone. Rainfall was associated with a significant increase of malaria risk in the Sahelian zone (OR= 1.22, 95% BCI:$(1.11 - 1.41)$). Day temperature was found to be the main risk factor in the Niger delta (OR= 1.66, 95% BCI:$(1.49 - 1.86)$). Nonlinearity was detected in the relationship between NDVI and the malaria risk in the Sudannian zone.

To study the predictive ability of the model, we divided the data into a training set used to fit the model and a test set for evaluating predictions. The training set consists of 80% of survey locations randomly sampled for each ecological zone and the test set includes the remaining points. The procedure was repeated 5 times (with 5 different training/testing sets) and the model predictive ability was assessed using a log-score criterion, defined as the negative log likelihood evaluated at the testing locations. For the purpose of comparison,

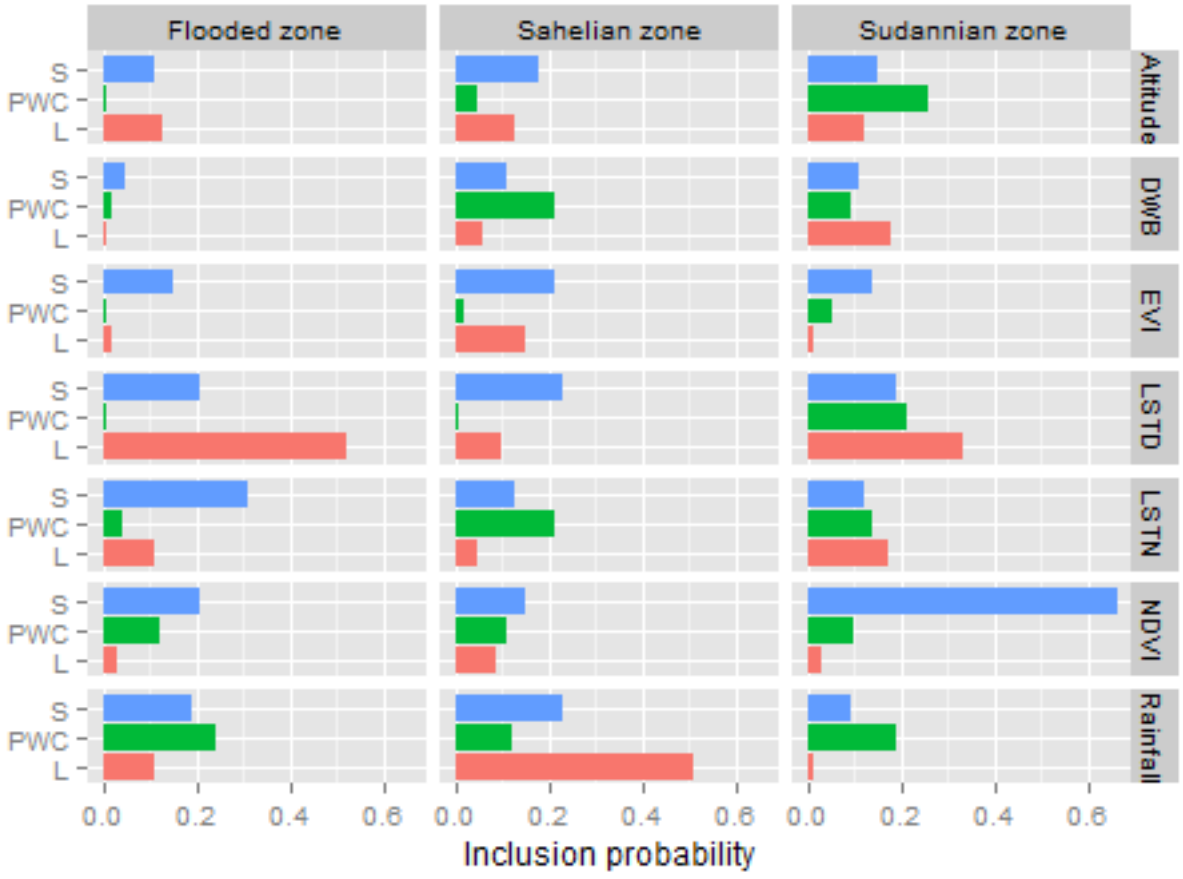


Figure 4.3: Inclusion probabilities per ecological zone and functional form (L=linear, PWC=piece-wise constant, S=spline).

the model proposed by Gosoniu et al. (2009) (full B-spline model) as well as the same model with a stationary covariance matrix was used for fitting and prediction on the same sets. Figure 4.5 compares the averaged log-likelihood between the three models. The plot indicates that Model 1 had a lower median log-score and smaller variability.

The three models were used to perform spatial prediction throughout the study area at 2 km resolution. The predicted parasitaemia risk in Figure 4.6 obtained using Model1 suggests an overall trend of increasing risk from the North to the South. The regions with the highest risk are Sikasso and Segou in the South of the country and Kayes at the border with Senegal and Mauritania. Figure 4.7 and 4.8 show a similar geographical pattern but larger uncertainties. Moreover, the different mean structures in each ecological zones produce discontinuities at the borders in absence of spatially varying weights.

A sensitivity analysis was performed to study the effect of the spatially varying weights in the selection of covariates running MCMC under different specification (different values of r keeping fixed the spacing between the grid points g). Results are shown in Table 4.2 and expressed in terms of the ratio between the radius and the grid spacing. Under the three settings (radius smaller than the spacing, equals or bigger) the model with the highest posterior probability remains the same, but the probabilities are different. In our analysis, we have defined the radius equals to the grid spacing. When the radius is smaller than the spacing, Model 1 was selected with a posterior probability of 0.51; very few points were affected by the covariates selected in the neighboring zones and this results in discontinuities in the prediction map. When the radius is higher than the spacing, Model 1 was selected with a low posterior probability (0.31) and several other competing models appear to be selected with probability of around 10% . This specification produced an oversmoothed prediction map. The inferences were computed by using JAGS (Plummer, 2003); the code for both models is available from the authors on request.

4.5 Discussion

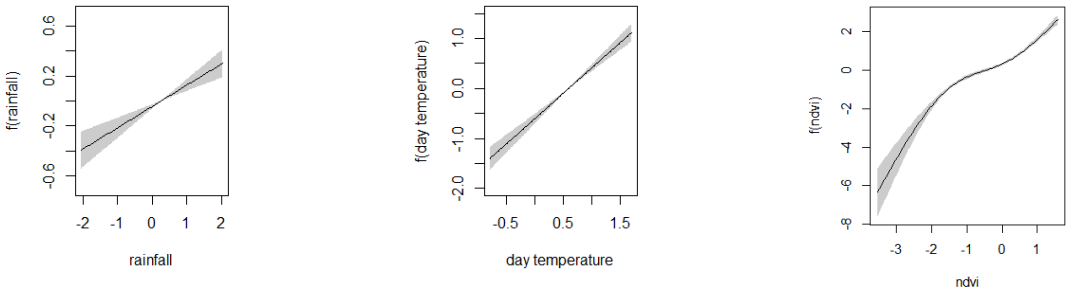
We have developed Bayesian methodology to model non-stationary geostatistical data when the study area consists of irregularly shaped zones with different characteristics. The methods described allow the choice of covariates and their corresponding functional forms by zone via a Bayesian variable selection procedure. Spatially varying weights were used in the regression model to take into account the dependence of the covariates affecting the disease outcome at a given location not only on the zone associated to the location but also on the neighboring regions within a certain radius. The weights introduced into the model

Table 4.1: Posterior model probabilities.

Model	Mean structure	p
1.	Rainfall (Sahelian zone, *linear*) + Day Temperature (Flooded zone, *linear*) + Area type (Flooded zone, *piece-wise constant*) + NDVI (Sudannian zone, *spline*) + Area type (Sudannian zone, *piece-wise constant*)	0.54
2.	Rainfall (Sahelian zone, *linear*) + Night Temperature (Flooded zone, *spline*) + Area type (Flooded zone, *piece-wise constant*) + NDVI (Sudannian zone, *spline*) + Area type (Sudannian zone, *piece-wise constant*)	0.12
3.	Rainfall (Sahelian zone, *spline*) + Day Temperature (Flooded zone, *linear*) + Area type (Flooded zone, *piece-wise constant*) + NDVI (Sudannian zone, *spline*) + Area type (Sudannian zone, *piece-wise constant*)	0.11

Table 4.2: Posterior model probabilities of the first selected model with different values of the ratio between the radius and the grid spacing.

Ratio	Sahelian zone	Flooded zone	Sudannian zone	p
r/g=1	Rainfall (*linear*)	Day temperature (*linear*) Area type (*piece-wise constant*)	NDVI (*spline*) Area type (*piece-wise constant*)	0.54
r/g=0.5	Rainfall (*linear*)	Day temperature (*linear*) Area type (*piece-wise constant*)	NDVI (*spline*) Area type (*piece-wise constant*)	0.51
r/g=1.5	Rainfall (*linear*)	Day temperature (*linear*) Area type (*piece-wise constant*)	NDVI (*spline*) Area type (*piece-wise constant*)	0.31



(a) **(b)** **(c)**

Figure 4.4: Estimated relationship between predictors and malaria risk in the three different ecological zones: (a) Sahelian zone, (b) Flooded zone, (c) Sudannian zone.

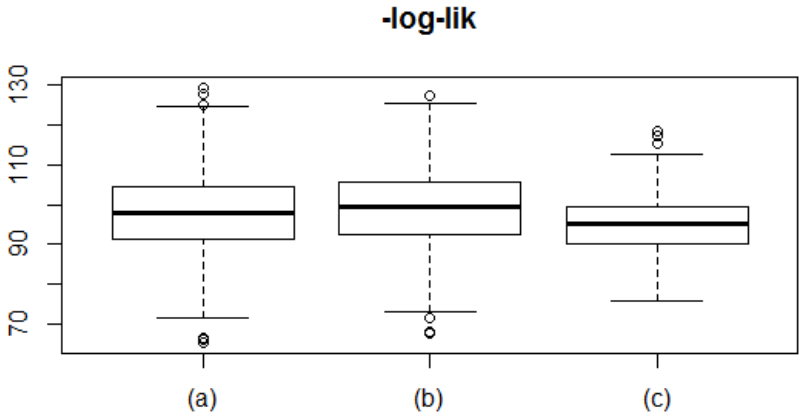


Figure 4.5: Log-score comparison between the full B-spline model, as in Gosoniu et al. (2009) (a), the full B-spline with a stationary covariance structure (b) and Model 1 (c).

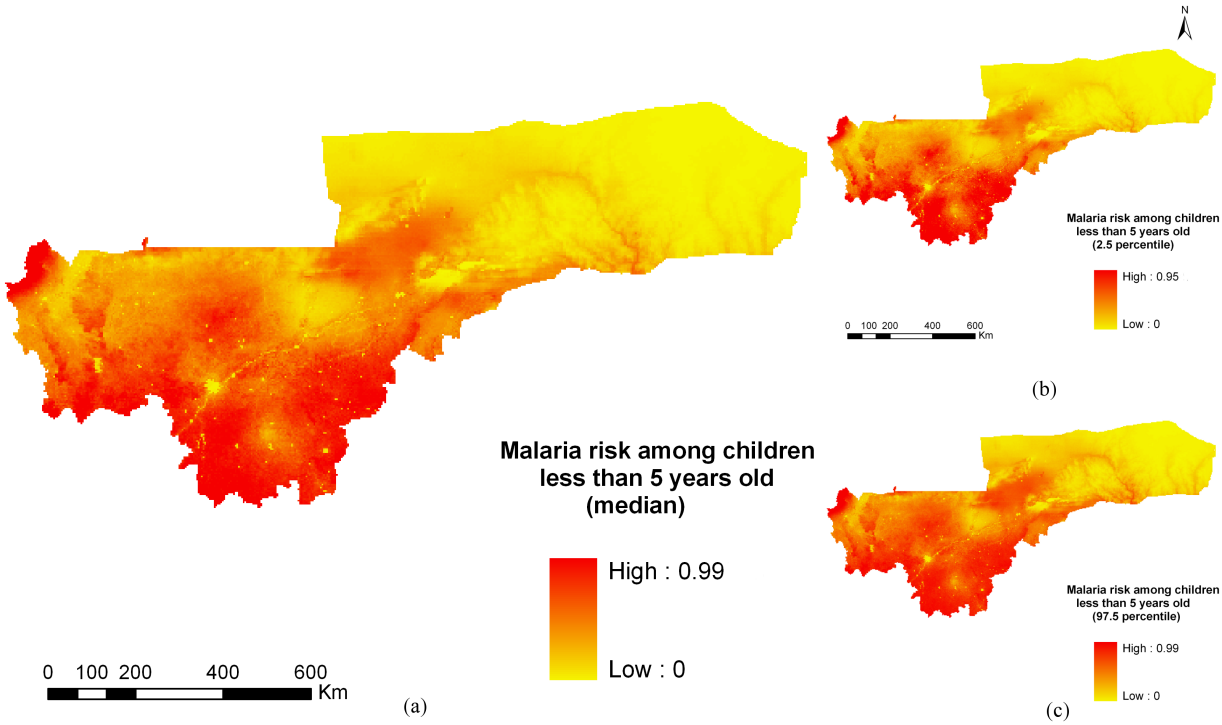


Figure 4.6: Predicted parasitaemia risk in children under 5 years. Map produced using the non-stationary model (Model 1) with different predictors in each ecological zone and spatially varying weights. Median (a) and Credible Intervals (b) and (c).

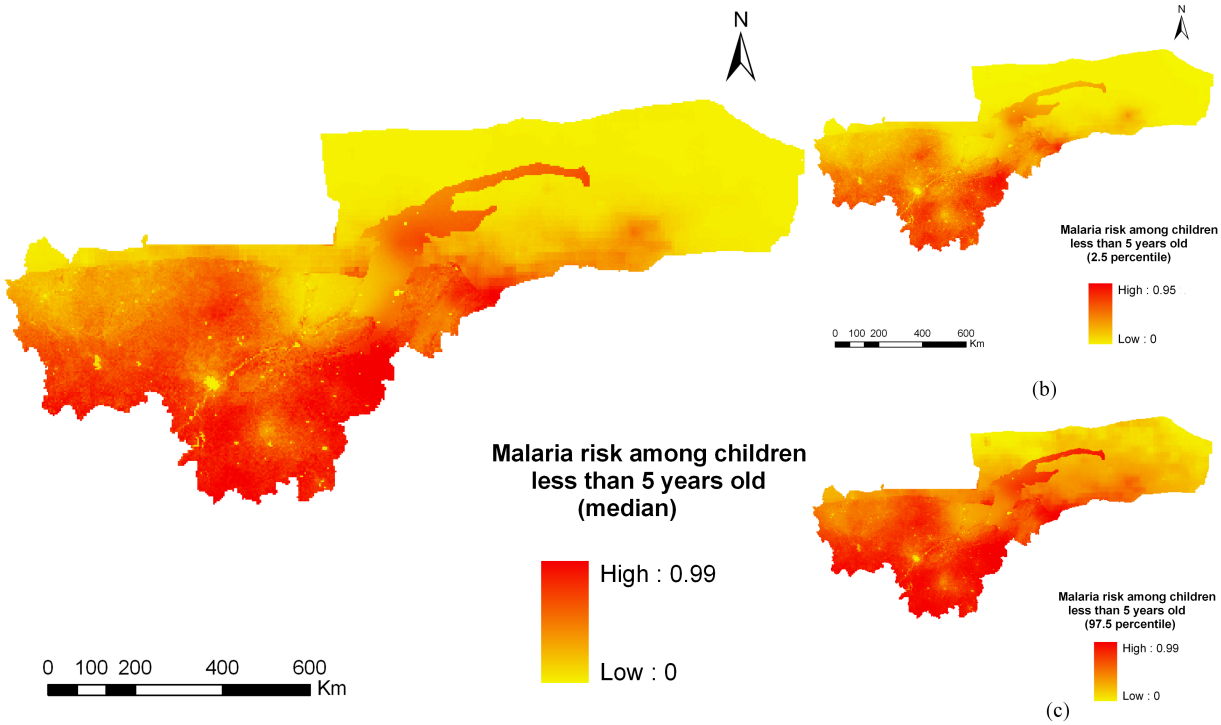


Figure 4.7: Predicted parasitaemia risk in children under 5 years. Map produced using a non-stationary full B-spline model, as in Gosoniu et al. (2009). Median (a) and Credible Intervals (b) and (c).

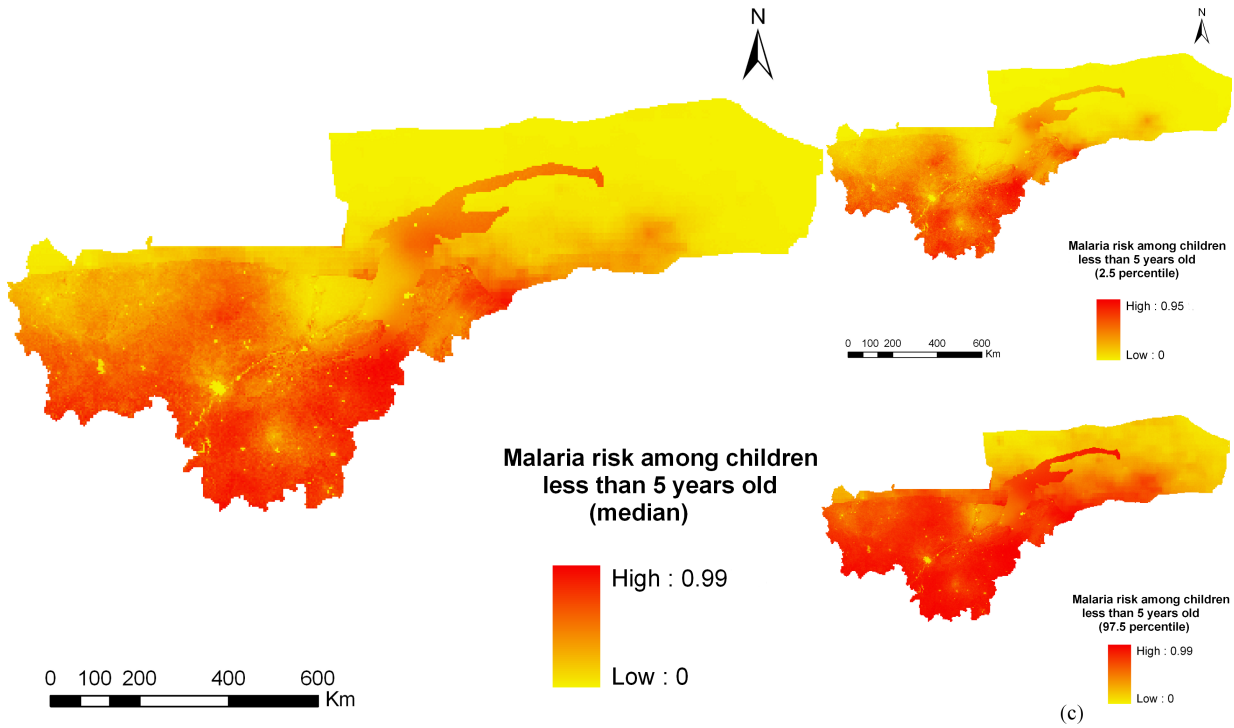


Figure 4.8: Predicted parasitaemia risk in children under 5 years. Map produced using a full B-spline model with stationary covariance structure. Median (a) and Credible Intervals (b) and (c).

smooth the predicted surface at the borders of the zones. Modeling a non-stationary spatial process enables the incorporation of the heterogeneity generated by effects of covariates as well as unmeasured factors that vary geographically in the study area.

The choice of the radius might influence posterior inference even though the weights were normalized. In particular, a large radius could lead to oversmoothing, while a small one may introduce discontinuities. In our model formulation, the radius was fixed during the estimation process; alternatively, it could be considered as a parameter estimated by the data.

Our modelling approach share similarities with other approaches to model non-stationarity such as spatially varying coefficients models (SVC) (Gelfand et al., 2003) and the geographically weighted regression (GWR) (Fotheringham et al., 2003). GWR is commonly seen as a descriptive approach that uses spatial weights to estimate spatially adaptive coefficients whereas SVC places either a univariate or multivariate spatial process on those regression coefficients that are thought to vary spatially (Finley, 2011). GWR has recently been shown to produce biased estimates and its application is not straightforward for generalized models. SVC offer a richer inferential framework at the cost of being computationally demanding. Moreover, identifiability issues may arise from the estimation of several spatially structured covariates.

Our model can be easily implemented in standard software for Bayesian inference (e.g.BUGS) and allows a parsimonious model definition yet leading to best predictive performance. The model in its current formulation does not take into account potential interaction terms between the covariates. The variable selection procedure can be easily extended to identify important interactions.

A natural field of application of the proposed methods is that of spatial epidemiology of environmentally driven diseases, where the study area is often large, contains different ecological zones and the effects of predictors may depend on the zone. Our example is focused on a study of malaria risk in Mali. The malaria endemic area in the country is divided into three different ecological zones. Malaria transmission is influenced by suitable rainfall and temperature that affect mosquito survival and longevity and therefore contribute to abundance of the mosquito population. The Bayesian variable selection procedure identified the most important environmental predictors of parasitaemia risk in each ecological zone. These predictors had meaningful biological interpretation. In particular, our analysis showed that in the Sahel where the amount of precipitation is very low, an increase in the amount of rainfall was associated with an increase in malaria risk. In the flooded region in

the centre of the country, temperature was the most important predictor. In the Sudannian ecological zone vegetation index, which is a proxy of humidity, was identified as the main factor affecting the disease risk. Malaria in Africa is present in both rural and urban areas (Machault et al., 2011) but, as confirmed by our analysis, levels of transmission in urban areas are usually lower than those in peri-urban and rural places. The estimated malaria prevalence map identified high risk areas in the centre (Sigou region) and South of the country (Sikasso region).

Earlier mapping efforts of malaria risk in Mali are based on compilation of historical survey data. Our results are consistent with the ones obtained by Gemperli et al. (2006a) and Gemperli et al. (2006b). A similar pattern is also observed comparing our map with the one produced by Gosoniu et al. (2009) with exception of the parts of the country at the border with Burkina Faso and Côte d'Ivoire. The map of Mali produced by the Malaria Atlas Project (Hay and Snow, 2006) shows similar values of predicted risk in the areas of Sikasso and Sigou, but much lower in Kayes and in the South-East region.

The proposed methodology improves disease risk prediction over large areas compared to commonly used stationary geostatistical models. The described models can be used to address the current needs of international agencies (e.g. World Health Organization, The Global Fund) which are interested in global atlases of infectious disease burden and estimates of the required amount of preventive and curative treatments.

4.6 Aknowledgments

This work was supported by the Swiss-South Africa Joint Research Project No. IZLSZ3 122926. The authors thank Measure DHS for providing the malaria data. Thanks also to Dr. Seydou Doumbia and Dr. Mahamadou Diakité for helpful comments.

Chapter 5

Geostatistical modeling of malaria risk in Mozambique: assessing the effect of spatial resolution of remote sensing-derived environmental variables

Giardina F.[1,2], Franke J.[3], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Remote Sensing Solutions GmbH, Baierbrunn, Germany

This paper has been published in *Geospatial Health* 10 (2).

Abstract

The study of malaria spatial epidemiology has benefited from significant progress in geostatistical modelling as well as recent advances in Geographic Information System (GIS) and (EO) systems that have led to the development of high (HR) and very high (VHR) resolution products. However, few studies have linked malaria survey data with RS-derived land cover/use (LC) variables. In this study, we assess the effect of the spatial resolution of RS-derived environmental variables on malaria risk estimation in Mozambique. We propose a proximity measure to define LC variables to be included as covariate in a geostatistical model. We use data collected in a Demographic and Health survey (DHS) carried out in 2011 throughout the country. We compare the risk predicted using HR (Modis) covariates with the one obtained employing VHR based on elevation measures by the Digital Elevation Model and a LC map produced by MALAREO, a FP7 funded project which covered part of Mozambique during 2010-2012. The number of infected children was predicted using Afripop population data and compared with an "enhanced" population layer derived by the MALAREO LC map.

5.1 Introduction

Malaria remains one of the most important parasitic disease of humans, and a leading cause of morbidity and mortality in the developing world, especially sub-Saharan Africa, where it constitutes a major impediment to economic development.

The study of malaria spatial epidemiology has benefited from the significant progress in the development of Geographic Information System (GIS), computerized systems capable of collecting, storing, handling, analyzing and displaying all forms of geographically referenced information, usually achieved by the Global Positioning System (GPS).

Advances in earth observation (EO) systems, gathering of information about Earth via remote sensing (RS) technologies, have led to the development of high spatial resolution products. The growing availability of RS data, some of them accessible free of charge via the Internet, played a crucial role in determining the environmental predictors of malaria transmission (Ceccato et al., 2005). RS data and spatial statistics have been used for mapping malariometric indices as presence and persistence of vectors' (mosquitoes of the species *Anopheles*) breeding sites, larval densities, the entomological inoculation rate (EIR) as well as malaria prevalence, morbidity and mortality in the human (Machault et al., 2011). The readily available up-to-date information on environmental variables pertinent to malaria transmission over large and remote regions makes RS a useful source of information for identification of pockets of transmission and epidemic early warning systems (EWS). RS can assist malaria control and elimination programs, through the development of spatial decision support systems enabling accurate and timely resource allocation (Clements et al., 2013).

The MALAREO project, (`www.malareo.eu`), supported by the Seventh Framework Programme (FP7) space research program, aimed at building GIS, EO and spatial statistics capacities and implement the use of EO products directly supporting the malaria control programmes (MCP) in Southern Africa. The project focused on the area that corresponds to the geographic region targeted by the Lubombo Spatial Development Initiative (LSDI) launched in 1999 for accelerating development, particularly with regard to agriculture and tourism within an area of approximately 30,000 km^2, covering southern Mozambique, eastern Swaziland, and north-eastern South Africa. The main product created within the MALAREO project is a high resolution (5m) land cover/land use (LC)[1] map based on RapidEye (Blackbridge now) technology. RapidEye indicates both the Earth

[1] Land cover and land use are often mapped together as a result from remotely sensed image, although land cover refers to characteristics of the biophysical Earth surface (e.g. water, vegetation, bare soil, artificial structures, while land use reflects human activities such as agriculture, forestry and urban development (Machault et al., 2011)

observation imagery (Franke et al., 2013) and the information provider focused on assisting in management decision-making.

The LC layer was further classified into malaria-relevant LC classes including wetlands, permanent and flowing water bodies, large scale agriculture, savanna and forests. A high resolution population density map was obtained by the combination of the LC data and census estimates following the approach used for the production of population layers in the Afripop project (Tatem et al., 2007), described in detail in Linard et al. (2011).

LC types have been associated with vector habitats based on simple classification techniques, as well as more sophisticated statistical models that link satellite-derived multitemporal meteorological data and earth observations with vector biology and abundance (Kalluri et al., 2007). A review of studies characterizing LC features and their roles in malaria transmission can be found in Stefani et al. (2013).

Very few studies used LC in mapping of malaria prevalence from survey data. Omumbo et al. (2005) used an LC layer produced by the Africover project (`http://www.africover.org`) from visual interpretation of Landsat digitally enhanced Thematic Mapper(ETM) satellite imagery to map malaria risk in East Africa. The authors defined two ecological zones and used LC classes "water bodies" and "area type" (urban/rural), defining them as the percentage area of each pixel occupied by each class. Craig et al. (2007) regrouped the thirteen United States Geological Survey land cover classes (Anderson, 1976) into two categories, broadly corresponding to drier and moister land cover types in Botswana. Gosoniu et al. (2009) employed LC data from the United States Geological Survey (USGS) and grouped them into the following six categories: urban area, cropland, grass/shrub land/savanna, water bodies, wetland and forest. Both Craig et al. (2007) and Gosoniu et al. (2009) used LC as categorical variable in their models. Riedel et al. (2010) assessed the role of LC, from Moderate-resolution Imaging Spectroradiometer (Modis), in the analysis of malaria indicator survey data (MIS) in Zambia. Five categories were defined: wetlands, forests, urban areas, shrublands and others. At each cluster location, the land cover covariate was summarized by the proportion of each land category within a radius of 3 km. In the above works, associations were found in particular with the "urban" LC class, where the odds of malaria were significantly lower, but the results in general varied by studies.

In this work, we study the effect of the spatial resolution of RS-derived environmental covariates (LC and elevation) and population density on the estimation of malaria risk and number of infected children. Furthermore, we propose a modelling strategy for the LC covariate that allows direct estimation of the effect of each LC class type and we study

associations with malaria risk in a geostatistical model. The data used in the analysis were collected in the malaria module of the Demographic and Health survey (DHS) conducted in 2011 in Mozambique[2] and HR environmental variables were freely available on the Internet. In the area of Mozambique belonging to the LSDI area (approximately 11 km^2 in the southern part of Maputo province), the LC and population density layers were used for model validation. We produce spatially explicit estimates of parasitaemia risk and number of infected children in the whole country and we perform a predictive analysis using very high resolution data (MALAREO products and elevation from DEM) and compare the estimates in terms of log-score (predictive performance) with the lower resolution products. Malaria risk and number of infected children below the age of 5 years were produced in the MALAREO area over grids of 1km, 500m and 100m spatial resolution.

5.2 Materials and methods

5.2.1 Study area

The Republic of Mozambique is bordered by the Indian Ocean to the east, Tanzania to the north, Malawi and Zambia to the northwest, Zimbabwe to the west and Swaziland and South Africa to the southwest. Malaria remains a major cause of morbidity and mortality in the country. It is endemic throughout the country, with regions ranging from mesoendemic to hyperendemic. The climate creates a favourable environment for the main malaria vectors: *Anopheles gambiae*, *arabiensi*, and *funestus* species. *P. falciparum* is the most common parasite and it is responsible for approximately 90% of all malaria infections. The peak of transmission occur during and after the rainy season, between December and April, although malaria is transmitted year round. In the last decade the MCP has implemented large scale IRS programs in several areas of 42 districts (Ministerio da Saúde e Instituto Nacional de Estatística e ICF International, 2013). IRS was also the major component of the Lubombo Spatial Development Initiative (LSDI). Distribution of insecticide treated nets (ITN) and long lasting insecticidal nets (LLIN) targeted all age groups since 2009. Bednet coverage is estimated to have reached almost 40% by 2011 (WHO Malaria report 2012).

[2]Data have been already analyzed elsewhere, without the inclusion of LC classes, see Giardina et al. (2013a)

5.2.2 Data

Malaria data

The DHS in Mozambique was carried out between June and November 2011 and involved around 13000 households. A total of 4885 children was tested for parasitemia with rapid diagnostic test (RDT) and microscopy. Geo-reference and parasitaemia measuraments were available for 603 clusters (groups of households) in the survey.

Remote sensing data

Land surface temperature (LST) data for our analysis were obtained from Modis at 1 km spatial resolution. Dekadal rainfall data were available at 8 km resolution via Africa Data Dissemination Service. Elevation data were obtained from an interpolated digital elevation model from the U.S. Geological Survey - Earth Resources Observation and Science Data Center at a spatial resolution of 1 km. and from the Digital Elevation Modeling (NASA) at very high spatial resolution (30m). The environmental factors with available temporal resolution (LST and rainfall) were acquired for the 3-month period prior to the survey and the average was calculated and extracted for each data location. Afripop and MALAREO population density estimates at 100m resolution were used.

Land cover

The Modis product for LC was aligned to rapid-eye (MALAREO) categories. The allocation was done on the basis of the available description of the layers as well as a graphical assessment. The final categories are summarized in Figure 5.1.

5.2.3 Statistical analysis

LC proximity measure

While for other environmental factors we considered RS-derived values at locations only, we assume that LC classes may affect parasitaemia levels within larger areas surrounding the location. For this purpose, a measure of proximity was used to link LC type with the observed (DHS cluster) and predicted location. In particular, we defined $LC_{ij} = \exp\left(-d^*_{i,LC_j^{cat}}\right), \forall j = 1, \ldots, k$ where $d^*_{i,LC_j^{cat}}$ indicates the minimum Euclidean distance between location i and the LC category j.

Geostatistical model

Let Y_i and N_i be the number of malaria-infected and screened individuals at location i $(i = 1, \ldots, n)$ and p_i the probability of infection. We assume that Y_i arises from a

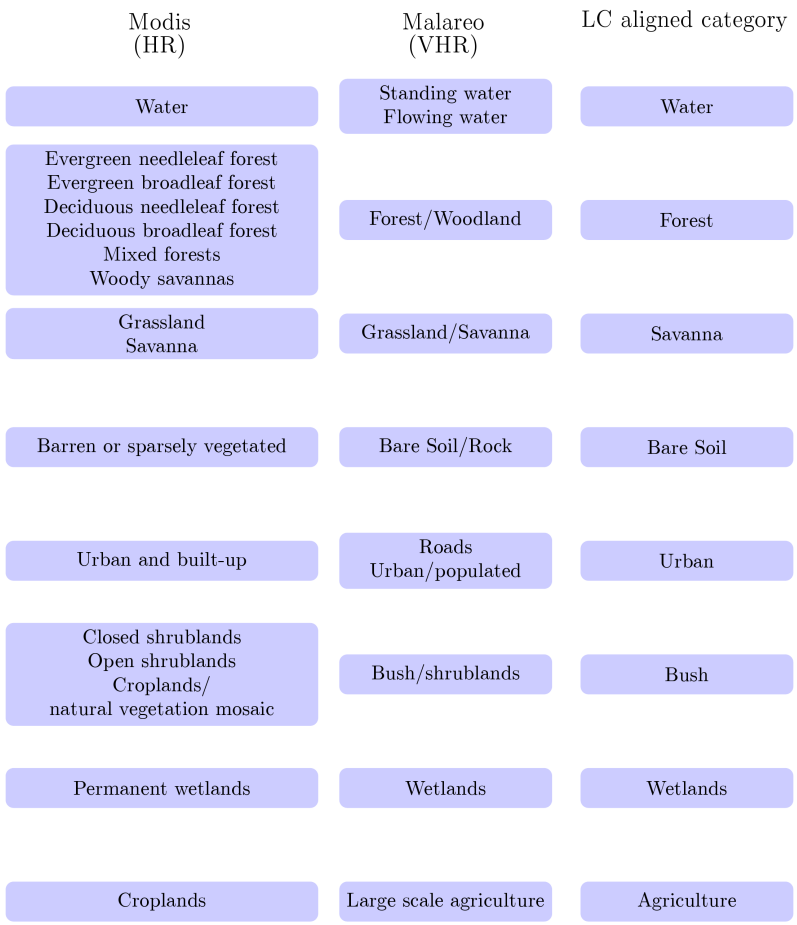


Figure 5.1: LC classes alignment. Classes defined in Modis (first column), classes defined in MALAREO (second column), classes used for the analysis (third column).

Binomial distribution, $Y_i \sim Bin(p_i, N_i)$. The influence of environmental covariates $\boldsymbol{X}_i$ and location-specific spatial random effects ω_i are modelled on the logit scale, i.e. $log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \omega_i$, where $\boldsymbol{\beta}$ is the vector of regression coefficients. Unobserved spatial variation is introduced on ω_i by assuming that $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ follows a latent stationary Gaussian process over the study region, $\boldsymbol{\omega} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. The matrix $\boldsymbol{\Sigma}$ has elements Σ_{ij} and represents the covariance between any pair of locations i and j. Assuming an isotropic exponential correlation function, the matrix elements Σ_{ij} are defined by $\Sigma_{ij} = \sigma^2 exp(-\rho d_{ij})$ with spatial variance σ^2, rate of correlation decay ρ with Euclidean distance between locations d_{ij}. The minimum distance for which the spatial correlation is less than 5% is referred to as range and can be calculated by $3/\rho$ in the exponential correlation function setting.

A Bayesian model formulation requires the specification of prior distributions of all model parameters. For the regression coefficients $\boldsymbol{\beta}$, we assumed Normal prior distributions with mean 0 and large variance. For the spatial parameters σ^2 and ρ, we chose non-informative inverse Gamma and Gamma distributions, respectively.

The model was fitted using MCMC simulation implemented in the software JAGS (Just Another Gibbs Sampler, Plummer (2003)). Spatially explicit estimates of the malaria risk and number of infected were obtained through the predictive distributions over a grid formed by pixels of 3km resolution.

Assessing the effect of spatial resolution on model-based predictions

The model was validated using as *training set* all DHS data except the 35 locations belonging to the MALAREO area (Figure 5.2), which formed the *testing set*. The model used HR variables in the fitting part and HR as well as VHR variables in the prediction, see Table 5.1. Model performance was compared in terms of log-predictive density (Robert, 1996). Spatially explicit predictions (malaria risk and number of infected) were obtained over grids covering this area with spatial resolution of 1km, 500m and 100m using both HR and VHR variables.

5.3 Results

The effect of the environmental and climatic factors on parasitaemia risk estimated from the full DHS dataset is shown in Table 5.2. The main determinants of malaria were rainfall and LSTD. Among the LC classes, the presence of large scale agriculture and bare soil reduced the odds of parasitaemia by 8% (95%BCI: 0-15%) and 44% (95%BCI: 26-60%),

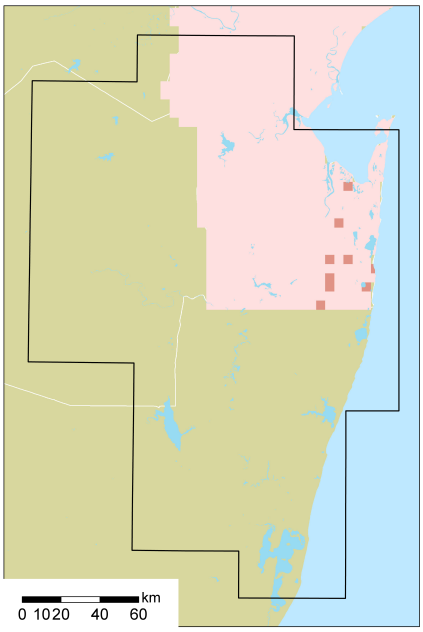


Figure 5.2: MALAREO project area. The area includes the Northern part of South Africa (KwaZulu-Natal province), eastern Swaziland and the Southern part of Mozambique.

Table 5.1: RS-derived environmental variables. Sources and spatial resolution for HR and VHR covariates used.

Variable	Source/Product (HR)	Sp. resolution	Source/Product (VHR)	Sp. resolution
LC	Modis (mcd12q1)	500m	Rapid Eye	5m
Elevation	Modis	100m	DEM	30m
LST	Modis (mod13a2)	1km	–	–
Rainfall	MEFW (ADDS)	8km	–	–
Population	–	–	Afripop (Landsat)	100m
Population	–	–	MALAREO (RapidEye)	100m

respectively. The presence of bush, forest, savanna and wetlands increased the odds of parasitaemia by 31% (95%BCI: 21-42%), 11% (95%BCI: 4-19%), 34% (95%BCI: 18-46%) and 37% (95%BCI: 55-75%). The estimates of the spatial parameters revealed a variance of 2.61 (95%BCI: 1.64-2.82) and a spatial range (the distance at which the correlation becomes negligible) of around 85.56 km (56.22-127.32).

The same model was used to predict malaria risk among children between the age of 0 and 5 years, over a grid of 3km resolution. Figure 5.3 shows that the two provinces with the highest malaria risk were Nampula and Zambezia, in the northern part of the country. The southern parts of the country were characterized by lower risk compared to the rest of the country (<10%), especially Maputo (city and province) and Gaza province.

Table 5.2: Posterior estimates arising from the geostatistical model fitted on the full DHS dataset with Modis LC. LC categories refer to the aligned variable.

Covariate	Median (95% BCI)
Rainfall	0.14(0.07, 0.22)
LSTN	-0.11(-0.40,0.16)
LSTD	0.31(0.09,0.54)
Elevation	-0.03(-0.14,0.07)
LC category	
Agriculture	-0.09(-0.17,-0.01)
Bush	0.27 (0.19,0.35)
Forest	0.11 (0.04,0.18)
Savanna	0.30(0.17,0.45)
Urban	0.05(-0.16,0.41)
Water	0.09(-0.2,0.40)
Bare soil	-0.59(-0.91,-0.30)
Wetlands	0.44(0.32,0.56)
Spatial parameter	Median (95% BCI)
σ^2	2.61(1.64,2.82)
ρ	2.31(1.51,3.43)

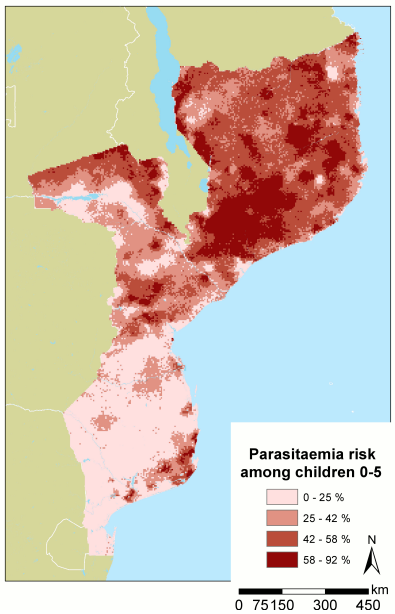


Figure 5.3: Predicted malaria risk among children under the age of 5 years. Median estimates are plotted at 3km resolution.

Estimates of the number of children infected by malaria parasites were obtained from the predictive distribution of the risk and population data at 100m spatial resolution provided by Afripop (Figure 5.4). In most of the country the number of infected children per 9 km^2 ranges from 1 to 10. In some densely populated areas, e.g. Maputo and Matola cities, and in very high risk areas, e.g. Zambezia province, the number can reach the value of 1800 children.

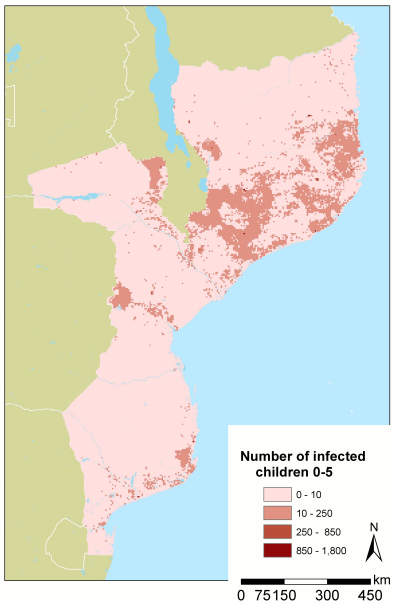


Figure 5.4: Predicted number of malaria infected children under the age of 5 yers. Median estimates are plotted at 3km resolution.

The model validation revealed that the use of VHR covariates in the 35 testing locations improved prediction performance. In particular, the model that employed the MALAREO layer for LC and the DEM values for altitude had a log-predictive density of -115.12 (95%BCI: -122.32,-104.21) whilst the model that used HR covariates (-132.22 (95%BCI: -143.11,-121.17)).

Predictions in the same area were carried out at several spatial resolutions. Figure 5.5 depicts the predicted malaria risk among children aged 0-59 months at 1km, 500m and 100m resolution using HR and VHR data. Table 5.3 shows how the estimated number of infected children is affected by the population layer (and, indirectly by the spatial resolution of the environmental covariates). On average, the total number of infected children estimated by the models increased with increasing resolution of the predictive grid. The use of HR variables tended to result in an overestimation in the number of infections.

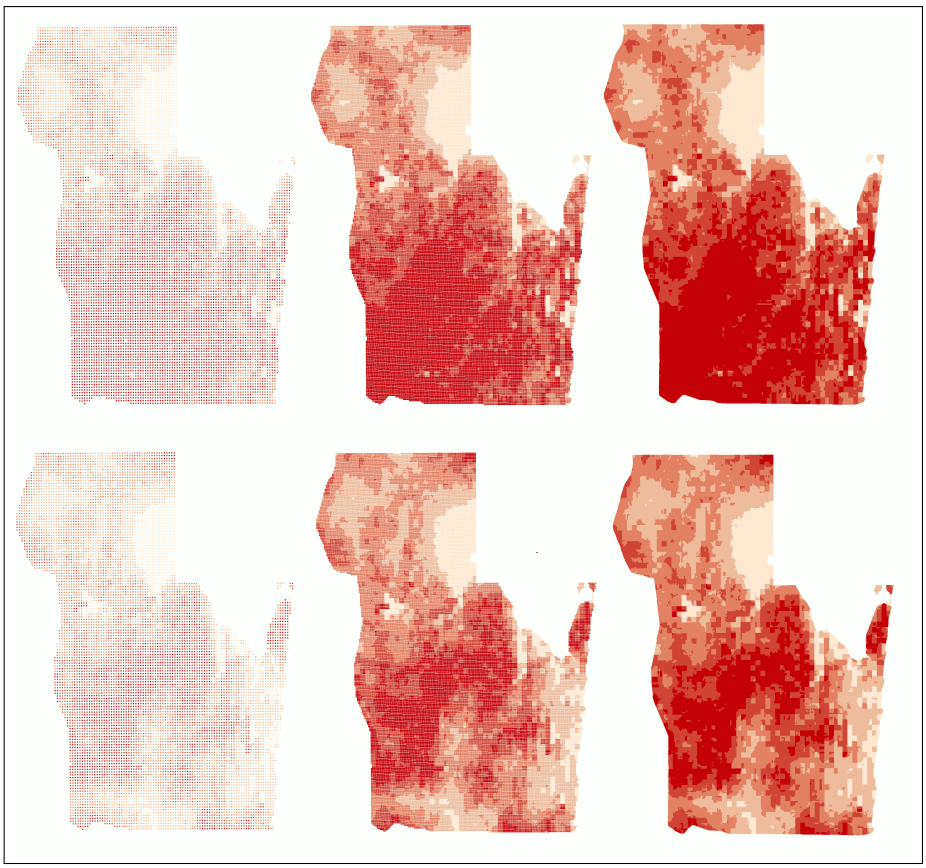


Figure 5.5: Predicted malaria risk (median) obtained by model with HR covariate (first row) and VHR covariates (second row). Spatial resolutions: 1km resolution (first column), 500m resolution (second column) and 100m resolution (third column).

Table 5.3: Estimated total number of infected children in the MALAREO area (median and 95% BCI) using HR and VHR products.

	1km	500m	100m
HR	43,554.66	45,170.95	45,605.47
	(42,334.51 - 44,234.32)	(44,524.77 - 46,123.47)	(44,532.44 - 46,892.33)
VHR	37,900.79	37,919.03	38,110.79
	(36,884.44 - 38,423.94)	(37,011.02 - 38,625.77)	(37,773.33 - 39,100.43)

5.4 Discussion

This study focuses on the use of HR and VHR RS-derived variables to obtain spatially explicit malaria burden estimates in geostatistical models. In particular, the work shows the effect of different spatial resolutions of elevation and LC layers (and derived population estimates) on the estimation of risk and number of infected children below the age of 5 years. Moreover, an alternative definition of the LC covariate based on a proximity measure is

proposed to study associations of 8 different LC types with malaria risk and obtain explicit effect estimation.

The analysis has been performed on data collected by the Mozambican DHS in 2011 with a geostatistical model utilizing HR and VHR RS environmental variables. The model was fitted with HR variables and a grid of 3km resolution was chosen in the prediction for comparison purposes with a previous work (Giardina et al., 2013a). The coefficients' estimates of the common variables (Rainfall, LSTN and LSTD) were in agreement with Giardina et al. (2013a) as well as the total malaria burden measure (number of infected children). The spatial parameters estimates (variance and decay parameter) also showed similar values.

A relatively small number of studies have included LC classes in geostatistical models for malaria risk mapping despite its important role in determining the suitability for transmission of the disease. This may be due to difficulties in the definition of the variable to be used in the models. Introducing LC covariate as relative frequencies of each group category within a buffer, is perhaps conceptually the best way of defining the variable but it has some drawbacks. For example, parameter estimates have to be expressed relatively to a baseline category and there is a certain arbitrariness in the choice of the reference category as well as the size of the buffer. Here we have proposed a proximity measure that does not account for the area covered by a specific LC class surrounding the locations, but it is based on the distance between locations and each LC classes. This work showed that "wetlands" and "bare soil", were important risk and protective factors in malaria modeling, respectively. The effect of large scale agriculture on malaria risk has always been controversial: it has often been assumed that a high number of malaria vectors resulting from irrigation schemes lead to increased malaria in local communities. However, recent studies in Africa have revealed that for many sites there is less malaria in irrigated communities than surrounding areas. It has been suggested that many communities near irrigation schemes benefit from the greater wealth created and consequently they have greater use of bednets, better access to improved healthcare and receive fewer infective bites compared with those outside such development schemes (Ijumba and Lindsay, 2001).

The MALAREO project took place during the period 2010-2012 in the LSDI area, therefore covering part of Mozambique. Within the MALAREO project a VHR LC map covering the study area at 5m resolution was produced. A secondary outcome was an "enhanced" population map, obtained by the combination of the LC layer with census data, aggregated at 100m resolution. Modis LC categories (HR) were aligned with the

MALAREO LC categories and used for validation purposes in the prediction of the parasitaemia risk at locations belonging to the MALAREO area. The comparison showed that the model which used VHR products (MALAREO LC and DEM elevation) had a higher predictive ability than the one that used HR data. Spatially explicit estimates over the grids of 1km, 500m and 100m showed large differences in the risk and in its spatial pattern. However, our results may be sensitive to different allocation of Modis categories to the final variable used for the model. The Modis LC layer is based on a global classification methodology and may miss some local features: in particular, the "wetland" category showed the largest differences in the comparison with the MALAREO layer. On the other end, the MALAREO LC categories were assigned by a "supervised" algorithm (people drove through the mapped areas) that allowed a more detailed description of the soil, that would not possible with Modis. However, VHR products like the MALAREO LC are very expensive and may not be feasible over large areas.

In this study, the estimated total number of infected children increased with increasing resolution of the predictive grid independently on the spatial resolution of the covariates used for prediction. The use of HR variables tended to result in an overestimation of the number of infections. Observed differences between the 1km resolution and the 500m resolution grid using HR covariates are the result of aggregation of environmental covariates as well as population density over larger areas, however the differences between the 500m and 100m resolution grid were only due to population density, as the Modis LC original resolution was 500m.

Accurate estimation of malaria parasitaemia risk has important implications on the planning of cost-effective control measures such as distribution of insecticides treated nets and indoor residual spraying. The estimation of number of infected can support NMCPs in the determination of treatment needs.

Chapter 6

Bayesian meta-analysis of heterogeneous geo-referenced disease survey data via transmission models

Giardina F.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland

This manuscript has been prepared in collaboration with MACEPA.

Abstract

Spatially explicit disease burden estimation is essential for public health policy. Information on disease burden is often combined from multiple sources, such as research studies that may differ from one another in their design. For example, disease surveys that aim at estimating prevalence may be heterogeneous in the sampling methods (e.g. random or preferential), diagnostic tools used, age groups of sampled individuals, spatial and temporal factors. Little has been done in the context of prevalence estimation from the combination of heterogeneous geo-referenced surveys. In this work, we address age and time heterogeneity between surveys by proposing a general formulation that couples spatial statistical models and mathematical transmission models. Our methodology is applied in the area of malaria mapping to obtain age and season specific high resolution disease risk estimates in Zambia.

6.1 Introduction

Estimating disease burden at high spatial resolution has become of increasing importance for decision and policy makers. Accurate morbidity and mortality figures disaggregated in space and time allow the identification of geographical and seasonal pattern of the disease and result in an enhanced understanding of the latter, guiding allocation of public health resources.

Information on the disease burden is often combined from multiple sources, such as research studies that may differ from one another in their design. For example, disease surveys that aim at estimating prevalence may be heterogeneous in the sampling methods (e.g. random or preferential), diagnostic tools used, age groups of sampled individuals, spatial and temporal factors. However, combining information from multiple surveys can improve estimates quality (i.e. reduce bias and improve precision); the work by Turner et al. (2009) discussed potential sources of internal and external bias and illustrated methodological development in the meta-analysis of multiple survey data. Combination of randomized and non-randomized surveys were proposed by Hedt and Pagano (2011) in the estimation of prevalence with a simple annealing methodology. In spatial settings Manzi et al. (2011) combined surveys to obtain improved small area statistics with a Bayesian hierarchical model which allows for additive bias.

Little has been done in the context of geostatistics where it is often necessary to combine information from multiple prevalence surveys in order to obtain high resolution risk estimates. The work by Crainiceanu et al. (2008) estimated *Loa loa* risk in space by combining morbidity questionnaires with parasitological surveys after calibrating the relationship between the diagnostic tool specific prevalences. Wang et al. (2008) adjusted for the bias arising from surveys that used different diagnostic tools by incorporating specificity and sensitivity parameters in a geostatistical model. Giorgi et al. (2013) described a general class of models to correct for spatially structured bias in random and preferential sampled surveys allowing for temporal variation in prevalence between consecutive survey-periods.

In this work, we address age and time heterogeneity between surveys by proposing a general formulation that couples geostatistics and mathematical transmission models. Our methodology is applied in the area of malaria mapping to obtain high resolution disease risk estimates in Zambia.

Malaria is a mosquito-borne disease of major public health importance, widespread throughout the tropical and subtropical regions, including parts of Africa, Asia and Americas. It is a leading cause of illness and death in large areas of the developing world,

especially Africa (WHO, 2012). Prevalence is the most widely available measure of malaria endemicity: a large number of community-based surveys assessing the presence of parasites in the blood have been conducted over the past decades. However, parasite prevalence surveys have been carried out in different areas and seasons sampling different and sometimes overlapping age groups population making their estimates not directly comparable. Space and time represent an important source of heterogeneity in prevalence measures, since malaria's main drivers are environmental and climatic conditions. In fact, factors like temperature and rainfall determine habitat suitability for the vectors, mosquitoes of the *Anopheles* species. In some African countries, the extreme changes due to the alternation of dry and wet periods, determine the seasonal nature of the disease. Moreover, the acquisition of partial immunity in older children and adults in endemic areas leads to age-dependence of prevalence measures.

Early work on geostatistical models of historical survey data for malaria was boosted by the Mapping Malaria Risk in Africa (MARA) project (Craig et al., 1999), a multi-country partnership that provided the most comprehensive source of malariometric data across Africa, containing age-specific geo-referenced prevalence data assembled from published and gray literature since the early 1960s until 2009. Kleinschmidt et al. (2001),Gemperli et al. (2006a) and Gosoniu et al. (2006) focused on a specific age group, discarding surveys with different or overlapping ages of the population resulting in unreliable malaria transmission estimates. The Garki malaria transmission model (Dietz et al., 1974) was employed by Gemperli et al. (2006b) and Gosoniu (2008) to convert observed prevalence data gathered from heterogeneous surveys into an estimated age-independent entomological measure of transmission intensity, which was further used for mapping purposes.

As part of the Malaria Atlas project (MAP), Hay et al. (2009) produced a continuous, global, malaria endemicity surface due to the main parasite species, *P. falciparum*, in 2007 using historical survey data (Moyes et al., 2013). The authors employed a Bayesian formulation of the catalytic model by Pull and Grab (1974), which allowed for age-specific prevalence estimation (Smith et al., 2007). The model assumed a space/time independent force of infection. Age-adjusted prevalence estimates were fitted separately in a geostatistical model. Therefore, the works Gemperli et al. (2006b), Gosoniu (2008) an Hay et al. (2009) were based on a 2-step procedure that did not account for the uncertainty arising from the age-standardization model. Moreover, these approaches ignored the survey period as source of heterogeneity in the prevalence measures.

Here we propose a Bayesian joint formulation for prevalence and incidence data by

embedding a modified version of the catalytic model by Pull and Grab (1974) into a geostatistical model. Furthermore, we extend the age-standardization procedure outlined in Smith et al. (2007) and Hay et al. (2009) into an age/season-standardization by allowing the force of infection to vary in space and time. In particular, the model estimates a latent force of infection that can vary geographically on a monthly basis, and links it to the fraction of population with parasitaemia through a function of epidemiological and local detection parameters. The estimation of these parameters permits the calculation of spatially varying age-prevalence curves. Geo-referenced malaria prevalence surveys can therefore be combined accounting for season and age heterogeneity and the spatial residual structure can be estimated. In summary, our approach combines a mechanistic model for malaria transmission and empirical estimation from data in a Bayesian framework, allowing the standardization of surveys to a unique (or a combination of defined) age category and season and the estimation of malaria burden at high spatial resolution.

We illustrate the methods using prevalence data from Zambia extracted by the MARA database accessible online (`http://www.mara-database.org/`) as well as data collected in the national malaria indicator survey (MIS) in 2006 that assessed malaria parasitaemia in children under five years of age. The latter has been already analyzed elsewhere (Riedel et al., 2010) using a geostatistical model. Confirmed malaria cases gathered at health district level by the health management information system (HMIS) in Zambia (Chanda et al., 2012) were used to estimate the force of infection.

6.2 Bayesian Hierarchical model

6.2.1 Modelling prevalence

Let $Y^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$ and $N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$ denote, respectively, the number of positives and the total number of screened at location $s \in \mathbb{S} \subseteq \mathbb{R}^2$ for each survey $i, \forall i = 1, \ldots, n$ carried out during months $[t_{i1}, t_{i2}]$ whose age target population ranged from a_{i1} to a_{i2}. We assume that $Y^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$ follows a Binomial distribution with parameters $N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$ and $\pi^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$, thus

$$Y^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s) \sim Binomial\left(N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s), \pi^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)\right)$$

Smith et al. (2007) expressed the disease risk $\pi^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$ for the specific age group $[a_{i1}, a_{i2}]$ by the ratio $\frac{\sum_{a=a_{i1}}^{a_{i2}} p^a_{it}(s)}{F(a)q_{is}(a)}$which is the product of $p^a_{it}(s)$, i.e. the true proportion of

infected people of age a at time t and location s, the age-dependent probability of parasitaemia detection, $F(a)$, and the number of sampled people of age a in survey i, $q_{is}(a)$, over the total number of screened $\sum_{a=a_{i1}}^{a_{i2}} q_{is}(a) = N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s)$. In the original formulation by Pull and Grab (1974) $p^a_{it}(s)$ is obtained by

$$p^a_{it}(s) = \frac{m}{(m+r)}\left(1 - \exp\left\{-a(m+r)\right\}\right) \tag{6.1}$$

where m is a constant force of infection and r represents the recovery/clearance rate of infection. Equation (6.1) is the solution of the catalytic model expressed by the differential equation $dp^a/da = m(1-p^a) - rp^a$ with initial conditions $p^0 = 0$. This choice implies $p^a_{it}(s) = p^a, \forall i \in 1,\ldots,n, \forall s \in \mathbb{S}, \forall t \in 1,\ldots 12$. Smith et al. (2007) proposed to model the age-dependent probability of parasitaemia detection (sensitivity) with a function of the following form:

$$F(a) = \begin{cases} 1 & a < \alpha_c \\ 1 - k(1 - \exp\left\{-c(a-\alpha_c)\right\}), & a \geq \alpha_c \end{cases}$$

where α_c represents the threshold age after which the sensitivity starts decreasing from 1 to the associated asymptotic value $1-k$ with a decline described by c. The function $F(\cdot)$, was motivated by the notion that sensitivity declines with age as blood-stage immunity reduces parasite densities to a point where they are often below the detection thresholds of microscopy (McKenzie et al., 2003). The total sampled population in survey i at location s carried out during months $[t_{i1}, t_{i2}]$ with age target population from a_{i1} to a_{i2}. is $N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}$ but the age-specific sample distribution $q_{is}(a)$ is usually unknown. Hay et al. (2009) assign $\boldsymbol{q}_{is}$ a Dirichlet-Multinomial distribution where a probability vector $\boldsymbol{f}_i$ is drawn from a Dirichlet distribution with known parameter vector $\boldsymbol{\theta}_i$ and $\boldsymbol{q}_{is}$ is the discrete sample drawn from the multinomial distribution of probability vector $\boldsymbol{f}_i$

$$\boldsymbol{q}_{is}|\boldsymbol{f}_i \sim Multinomial(N^{[a_{i1},a_{i2}]}_{[t_{i1},t_{i2}]}(s), \boldsymbol{f}_i)$$

$$\boldsymbol{f}_i|\boldsymbol{\theta}_i \sim Dirichlet(\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ assigns weights to the age classes for survey i that range from a_{i1} to a_{i2}, (the length of vector $\boldsymbol{q}$ and $\boldsymbol{f}$ depends on the survey).

We extend the above formulation by letting $p^a_{it}(s)$ vary over space and time, as a function

of the force of infection $m_t(s)$. In particular, Equation (6.1) becomes

$$p_{it}^{a}(s) = \frac{m_t(s)}{(m_t(s)+r)}\left(1-\exp\left\{-a(m_t(s)+r)\right\}\right) \tag{6.2}$$

and we assume

$$logit\left(\pi_{[t_{i1},t_{i2}]}^{[a_{i1},a_{i2}]}(s)\right) = logit\left(\frac{\sum_{t=t_{i1}}^{t_{i2}}\sum_{a=a_{i1}}^{a_{i2}} p_{it}^{a}(s)F(a)q_{is}(a)}{(t_{i2}-t_{i1}+1)\sum_{a=a_{i1}}^{a_{i2}} q_{is}(a)}\right)+\omega(s)$$

where $\omega(s)$ is a zero-mean latent Gaussian process with covariance function $Cov(\omega(s),\omega(s')) = \psi^2\exp(-\rho d(s,s'))$ being $d(s,s')$ the Euclidean distance between locations s and s' and ρ decay parameter and ψ^2a variance parameter.

6.2.2 Modelling incidence

We indicate with Z_{jtk} the number of confirmed malaria cases in the general population living in district j (second administrative level) $\forall j \in 1\ldots,n_d$ in month $t, t = 1,\ldots,12$ and year $k, k = 1,\ldots,K$. The environmental covariates are available on a grid that is a refinement of the administrative division for which the response variable Z_{jtk} is available. Following the approach proposed by Zhu et al. (2000), we model the misalignment arising from the different geographical scales between the disease outcome and the covariates, as follows:

$$Z_{jtk}|\beta,a,b,\omega_i,\phi_t,\epsilon_k \sim Poisson\left(\sum_{l=1}^{L_j} P_{jltk}\exp(\eta_{jltk})\right)$$

where l index the subregions of district j and the mean η_{jltk} can be decomposed in the sum of a monthly average μ_{jlt} and a set of year-specific random effects ϵ_k, $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_K)$, i.e. $\eta_{jlkt} = \mu_{jlt} + \epsilon_k$ and

$$\mu_{jlt} = \boldsymbol{X}'_{jlt}\boldsymbol{\beta} + \sum_{t=1}^{12}\gamma\sin\left(\frac{2\pi t}{T}\right)+\delta\cos\left(\frac{2\pi t}{T}\right)+\phi_j+\zeta_t$$

The set of p covariates $\boldsymbol{X}'$ is composed by monthly averages over the years in each subregion l of district j and $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients; the sum of periodic functions models the seasonality in the reported cases ($T = 12$ corresponds to one cycle

of transmission within each year). We consider the population P_{jltk} constant across the years, i.e. $P_{jltk} = P_{jlt}$.

We specify independent Gaussian priors for each regression parameter β, and seasonal terms γ and δ. i.e. $\beta, \gamma, \delta \sim N(0, h_1)$. The set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{72})$ are defined as the sum of a spatially unstructured term $\boldsymbol{\gamma}$ such that $\forall j, \gamma_j \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 \sim U(0, h_2)$ and a conditionally auto-regressive term $\boldsymbol{\theta}$, i.e., $\forall j, \theta_j | \boldsymbol{\theta}_{-j} \sim N\left(\frac{\sum_{j'=1}^{72} v_{jj'}\theta_j}{\sum_{j'=1}^{72} v_{jj'}}, \frac{\tau^2}{\sum_{j'=1}^{72} v_{jj'}}\right)$ where $\tau^2 \sim U(0, h_3)$ and v are binary weights based on geographical contiguity: $v_{jj'} = 1$ if districts (j, j') share a common border (denoted $j \sim j'$), and zero otherwise. We adopt Gaussian prior distributions on month- and year- specific random effects, that is $\zeta_t \overset{iid}{\sim} N(0, \nu^2), \forall t = 1, \ldots, 12$ and $\epsilon_k \overset{iid}{\sim} N(0, \lambda^2), \forall k = 1, \ldots, K$ respectively (where $\nu^2, \lambda^2 \sim U(0, h_4)$).

The force of infection $m_t(s)$ in Equation (6.2) is approximated by the year-independent mean incidence μ_{jlt} where l is the pixel that includes s.

6.2.3 Age/season specific prevalence estimation and spatial kriging

Simulating from the posterior distributions of the parameters, prevalence estimates can be aligned to a unique age group $[a_1^*, a_2^*]$ and months $[t_1^*, t_2^*]$ with the following relation:

$$\pi_{[t_1^*, t_2^*]}^{[a_1^*, a_2^*]}(s) = \left(\frac{\sum_{t=t_1^*}^{t_2^*} \sum_{a=a_1^*}^{a_2^*} p_t^a(s) F(a) q_{is}(a)}{(t_2^* - t_1^* + 1) \sum_{a=a_1^*}^{a_2^*} q_{is}(a)} \right) logit^{-1}(\omega(s))$$

where $\boldsymbol{q}_{is}$ is the vector of size $[a_2^* - a_1^* + 1]$ sampled from th Multinomial distribution of paramter $\boldsymbol{\theta}$. Bayesian kriging is performed to obtain age and season adjusted prevalence maps.

6.3 Application to malaria prevalence surveys

6.3.1 Malaria data

The MARA database contained 92 distinct geo-referenced parasite prevalence random surveys since 1978. Around the 20% of the surveys were conducted during the first months

of the year, where the rains are abundant and the peak of transmission occurs. The rest of surveys was spread throughout the year, including very dry months. The majority of surveys (53%) targeted school age children but wide information is available on adults as well. Parasitaemia was assessed with microscopy. The MIS was carried out from May to June 2006, after the end of the rainy season, and sampled 109 geo-referenced clusters (group of households). To estimate the force of infection we have used malaria incidence information. These surveillance data consist of laboratory confirmed malaria cases and are gathered routinely by the national HMIS in Zambia since 2009. Data are reported monthly in each one of the 72 districts. We have considered 3 years (from January 2009 to December 2011).

6.3.2 Environmental data derived from remote sensing sources

Data on potential environmental predictors for malaria are available from remote sensing (RS) sources at high spatial and temporal resolution. Day and night land surface temperature (LST) as well as normalized difference vegetation index (NDVI) were downloaded from the Moderate Resolution Imaging Spectroradiometer (MODIS), maintained by the United States Geological Survey (USGS) Land Processes Distributed Active Archive Center (LP DAAC). Precipitation data are available through the Africa Data Dissemination Service (ADDS), an operational part of the Famine Early Warning Systems Network (FEWS NET). Altitude data were downloaded by the digital elevation model derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery. Water bodies were obtained by freely available ARCGIS layers. Population data at high spatial resolution are available from a combination of RS sources and census data (Tatem et al., 2007).

6.3.3 Implementation

Environmental and climatic variables were re-sampled over a grid of 3 km resolution and summarized as monthly averages. To account for the elapsing time between the climatic suitability for malaria transmission and number of reported cases, precipitation was considered as the cumulative amount in the 2 months prior to the reported cases. The 3 km resolution grid, resulting in around 100000 pixels, was used to impute the mean incidence η_{jltk} and considered adequate to approximate the value of the same at location s falling in pixel l of district j. The same grid was used to obtain age/season-adjusted prevalence predictions for mapping.

The age a was discretized in 1-year intervals from 0 to the age of 29 years and 5-year intervals from 30 to 75 years. The hyperparameter vector $\boldsymbol{\theta}$ was set to the values reported by the United Nations population division estimates (UN, 2012b). The prior distributions that were assigned to the epidemiological and local detection parameters were fairly informative since they represent characteristic quantities in malaria and there is available literature on their range of values, e.g. (Smith et al., 2007; Ross and Smith, 2010). In particular, the asymptotic sensitivity was assigned a Beta prior distribution with shape and scale parameters in order to set the median at 0.50 and the 0.975-th quantile at 0.90, the age of immunity acquisition was assumed uniformly distributed between 5 and 12 (years), and c was given a Gamma distribution of prior mean 0.05 (years^{-1}) and variance ten times larger than the mean. The recovery rate r was centered around 1.8 (years^{-1}) (it takes around 200 days to clear an infection (Smith et al., 2005)).

The hyperparameters h_1, h_2, h_3, h_4 were set to the value 100 to impose non-informative priors on the variance parameters and coefficients. A Markov chain Monte Carlo (MCMC) algorithm was coded in `R` to sample from the unknown quantities and two parallel chains were run for 100000 iterations. Convergence was monitored with the Geweke test and the Gelman's and Rubin's test implemented in the `CODA` package (Plummer et al., 2006).

6.3.4 Model fitting

Fitting the joint model provides us with the estimation of the temporal and spatial pattern of malaria incidence in Zambia as well as related risk factors. Furthermore, epidemiological and local detection parameters characterizing, respectively, the catalytic model and the sensitivity function are estimated allowing the link with the prevalence model. Jointly with the spatial structure governing the prevalence data, risk estimates adjusted for age and season heterogeneity are obtained.

Malaria incidence in Zambia is highly governed by environmental and climatic factors, in particular NDVI and elevation, as given by Table 6.1.

Seasonality in transmission is strong, as can be seen in Figure 6.1 which shows the fitted monthly mean incidence at district level. The peak of transmission estimated by the model is between November and April. Figure 6.1 depicts the spatial distribution of malaria incidence for selected 4 months. Posterior estimates of epidemiological and local detection parameters are reported in Table 6.2. Spatial variance and decay parameter estimates for the geostatistical prevalence model can be seen in Table 6.3.

Table 6.1: Posterior estimates: environmental factors and spatio/temporal parameters affecting malaria incidence. Covariates were standardized for comparison purposes.

Covariate	Posterior estimate (median, 95%CI)
NDVI	$2.01(1.45, 2.32)$
Rainfall	$0.24(-0.09, 0.42)$
Distance to water bodies	$-0.15(0.11, 0.21)$
LST	$0.12(-0.03, 0.23)$
Elevation	$-0.44(-0.51, -0.31)$
cos	$0.11(-0.05, 0.27)$
sin	$0.55(0.38, 0.73)$
Spatio-temporal parameters	Posterior estimate (median, 95%CI)
$\sigma^2(iid)$	$0.53(0.12, 0.76)$
$\tau^2(CAR)$	$0.85(0.21, 1.43)$
$\nu^2(month)$	$1.43(0.82, 2.31)$
$\lambda^2(year)$	$0.92(0.45, 1.12)$

Table 6.2: Epidemiological and local detection parameters.

Parameter	Posterior estimate (median, 95%CI)
α_c	$8.41(5.13, 9.72)$
$1-k$	$0.42(0.28, 0.61)$
c	$0.11(0.08, 0.13)$
r	$1.71(1.60, 2.01)$

6.3.5 Age/season-specific and spatially-explicit risk estimation

With the purpose of illustrating the results obtained with the proposed methodology, we used the posterior samples of the parameters to obtain season and age specific prevalence estimates during September (low transmission) and December (high transmission) for two age groups (1-4 years and 5-14 years). Figure 6.3 (a) and Figure 6.4 (a) depict the mean incidence at district level, Figure 6.3 (b) and Figure 6.4 (b) show the estimated mean incidence at pixel level, Figure 6.3 and Figure 6.4 (c)-(d) are adjusted risk estimates by using the fitted age-prevalence curves and spatial kriging on the same grid.

Figure 6.2 depicts the age-prevalence curves obtained from posterior samples in two different pixels at different mean incidence levels with uncertainty bounds.

Figure 6.1: Model-based mean incidence at district level. 3. March, 6. June, 9. September and 12. December.

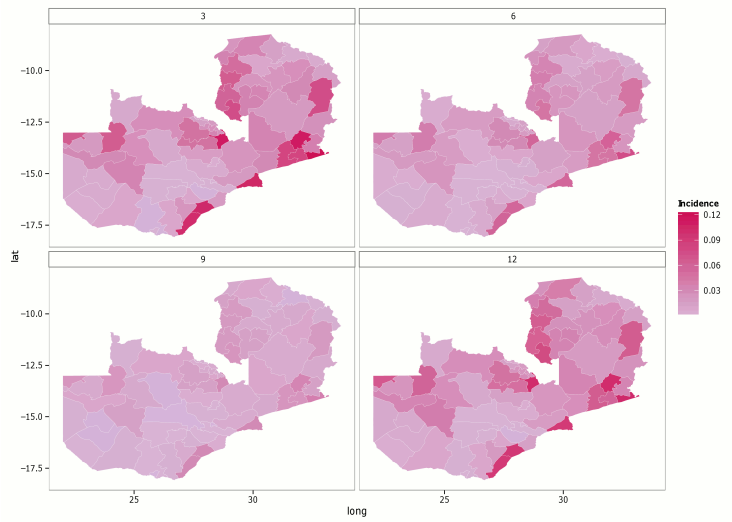


Table 6.3: Spatial parameters estimated by the geostatistical prevalence model.

Spatial parameters	Posterior estimate (median, 95%CI)
σ^2	$1.23(1.10, 1.45)$
ρ	$2.43(2.11, 3.44)$

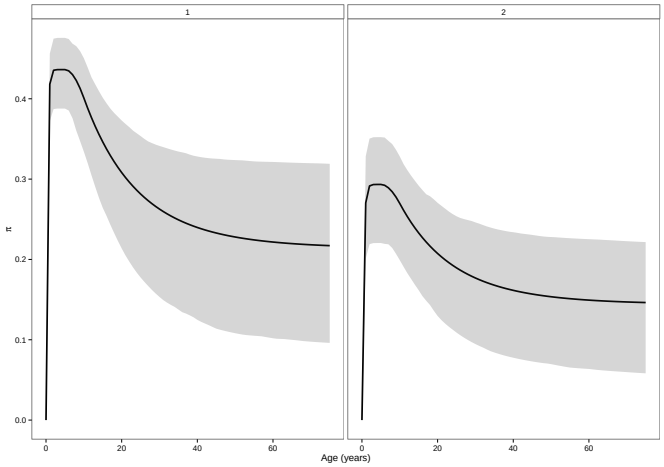
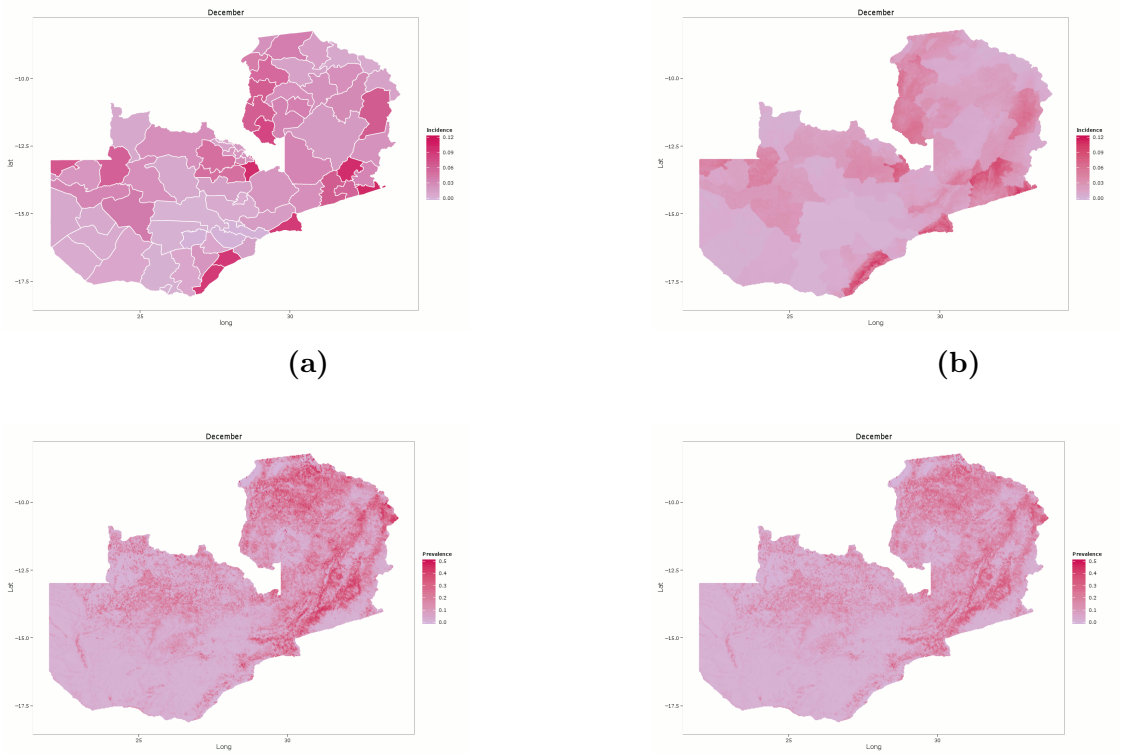


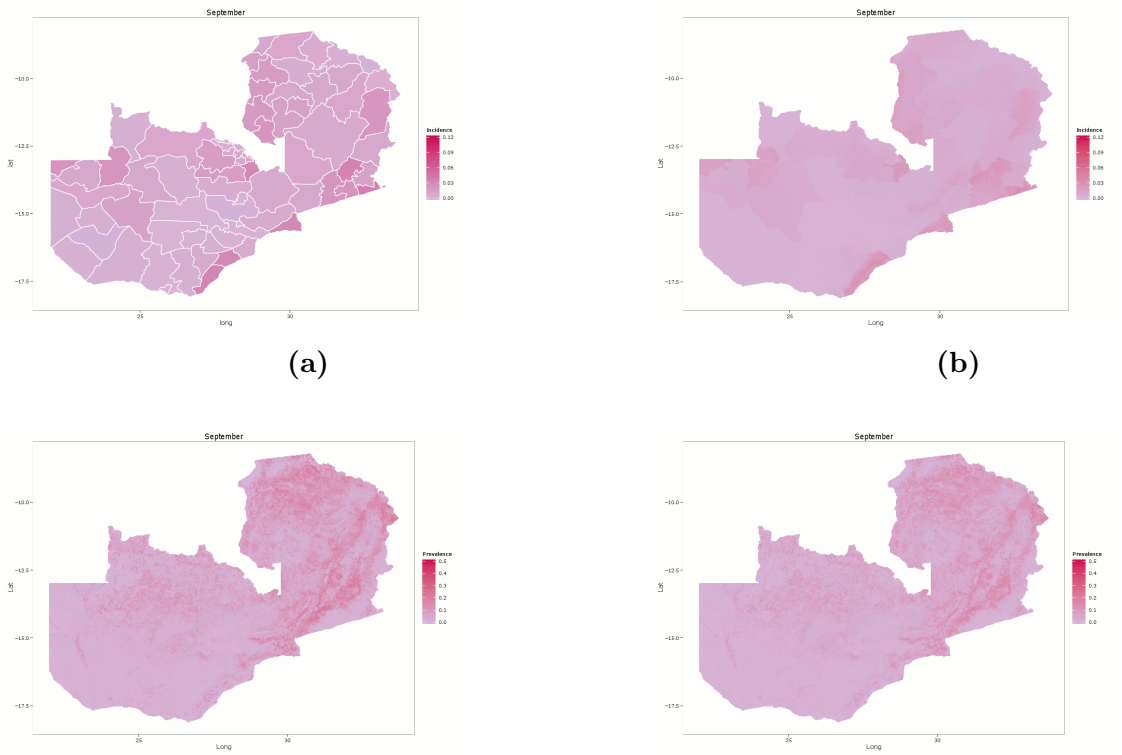
Figure 6.2: Age prevalence curves for two different forces of infection.



(a) **(b)**

(c) **(d)**

Figure 6.3: High transmission season (December) (a) Fitted mean incidence at district level, (b) Imputed mean incidence at 3 km resolution, (c) Prevalence among age category 1-4, (d) Prevalence among age category 5-14.



(a) **(b)**

(c) **(d)**

Figure 6.4: Low transmission season (September) (a) Fitted mean incidence at district level,(b) Imputed mean incidence at 3 km resolution, (c) Prevalence among age category 1-4, (d) Prevalence among age category 5-14.

6.4 Discussion

In this paper, we address age and time heterogeneity in prevalence modeling when information form multiple geo-referenced surveys is combined. Our approach couples a mathematical transmission model with spatial statistical models in a Bayesian framework, allowing the estimation of age and season specific disease risk at high spatial resolution. We develop a joint model formulation for prevalence and incidence data by embedding an extended version of the catalytic model by Pull and Grab (1974) into the hierarchical Bayesian structure, thus accounting for the uncertainty arising form the age/season standardization in the risk estimation.

We have illustrated the proposed methodology with an application to geo-referenced malaria prevalence survey data collected in Zambia in the period 1978-2006. Previous work on modeling heterogeneity in geo-referenced surveys for malaria mapping (Gemperli et al., 2006b; Gosoniu, 2008; Hay et al., 2009) were based on a 2-step procedure to (i) obtain age-correction factors and (ii) separately fit age-adjusted prevalence estimates in a geostatistical model, ignoring adjustment uncertainty. Moreover, the heterogeneity due to the different survey periods was not considered. However, seasonality is one of the most important sources of heterogeneity in malaria prevalence estimates: we incorporate season-dependence in our model by allowing space and time variations in the force of infection.

We obtain spatially varying age-prevalence curves that included a diagnostic sensitivity as a decreasing function of age. Conceptually, this function can be thought of as the decline in the probability of detecting an active infection, although the real reason for the decline in prevalence might be that immunity leads to real declines in the force of infection or real increases in the clearance rate. For the purpose of our analysis, we consider the biological reasons for the decline not relevant. Moreover, in our model the force of infection is estimated by the rate of reported cases at health districts in the general population (age-disaggregated data were not available). The threshold age after which the sensitivity (and therefore the prevalence) starts declining is estimated by one single parameter; however it is known that the age of immunity development decreases for increasing forces of infection Filipe et al. (2007), therefore a spatially varying prior on the parameter α_c could capture variability in the peak age of transmission. Furthermore, a more flexible space-time structure by modelling interactions (Knorr-Held, 2000; Lagazio et al., 2001) on incidence data could be considered.

The present work aims mainly at illustrating the methodology that can be applied to other settings, perhaps with scarcity of resources. Zambia has in fact achieved significant

reductions in the malaria burden in the last years, especially due to the continuous control interventions. For this reason, we have focused on prevalence surveys prior to the massive intervention scale-up. On the other hand, more recent incidence data are used to estimate the seasonality pattern in the force of infection. However, we do not expect large changes in the latter.

The proposed methodology can be extended to include more hetereogenity factors, (i.e. varying sensitivity function due to different diagnostic tools, preferential sampling, etc...) and could be applied in the current development to other diseases that show age-dependence (possibly linked to the acquisition of immunity) and/or seasonality. For example, schistosomiasis, a chronic and poverty-promoting disease caused by trematodes of the genus *Schistosoma*, is characterized by typical age-prevalence curves in endemic setting (Raso et al., 2007).

Chapter 7

Assessing the Spatial Effects of Vector Control Interventions on Changes of Malaria Parasitemia Risk in Africa

Giardina F.[1,2], Kasasa S.[3], Sié A.[4], Utzinger J.[1,2], Tanner M.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

[3] School of Public Health, Makerere University College of Health Sciences, Kampala, Uganda

[4] Centre de Recherche en Santé de Nouna, Nouna, Burkina Faso

This paper has been published in *The Lancet Global Health* 2 (10), e601-e615.

Abstract

Background: Decreases in malaria over the past decade have been mainly associated with the expanded implementation of vector control measures such as insecticide-treated nets (ITNs) and indoor residual spraying (IRS). Malaria Indicator Surveys (MIS) collect information of key malaria indicators through national representative household surveys. The objective of this study was to estimate changes in malaria parasitemia risk at high spatial resolution in sub-Saharan Africa, and to quantify the effects of malaria interventions at national and sub-national level.

Methods and Findings: We analyzed MIS data from six sub-Saharan countries: Angola, Liberia, Mozambique, Senegal, Rwanda, and Tanzania. Bayesian geostatistical models were utilized to estimate the current malaria risk, and to determine the change relative to the period between the last two national surveys. We applied Bayesian variable selection procedures to select the most relevant ITN measure in reducing malaria risk and performed spatial kriging over the study area to produce intervention coverage maps. The contribution of ITN and IRS on the change of malaria risk was estimated after adjusting for climatic factors. Spatially varying coefficients of intervention coverage indicators allowed estimation of their effects at sub-national level. In all of the countries, the probability of decrease in parasitemia varied substantially from one area to another. ITN was an important factor in reducing malaria risk under different definitions of coverage. An overall ITN effect at country level was significant only in Angola and Senegal; however, in all countries significant effects for IRS and ITN were seen at regional level.

Conclusions: The described methodology is useful for identifying areas where changes of malaria risk occurred and for describing the geographical pattern of the disease. The effects of interventions varies in space, which might be driven by local endemicity levels. The produced maps provide a powerful visual tool for national malaria control programs to identify areas where targeted strategies and resources are most needed or likely to have the greatest impact on reducing the risk of parasitemia.

7.1 Introduction

In 2011, the number of people at risk of contracting malaria was estimated to be 3.3 billion. Individuals living in sub-Saharan Africa had the highest risk of acquiring the disease: approximately 80% of cases and 90% of deaths occurred in the World Health Organization (WHO) African Region, with children under 5 years of age and pregnant women most severely affected (WHO, 2012). However, the past decade has seen decreases in malaria caused by *Plasmodium falciparum*, the most deadly and predominant parasite species in Africa. These reductions are going hand-in-hand with increases in international funding for malaria prevention, control, and elimination, which have led to tremendous expansion in implementing national malaria control programs (NMCPs) (Alonso and Tanner, 2013). The NMCPs' main strategies include (i) vector control through the use of insecticide-treated nets (ITNs), indoor residual spraying (IRS) and, in some specific settings, larval control; (ii) chemoprevention for the most vulnerable populations; (iii) confirming malaria diagnosis through microscopy or rapid diagnostic tests (RDTs) for every suspected case; and (iv) timely treatment with appropriate antimalarials (O'Meara et al., 2010).

Malaria reduction is part of the Millennium Development Goals (MDGs), aiming to halve malaria incidence by 2015 as compared to 1990 (UN, 2012a). To monitor and evaluate progress toward this target, the following set of indicators are proposed by the United Nations (UN): incidence and death rates associated with malaria, proportion of children under 5 sleeping under ITNs, and proportion of children under 5 with fever who are treated with appropriate antimalarial drugs (UN, 2012a). Renewed interest in malaria elimination and eradication has led to the definition of new targets. In 2008, the Global Malaria Action Plan (GMAP), put forward by the Roll Back Malaria (RBM) Partnership (Roll Back Malaria, 2008), advocated reducing malaria cases by 75% (from 2000 levels) and malaria deaths to near zero, by 2015. Since 2007, WHO has recommended universal coverage with ITNs (preferably long lasting insecticide-treated nets (LLINs)), rather than a pre-determined number of nets per household or exclusively targeting household members at high risk, i.e., pregnant women and children under 5 years of age (WHO, 2012). In 2010, the GMAP (Global Partnership to Roll Back Malaria, 2010) called for rapid scaling-up to achieve universal coverage with some form of vector control.

Malaria Indicator Surveys (MIS) were developed by RBM to coordinate global efforts to fight malaria. MIS collect national and regional or provincial data from a representative sample of respondents. Surveys collect information about ITN ownership and use, IRS, prompt and effective treatment of fever in young children, and the prevention of malaria

in pregnant women. Most MIS also include measurement of malaria parasites and anemia among children under 5 years and pregnant women. MIS are usually carried out during high malaria transmission seasons. These nationally representative household surveys provide the most precise benchmark of progress toward internationally agreed upon targets.

MIS and other household surveys, such as Demographic and Health Surveys (DHS) and Multiple Indicator Surveys (MICS), have been used to estimate ITN coverage in several sub-Saharan countries (Miller et al., 2007; Noor et al., 2009; Burgert et al., 2012). Ownership indicators have been considered, such as the proportion of households with at least one ITN (or one ITN for every two people), as have use indicators, i.e., the proportion of the population (or children or pregnant women) who slept under an ITN the night before the survey. The number of nets in Africa has been estimated by analyzing household surveys, including MIS, between 1999 and 2005. On this basis, the number of ITNs needed to achieve high coverage has been calculated (Miller et al., 2007). The changes in ITN coverage among children under the age of 5 years, reported between 1999–2003 and 2004–2007, have also been examined and thematic maps as well as projections of ITN coverage in 2007 were produced (Noor et al., 2009). However, estimating ITN coverage through national household surveys presents some challenges. For instance, national estimates could be underestimated or overestimated if the actual population at risk of developing malaria, that is if the level of endemicity is not taken into account. Seasonality may represent an additional source of bias if surveys are carried out during dry and hot months when people are less likely to sleep under bednets (Burgert et al., 2012).

By including parasitemia data, MIS permit the assessment of the impact of several factors, including malaria control strategies, on health outcomes under "real-world conditions" (Lim et al., 2011). Efficacy of sleeping under ITNs in preventing malaria transmission has been evidenced by a systematic review and meta-analysis of randomized controlled trials (Lengeler et al., 2004) showing that regular ITN use can reduce all-cause child mortality by around 20% in malaria-endemic areas and cut malaria episodes by half.

Geostatistical models have been used to analyze MIS in different African countries and to estimate the spatial effects of bednet ownership and use after adjusting for climatic factors and socio-economic indicators. An analysis of the Angolan MIS in 2006 reported a reduction in risk in areas with at least 0.2 ITNs per person (Gosoniu et al., 2010). In a more recent analysis of the Tanzanian MIS carried out from 2007-2008, ownership of at least one ITN per household was the indicator used to assess coverage impact, although it showed no protective effect (Gosoniu et al., 2012). However, the analysis of the Zambian MIS in 2007

found a preventive effect on malaria risk using the same ownership indicator (Riedel et al., 2010). Different bednet coverage indicators, as suggested by reviews of bednet delivery strategies (Thwing et al., 2011; Kilian et al., 2010) were considered in a variable selection procedure performed in a geostatistical analysis of the DHS 2008 in Senegal (Giardina et al., 2012) to identify the variable most associated with decreased malaria risk. The analysis concluded that the presence of at least one ITN per every two household members reduced the odds of parasitemia by 86%. Pooled analyzes of MIS were further conducted to assess the effect of maternal education and household wealth on malaria risk in children (Siri and Lutz, 2012). The effectiveness of different control strategies in preventing malaria has also been assessed through simulating different settings and scenarios using mathematical models (Chitnis et al., 2010; Griffin et al., 2010; White et al., 2009).

Geostatistical models represent the most appropriate way of analyzing MIS data. They enable the relation between malaria prevalence and intervention strategies to be quantified, after adjusting for environmental factors and socioeconomic status, while allowing correlation among spatial locations. Geostatistical analyzes of MIS have produced spatially explicit estimates of disease risk and of the number of infected children below the age of 5 years (Gosoniu et al., 2010, 2012; Riedel et al., 2010; Giardina et al., 2012; Hwang et al., 2010; Jima et al., 2010). Such maps provide a powerful visual tool for NMCPs, identifying areas where targeted strategies and resources are most needed or most likely to have the greatest impact. By 2013, six African countries had completed two rounds of MIS: Angola, Liberia, Mozambique, Rwanda, Senegal, and Tanzania. We performed a spatio-temporal analysis to estimate changes in malaria parasitemia risk across these countries. Additionally, we quantified the spatial effects of control measures (i.e., ITN and IRS coverage) at national and subnational level in reducing malaria risk, after taking into account climatic factors.

7.2 Methods

7.2.1 Data Sources

MIS and DHS data

To assess the change in malaria parasitemia risk over time, the analysis included all the countries in sub-Saharan Africa with publicly available data from at least two MIS carried out at different times (2006-2008 and 2010-2012). DHS that collected relevant health and intervention outcomes (measurements of malaria parasites and ITN or IRS coverage

assessment) were included in the analysis, as well. The protocol for each survey was submitted to and approved by the Ethical Review Committee at the NMCPs and the Institutional Review Board (IRB) of Macro International. Written informed consent was obtained from the respondents participating in the survey.

Only surveys with global positioning system (GPS) information gathered at cluster level were considered. A summary of the countries included and the timing of their surveys can be found in Table 7.1. A more detailed description including main vectors, transmission season, and intervention implementation of the countries considered is provided in the Supporting Information.

Health Outcomes

In both MIS and DHS, the presence of malaria parasites is determined by RDT or by analyzing thick or thin blood films on microscope slides. Most surveys used both diagnostic approaches. However, for our analysis, positivity was defined only via blood films as they are more reliable than RDT performed in the field (Wongsrichanalai et al., 2007).

Interventions

ITN coverage was assessed by defining several indicators derived from variables collected through survey questionnaires. Intervention coverage measures can be defined at different levels: individual, household, or cluster. However, to evaluate geographically the role that intervention scale-up played in reducing parasitemia risk, only cluster-level intervention coverage indicators were considered. The two surveys carried out in each country did not consider the same households, and hence, the spatial analysis of the change in risk was conducted at cluster level.

Following the review by Kilian et al. (2010), we defined the following indicators of ITN ownership: the proportion of households with at least one ITN (indicator currently used by RBM), the proportion of households with at least one ITN for every two people (new indicator considered by RBM) and the mean nets-to-people ratio. Use was defined as the proportion of children aged 0-59 months who slept under an ITN the night prior to the survey (MDG indicator of interest and currently used by RBM) and the proportion of people who slept under an ITN the night prior to the survey.

IRS coverage was obtained from DHS and MIS, reporting whether the house had been sprayed within the previous 12 months. The proportion of sprayed households within a cluster was used as a potential factor in reducing parasitemia risk.

Environmental Predictors Derived from Remote Sensing Data

The main drivers of malaria transmission are climatic and environmental factors. Advances in remote sensing and geographic information system (GIS) have permitted spatial analysis of the relation between malaria risk and environmental indices, as well as accurate predictions over large study areas. The survey locations where parasitemia is assessed need to be geo-referenced (commonly by GPS) so that environmental and climatic proxies can be extracted for each data location.

Normalized difference vegetation index (NDVI) and land surface temperature (LST) data for our analysis were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) at 1 km spatial resolution. Decadal rainfall data were available at 8 km resolution via Africa Data Dissemination Service. Elevation data were obtained from an interpolated digital elevation model from the U.S. Geological Survey - Earth Resources Observation and Science Data Center at a spatial resolution of 1 km. The environmental factors with available temporal resolution (LST, NDVI and rainfall) were acquired for the 6-month period prior to the survey and the average was calculated and extracted for each data location.

7.2.2 Models

A geostatistical model was developed and fitted to assess the effect of climatic and environmental conditions on parasitemia risk in each country using survey data from 2006-2008 and 2010-2012. Bayesian kriging was employed to predict malaria risk at high spatial resolution at the two time periods. Furthermore, the probability of parasitemia risk reduction at each pixel was estimated as well as the total number of children infected, stratified by country and survey period, calculating their difference. We made use of population data provided by Afripop (Tatem et al., 2013a) that consist of spatial estimations of number of children below the age of 5 years per km^2 in 2010.

To estimate the effect of interventions, we modeled the change of parasitemia risk (on the logit scale) as a function of the difference in climatic conditions between the two time points (surveys) and intervention coverage (i.e., ITN and IRS). To account for potential interactions with endemicity levels, we fitted a second model to estimate different intervention effects for each regional unit (first administrative division).

The bednet coverage indicators presented in the previous section are highly correlated; therefore we have defined a Bayesian variable selection procedure that selects only one (or

none) bednet ownership indicator and one (or none) bednet use indicator. Further details on the models can be found in the Supporting Information.

7.2.3 Software

'Just Another Gibbs Sampler' (JAGS) (Plummer, 2003) was used to implement the variable selection approach that allowed us to choose among the different bednet coverage indicators. 'Integrated nested Laplace approximation' (INLA) (Rue et al., 2009) with the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011) was used to perform model fit and prediction.

7.3 Results

A summary of the six African countries analyzed in this study, including a descriptive analysis of the data collected in the two malaria national surveys, is given in Table 7.1. An overall decreasing trend in parasitemia prevalence can be seen, with the exception of Liberia and Mozambique. Bednet and IRS coverage has remained constant or slightly higher in the second survey, compared to the first in all countries whilst a decrease in Angola was observed. The results of the spatial analysis are illustrated on a country basis. We will refer to an effect as "significant" throughout the manuscript whenever the credible intervals do not include 0 or the odds ratio credible intervals do not include 1. Coefficient estimates reported in the text are obtained using actual ITN and IRS coverage values and differ from the ones presented in the tables, which are based on standardized covariates to allow for comparison among predictor effects.

7.3.1 Angola

A change in parasitemia risk among children in Angola can be seen over the period 2006-2011. While the first survey was carried out partially during the long rainy season, the second survey covered almost completely the high transmission period. With the exception of the coast and some areas in the south, in 2006/2007, parasitemia risk was high and almost evenly spread, reaching peaks of 80% (Figure 7.1a). In 2011, parasitemia was concentrated in specific areas, particularly in the northern part of the country (i.e., Zaire, Bengo, and Cuanza Norte provinces; Figure 7.1b). Overall, the probability of observing a reduction in parasitemia risk is higher than 50% throughout the country (Figure 7.1d). Although a decrease can also be seen in Huila province, this area remains at high risk. Parasitemia remained stable in Cabinda province.

Environmental factors similarly affected parasitemia risk in the two surveys and the spatial parameters showed comparable estimates (Table 7.2). However, the spatial variance estimate was higher in the model for the second survey. The ITN coverage measure, chosen via the Bayesian variable selection procedure to model the effect of control on the change of malaria risk, was the proportion of people who slept under an ITN the night prior to the survey (Table 7.3). Table 7.4 shows that the change in environmental and climatic conditions was not significantly associated with the change in risk. Intervention measures were highly associated with parasitemia decreases in the country. For every 1% increase in ITN coverage, the odds ratio of parasitemia (second survey versus first survey) decreases by 5% (95% CI: 3-7%). The IRS effect was not significant when modeled at country level (odds ratio: 0.23, 95% CI: 0.02-1.91). Angola presents high to moderate ITN coverage in the outer part of the country and very low coverage in the interior, while IRS coverage is still quite low with the exception of a few areas (see interventions coverage maps in Figures 7.1e and 7.1f). ITN and IRS effects at area level are shown in Figures 7.1g and 7.1h): ITN had a significant effect in protecting against malaria in the south-eastern province of Cuando Cubango and in the coastal province of Benguela, while IRS showed a significant effect mainly in the central part of the country (Malanje, Bengo, and Cuanza Norte provinces).

The overall estimated decline in the number of infections among children aged between 0 and 59 months dropped by 52.0% (95% CI: 50.7-52.5%), Table 7.5.

7.3.2 Liberia

Analysis of the two surveys conducted in Liberia revealed that parasitemia risk is spread throughout the country. Environmental conditions do not appear to be clear drivers of malaria; none of them are significantly associated with the outcome at each time point. Spatial parameter estimates show similar values of variance and range (Table 7.2) in both surveys. The first survey (MIS 2008/2009) was carried out after the peak transmission period, while the second survey was implemented during the high transmission period (rainy season). Table 7.4 shows that differences between the two surveys in terms of rainfall, NDVI, and LST (night) are positively associated with the change in the log odds of parasitemia. Figure 7.2a shows that parasitemia risk was high (60-70%) in the continental part of the country in 2008 and slightly lower in the coastal counties of Grand Cape Mount, Bomi, Montserrado, and Margibi (40-50%). The current situation, as shown in Figure 7.2b, shows the highest risk to be in Grand Gedeh, River Gee, and Grand Kru counties in the south. Only the capital, Monrovia, shows low risk (<10%). Figure 7.2d illustrates the

geographic pattern of change in risk: while the probability of observing a decrease in parasitemia risk in the north-western part of the country is above 50%, the probability is very low in the south.

Malaria control interventions in Liberia have revealed low coverage. IRS, for example, has only been implemented in a few targeted areas and coverage data were not collected in the surveys, while ITN coverage (selected as the proportion of households in the cluster with at least one ITN; Table 7.3) is high (˜60%), mainly in the area of the capital (Figure 7.2e). The overall effect of ITN was moderate and not significant (odds ratio: 0.77, 95% CI: 0.55-1.32), however the model with spatially varying ITN coefficients (Figure 7.2f) showed a significant protective effect in the northern counties (Montserrado and Gbarpolu).

As shown by Table 7.5, the estimated number of infections among children below the age of 5 years has reduced by 14.8% (95% CI: 13.4-15.8%).

7.3.3 Mozambique

The two national surveys in Mozambique that included parasitemia data were conducted after the rainy season, although the second survey (DHS 2011) collected data over a somewhat longer period, partially explained by the larger number of surveyed clusters. The two provinces with the highest malaria risk were Nampula and Zambezia, in the northern part of the country. The southern parts of the country were characterized by lower risk compared to the rest of the country (<10%), especially Maputo (city and province) and Gaza province (Figure 7.3b). The overall prevalence estimated from the second survey was similar to the one estimated from the first survey.

The main drivers of malaria parasitemia are rainfall and NDVI. The average LST (day) showed a significant association with parasitemia in the analysis of the second survey, as summarized in Table 7.2. Higher estimates of spatial variance reflect higher variation of parasitemia risk compared to 2007. Differences in climatic conditions such as LST (day) and NDVI between the two surveys were associated with the change in malaria risk (Table 7.4).

Table 7.3 shows that, at country level, ITN coverage (i.e., proportion of households in the cluster having at least one ITN) and IRS coverage effects were not significant (ITN odds ratio 0.91, (95% CI: 0.53-1.55), (IRS odds ratio 0.72 (95% CI: 0.42-1.23)). IRS was implemented mainly in the southern part of the country, but some other areas in the centre, particularly Zambezia province, showed coverage of 50%. ITN is more common throughout the country, reaching 70% coverage in most provinces (see coverage maps in Figures 7.3e

and 7.3f). Both ITN and IRS effects were found to be significant in one northern province (Niassa); ITN effect was also significant in Tete province and IRS effect was significant in Cabo Delgado in north-eastern Mozambique (see Figures 7.3g and 7.3h).

7.3.4 Rwanda

The 2007 DHS was carried out in a period when malaria transmission is low, whereas the second DHS (in 2011) included the short rainy season. Nevertheless, comparing the maps in Figures 7.4a and 7.4b, there is a decreasing trend in malaria risk in some regions. Parasitemia risk ranges from 0 to 20%. The probability of observing a decline in parasitemia was low (since risk remains low/steady in most of the country) except in Nyagatare and Gatsibo districts in the north of the East province and Gisagara district in the Southern province, where it is higher than 50% (Figure 7.4d).

The impact of environmental factors on the geographical spread of parasitemia risk is not significant in most cases. Only elevation was negatively associated with malaria risk in the analysis of data collected during the first survey. Information on IRS was not collected during the surveys. ITN coverage is very high, in terms of proportion of children in the cluster sleeping under an ITN. It is distributed almost uniformly across the country, with highest coverage in the Lake Kivu area and the capital, Kigali (Figure 7.4e). The effects of the ITN intervention is evident at sub-national level, particularly in the East and South provinces (Figure 7.4f). However, the overall effect at country level lacked statistical significance (Table 7.4).

The overall estimated number of infections among children below the age of 5 years has decreased by 41.9% (95% CI: 39.4-44.5%), Table 7.5.

7.3.5 Senegal

Malaria in Senegal is concentrated in the central and southern parts of the country. The first survey (MIS 2008) was conducted after the rainy season, in a period in which transmission starts decreasing from high levels. The field work for the second survey (DHS 2010) included one month of high transmission. However, parasitemia risk in some known "hot spots" of transmission, like the regions of Tambacounda, Kaffrine and Kolda in the south, has dramatically decreased (Figures 7.5a and 7.5b). The map in Figure 7.5d shows that the probability of observing a decline in parasitemia risk is more than 50% in most areas of the country, with the exception of Kedougou region in the south and the Sant-Louis region in the north, where the risk has remained fairly constant though at low levels.

The main environmental drivers are rainfall and NDVI. The relationship between these two climatic features and parasitemia was much stronger in the first survey, whereas NDVI was not significantly associated with malaria parasitemia in the second survey. We found a smaller spatial variance, but higher range in the second survey (Table 7.2). Table 7.3 indicates that the proportion of people in the cluster sleeping under a net is among the ITN coverage measures that best describe the change in malaria risk. Most of the clusters sampled along the Senegal River (at the border with Mauritania) show high coverage, as do most of the clusters in Kolda and Kaolack-Kaffrine. IRS is still limited to some areas of Saint-Louis in the north and at the border between Kolda and Kedougou, as shown by the IRS coverage map (Figure 7.5f). A 1% increase in ITN coverage was associated with a reduction in the odds of parasitemia of 1% (95% CI: 0-3%). The effect of IRS at country level was not significant. However, if estimated at area level, IRS shows a higher and significant effect in the Tambacounda region (Figure 7.5h). Similarly, the effect of ITN coverage was significant in reducing parasitemia risk in both the Tambacounda and neighboring Kolda regions, as shown in Figure 7.5g.

As shown by Table 7.5, the estimated number of infections among children below the age of 5 years has fallen by 40.3% (95% CI: 39.2-41.4%).

7.3.6 Tanzania

Parasitemia in the country was modeled in the mainland and, separately, in the islands. Risk on the islands was very low (0-5%) in 2008 and it was slightly lower in 2012, as can be seen by comparing the predicted prevalence in the two surveys (Figures 7.6a and 7.6b). In the mainland, the risk of parasitemia is high, reaching up to 70-80% in the northern and southern regions. Figure 7.6e shows that the probability of a decrease in parasitemia risk during the second survey is above 50% on the islands of Zanzibar as well as in most areas of the mainland. In some other areas, an increase in the risk can be detected. A positive association between change in malaria risk and rainfall was also estimated (Table 7.4).

ITN coverage ranges from 25% to 95%, with the highest estimates obtained in the urban area of Dar es Salaam and the islands of Zanzibar (Unguya and Pemba; no cluster was sampled on Mafia Island in the second survey). IRS is mainly implemented in the Lake Victoria area, the islands of Zanzibar and Dar es Salaam (Figures 7.6e and 7.6f). IRS was effective in the lake area and in two southern provinces (Morogoro and Iringa) as well as on the islands (Figure 7.6h). Dar es Salaam and Shinyanga provinces mostly benefited from ITN intervention. At country level, the overall effects of ITN and IRS interventions

were not significant.

The estimated number of infections among children belo the age of 5 years reduced by 30.1% (95% CI: 27.4-30.4%) in the country, as shown by Table 7.5.

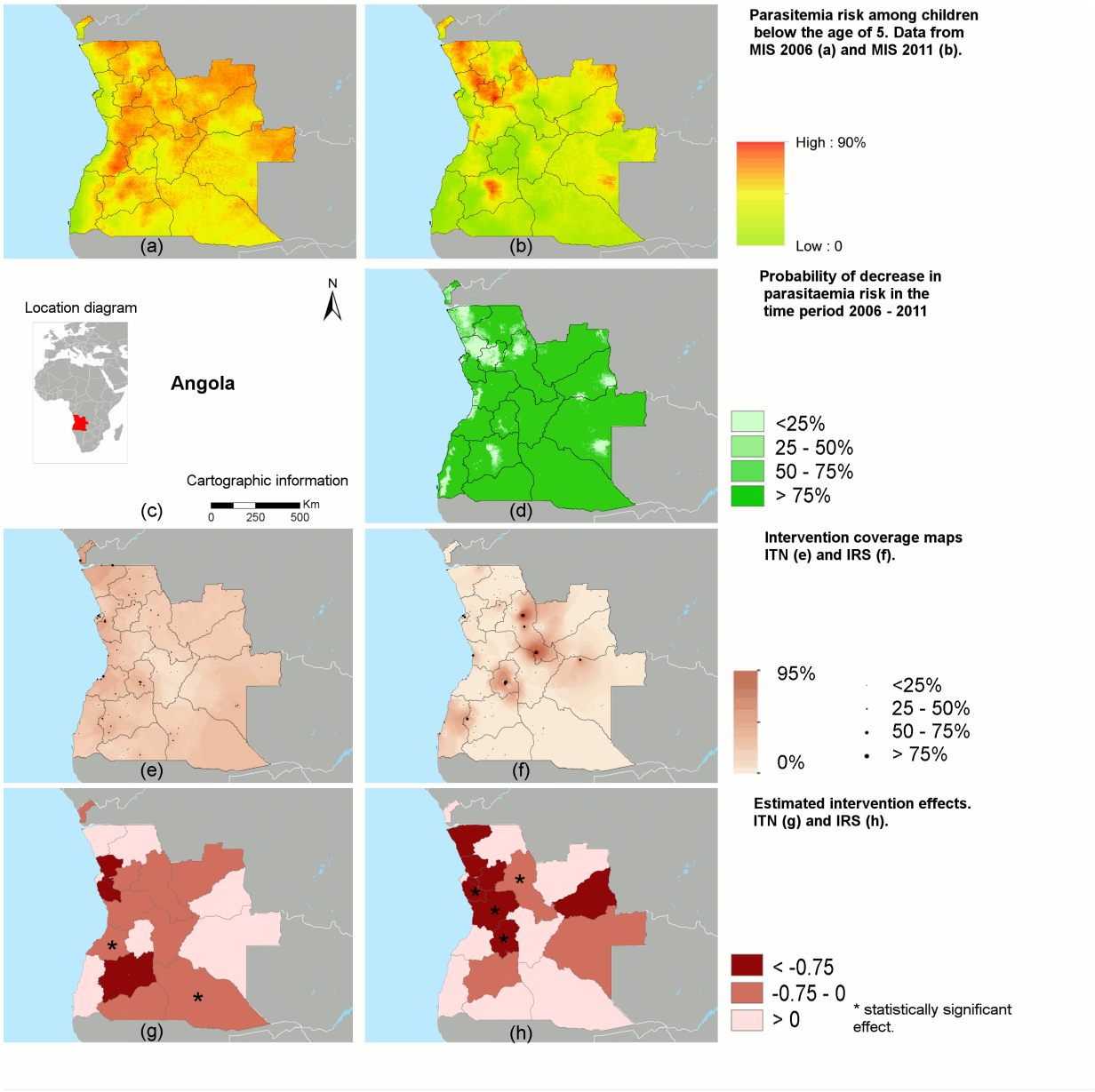


Figure 7.1: Angola. Predicted parasitemia risk in 2006 (a) and 2010 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2006-2010 (d), ITN (e) and IRS (f) coverage maps, estimated effects of interventions: ITN (g) and IRS (h) (median plotted).

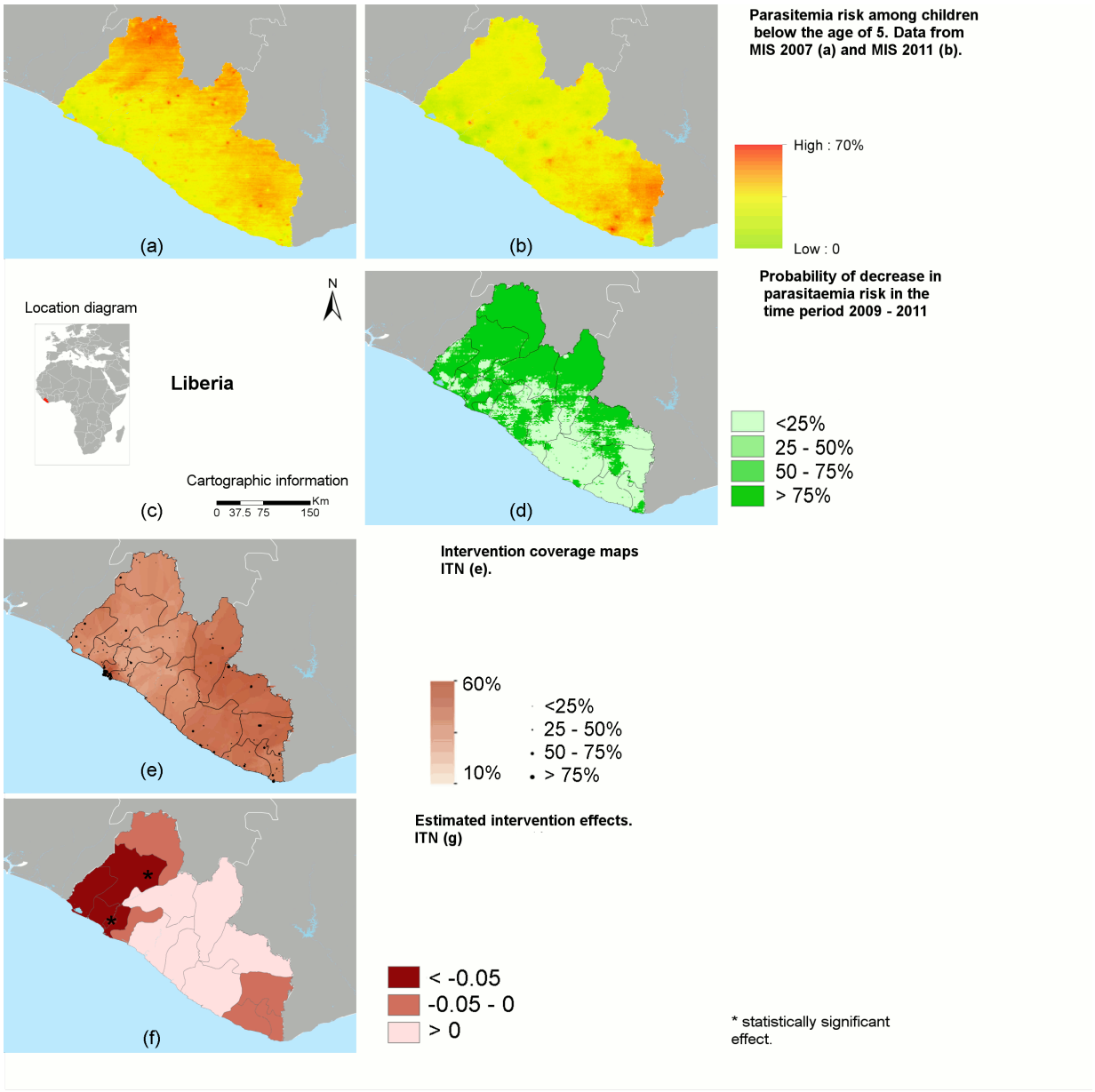


Figure 7.2: Liberia. Predicted parasitemia risk in 2007 (a) and 2011 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2007-2011 (d), ITN (e) coverage map, estimated effects of interventions: ITN (f) (median plotted).

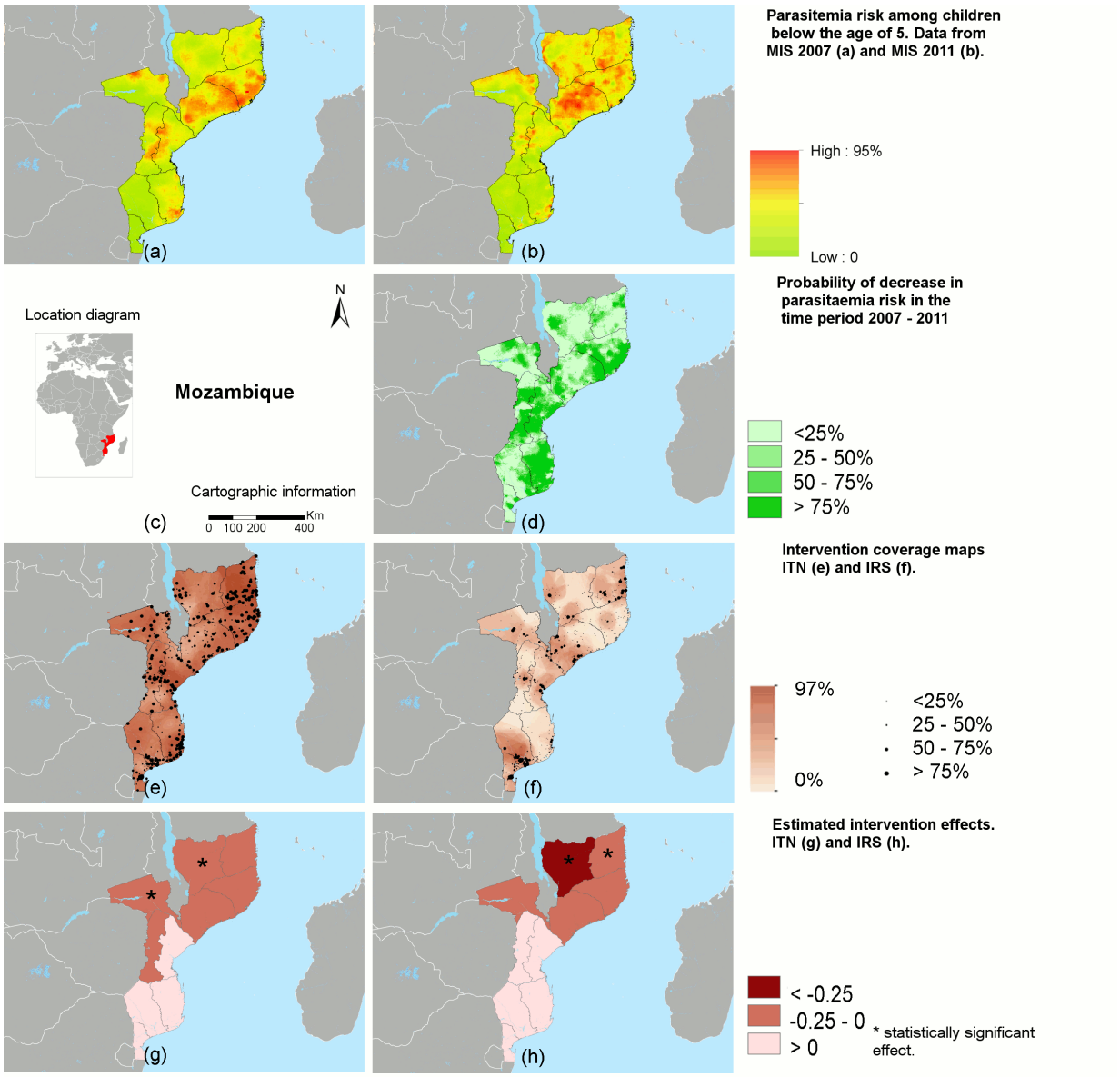


Figure 7.3: Mozambique. Predicted parasitemia risk in 2007 (a) and 2011 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2007-2011 (d), ITN (e) and IRS (f) coverage maps, estimated effects of interventions: ITN (g) and IRS (h) (median plotted).

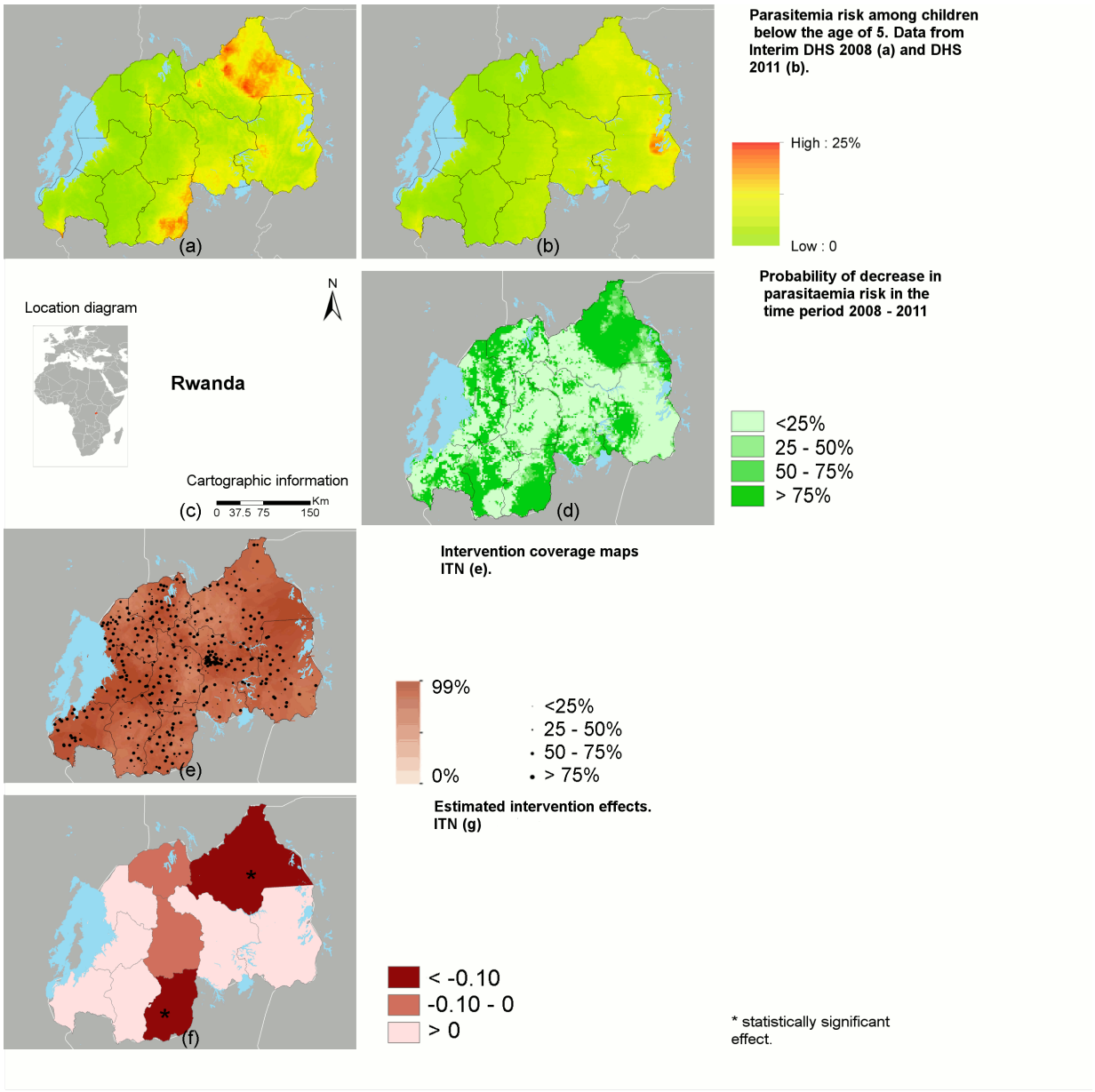


Figure 7.4: Rwanda. Predicted parasitemia risk in 2008 (a) and 2011 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2008-2011 (d), ITN (e) coverage map, estimated effects of interventions: ITN (g) (median plotted).

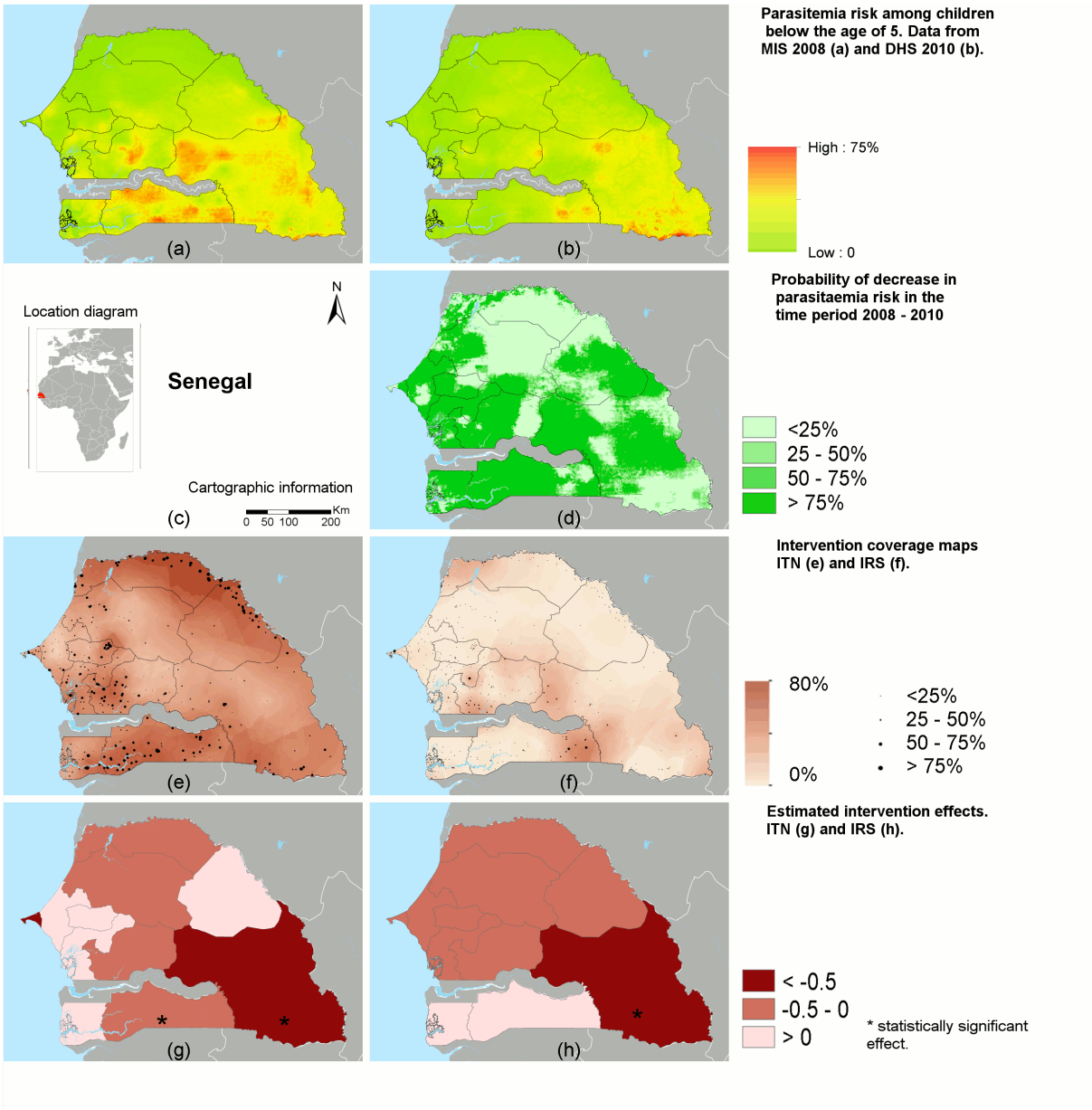


Figure 7.5: Senegal. Predicted parasitemia risk in 2008 (a) and 2010 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2008-2010 (d), ITN (e) and IRS (f) coverage maps, estimated effects of interventions: ITN (g) and IRS (h) (median plotted).

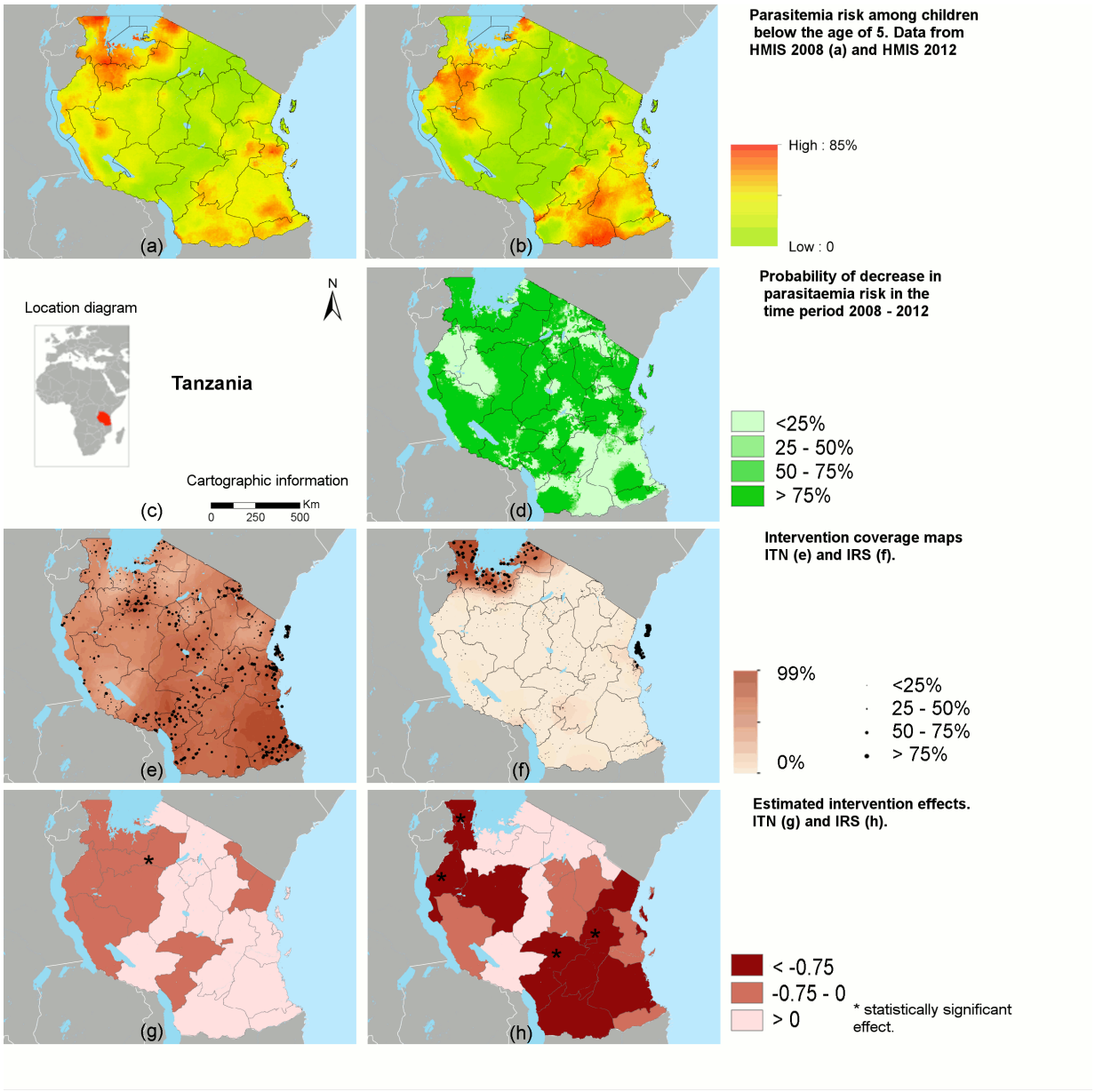


Figure 7.6: Tanzania. Predicted parasitemia risk in 2008 (a) and 2012 (b), location diagram and cartographic information (c), probability of observing a decline in the time period 2008-2012 (d), ITN (e) and IRS (f) coverage maps, estimated effects of interventions: ITN (g) and IRS (h) (median plotted).

Table 7.1: The six sub-Saharan countries and surveys included in the spatial analysis of malaria parasitemia risk and effects of interventions. Prevalence and intervention coverage estimates are expressed in terms of median and 95% CI. ITN_1^0: proportion of households with at least one ITN, ITN_2^0: proportion of households with at least one ITN for every two people, ITN_3^0: mean nets-to-people ratio, ITN_1^u: proportion of children aged 0-59 months who slept under an ITN the night prior to the survey, ITN_2^u: proportion of people who slept under a ITN the night prior to the survey.

Indicator	**Country**					
	Angola	**Liberia**	**Mozambique**	**Rwanda**	**Senegal**	**Tanzania**
Main malaria transmission period	March-May, October-November	September-October	December-April	May-June, November-December	September-December	March-May October-December
1st Survey						
Survey type	MIS	MIS	MIS	Interim DHS	MIS	HMIS
Survey period	Nov 2006 - Apr 2007	December 2008 - March 2009	June 2007 - July 2007	December 2007 - April 2008	November 2008 - January 2009	October 2007 - February 2008
No. of locations	115	150	345	246	315	465
No. of households	1,518	2,712	–	3,508	4,523	9,142
Parasitemia prevalence	0.22 (0.17, 0.27)	0.31 (0.28, 0.34)	0.30 (0.27, 0.33)	0.02 (0.01, 0.02)	0.06 (0.05, 0.07)	0.12 (0.11, 0.14)
ITN ownership						
ITN_1^0	0.40 (0.35, 0.46)	0.32 (0.29, 0.33)	–	0.78 (0.76, 0.80)	0.36 (0.33, 0.38)	0.68 (0.66, 0.71)
ITN_2^0	0.15 (0.12, 0.17)	0.12 (0.10, 0.13)	–	0.23 (0.21, 0.25)	0.34 (0.32, 0.35)	0.42 (0.38, 0.45)
ITN_3^0	0.21 (0.19, 0.22)	0.16 (0.14, 0.17)	–	0.24 (0.22, 0.28)	0.38 (0.31, 0.44)	0.38 (0.33, 0.41)
ITN use						
ITN_1^u	0.37 (0.32, 0.39)	0.35 (0.33, 0.36)	–	0.33 (0.31, 0.35)	0.42 (0.38, 0.43)	0.55 (0.51, 0.57)
ITN_2^u	0.23 (0.21, 0.25)	0.26 (0.24, 0.27)	–	0.21 (0.19, 0.23)	0.38 (0.33, 0.39)	0.51 (0.49, 0.53)
IRS	0.04 (0.01, 0.06)	–	–	–	–	0.19 (0.09, 0.27)
2nd Survey						
Survey type	MIS	MIS	DHS	DHS	DHS	HMIS
Survey period	January 2011 - May 2011	September 2011 - December 2011	June 2011 - November 2011	September 2010 - March 2011	October 2010 - April 2011	December 2011 - May 2012
No. of locations	228	150	603	470	383	573
No. of households	8,030	2,140	3,461	1,472	7,902	10,040
Parasitemia prevalence	0.11 (0.08, 0.14)	0.28 (0.25, 0.32)	0.30 (0.28, 0.33)	0.01 (0.00, 0.02)	0.04 (0.03, 0.05)	0.08 (0.07, 0.09)
ITN ownership						
ITN_1^0	0.42 (0.38, 0.44)	0.41 (0.37, 0.43)	0.65(0.61, 0.68)	0.93 (0.91, 0.95)	0.42 (0.38, 0.43)	0.95 (0.88, 0.96)
ITN_2^0	0.06 (0.04, 0.08)	0.17 (0.15, 0.22)	0.26 (0.25, 0.28)	0.33 (0.31, 0.35)	0.26 (0.25, 0.27)	0.57 (0.51, 0.59)
ITN_3^0	0.09 (0.07, 0.11)	0.17 (0.15, 0.20)	0.24 (0.22, 0.27)	0.37 (0.32, 0.39)	0.29 (0.27, 0.31)	0.48 (0.47, 0.51)
ITN use						
ITN_1^u	0.26 (0.21, 0.28)	0.38 (0.37, 0.41)	0.40 (0.38, 0.42)	0.75 (0.71, 0.77)	0.46 (0.42, 0.47)	0.77 (0.73, 0.78)
ITN_2^u	0.19 (0.15, 0.22)	0.32 (0.31, 0.35)	0.32 (0.31, 0.34)	0.15 (0.12, 0.18)	0.41 (0.38, 0.43)	0.69 (0.63, 0.71)
IRS	0.07 (0.05, 0.12)	–	0.24 (0.21, 0.28)	–	0.05 (0.03, 0.21)	0.22 (0.10, 0.35)

Table 7.2: Environmental factors and spatial parameter estimates stratified by country and survey. The estimates are expressed in terms of Medians and 95% credible intervals. Covariates have been standardised to make estimates comparable.

Country	**Angola**	**Liberia**	**Mozambique**	**Rwanda**	**Senegal**	**Tanzania**
			1st Survey			
Covariate	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)
Rainfall	1.09 (0.01, 2.18)	0.06 (-0.10, 0.23)	0.43 (0.19, 0.76)	-0.06 (-0.57, 0.49)	0.91 (0.24, 1.69)	0.49 (0.15, 0.82)
NDVI	0.64 (-0.06, 1.34)	0.08 (-0.10, 0.25)	0.49 (0.21, 0.65)	0.00 (-0.43, 0.40)	0.59 (0.02, 1.13)	0.39 (0.16, 0.64)
Altitude	0.53 (-0.64, 1.77)	0.11 (-0.11, 0.32)	-0.21 (-0.69, 0.23)	-1.05 (-2.14, -0.07)	0.18 (-0.14, 0.49)	-0.30 (-1.03, 0.39)
LST night	0.95 (0.13, 1.85)	-0.12 (-0.35, 0.10)	-0.28 (-0.64, 0.07)	-0.29 (-1.22, 0.60)	0.24 (-0.36, 0.78)	0.18 (-0.38, 0.76)
LST day	0.55 (1.85, 1.25)	-0.06 (-0.26, 0.13)	0.18 (-0.05, 0.42)	0.33 (-0.41, 1.10)	0.16 (-0.45, 0.77)	0.01 (-0.24, 0.26)
Spatial parameter						
σ_1^2	3.19 (2.01,3.99)	0.73 (0.42, 0.89)	1.35 (0.99, 1.62)	2.45 (1.81, 3.03)	3.46 (2.52, 3.66)	2.00 (1.22, 2.26)
r_1 (degr)	0.91 (0.70,1.01)	0.06 (0.02, 0.11)	0.88 (0.75, 1.02)	0.15 (0.10, 0.22)	0.44 (0.21, 0.64)	1.30 (0.95, 1.44)
			2nd Survey			
Rainfall	0.88 (-0.03, 1.86)	0.24 (0.04, 0.44)	0.45 (0.24, 0.66)	-0.15 (-0.79, 0.42)	0.50 (0.00, 1.09)	0.33 (-0.14, 0.80)
NDVI	1.00 (0.35, 1.66)	0.15 (-0.06, 0.37)	0.57 (0.33, 0.82)	0.07 (-0.48, 0.63)	0.27 (-0.22, 0.73)	0.60 (0.32, 0.89)
Altitude	-1.53 (-3.05, -0.08)	-0.03 (-0.24, 0.18)	0.10 (-0.21, 0.40)	0.05 (-1.34, 1.36)	0.23 (-0.02, 0.47)	-1.59 (-2.45, -0.78)
LST night	0.05 (-1.07, 1.18)	0.08 (-0.15, 0.32)	-0.05 (-0.32, 0.22)	0.72 (-0.64, 2.02)	0.20 (-0.19, 0.58)	0.10 (-0.57, 0.78)
LST day	-0.36 (-0.83, 0.09)	-0.16 (-0.38, 0.05)	0.41 (0.20, 0.63)	0.62 (-0.42, 1.81)	0.47 (-0.01,1.01)	0.42 (0.09, 0.69)
Spatial parameter						
σ_2^2	6.39 (5.25,6.57)	0.47 (0.12, 0.63)	2.18 (1.57, 2.50)	0.75 (0.32, 1.42)	2.12 (1.95, 2.43)	5.22 (3.89, 5.66)
r_2 (degr)	1.02 (0.83,1.17)	0.12 (0.07, 0.25)	0.31 (0.23, 0.55)	0.90 (0.84, 1.01)	0.47 (0.33, 0.54)	2.07 (1.87, 2.22)

Table 7.3: Posterior inclusion probability of bednet coverage indicators per country.

Bednet indicator	**Angola**	**Liberia**	**Mozambique**	**Rwanda**	**Senegal**	**Tanzania**
Proportion of households with at least one ITN	0.02	0.55	0.51	0.05	0.02	0.57
Proportion of households with at least one ITN for every two people	0.09	0.05	0.15	0.11	0.09	0.11
Mean nets-to-people ratio	0.01	0.10	0.05	0.10	0.01	0.01
Proportion of children aged 0-59 months who slept under an ITN the night prior to the survey	0.11	0.02	0.08	0.61	0.02	0.10
Proportion of people who slept under an ITN the night prior to the survey	0.73	0.15	0.15	0.11	0.81	0.08

Table 7.4: Environmental impact and interventions effect on the change of risk. Median and 95% credible intervals. Covariates have been standardised to make estimates comparable. [a] Proportion of people in the cluster who slept under an ITN the night prior to the survey, [b] Proportion of households in the cluster with at least one ITN, [c] Proportion of children aged 0-59 months who slept under a ITN the night prior to the survey, [d] Proportion of households in the cluster with at least one ITN per every two household members.

	Angola	**Liberia**	**Mozambique**	**Rwanda**	**Senegal**	**Tanzania**
Covariate	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)	M (95% CI)
dRainfall	0.08 $(-0.58, 0.73)$	0.15 $(0.07, 0.24)$	0.20 $(-0.01, 0.40)$	-0.27 $(-1.02, 0.31)$	0.51 $(0.09, 0.93)$	0.36 $(0.01, 0.73)$
dNDVI	0.05 $(-0.20, 0.33)$	0.10 $(0.01, 0.19)$	0.27 $(0.11, 0.43)$	0.46 $(-0.26, 1.13)$	-0.06 $(-0.34, 0.24)$	0.15 $(-0.06, 0.36)$
dLST night	0.39 $(-0.01, 0.77)$	0.09 $(0.01, 0.17)$	0.04 $(-0.13, 0.23)$	0.12 $(-0.46, 0.71)$	0.15 $(-0.23, 0.50)$	0.15 $(-0.08, 0.39)$
dLST day	-0.05 $(-0.42, 0.35)$	-0.07 $(-0.16, 0.01)$	0.44 $(0.21, 0.67)$	0.16 $(-0.39, 0.74)$	0.17 $(-0.18, 0.57)$	0.03 $(-0.18, 0.23)$
dITN	-0.64^{a} $(-0.96, -0.33)$	-0.05^{b} $(-0.14, 0.05)$	-0.09^{b} $(-0.26, 0.06)$	-0.18^{c} $(-0.71, 0.42)$	-0.33^{a} $(-0.62, -0.04)$	-0.15^{d} $(-0.40, 0.07)$
dIRS	-0.29 $(-0.77, 0.13)$	$--$	-0.02 $(-0.17, 0.12)$	$--$	-0.19 $(-0.49, 0.08)$	-0.04 $(-0.37, 0.29)$

Table 7.5: Estimated number of infected children related to the second survey period (second column) and estimated number of infection reductions (third column). Model based estimates of reduction in national level prevalence (forth column). Estimates are expressed in terms of medians and 95% credible intervals.

Country	**Number of infected children (last survey)**	**Number of infected children reduction**	**Prevalence reduction**
Angola	$290,816.74$ $(285,030.31 - 296,603.22)$	$315,451.32$ $(307,507.51 - 318,395.22)$	0.09 $(0.09 - 0.10)$
Liberia	$152,317.7$ $(149,148.81 - 153,486.53)$	$26,534.99$ $(23,930.58 - 28,139.39)$	0.04 $(0.03 - 0.04)$
Mozambique	$1,222,360.42$ $(1,159,052.54 - 1,225,668.23)$	$-3,981.212$ $(-9,258.26 - 1,295.84)$	-0.00 $(-0.00 - 0.01)$
Rwanda	$18,638.52$ $(18,199.56 - 19,077.48)$	$13,457.34$ $(12,632.58 - 14,282.01)$	0.01 $(0.00 - 0.01)$
Senegal	$53,934.71$ $(53,343.72 - 54,525.7)$	$36,433.38$ $(35,445.42 - 37,421.33)$	$0.02(0.02 - 0.02)$
Tanzania	$1,168,437.43$ $(1,121,895.33 - 1,172,980.21)$	$503,169.41$ $(457,736.31 - 508,602.62)$	0.06 $(0.05 - 0.06)$

7.4 Discussion

The risk of, and burden due to malaria has considerably declined over the past decade in sub-Saharan Africa. Spatial analysis to determine the risk of malaria and guide interventions proved to be a productive research area in recent years (Mapping Malaria Risk in Africa (Craig et al., 1999), Malaria Atlas Project (Guerra et al., 2007)). Initial mapping efforts were based on historical data that were heterogeneous in survey seasons and in age groups sampled across locations. MIS are nationally representative surveys, conducted during medium/high transmission seasons and they focus on a specific age group (children below the age of 5 years). Hence, MIS generate reliable data for estimating the geographical distribution of parasitemia risk as well as malaria burden (Gosoniu et al., 2010, 2012; Riedel et al., 2010; Giardina et al., 2012; Jima et al., 2010; Agusto et al., 2012).

In the current piece, we analyzed malaria parasite prevalence data in six sub-Saharan countries that carried out two parasitemia national surveys, either MIS or DHS, with a malaria module. We provide spatially explicit estimates of parasitemia risk and analyze patterns of change both in space and over time. We studied the spatial effects of ITN and IRS in reducing parasitemia risk among children. Coverage maps were produced to assess progress toward universal coverage and to identify areas that have been successful in scaling-up. Furthermore, we provide a methodology to identify areas where changes (increases or decreases) in malaria risk occurred, and to estimate the spatial effects of interventions in space.

Our analysis revealed different spatial patterns of changes in parasitemia risk between the two surveys depending on the country. For example, Angola experienced a decline in risk throughout most of the country. In Senegal and Rwanda, the relationship between environmental factors and malaria risk has become less evident in the second survey, probably explained by the high coverage of malaria control interventions that blurs the strong links between malaria and climate. In other countries where the change in risk varies substantially from one area to another (e.g., Tanzania), the overall risk has decreased, but there are 'clusters' of high parasitemia, leading to an estimated higher spatial variance. Mozambique is the only country that did not show a significant reduction in the model-based estimates of prevalence and number of infections among children below the age of 5 years. In some areas of Tanzania, there was an increased parasitemia, possibly due to the scarce coverage of interventions as well as the positive association with the increased amount of rainfall during the second survey (Table 7.2), carried out during the long rainy season. Where the surveys were conducted during similar seasons (e.g., Angola and Rwanda), the

observed decline in risk is not presumed to be due to changes in environmental conditions. This is confirmed by the lack of associations shown in Table 7.4. It is recommended that surveys take place during the highest transmission season, but this is not always feasible due to logistical issues and lack of human resources to reach remote areas. For this reason, the surveys might lack comparability, unless idiosyncrasies of environmental conditions of the individual surveys are taken into account.

To estimate the number of children below the age of 5 years with parasitemia, based on both first and second round surveys, we used readily available Afripop population data for the year 2010. The use of year-specific population data would lead to more accurate estimates of children infected in the two surveys but this information is not available at high spatial resolution. Considering the average annual rate of population growth of 2.63-3.82% (UN, 2012b), our figures tend to underestimate the burden reduction.

IRS coverage was assessed based on the proportion of households within a cluster that were sprayed in the last 6 months. ITN coverage, however, was measured by several indicators that are highly correlated. The employed Bayesian variable selection procedure allows one to choose the ITN coverage measure with the highest posterior probability to be included in the model. The effects of interventions at sub-national level was assessed by a geostatistical model with spatially varying coefficients because of the known "spatial" nature of intervention (i.e., community) effects and because we assume that neighboring areas are affected similarly by a specific intervention. In contrast, far away areas may show different effects because of different endemicity levels. In fact, interventions are an endogenous variable in modeling risk because they are known to reduce parasitemia risk, whereas areas with higher endemicity are more likely to show high intervention coverage.

We used the administrative division as the unit of analysis for estimating intervention spatial effects. In most of the countries, this division corresponds to the health division at which decisions can be made. Furthermore, it represents the smallest areas, with at least three observations, sufficient for estimating intervention spatial effects. Intervention effects varied from country to country and geographically within countries. In some regions, intervention effects were significant but no decline in prevalence was detected. However, there could be a decline in transmission that, in contexts of high transmission intensity, does not immediately correspond to a decline in risk. In other regions, no significant association between intervention measures and malaria risk was observed. This finding might be explained by other factors like resistance of vectors to insecticides, ITN condition (e.g., holes), or response bias. We are currently investigating other methods for including

additional information about bednet type and condition in the evaluation of intervention effects. The variability observed in ITN and IRS effects supports the need to evaluate intervention programs at local scales.

Intervention coverage data were considered only from the second survey, as we were interested in quantifying the odds of changes in parasitemia over time, for given values of coverage. We believe that the current intervention coverage is the consequence of the most recent distribution of ITN or implementation of IRS. Furthermore, the comparison of coverage levels in the first and second round surveys did not show large variations. The coverage maps produced can aid control programs in future interventions delivery. We considered countries where 100% of the population is at risk; however the heterogeneity in intervention coverage shows the need to 'spatially' assess progress toward universal coverage, identifying areas where coverage is currently particularly low. We are currently developing a more comprehensive intervention coverage map, combining different sources of information on ITN use and ownership (e.g. MICS, World Health Surveys, Living Standards Measurement Study-Integrated Survey), to provide more accurate estimates and to study their effect on risk.

We have defined positivity to parasitemia in our analysis based on microscopy (blood films examined under a microscope by trained personnel). Children were also tested with RDT in the field. The latter data were discarded from the analysis, as they conflicted with microscopy results. Several factors in the manufacturing process as well as environmental conditions may affect RDT performance in areas with high temperatures during transport and storage (Wongsrichanalai et al., 2007). Nevertheless, RDT results could be included in the analysis, taking into account the sensitivity and specificity of both tests.

Prevalence survey data are valuable for assessing progress toward disease elimination because they facilitate studying the pattern of change in malaria in space and time as well as estimate the effects of interventions on parasitemia risk changes. Prevalence is easily measured and its widespread use makes it suitable for monitoring and evaluation (Smith et al., 2009). However, evaluating changes in disease burden includes other indicators that measure incidence and mortality; this is not an easy task in many sub-Saharan countries, where health information systems are still very weak and death records are incomplete.

Intervention effects can be better estimated by entomological parameters because IRS and ITN primarily affect the mosquitoes' feeding cycle and death rates, leading to changes in the vectorial capacity. Population-level intervention benefits (such as reduction in prevalence) occur eventually as a result of changes in the transmission cycle. However, these

entomological parameters are not easy to collect over large areas.

Malaria modules, collecting data on behaviors and knowledge related to malaria, were introduced in 1999 as part of DHS. MIS with parasitemia testing were only implemented in 2006 (Fabic et al., 2012). Currently, MIS are either run as stand-alone surveys or incorporated into DHS. While reducing the costs of implementing the surveys, combining DHS and MIS in a unique survey implies extending the field study period to seasons of low or no transmission for malaria. This may result in an underestimation of the prevalence and may introduce bias in the spatial pattern of malaria risk. When MIS are run independently, they are usually carried out during the high transmission season in order to capture the highest number of infections. Sometimes this is not feasible. Nevertheless, the 'peak of transmission' is subject to geographic as well as annual variability. The use of 'rolling' cross-sectional surveys can provide a potential solution to these issues. Already proposed and implemented in the context of DHS and nutritional surveys (Rowe et al., 2009), a year-long rolling MIS was also implemented in one district of southern Malawi (Roca-Feltrer et al., 2012). Limiting the collection of household data to a few key malaria indicators reduces the time and costs implied by a rolling survey. However, their feasibility at national scale remains to be investigated.

To our knowledge, this is the first study presenting spatially explicit estimates of the probability of malaria risk reductions and of the effects of interventions in an ensemble of six sub-Saharan countries after adjusting for climatic factors. These estimates can be used to evaluate the accuracy of mathematical models that predict malaria risk under different levels of intervention coverage. Our maps provide important information for control program managers as they monitor and plan future interventions.

7.5 Supporting information

7.5.1 Profiles of the Countries Considered

Angola

Malaria transmission is highest in northern Angola, while the southern provinces have highly seasonal or epidemic malaria. Malaria is hyperendemic in northeastern Angola, including Cabinda province, a non-contiguous province in the north of the country. The central and coastal areas are largely mesoendemic with stable transmission. The four southern provinces bordering Namibia have highly seasonal transmission and are prone to epidemics. In the north, the peak malaria transmission season extends from March to May,

with a secondary peak in October or November. *P. falciparum* is responsible for more than 90% of all infections. The primary vectors in the high transmission areas are *Anopheles gambiae* and *An. funestus*.

Liberia

Malaria in Liberia is endemic throughout the country, with year-round transmission and a peak during September to October. All infections are attributable to *P. falciparum*. The climate is favorable to breeding the mosquitoes that are the major vectors of malaria, in particular *An. gambiae s.s.*

Mozambique

Although there are signs of declining malaria prevalence in Mozambique, the disease remains a major cause of morbidity and mortality. Malaria is endemic throughout the country, with regions ranging from mesoendemic to hyperendemic. Most residents live in areas where malaria is transmitted year round; peaks occur during and after the rainy season, between December and April. The climate in Mozambique creates a favorable environment for *An. gambiae*, *An. arabiensi*, and *An. funestus*. *P. falciparum* is the most common parasite and it is responsible for approximately 90% of all malaria infections.

Senegal

Senegal is one of the African countries in which dramatic progress has been made in malaria control since 2000. Transmission is seasonal, with high transmission occurring from September to December, toward the end of (and immediately following) the rainy season. While the south of Senegal is hyperendemic, the northern part of the country is hypo-endemic, with a low rate of malaria transmission. *P. falciparum* is the major malaria parasite species, accounting for more than 90% of all infections. The main vector species are *An. gambiae s.s.*, *An. arabiensis*, *An. funestus*, and *An. melas*.

Rwanda

Rwanda has made significant progress in scaling up malaria control interventions and decreases in malaria morbidity and mortality rates have been observed over the last years. Malaria is mesoendemic in the plains and epidemic-prone in the high plateaus and hills. Transmission occurs year-round, with two peaks (May-June, November-December) following distinct rainy seasons. Major vector species are *An. gambiae*, *An. arabiensis* and *An. funestus*.

Tanzania

Malaria is a major public health problem in Tanzania. On the mainland, 93% of the population lives in areas where malaria is transmitted. Unstable seasonal malaria transmission occurs in approximately 20% of the country, while stable malaria with seasonal variation occurs in another 20%. Transmission peaks during the long and short rainy seasons (May and December, respectively). The remaining malaria endemic areas in Tanzania (60%) are characterized by stable perennial transmission. *P. falciparum* accounts for 96% of malaria infections in Tanzania and the remaining 4% due to *P. malariae* and *P. ovale*. The principal vectors of malaria on the mainland are the *An. gambiae* complex (*An. gambiae s.l.* and *An. arabiensis*) and *An. funestus*. Zanzibar is characterized by very low levels of malaria transmission, although the islands remain vulnerable to outbreaks. Interventions have played an important role in achieving these levels, with high coverage of ITN and IRS.

7.5.2 Models

Estimating Parasitemia Risk at two Time Points

A geostatistical model was fitted to perform risk factor analysis and to assess the effect of climatic and environmental conditions on parasitemia risk for each country from 2007-2008. Let $Y_1(s_i)$ be the number of children who tested positive for parasitemia in cluster s_i in the first survey for each country and $N_1(s_i)$ the total number of children screened. We assume that each $Y_1(s_i)$ follows a Binomial distribution, i.e., $Y_1(s_i) \mid N_1(s_i), \pi_1(s_i) \sim Bin(N_1(s_i), \pi_1(s_i))$, $\forall i \in 1, ..., n_1$ where $\mathbf{s} = (s_1, s_2, ..., s_{n_1})$ is the set of locations surveyed and $\pi_1(\cdot)$ indicates the parasitemia risk. We formulate a Bayesian geostatistical model to analyze the parasitemia risk on the logit scale as follows:

$$logit(\pi_1(s_i)) = \boldsymbol{\beta}_1^T \mathbf{X}_1(s_i) + \omega_1(s_i)$$

where $\mathbf{X}_1(s_i)$ is the set of environmental predictors listed in the previous section evaluated at location s_i and $\boldsymbol{\beta}_1$ is the vector of regression coefficients. To account for spatial correlation in the response, we introduce the latent variables $\boldsymbol{\omega} = (\omega_1(s_1), \omega_1(s_2), ..., \omega_1(s_n))$ that follow a zero-mean multivariate normal distribution, i.e., $\boldsymbol{\omega}_1 \sim N(0, \boldsymbol{\Sigma}_1)$ with Matern covariance function between locations s_1 and s_2, that is, $\Sigma_1(s_1, s_2) = \frac{\sigma_1^2 (\kappa_1 d(s_1,s_2)^\nu K_\nu(\kappa_1 d(s_1,s_2))}{\Gamma(\nu) 2^{\nu-1}}$, where σ_1^2 is the spatial process variance, $d(s_1, s_2)$ is the distance between s_1 and s_2 and κ_1 is the scaling parameter. K_ν is the modified Bessel function of second kind and order ν. The Matern specification of the covariance matrix implies that the spatial range r_1, that

is the distance at which spatial correlation becomes negligible (i.e., smaller than 10%) is $r_1 = \frac{\sqrt{8}}{\kappa_1}$. We complete Bayesian model formulation by specifying prior distributions for the remaining parameters and hyperparameters. In particular, we choose log-gamma priors for $log(\sigma_1^2)$ and $log(r_1)$. Normal priors $N(0, 0.01)$ were assigned for the regression coefficients and intercept.

Bayesian kriging was used to predict malaria risk at high spatial resolution. Each country was divided into a grid formed by 1 km resolution pixels. Risk estimates at each pixel were obtained through the predictive distribution.

A geostatistical model similar to the one described above was employed to obtain contemporary estimates of malaria risk. In particular, we assumed a Binomial distribution for the number of positive children $Y_2(s_i')$, that is $Y_2(s_i') \mid N_2(s_i'), \pi_2(s_i') \sim Bin(N_2(s_i'), \pi_2(s_i'))$, $\forall i \epsilon 1, ..., n_1$ where $\mathbf{s}' = (s_1', s_2', ..., s_{n_1}')$ is the vector of locations sampled in the second survey, generally different than $\mathbf{s} = (s_1, s_2, ..., s_{n_1})$. We modeled $\pi_2(s_i')$ as a function of the environmental conditions in the second time point and a normally distributed spatial process $\boldsymbol{\omega}_2$, that is, $\boldsymbol{\omega}_2 \sim N(0, \boldsymbol{\Sigma}_2)$ with spatial variance σ_2^2 and scaling parameter κ_2. On the logit scale, the relation takes the form: $logit(\pi_2(s_i')) = \boldsymbol{\beta}_2^T \mathbf{X}_2(s_i') + \omega_2(s_i')$. The same grids were used for predictions in the second time point for each country.

Modeling the Effects of Interventions on the Change of Parasitemia Risk

We modeled the change of parasitemia risk (on the logit scale) as a function of the difference in climatic conditions between the two time points (surveys) and intervention coverage quantified as ITN and IRS as follows:

$$logit(\pi_2(s_i')) - logit(\pi_1(s_i')) = X_2(s_i')\boldsymbol{\beta}_2 - X_1(s_i')\boldsymbol{\beta}_1 + \alpha_1 ITN(s_i') + \alpha_2 IRS(s_i') + \omega_c(s_i') \quad (7.1)$$

where $ITN(s_i')$ is one of the bednet coverage measures discussed in the previous section, $IRS(s_i')$ indicates the proportion of sprayed households in cluster s_i' and $\omega_c(s_i')$ represents the latent process, modeling the spatial correlation of the parasitemia change. We assign $\boldsymbol{\omega}_c$ the prior distribution $\boldsymbol{\omega}_c \sim N(0, \boldsymbol{\Sigma}_c)$ with spatial variance σ_c^2 and scaling parameter κ_c. The coefficients α_1 and α_2 model the effect of intervention strategies on the change of parasitemia risk. In particular, $exp(\alpha_1)$ and $exp(\alpha_2)$ represent the expected change in the odds ratio of parasitemia (second survey *versus* first survey) associated with 1 unit variation in ITN and IRS coverage, respectively. We consider only the interventions in the second survey because we want to quantify the contribution of a certain level of ITN and IRS coverage in reducing malaria risk.

Since the value of parasitemia risk during the first survey $\pi_1(\cdot)$ was not directly available at locations $\mathbf{s}'$ of the second survey, we first aligned observations and derived the distribution of $Z(s'_i) = logit(\pi_1(s_i)$ conditionally on the spatial process and regression parameters, i.e., $Z(s'_i) \mid \boldsymbol{\beta}_1, \omega_1(s) \sim N(\boldsymbol{\beta}_1^T \mathbf{X}_1(s'_i) + \Sigma_{ss'}\Sigma_s^{-1}\omega_1(s), \Sigma_{s'} - \Sigma_{s's}\Sigma_{s'}^{-1}\Sigma_{ss'})$. The joint posterior distribution of the parameters is obtained by $[Y_2(\mathbf{s}') \mid \boldsymbol{\beta}_2, Z(\mathbf{s}'), \omega_2(\mathbf{s}')][Z(\mathbf{s}') \mid \boldsymbol{\beta}_1, \omega_1(\mathbf{s})][\omega_2(s') \mid \sigma_2^2, \kappa_2][\omega_1(s) \mid \sigma_1^2, \kappa_1][\boldsymbol{\beta}_1][\boldsymbol{\beta}_2]$ and Equation (7.1) can now be rewritten as $logit(\pi_2(s'_i)) = Z(s'_i) + \boldsymbol{\beta}_0(\mathbf{X}_2(s'_i) - \mathbf{X}_1(s'_i)) + \alpha_1 ITN(s'_i) + \alpha_2 IRS(s'_i) + \omega_c(s'_i)$, where we have expressed β_0 as $\beta_0 = \beta_2 - \beta_1$.

To account for potential interactions with endemicity levels, we fitted a second model to estimate different intervention effects for each regional unit (first administrative division). The model is expressed as follows

$$logit(\pi_2(s'_i)) = Z(s'_i) + \boldsymbol{\beta}_0(\mathbf{X}_2(s'_i) - \mathbf{X}_1(s'_i)) + \alpha_1(A_{s'_i}) ITN(s'_i) + \alpha_2(A_{s'_i}) IRS(s'_i) + \omega_c(s'_i).$$

The intervention effects are now denoted by $\alpha_k(A_{s'_i})$, $(k = 1, 2)$ and defined on a regional level and $A(s'_i)$ denotes the region where s'_i falls. Each $\alpha_k(A_i)$ is factorized by the sum of a conditional autoregressive effect that takes into account the similarity of the effects across regions and an independent random part, i.e., $\alpha_k(A_i) = \alpha_k^c(A_i) + \grave{o}_k(A_i)$, where $\alpha_k^c(A_i) \mid \alpha_k^c(A_j), i \neq j, \tau_{kc} \sim N(\frac{1}{n_i}\sum_{i\sim j}\alpha_k^c(A_j), \frac{1}{n_i\tau_{kc}})$ indicating with $i \sim j$ the neighborhood relation between area A_i and A_j, and $\grave{o}_k(A_i) \sim N(0, 1/\tau_{k\grave{o}})$.

Selecting Bednet Coverage Indicators

The bednet coverage indicators presented in the previous section are highly correlated; therefore we have defined a Bayesian variable selection procedure that selects only one (or none) bednet ownership indicator and one (or none) bednet use indicator. We denote each one of the three bednet ownership indicators with ITN_j^0 $(j = 1, ...3)$, while ITN_j^u $(j = 1, 2)$ represents the indicators of bednet use. In particular, we assume spike and slab *a priori* distributions (Ishwaran and Rao, 2005) for the coefficients

$$\alpha_j^0 \sim N(0, \sigma_j^2) + (1 - \gamma_j^0)N(0, \delta\sigma_j^2), \quad j = 1, 2, 3$$

$$\alpha_j^u \sim N(0, \sigma_j^2) + (1 - \gamma_j^u)N(0, \delta\sigma_j^2), \quad j = 1, 2$$

where σ_j^2 is assigned a gamma distribution and δ is a small constant that shrinks the variance toward very small values if γ_j is zero and ITN_j is not considered relevant to the model. A Bernoulli (0.5) prior distribution was assumed for γ_j^u, while a categorical prior

distribution was adopted for γ_j^0, i.e., $Categorical(\mathbf{p})$, with $\mathbf{p} = (1/3, 1/3, 1/3)$ to constrain the model in choosing only one indicator and assuming equal probability of inclusion *a priori*.

Estimating the Probability of Parasitemia Risk Reduction

Conditional on the data and the model parameters, the predictive density for each time point can be expressed as:

$$P(\mathbf{Y}_t^0 \mid \mathbf{Y}_t, \mathbf{N}_t) = \int P(\mathbf{Y}_t^0 \mid \boldsymbol{\beta}_t, \boldsymbol{\omega}_t^0) P(\boldsymbol{\omega}_t^0 \mid \boldsymbol{\omega}_t, \sigma_t^2, \kappa_t) P(\boldsymbol{\beta}_t, \boldsymbol{\omega}_t, \sigma_t^2, \kappa_t \mid \mathbf{Y}_t, \mathbf{N}_t) \, d\boldsymbol{\beta}_t d\boldsymbol{\omega}_t^0 d\boldsymbol{\omega}_t d\sigma_t^2 d\kappa_t$$

where $\mathbf{Y}_t^0 = (Y_t(s_1^0), Y_t(s_2^0), ..., Y_t(s_m^0))$ are the predicted number of positives in each pixel s_i^0 $\forall i \in 1, \ldots, m$, and time point t, $t = 1, 2$, and $P(\boldsymbol{\beta}_t, \boldsymbol{\omega}_t, \sigma_t^2, \kappa_t \mid \mathbf{Y}_t, \mathbf{N}_t)$ is the joint posterior distribution of parameters and hyperparameters while $\boldsymbol{\omega}_t = (\omega_t(s_1^0), \omega_t(s_2^0), ..., \omega_t(s_m^0))$ is the vector of the spatial process at new sites. Conditional on the spatial process and regression parameters, $Y_t^0(s_i^0) \sim Bin(N_t(s_i^0), \pi_t^0(s_i^0))$, with risk $\pi_t^0(s_i^0)$, given by $logit(\pi_t^0(s_i^0)) = \boldsymbol{\beta}_t^T \mathbf{X}_t(s_i^0) + \omega_t^0(s_i^0)$ and $N_t(s_i^0)$ indicates the population of children living in the pixel s_i^0.

To estimate the probability of risk reduction at each pixel we have compared $\pi_1^0(s_i^0)$ with $\pi_2^0(s_i^0)$, $\forall i \epsilon 1, ..., m$ and calculated $P(\pi_2^0(s_i^0) < \pi_1^0(s_i^0))$. Furthermore, we have estimated the total number of children infected by country during the first and second survey period ($\sum_{i=1}^m Y_1^0(s_i^0)$ and $\sum_{i=1}^m Y_2^0(s_i^0)$ respectively) and their difference. We made use of population data provided by Afripop (Tatem et al., 2013a) that consist of spatial estimations of number of children less than 5 years of age per 1 km^2 in 2010.

Acknowledgments

We would like to acknowledge Measure DHS for making the data available and the Swiss National Science Foundation (SNSF) Swiss Programme for Research on Global Issues for Development (R4D) project no. IZ01Z0-147286 for their financial support.

Chapter 8

Discussion

8.1 Significance

This research work contributes to the field of malaria epidemiology in sub-Saharan Africa and develops novel spatio-temporal statistical methodology to estimate disease risk from contemporary and historical survey data and assess the effectiveness of control interventions. The modelling strategies, methodologies and results of our research are described in the six manuscripts included as chapters in this thesis. Each chapter of the thesis includes a detailed discussion; this section provides an overview of the main research contributions and discusses implications in malaria control, limitations and extensions for future research.

8.1.1 Contributions in spatial statistics

Geo-referenced prevalence survey data are commonly modelled via a Binomial distribution. However, when the data consist of a large number of zeros, Binomial models are not appropriate and may underestimate the probabilty of zero-prevalence. In Chapter 2, we propose the use of zero-inflated Binomial models (Lambert, 1992), defined as two-component mixtures of a point mass at zero with a Binomial distribution, and show that they are more suitable in these situations by comparing their predictive ability with standard binomial analogues. Zero-inflated Binomial models for prevalence data have not been applied before in the context of geostatistical modelling of infectious disease data. To our knowledge, the only application is in the modeling of sparse malaria entomological data (Amek et al., 2011). The main research contribution of the thesis in this area was to explore the different modelling strategies for zero-inflated data and suggest model formulations in a geostatistical setting. In particular, hurdle models, that combine a point mass at zero with a truncated Binomial distribution for the non zero values (Mullahy, 1986) are studied and compared to zero-inflated Binomial models in geostatistical settings. Zero-inflated Binomial models allow two types of zeros: "structural" that arise from the point mass at zero and "chance" zeros modelled by the Binomial distribution; Binomial hurdle models treat zeros and non-zeros separately. We observe that zero-inflated models may suffer from weak identifiability of the spatial process characterizing the mixing probability, while hurdle models are in general more stable (Chapter 3).

Most applications of geostatistical models assume that the spatial correlation is a function of the distance and independent of locations, that is, the spatial process is stationary. This hypothesis is not appropriate when malaria data are analyzed over large areas, since local characteristics influence the spatial structure differently at various locations. We have developed Bayesian methodology to model non-stationary geostatistical data when

the study area consists of irregularly shaped zones with different characteristics. Non-stationarity was modelled as weighted combinations of stationary spatial covariance functions and specific spatially varying weights are proposed to account for irregularly shaped partitions of the study area. The proposed methodology improves disease risk prediction over large areas compared to commonly used stationary geostatistical models and the weights introduced into the model smooth the predicted surface at the borders of the zones (Chapter4).

Variable selection and model choice are essential components in statistical modelling. However, little has been done in geostatistical settings. Typically, spatial correlation is ignored in the selection of explanatory variables, influencing model selection as well as parameter estimation (Hoeting et al., 2006b). In this thesis we tackle the problem under several model specifications. In Chapter 2, we apply Bayesian variable selection methods to choose the environmental predictors determining the malaria risk in zero-inflated models and in Chapter 3 we propose Bayesian variable selection methods to allow the choice of both fixed and random effects in modelling the probability of (extra-) zeros as well as the rest of the data (arising from Binomial or truncated Binomial distributions). Specific prior distributions for spatial process selection based on non-zero random effects variances are proposed and analyzed. Over large areas, with a natural partition (e.g. ecological zones), the effects of environment and climate may depend on the regions and may be non-linear. In Chapter 4, we develop a Bayesian variable selection procedure for non-stationary models that allows the choice of covariates and their corresponding functional forms (e.g. linear, categorical, spline) by regions. Spatially varying weights were used in the regression model to take into account the dependence of the covariates affecting the disease outcome at a given location not only on the zone associated to the location but also on the neighboring regions within a certain radius. Variable selection methods were proved to be useful in the selection of one among correlated predictors, for example in the choice of intervention coverage indicators. ITN coverage was defined using different measures in Chapter 2 and Chapter 7 and a Multinomial (or categorical) prior was assigned to the inclusion probabilities.

Motivated by the question of assessing the effectiveness of intervention strategies on reducing malaria risk, spatially varying coefficients models were developed (Chapter 7). Using a conditional autoregressive prior on the coefficients, these models allowed the estimation of covariates' effects at sub-national levels.

When information form multiple geo-referenced surveys is combined, sources of heterogeneity should be properly accounted for and modelled. Age and season are important source of heterogeneity in prevalence modeling. In Chapter 6, we propose an approach that couples mathematical transmission models with spatial statistical models in a Bayesian framework, allowing the estimation of age and season specific risk at high spatial resolution. We achieve a joint model formulation for prevalence and incidence data by embedding an extended version of the catalytic model by Pull and Grab (1974) into our hierarchical Bayesian structure, thus accounting for the uncertainty arising form the age/season standardization in the risk estimation. Previous work on modeling heterogeneity in geo-referenced surveys for malaria mapping (Kleinschmidt et al., 2001; Gemperli et al., 2006a; Gosoniu et al., 2006) focused on a specific age group, discarding surveys with different or overlapping ages of the population resulting in unreliable malaria transmission estimates. More recent works (Gemperli et al., 2006b; Gosoniu, 2008; Hay et al., 2009) were based on a 2-step procedure to (i) obtain age-correction factors and (ii) separately fit age-adjusted prevalence estimates in a geostatistical model, ignoring adjustment uncertainty. Moreover, the heterogeneity due to the different survey periods was not considered. However, seasonality is one of the most important sources of heterogeneity in malaria prevalence estimates: we incorporate season-dependence in our model by allowing space and time variations in the force of infection. We obtain spatially varying age-prevalence curves that included a diagnostic sensitivity as a decreasing function of age.

8.1.2 Implications in malaria epidemiology and control

The past decade has seen decreases in malaria caused by *Plasmodium falciparum*, the most deadly and predominant parasite species in Africa. Malaria reduction is part of the Millennium Development Goals (MDGs), aiming to halve malaria incidence by 2015 as compared to 1990 (UN, 2012a). In 2008, the Global Malaria Action Plan (GMAP), put forward by the Roll Back Malaria (RBM) Partnership (Global Partnership to Roll Back Malaria, 2010), advocated reducing malaria cases by 75% (from 2000 levels) and malaria deaths to near zero, by 2015. Since 2007, WHO has recommended universal ITN coverage. Malaria Indicator Surveys (MIS) were developed by RBM to coordinate global efforts to fight malaria.

Analyzing MIS surveys, this thesis contributes to the monitoring and evaluation of the progress toward these targets. In particular, we provide improved estimates of malaria risk and intervention coverage at high spatial resolution. We have produced smooth maps of

parasitemia risk for Angola, Liberia, Mali, Mozambique, Rwanda, Senegal, Tanzania and Zambia and we have assessed the temporal trends along with the scale-up of intervention coverage in Angola, Liberia, Mozambique, Rwanda, Senegal and Tanzania. The maps produced are useful tools for identifying priority areas for disease control. The spatial variability observed in ITN and IRS effectiveness supports the need to evaluate intervention programs at local scales. We obtained spatially explicit estimates of the probability of malaria risk reductions. These estimates are of great value in validating mathematical models that predict malaria risk under different levels of intervention coverage. Our maps provide important information for control program managers as they monitor and plan future interventions.

The focus on malaria eradication in selected countries in sub-Saharan Africa, suggests that forthcoming surveys will include a large number of locations with zero prevalence and the zero-inflated models developed in this thesis would provide a suitable way to provide accurate risk estimates. Moreover, the factors leading to the onset/end of transmission in a specific area may differ from the ones causing an increase or decrease in malaria risk. Therefore the proposed variable selection strategy may be useful in identifying determinants of transmission suitability and conditional malaria risk.

Some countries in sub-Saharan Africa do not have a national survey and rely on historical survey data for spatial risk estimation. The methodology of age/season standardization, outlined in Chapter 6, can be applied in these settings.

Obtaining accurate risk estimates depends on the quality of the environmental predictors used for building predictive models. Chapter 5 focuses on assessing the effect of very high resolution environmental covariates derived by remote sensing sources on spatially explicit malaria burden estimates in geostatistical models. This work was carried out as part of the MALAREO project, (`www.malareo.eu`), supported by the Seventh Framework Programme (FP7) space research program, with the objective of building GIS, EO and spatial statistics capacities and implementing EO products supporting the malaria control programmes in Southern Africa. The main product created within the MALAREO project is a high resolution (5m) land cover/land use map based on RapidEye technology. A secondary outcome was an "enhanced" population map, obtained by the combination of the LC layer with census data, aggregated at 100m resolution. Both land cover and population layers were used in the geostatistical analysis of the Mozambican DHS in 2011 and showed higher predictive ability in the comparison with lower resolution products.

8.2 Limitations

In the estimation of changes in the number of children below the age of 5 years with parasitemia, during the time period considered (2006-2011), we used readily available Afripop population data for the year 2010. The use of year-specific population data would lead to more accurate estimates of children infected in the two surveys but this information is not available at high spatial resolution. Considering the average annual rate of population growth of 2.63-3.82% (UN, 2012b), our figures tend to underestimate the burden reduction. We used the administrative division as the unit of analysis for estimating intervention effectiveness. While this assumption is not justified by any biological reason, in most of the countries, this division corresponds to the health division at which decisions can be made. In some regions, no significant association between intervention measures and malaria risk was observed. This finding might be explained by other factors like resistance of vectors to insecticides, ITN condition (e.g., holes), or response bias. Prevalence survey data are valuable for assessing progress toward disease elimination because they facilitate studying the pattern of change in malaria in space and time as well as estimate the effectiveness of interventions on parasitemia risk changes. Prevalence is easily measured and its widespread use makes it suitable for monitoring and evaluation (Smith et al., 2009). However, evaluating changes in disease burden includes other indicators that measure incidence and mortality; this is not an easy task in many sub-Saharan countries, where health information systems are still very weak and death records are incomplete. Intervention effectiveness can be better estimated by entomological parameters because IRS and ITN primarily affect the mosquitoes' feeding cycle and death rates, leading to changes in the vectorial capacity. Population-level intervention benefits (such as reduction in prevalence) occur eventually as a result of changes in the transmission cycle. However, these entomological parameters are not easy to collect over large areas.

8.3 Extensions

Several methodological approaches proposed in this thesis can be applied to other environmentally driven diseases. For example, sparse geo-referenced survey data can arise from studies on other parasitic diseases (e.g. neglected diseases such as soil-transmitted helminths) making suitable the zero-inflated modelling strategy proposed in Chapter 2 and Chapter 3; the approach to model non-stationarity over large areas (Chapter 4) can be used to address the current needs of international agencies (e.g. World Health Organization, The Global Fund etc...) which are interested in global atlases of infectious disease burden and

estimates of the required amount of preventive and curative treatments; schistosomiasis, a chronic and poverty-promoting disease caused by trematodes of the genus *Schistosoma*, is characterized by typical age-prevalence curves in endemic setting Raso et al. (2007), therefore age-specific disease risk estimates can be modelled follwing the methodology outlined in Chapter 5.

The GPS latitude/longitude positions in MIS/DHS survey were randomly displaced so that clusters contain a minimum of 0 and a maximum of 5 kilometers of positional error. This misplacement was added for respondent confidentiality reasons. A general extension of this work would be to explicitly model the bias introduced by the positional error, for example following the approach proposed by Fanshawe and Diggle (2011).

Throughout this work, we have defined positivity to parasitemia based on microscopy (blood films examined under a microscope by trained personnel). Children were also tested with RDT in the field. The latter data were discarded from the analysis, as they conflicted with microscopy results. Several factors in the manufacturing process as well as environmental conditions may affect RDT performance in areas with high temperatures during transport and storage (Wongsrichanalai et al., 2007). Nevertheless, RDT results could be included in the analysis, taking into account the sensitivity and specificity of both tests.

We are currently investigating other methods for including additional information about bednet type and condition in the evaluation of intervention effectiveness. Furthermore, a more comprehensive intervention coverage map can be obtained combining different sources of information on ITN use and ownership (e.g. MICS, World Health Surveys, Living Standards Measurement Study-Integrated Survey), to provide more accurate estimates and to study their effect on risk.

Socio-economic status represents an important determinant of the disease (Tusting et al., 2013). However, few studies have linked it to malaria risk mapping. This can be due to the difficulty of defining a representative "poverty index" as well as to generating consistent spatial estimates. The Oxford Poverty and Human Development Initiative, for instance, proposed a Multidimensional Poverty Index (MPI, `http://www.ophi.org.uk/policy/multidimensional-poverty-index/`) to capture the multiple aspects that constitute poverty. Tatem et al. (2013b) produced surfaces of poverty for selected countries plotting the proportion of people per 100 square meters living in poverty, as defined by the MPI, with associated uncertainty metric. Our work can be extended to test the association of this poverty indicator on malaria risk with geostatistical models.

Overall, this thesis developed novel statistical methodology to improve malaria risk

estimates and assess effectiveness of interventions at high spatial resolution. The described methods may not be applied directly from field practitioners or NMCP personnel, since they require specialized knowledge. However, we are currently working on the implementation of the models with entirely free softwares and user-friendly interfaces to be distributed to the NMCPs and facilitate their work in monitoring and evaluating the progress in the fight of the disease.

Bibliography

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and ecological statistics*, 9(4):341–355.

Agusto, F. B., Del Valle, S. Y., Blayneh, K. W., Ngonghala, C. N., Goncalves, M. J., Li, N., Zhao, R., and Gong, H. (2012). The impact of bed-net use on malaria prevalence. *Journal of theoretical biology*, 320:58–65.

Alonso, P. L. and Tanner, M. (2013). Public health challenges and prospects for malaria control and elimination. *Nature medicine*, 19(2):150–155.

Amek, N., Bayoh, N., Hamel, M., Lindblade, K. A., Gimnig, J., Laserson, K. F., Slutsker, L., Smith, T., and Vounatsou, P. (2011). Spatio-temporal modeling of sparse geostatistical malaria sporozoite rate data using a zero inflated binomial model. *Spatial and Spatio-temporal Epidemiology*, 2(4):283–290.

Amek, N., Bayoh, N., Hamel, M., Lindblade, K. A., Gimnig, J. E., Odhiambo, F., Laserson, K. F., Slutsker, L., Smith, T., and Vounatsou, P. (2012). Spatial and temporal dynamics of malaria transmission in rural Western Kenya. *Parasites & Vectors*, 5(1):1–13.

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office.

Banerjee, S., Gelfand, A., Knight, J., and Sirmans, C. (2004). Spatial modeling of house prices using normalized distance-weighted sums of stationary processes. *Journal of Business and Economic Statistics*, 22:206–213.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32:870–897.

Bauwens, I., Franke, J., and Gebreslasie, M. (2011). Malareo - earth observation to support malaria control in southern Africa. *IEEE International Geoscience and Remote Sensing Symposium*, pages 3–6.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Bové, D. S., Held, L., and Kauermann, G. (2012). Mixtures of g-priors for generalised additive model selection with penalised splines. http://arxiv.org/pdf/1108.3520.pdf.

Bruce-Chwatt, L. J. et al. (1980). *Essential malariology.* William Heinemann Medical Books Ltd.

Burgert, C., Bradley, S., Eckert, E., and Arnold, F. (2012). Improving estimates of insecticide-treated mosquito net coverage from household surveys: Using geographic coordinates to account for endemicity and seasonality. Technical report, Calverton Maryland ICF International MEASURE DHS.

Ceccato, P., Connor, S., Jeanne, I., and Thomson, M. (2005). Application of geographical information systems and remote sensing technologies for assessing and monitoring malaria risk. *Parassitologia*, 47(1):81–96.

Centre de Recherche pour le Développement Humain (Sénégal) et ICF Macro (2009). Enquête nationale sur le paludisme au sénégal 2008-2009. Technical report.

Chammartin, F., Hürlimann, E., Raso, G., N'Goran, E. K., Utzinger, J., and Vounatsou, P. (2013a). Statistical methodological issues in mapping historical schistosomiasis survey data. *Acta tropica*, 128(2):345–352.

Chammartin, F., Scholte, R., Guimarães, L. H., Tanner, M., Utzinger, J., and Vounatsou, P. (2013b). Soil-transmitted helminth infection in South America: a systematic review and geostatistical meta-analysis. *The Lancet infectious diseases*, 13(6):507–518.

Chammartin, F., Scholte, R., Malone, J., Bavia, M. E., Nieto, P., Utzinger, J., and Vounatsou, P. (2013c). Modelling the geographical distribution of soil-transmitted helminth infections in bolivia. *Parasites & vectors*, 6(1):1–14.

Chanda, E., Coleman, M., Kleinschmidt, I., Hemingway, J., Hamainza, B., Masaninga, F., Chanda-Kapata, P., Baboo, K. S., Dürrheim, D. N., and Coleman, M. (2012). Impact assessment of malaria vector control using routine surveillance data in Zambia: implications for Monitoring and Evaluation. *Malaria Journal*, 11:437.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59:762–769.

Chitnis, N., Schapira, A., Smith, T., and Steketee, R. (2010). Comparing the effectiveness of malaria vector-control interventions through a mathematical model. *The American journal of tropical medicine and hygiene*, 83(2):230.

Cibulskis, R. E., Aregawi, M., Williams, R., Otten, M., and Dye, C. (2011). Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods. *PLoS medicine*, 8(12):e1001142.

Clements, A., Garba, A., Sacko, M., Touré, S., Dembelé, R., Landouré, A., Bosque-Oliva, E., Gabrielli, A. F., and Fenwick, A. (2008). Mapping the probability of schistosomiasis and associated uncertainty, West Africa. *Emerging Infectious Diseases*, 14:1629–1632.

Clements, A. C., Reid, H. L., Kelly, G. C., and Hay, S. I. (2013). Further shrinking the malaria map: how can geospatial science help to achieve malaria elimination? *The Lancet infectious diseases*, 13(8):709–718.

Craig, M., Sharp, B., Mabaso, M., and Kleinschmidt, I. (2007). Developing a spatial-statistical model and map of historical malaria prevalence in botswana using a staged variable selection procedure. *International journal of health geographics*, 6(1):44.

Craig, M., Snow, R., and Le Sueur, D. (1999). A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitology today*, 15(3):105–111.

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008). Bivariate binomial spatial modeling of loa loa prevalence in tropical Africa. *Journal of the American Statistical Association*, 103(481):21–37.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley-Interscience, revised edition.

Curtis, S., Banerjee, S., and Ghosal, S. (2014). Fast bayesian model assessment for non-parametric additive regression. *Computational Statistics & Data Analysis*, 71:347–358.

De Meillon, B. (1951). Malaria survey of south-West Africa. *Bulletin of the World Health Organization*, 4(3):333.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36.

Dietz, K., Molineaux, L., and Thomas, A. (1974). A malaria model tested in the African savannah. *Bulletin of the World Health Organization*, 50(3-4):347.

Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M. (2002). Childhood malaria in the gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506.

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

Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

ESRI (2011). Redlands, CA: Environmental Systems Research Institute.

Fabic, M. S., Choi, Y., and Bird, S. (2012). A systematic review of demographic and health surveys: data availability and utilization for research. *Bulletin of the World Health Organization*, 90(8):604–612.

Fanshawe, T. and Diggle, P. (2011). Spatial prediction in the presence of positional error. *Environmetrics*, 22(2):109–122.

FAO (2000). Global forest resources assessment. http://www.fao.org/forestry/fra/2000/report/en/.

Fernandes, M. V. M., Schmidt, A. M., and Migon, H. S. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, 9(1):3–25.

Filipe, J. A., Riley, E. M., Drakeley, C. J., Sutherland, C. J., and Ghani, A. C. (2007). Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. *PLoS computational biology*, 3(12):e255.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154.

Finley, A. O., Banerjee, S., and MacFarlane, D. W. (2011). A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, 106(493):31–48.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Franke, J., Bauwens, I., Deleu, J., de Montpellier, C., Dlamini, S., Gebreslasie, M., Giardina, F., Siegert, F., and Vounatsou, P. (2013). *MALAREO MapAtlas - Exploring the spatial dimension of malaria and its explaining environmental factors in Southern Africa by Earth Observation*.

Fuentes, M. (2001). A new high frequency kriging approach for nonstationarity environmental processes. *Environmetrics*, 12:469–483.

Garske, T., Ferguson, N. M., and Ghani, A. C. (2013). Estimating air temperature and its influence on malaria transmission across Africa. *PloS ONE*, 8(2):e56487.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146.

Gelfand, A. E., Banerjee, S., and Carlin, B. P. (2004a). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman Hall, Boca Raton.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. (2004b). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312.

Gelfand, A. E., Sirmans, H. K. K. C. F., and Banerjee, S. (2003). Spatial modelling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387–396.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Dyk, D. A. V., Huang, Z., and Boscardin, W. J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal Of Computational And Graphical Statistics*, 17:95–122.

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models.

Gemperli, A., Sogoba, N., Fondjo, E., Mabaso, M., Bagayoko, M., Briët, O. J., Anderegg, D., Liebe, J., Smith, T., and Vounatsou, P. (2006a). Mapping malaria transmission in West and Central Africa. *Tropical Medicine & International Health*, 11(7):1032–1046.

Gemperli, A. and Vounatsou, P. (2006). Strategies for fitting large, geostatistical data in mcmc simulation. *Communications in Statistics-Simulation and Computation*, 35:331–345.

Gemperli, A., Vounatsou, P., Sogoba, N., and Smith, T. (2006b). Malaria mapping using transmission models: application to survey data from Mali. *American Journal of Epidemiology*, 163(3):289–297.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gething, P. W., Elyazar, I. R., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., Patil, A. P., Tatem, A. J., Howes, R. E., Myers, M. F., et al. (2012). A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS neglected tropical diseases*, 6(9):e1814.

Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I., Johnston, G. L., Tatem, A. J., Hay, S. I., et al. (2011). A new world malaria map: *Plasmodium falciparum* endemicity in 2010.

Giardina, F., Gosoniu, L., Konate, L., Diouf, M. B., Perry, R., Gaye, O., Faye, O., and Vounatsou, P. (2012). Estimating the burden of malaria in senegal: Bayesian zero-inflated binomial geostatistical modeling of the mis 2008 data. *PLoS ONE*, 7(3):e32625.

Giardina, F., Kasasa, S., Sié, A., Utzinger, J., Tanner, M., and Vounatsou, P. (2013a). Assessing the impact of interventions on malaria parasitemia risk in Africa: A spatio-temporal analysis of malaria indicator survey data. *Submitted*.

Giardina, F., Sogoba, N., and Vounatsou, P. (2013b). Bayesian variable selection in semi-parametric and non-stationary geostatistical models: an application to mapping malaria risk in Mali. (submitted).

Giorgi, E., Sesay, S. S., Terlouw, D. J., and Diggle, P. J. (2013). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *arXiv preprint arXiv:1308.2790*.

Global Partnership to Roll Back Malaria (2010). Roll back malaria progress & impact series: Focus on senegal. Technical report, World Health Organization: Geneva.

Gosoniu, L. (2008). *Development of Bayesian geostatistical models with applications in malaria epidemiology*. PhD thesis, University of Basel.

Gosoniu, L., Msengwa, A., Lengeler, C., and Vounatsou, P. (2012). Spatially explicit burden estimates of malaria in tanzania: Bayesian geostatistical modeling of the malaria indicator survey data. *PloS ONE*, 7(5):e23966.

Gosoniu, L., Veta, A. M., and Vounatsou, P. (2010). Bayesian geostatistical modeling of Malaria Indicator Survey data in Angola. *PLoS ONE*, 5(3):e9322.

Gosoniu, L. and Vounatsou, P. (2011). Geostatistical variable selection in malaria risk mapping: an application to the Liberia Malaria Indicator Survey data. *Submitted*.

Gosoniu, L., Vounatsou, P., Sogoba, N., Maire, N., and Smith, T. (2009). Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. *Computational Statistics & Data Analysis*, 53(9):3358–3371.

Gosoniu, L., Vounatsou, P., Sogoba, N., and Smith, T. (2006). Bayesian modelling of geostatistical malaria risk data. *Geospatial Health*, 1(1):127–139.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Griffin, J. T., Hollingsworth, T. D., Okell, L. C., Churcher, T. S., White, M., Hinsley, W., Bousema, T., Drakeley, C. J., Ferguson, N. M., Basáñez, M.-G., et al. (2010). Reducing *plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS medicine*, 7(8):e1000324.

Guerra, C. A., Hay, S. I., Lucioparedes, L. S., Gikandi, P. W., Tatem, A. J., Noor, A. M., and Snow, R. W. (2007). Assembling a global database of malaria parasite prevalence for the malaria atlas project. *Malaria journal*, 6(1):17.

Hall, D. B. (2000). Zero-inflated Poisson and Binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R., Brooker, S., et al. (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS medicine*, 6(3):e1000048.

Hay, S. I. and Snow, R. W. (2006). The malaria atlas project: developing global maps of malaria risk. *PLoS Medicine*, 3(12):e473.

Hedt, B. L. and Pagano, M. (2011). Health indicators: eliminating bias from convenience sampling estimators. *Statistics in Medicine*, 30(5):560–568.

Higdon, D. (1998). A process-convolution approach to modeling temperatures in the north atlantic ocean. *Journal of Environmental and Ecological Statistics*, 5:173–190.

Hoeting, J. A., Davis, R., Merton, A., and Thomspon, S. (2006a). Model selection for geostatistical models. *Ecological Applications*, 16:87–98.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E. (2006b). Model selection for geostatistical models. *Ecological Applications*, 16(1):87–98.

Hwang, J., Graves, P. M., Jima, D., Reithinger, R., Kachur, S. P., et al. (2010). Knowledge of malaria and its association with malaria-related behaviorsresults from the malaria indicator survey, ethiopia, 2007. *PLoS ONE*, 5(7):e11692.

Ijumba, J. and Lindsay, S. (2001). Impact of irrigation on malaria in Africa: paddies paradox. *Medical and veterinary entomology*, 15(1):1–11.

Ishwaran, H. and Rao, J. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33:730–773.

Jima, D., Getachew, A., Bilak, H., Steketee, R. W., Emerson, P. M., Graves, P. M., Gebre, T., Reithinger, R., Hwang, J., et al. (2010). Malaria indicator survey 2007, ethiopia:

coverage and use of major malaria prevention and control interventions. *Malaria Journal*, 9(58):10–1186.

Jochmann, M. (2009). What belongs where? variable selection for zero-inflated count models with an application to the demand for health care. Working papers 0923, University of Strathclyde Business School, Department of Economics.

Kalluri, S., Gilruth, P., Rogers, D., and Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. *PLoS pathogens*, 3(10):e116.

Kasasa, S., Asoala, V., Gosoniu, L., Anto, F., Adjuik, M., Tindana, C., Smith, T., Owusu-Agyei, S., Vounatsou, P., et al. (2013). Spatio-temporal malaria transmission patterns in navrongo demographic surveillance site, northern ghana. *Malaria journal*, 12(1):63.

Kilian, A., Wijayanandana, N., and Ssekitoleeko, J. (2010). Review of delivery strategies for insecticide treated mosquito nets: are we ready for the next phase of malaria control efforts? *TropIKA. net*, 1(1).

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63:690–698.

Kleinschmidt, I., Bagayoko, M., Clarke, G., Craig, M., and Le Sueur, D. (2000). A spatial statistical approach to malaria mapping. *International Journal of Epidemiology*, 29(2):355–361.

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

Kleinschmidt, I. and Sharp, B. (2001). Patterns in age-specific malaria incidence in a population exposed to low levels of malaria transmission intensity. *Tropical Medicine & International Health*, 6(12):986–991.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lagazio, C., Dreassi, E., and Biggeri, A. (2001). A hierarchical Bayesian model for space-time variation of disease risk. *Statistical Modelling*, 1(1):17–29.

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

Lawson, A. (2013). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Second Edition*. Chapman & Hall/Crc Interdisciplinary Statistics. Taylor & Francis Group.

Lengeler, C. et al. (2004). Insecticide-treated bed nets and curtains for preventing malaria. *Cochrane Database Systematic Review*, 2(2).

Lim, S. S., Fullman, N., Stokes, A., Ravishankar, N., Masiye, F., Murray, C. J., and Gakidou, E. (2011). Net benefits: a multicountry analysis of observational data examining associations between insecticide-treated mosquito nets and health outcomes. *PLoS medicine*, 8(9):e1001091.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., and Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE*, 7(2):e31743.

Linard, C., Gilbert, M., and Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, 76(5):525–538.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067.

Lysenko, A. and Semashko, I. (1968). Geography of malaria (a medical-geographical study of an ancient disease). *Medical Geography*, pages 25–146.

Mabaso, M., Craig, M., Vounatsou, P., and Smith, T. (2005). Towards empirical description of malaria seasonality in southern Africa: the example of Zimbabwe. *Tropical Medicine & International Health*, 10(9):909–918.

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Briolant, S., Lacaux, J.-P., Rogier, C., et al. (2011). The use of remotely sensed environmental data in the study of malaria. *Geospatial Health*, 5(2):151–168.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., and Thompson, S. G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1):31–50.

MARA/ARMA (1998). Towards an atlas of malaria risk in Africa. first technical report of the mara/arma collaboration. *Durban, South Africa.*

Martino, S. and Rue, H. (2010). Case studies in bayesian computation using inla. In *Complex data modeling and computationally intensive statistical methods*, pages 99–114. Springer.

McKenzie, F. E., Sirichaisnthop, J., Miller, R. S., Gasser Jr, R. A., and Wongsrichanalai, C. (2003). Dependence of malaria detection and species diagnosis by microscopy on parasite density. *The American journal of tropical medicine and hygiene*, 69(4):372.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087.

Miller, J. M., Korenromp, E. L., Nahlen, B. L., and Steketee, R. W. (2007). Estimating the number of insecticide-treated nets required by African households to reach continent-wide malaria coverage targets. *JAMA: the journal of the American Medical Association*, 297(20):2241–2250.

Ministère de la Santé, Programme National de Lutte contre le Paludisme, INFO-STAT, ICF Macro (2010). Enquête sur la prévalence de l'anémie et de la parasitémie palustre chez les enfants au Mali. Technical report, Bamako, Mali and Calverton, Maryland, USA.

Ministerio da Saúde e Instituto Nacional de Estatística e ICF International (2013). Moçambique inquérito demográfico e de saúde 2011. Technical report, Calverton, Maryland, USA.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Moyes, C. L., Temperley, W. H., Henry, A. J., Burgert, C. R., and Hay, S. I. (2013). Providing open access data online to advance malaria research and control. *world*, 17:20.

Mueller, I., Schoepflin, S., Smith, T. A., Benton, K. L., Bretscher, M. T., Lin, E., Kiniboro, B., Zimmerman, P. A., Speed, T. P., Siba, P., et al. (2012). Force of infection is key to understanding the epidemiology of *plasmodium falciparum* malaria in papua new guinean children. *Proceedings of the National Academy of Sciences*, 109(25):10030–10035.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S., et al. (2013). Disability-adjusted life years

(DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2197–2223.

Musenge, E., Vounatsou, P., and Kahn, K. (2011). Space-time confounding adjusted determinants of child hiv/tb mortality for large zero-inflated data in rural south Africa. *Spatial and Spatio-temporal Epidemiology*, 2(4):205–217.

Neelon, B., Ghosh, P., and Loebs, P. F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):389–413.

Nelson, G. (1959). Atlas of Kenya. *Surveys of Kenya, Crown Printers*.

Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., Fall, I. S., and Snow, R. W. (2014). The changing risk of¡ i¿ plasmodium falciparum¡/i¿ malaria infection in africa: 2000–10: a spatial and temporal analysis of transmission intensity. *The Lancet*.

Noor, A. M., Mutheu, J. J., Tatem, A. J., Hay, S. I., and Snow, R. W. (2009). Insecticide-treated net coverage in Africa: mapping progress in 2000–07. *The Lancet*, 373(9657):58–67.

Nychka, D., Wikle, C., and Royle, J. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331.

O'Hara, R. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4:85–118.

O'Meara, W. P., Mangeni, J. N., Steketee, R., and Greenwood, B. (2010). Changes in the burden of malaria in sub-Saharan Africa. *The Lancet infectious diseases*, 10(8):545–555.

Omumbo, J., Hay, S., Snow, R., Tatem, A., and Rogers, D. (2005). Modelling malaria risk in east Africa at high-spatial resolution. *Tropical Medicine & International Health*, 10(6):557–566.

Patz, J. A., Campbell-Lendrum, D., Holloway, T., and Foley, J. A. (2005). Impact of regional climate change on human health. *Nature*, 438(7066):310–317.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pages 20–22.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Pull, J. and Grab, B. (1974). A simple epidemiological model for evaluating the malaria inoculation rate and the risk of infection in infants. *Bulletin of the World Health Organization*, 51(5):507.

Pullan, R. L., Bethony, J. M., Geiger, S., Cundill, B., Correa-Oliveira, R., Quinnell, R. J., and Brooker, S. (2008). Human helminth co-infection: analysis of spatial patterns and risk factors in a brazilian community. *PLoS Neglected Tropical Disease*, 6:2(12):e352.

Raso, G., Vounatsou, P., Gosoniou, L., Tanner, M., N'Goran, E. K., and Utzinger, J. (2006a). Risk factors and spatial patterns of hookworm infection among schoolchildren in a rural area of Western Côte d'Ivoire. *International Journal for Parasitology*, 36:201–210.

Raso, G., Vounatsou, P., McManus, D. P., N'Goran, E. K., and Utzinger, J. (2007). A bayesian approach to estimate the age-specific prevalence of *schistosoma mansoni* and implications for schistosomiasis control. *International journal for parasitology*, 37(13):1491–1500.

Raso, G., Vounatsou, P., Singer, B. H., N'Goran, E. K., Tanner, M., and Utzinger, J. (2006b). An integrated approach for risk assessment and spatial prediction for *schistosoma mansoni*-hookworm co-infection. *Proceedings of the National Academy of Sciences of the Unites States of America*, 103:6934–6939.

Rathbun, S. L. and Fei, S. (2006). A spatial zero-inflated Poisson regression model for oak regeneration. *Environmental and Ecological Statistics*, 13(4):409–426.

Recta, V., Haran, M., and Rosenberger, V. L. (2012). A two-stage model for incidence and prevalence in point-level spatial count data. *Environmetrics*, 23:162–174.

Reich, B. J., Fuentes, M., Herring, A. H., and Evenson, K. R. (2010). Bayesian variable selection for multivariate spatially varying coefficient regression. *Biometrics*, 66:772–782.

Riedel, N., Vounatsou, P., Miller, J. M., Gosoniu, L., Chizema-Kawesha, E., Mukonka, V., Steketee, R. W., et al. (2010). Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modelling of the 2006 Zambia national Malaria Indicator Survey (zmis). *Malaria Journal*, 9:37.

Robert, C. P. (1996). Intrinsic losses. *Theory and decision*, 40(2):191–214.

Roca-Feltrer, A., Lalloo, D. G., Phiri, K., and Terlouw, D. J. (2012). Rolling Malaria Indicator Surveys (rMIS): A potential district-level malaria Monitoring and Evaluation (M&E) tool for program managers. *The American journal of tropical medicine and hygiene*, 86(1):96–98.

Roll Back Malaria (2008). The global malaria action plan - for a malaria free world. Technical report, World Health Organization.

Ross, A. and Smith, T. (2010). Interpreting malaria age-prevalence and incidence curves: a simulation study of the effects of different types of heterogeneity. *Malaria Journal*, 9:132.

Rowe, A. K. et al. (2009). Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries. *American Journal of Tropical Medicine and Hygiene*, 80(6):971.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Rumisha, S. F. (2013). *Modelling the seasonal and spatial variation of malaria transmission in relation to mortality in Africa*. PhD thesis, University of Basel.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, cambridge, uk edition.

Rutstein, S. and Johnson, K. (2011). The dhs wealth index. Technical Report 6, Calverton, Maryland: ORC Macro.

Sachs, J. and Malaney, P. (2002). The economic and social burden of malaria. *Nature*, 415(6872):680–685.

Sampson, P. D. (2010). Constructions for nonstationary spatial processes. In A.E. Gelfand, P.J. Diggle, M. F. and Guttorp, P., editors, *Handbook of Spatial Statistics*, pages 119–130. CRC Press.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2013). A bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection. *Journal of the American Statistical Association*, 107:1518–1532.

Schlagenhauf-Lawlor, P. (2008). *Travelers' malaria*. PMPH-USA.

Schmidt, A. M. and Gelfand, A. E. (2003). A bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 108(D24).

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

Siri, J. G. and Lutz, W. (2012). The independent effects of maternal education and household wealth on malaria risk in children.

Smith, D., Dushoff, J., Snow, R., and Hay, S. (2005). The entomological inoculation rate and *plasmodium falciparum* infection in African children. *Nature*, 438(7067):492–495.

Smith, D. L., Guerra, C. A., Snow, R. W., and Hay, S. I. (2007). Standardizing estimates of the *plasmodium falciparum* parasite rate. *Malaria journal*, 6(1):131.

Smith, D. L., Hay, S. I., Noor, A. M., and Snow, R. W. (2009). Predicting changing malaria risk after expanded insecticide-treated net coverage in Africa. *Trends in parasitology*, 25(11):511–516.

Smith, M. and Fahrmeir, L. (2007). Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102:417–431.

Smith, T., Charlwood, J., Takken, W., Tanner, M., and Spiegelhalter, D. (1995). Mapping the densities of malaria vectors within a single village. *Acta tropica*, 59(1):1–18.

Sogoba, N., Vounatsou, P., Bagayoko, M., Doumbia, S., Dolo, G., Gosoniu, L., Traore, S., Toure, Y., and Smith, T. (2007). The spatial distribution of *anopheles gambiae sensu stricto* and *an. arabiensis* in Mali. *Geospatial Health*, 1(2):213–222.

Stefani, A., Dusfour, I., Corrêa, A. P. S., Cruz, M. C., Dessay, N., Galardo, A. K., Galardo, C. D., Girod, R., Gomes, M. S., Gurgel, H., et al. (2013). Land cover, land use and malaria in the amazon: a systematic literature review of studies using remotely sensed data. *Malaria journal*, 12(1):1–8.

Tatem, A. J., Garcia, A. J., Snow, R. W., Noor, A. M., Gaughan, A. E., Gilbert, M., and Linard, C. (2013a). Millennium development health metrics: where do Africas children and women of childbearing age live? *Population health metrics*, 11(1):11.

Tatem, A. J., Gething, P. W., Bhatt, S., Weiss, D., and Pezzulo, C. (2013b). Pilot high resolution poverty maps. Technical report, University of Southampton/Oxford.

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., and Hay, S. I. (2007). High resolution population maps for low income nations: combining land cover and census in east Africa. *PLoS ONE*, 2(12):e1298.

Thiam, S., Thior, M., Faye, B., Ndiop, M., Diouf, M. L., Diouf, M. B., Diallo, I., Fall, F. B., Ndiaye, J. L., Albertini, A., et al. (2011). Major reduction in anti-malarial drug consumption in senegal after nation-wide introduction of malaria rapid diagnostic tests. *PLoS ONE*, 6(4):e18419.

Thwing, J. I., Perry, R. T., Townes, D. A., Diouf, M. B., Ndiaye, S., and Thior, M. (2011). Success of senegal's first nationwide distribution of long-lasting insecticide-treated nets to children under five-contribution toward universal coverage. *Malaria Journal*, 10:86.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, 17:76–94.

Turner, R. M., Spiegelhalter, D. J., Smith, G., and Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47.

Tusting, L. S., Willey, B., Lucas, H., Thompson, J., Kafy, H. T., Smith, R., and Lindsay, S. W. (2013). Socioeconomic development as an intervention against malaria: a systematic review and meta-analysis. *The Lancet*.

UN (2006). Collection and dissemination of data from the demographic yearbook. Technical report, New York: United Nations, Statistics division website.

UN (2012a). The millennium development goals report. Technical report, United Nations: New York.

UN (2012b). World population prospects: The 2012 revision. Technical report, New York: United Nations, Department of Economic and Social Affairs, Population Division.

Wagner, H. and Duller, C. (2012). Bayesian model selection for logistic regression models with random intercept. *Computational Statistics and Data Analysis*, 56:1256–1274.

Wang, X.-H., Zhou, X.-N., Vounatsou, P., Chen, Z., Utzinger, J., Yang, K., Steinmann, P., and Wu, X.-H. (2008). Bayesian spatio-temporal modeling of *schistosoma japonicum* prevalence data in the absence of a diagnostic gold standard. *PLoS neglected tropical diseases*, 2(6):e250.

White, L. J., Maude, R. J., Pongtavornpinyo, W., Saralamba, S., Aguas, R., Van Effelterre, T., Day, N. P., and White, N. J. (2009). The role of simple mathematical models in malaria elimination strategy design. *Malaria journal*, 8(1):212.

WHO (2008). World malaria report 2008. Technical report, World Health Organization: Geneva.

WHO (2009). World malaria report 2009. Technical report, World Health Organization: Geneva.

WHO (2012). World malaria report 2012. Technical report, World Health Organization: Geneva.

Wongsrichanalai, C., Barcus, M. J., Muth, S., Sutamihardja, A., and Wernsdorfer, W. H. (2007). A review of malaria diagnostic tools: microscopy and rapid diagnostic test (rdt). *The American journal of tropical medicine and hygiene*, 77(6 Suppl):119–127.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design bayesian generalized linear mixed models. *Statistical Science*, 21:35–51.

Zhu, L., Carlin, B. P., English, P., and Scalf, R. (2000). Hierarchical modeling of spatio-temporally misaligned data: relating traffic density to pediatric asthma hospitalizations. *Environmetrics*, 11(1):43–61.