Development of Bayesian geostatistical models with applications in malaria epidemiology

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Laura Goşoniu
aus Bukarest, Rumänien

Basel, December 2008

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. M. Tanner, Dr. P. Vounatsou, Prof. Dr. T. Smith und Prof. Dr. M. Schumacher.

Basel, den 19. Dezember 2007

Prof. Dr. Hans-Peter Hauri
Dekan

Dedicated to my parents
and my beloved husband, Dominic

Summary

Plasmodium falciparum malaria is a leading infectious disease and a major cause of morbidity and mortality in large areas of the developing world, especially Africa. Accurate estimates of the burden of the disease are useful for planning and implementing malaria control interventions and for monitoring the impact of prevention and control activities. Information on the population at risk of malaria can be compared to existing levels of service provision to identify underserved populations and to target interventions to high priority areas. The current available statistics for malaria burden are not reliable because of the poor malaria case reporting systems in most African countries and the lack of national representative malaria surveys. Accurate maps of malaria distribution together with human population totals are valuable tools for generating valid estimates of population at risk.

Empirical mapping of the geographical patterns of malaria transmission in Africa requires field survey data on prevalence of infection. The Mapping Malaria Risk in Africa (MARA) is the most comprehensive database on malariological survey data across all sub-Sahara Africa. Transmission of malaria is environmentally driven because it depends on the distribution and abundance of mosquitoes, which are sensitive to environmental and climatic conditions. Estimating the environment-disease relation, the burden of malaria can be predicted at places where data on transmission are not available. Malaria data collected at fixed locations over a continuous study area (geostatistical 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. Geographical dependence is considered as a function of the distance between locations. These models are highly parametrized. State-of-the-art Bayesian computation implemented via Markov chain Monte Carlo (MCMC) simulation methods enables model fit. A common assumption in geostatistical modeling of malaria

data is the stationarity, that is the spatial correlation is a function of distance between locations and not of the locations themselves. This hypothesis does not always hold, especially when modeling malaria over large areas, hence geostatistical models that take into account non-stationarity need to be assessed. Fitting geostatistical models requires repeated inversions of the variance-covariance matrix modeling geographical dependence. For very large number of data locations matrix inversion is considered infeasible. Methods for optimizing this computation are needed. In addition, the relation between environmental factors and malaria risk is often not linear and parametric functions may not be able to determine the shape of the relationship. Nonparametric geostatistical regression models that allow the data to determine the form of the environment-malaria relation need to be further developed and applied in malaria mapping.

The aim of this thesis was to develop appropriate models for non-stationary and large geostatistical data that can be applied in the field of malaria epidemiology to produce accurate maps of malaria distribution. The main contributions of this thesis are the development of methods for: (i) analyzing non-stationary malaria survey data; (ii) modeling the non-linear relation between malaria risk and environment/climatic conditions; (iii) modeling geostatistical mortality data collected at very large number of locations and (iv) adjusting for seasonality and age in mapping heterogeneous malaria survey data.

Chapter 2 assessed the spatial effect of bednets on all-cause child mortality by analyzing data from a large follow-up study in an area of high perennial malaria transmission in Kilombero Valley, southern Tanzania. The results indicated a lack of community effect of bednets density possibly because of the homogeneous characteristic of nets coverage and the small proportion of re-treated nets in the study area. The mortality data of this application were collected over $7,403$ locations. To overcome large matrix inversion a Bayesian geostatistical model was developed. This model estimates the spatial process by a subset of locations and approximates the location-specific random effects by a weighted sum of the subset of location-specific random effects with the weights inversely proportional to the separation distance.

In Chapter 3 a Bayesian non-stationary model was developed by partitioning the study region into fixed subregions, assuming a separate stationary spatial process in each tile and taking into account between-tile correlation. This methodology was applied on malaria survey data extracted from the MARA database and produced parasitaemia risk maps in Mali. The predictive ability of the non-stationary model was compared with the stationary

analogue and the results showed that the stationarity assumption influenced the significance of environmental predictors as well as the the estimation of the spatial parameters. This indicates that the assumptions about the spatial process play an important role in inference. Model validation showed that the non-stationary model had better predictive ability. In addition, experts opinion suggested that the parasitaemia risk map based on the non-stationary model reflects better the malaria situation in Mali. This work revealed that non-stationarity is an essential characteristic which should be considered when mapping malaria.

Chapter 4 employed the above non-stationary model to produce maps of malaria risk in West Africa considering as fixed tiles the four agro-ecological zones that partition the region. Non-linearity in the relation between parasitaemia risk and environmental conditions was assessed and it was addressed via P-splines within a Bayesian geostatistical model formulation. The model allowed a separate malaria-environment relation in each zone. The discontinuities at the borders between the zones were avoided since the spatial correlation was modeled by a mixture of spatial processes over the entire study area, with the weights chosen to be exponential functions of the distance between the locations and the centers of the zones corresponding to each of the spatial processes.

The above modeling approach is suitable for mapping malaria over areas with an obvious fixed partitioning (i.e. ecological zones). For areas where this is not possible, a non-stationary model was developed in Chapter 5 by allowing the data to decide on the number and shape of the tiles and thus to determine the different spatial processes. The partitioning of the study area was based on random Voronoi tessellations and model parameters were estimated via reversible jump Markov chain Monte Carlo (RJMCMC) due to the variable dimension of the parameter space.

In Chapter 6 the feasibility of using the recently developed mathematical malaria transmission models to adjust for age and seasonality in mapping historical malaria survey data was investigated. In particular, the transmission model was employed to translate age heterogeneous survey data from Mali into a common measure of transmission intensity. A Bayesian geostatistical model was fitted on the transmission intensity estimates using as covariates a number of environmental/climatic variables. Bayesian kriging was employed to produce smooth maps of transmission intensity, which were further converted to age specific parasitaemia risk maps. Model validation on a number of test locations showed that this transmission model gives better predictions than modeling directly the prevalence

data. This approach was further validated by analyzing the nationally representative malaria surveys data derived from the Malaria Indicator surveys (MIS) in Zambia. Although MIS data do not have the same limitations with the historical data, the purpose of the analyzes was to compare the maps obtained by modeling 1) directly the raw prevalence data and 2) transmission intensity data derived via the transmission model. Both maps predicted similar patterns of malaria risk, however the map based on the transmission model predicted a slightly higher lever of endemicity. The use of transmission models on malaria mapping enables adjusting for seasonality and age dependence of malaria prevalence and it includes all available historical data collected at different age groups.

Zusammenfassung

Plasmodium falciparum Malaria ist eine der häufigsten infektiösen Krankheiten und der Hauptverursacher von Morbidität und Sterblichkeit in weiten Teilen der Dritten Welt, besonders in Afrika. Genaue Abschätzungen zur Belastung durch die Erkrankung sind hilfreich für die Planung und Durchführung von Malaria Interventionen und für die überwachung von Präventions- und Kontrolleinflüssen. Informationen über die gefährdeten Bevölkerungen durch Malaria können mit verschiedenen existierenden Modellen verglichen werden um benachteiligte Gruppen zu identifizieren und gezielt Verbeugungen in den wichtigsten Gebieten zu ergreifen. Die derzeitig verfügbaren statistischen Methoden zur Bestimmung der Belastung durch Malaria sind allerdings nicht sehr zuverlässig, da in den meisten afrikanischen Ländern nur ein dürftiges Meldesystem für Malariafälle besteht. Ausserdem fehlen landesweite repräsentative Studien. Korrekte Karten zur Malariatransmission zusammen mit den Gesamtzahlen der Bevölkerung sind nützliche Werkzeuge um Abschätzungen über die gefährdete Bevölkerung zu erhalten.

Empirische Karten über die geografischen Muster der Malaria Verbreitung in Afrika benötigen Prävalenzdaten aus Studien über die Erkrankung. Die Datenbank "Mapping Malaria Risk in Africa (MARA)" ist die umfangreichste ihrer Art welche Daten über Malaria bezogene Studien in Afrika südlich der Sahara sammelt. Die Ausbreitung von Malaria wird durch ökologische Faktoren beeinflusst, weil die Erkrankung von der Verteilung und Menge von Moskitos abhängig ist, welche empfindsam auf Umwelt und Klima reagieren. Durch die Einbeziehung der Korrelation von Umwelt und Erkrankung kann die Belastung durch Malaria selbst an Plätzen abgeschätzt werden über die ansonsten keine weiteren Daten zur Verbreitung von Malaria zur Verfügung stehen. Daten die an einer bestimmten Anzahl von Orten gesammelt wurden (geostatistische Daten) sind räumlich korreliert, da die bekannten Einflussfaktoren die Malariatransmission zueinander ähnlich in benachbarten Gebieten beeinflussen. Geostatistische Modelle berücksichtigen diese räumlichen Beziehungen,

indem sie einen ortsspezifischen Fehlerterm einführen und die geographische Abhängigkeit durch eine Funktion der Distanz zwischen den einzelnen Orten wiedergegeben. Diese Modelle sind allerdings hoch parametrisiert. Hochmoderne Bayes'sche Berechnungen, welche durch "Markov chain Monte Carlo" Simulationen implementiert werden, erlauben allerdings deren Modellierung. Eine gebräuchliche Annahme beim geostatistischen Modellieren ist die der Stationarität. Das bedeutet, dass die räumliche Korrelation eine Funktion der Distanz zwischen den Orten ist und nicht der Orte selber. Diese Behauptung gilt allerdings nicht immer, besonders dann nicht wenn die Malariatransmission über grosse Entfernungen modelliert werden soll. Deshalb müssen geostatistische Modelle benutzt werden die zusätzlich die Nicht-Stationarität berücksichtigen. Das Durchlaufen von geostatistischen Modellen erfordert mehrfache Inversionen der Varianz-Kovarianz Matrix die die geografische Abhängigkeit darstellt. Für eine sehr hohe Anzahl von Orten wird dies allerdings als nicht machbar eingestuft, deshalb werden Methoden zur Optimierung dieser Berechnungen benötigt. Ein weiteres Problem ist, dass die Abhängigkeiten zwischen den ökologischen Faktoren und des Malariarisikos häufig nicht linear sind und daher parametrische Funktionen nicht in der Lage sind die Form dieser Beziehung wiederzugeben. Daher müssen nichtparametrische geostatistische Regressionsmodelle, welche den Daten erlauben die Form der Umwelt-Malaria Beziehung anzunehmen, weiterentwickelt werden um sie für die Kartierung von Malaria verfügbar zu machen.

Das Ziel diese Arbeit war es geeignete Modelle für die Nicht-Stationarität und grosse Mengen an geostatistischen Daten zu entwickeln, die sich für die Malariaepidemiologie nutzen lassen um exakte Karten der Malariaverteilung zu erstellen. Das Hauptaugenmerk lag dabei auf der Entwicklung von Methoden für: (i) die Analyse von nicht stationären Malaria Studien; (ii) die Modellierung der nicht linearen Beziehung zwischen Malariarisiko und Umwelt-/klimatischen Bedingungen; (iii) die Modellierung von geostatistischen Sterblichkeitsdaten welche an sehr vielen Orten gesammelt wurden und (iv) die Einbeziehung von Unterschieden in der Jahreszeit und dem Alter der Studienteilnehmer bei der Kartierung von verschiedenartigen Malariastudien.

Kapitel 2 untersucht den räumlichen Effekt von Bettnetzen auf die allgemeine Sterblichkeit von Kindern durch die Analyse von Daten einer grossen Follow-Up-Studie in einem Gebiet mit über das Jahr konstant hoher Malariaverbreitung in Kilombero Valley, südlich

von Tansania. Die Resultate weisen auf einen fehlenden Gemeinschaftseffekt der Bettnetzdichte hin, vermutlich aufgrund der einheitlichen Abdeckung des Studiengebiets mit Netzen und der geringen Proportion von erneut behandelten Netzen. Die Sterblichkeitsdaten dieser Applikation wurden in 7403 Orten gesammelt. Um das Problem der Matrix Inversion zu umgehen wurde ein Bayes'sches geostatistisches Modell entwickelt. Dieses Modell berechnet den räumlichen Prozess durch eine Untergruppe von Orten und schätzt die ortsspezifischen Fehlerterme durch die gewichtete Summe der ortsspezifischen Fehlerterme der Untergruppen durch Wichtungen ab, welche umgekehrt proportional zu der Distanz der Orte sind.

In Kapitel 3 wurde ein Bayes'sches nicht stationäres Model durch die Aufteilung der Studienregion in feste Unterregionen entwickelt. Das Modell beruht auf der Annahme, dass in jedem Teilstück ein fester stationärer räumlicher Prozess abläuft, wobei die Korrelation zwischen den einzelnen Teilstücken mit berücksichtigt wurde. Diese Methode wurde an Daten zu Malaria aus der MARA Datenbank angewandt und es wurde daraus eine Risikokarte für Mali erstellt. Die Vorhersagekraft des nicht stationären Modells wurde mit der des analogen stationären Modells verglichen. Die Resultate zeigten dass die Stationaritätsannahme die Signifikanz ökologischer Prädiktoren ebenso wie Abschätzung der räumlichen Faktoren beeinflusst. Dies deutet im Umkehrschluss an, dass die Annahmen über den räumlichen Prozess eine bedeutende Rolle spielen. Modellbewertungen zeigten dabei eine bessere Vorhersagekraft für das nicht stationäre Modell an. Zusätzlich bestätigten Expertenmeinungen, dass die Risikokarte für Malaria im Falle des nicht stationären Modells besser die Lage in Mali widerspiegelt. Diese Arbeit enthüllte dass Nicht-Stationarität eine essentielle Eigenschaft ist, welche unbedingt bei der Erstellung von weiterem Kartenmaterial für Malaria berücksichtigt werden sollte.

Kapitel 4 wendet das obere nicht stationäre Modell für die Kartenerstellung zum Malariarisiko in Westafrika an, wobei als feste Teilstücke die vier Agrarkulturregionen dienten die die Region aufteilen. Die Nicht-Linearität zwischen dem Risiko Malariaparasiten zu haben und den ökologischen Bedingungen wurde beachtet und durch die Benutzung von P-Splines in der Bayes'schen geostatistischen Modellformulierung vermieden. Das Modell erlaubte eine separate Malaria-Umwelt Beziehung in jeder Zone. Die Diskontinuität an den Grenzen der Zonen wurde vermieden, indem die räumliche Korrelation durch eine Mischung von räumlichen Prozessen über die gesamte Studienfläche modelliert wurde. Dabei wurden die Gewichtungen als exponentielle Funktionen der Distanz zwischen den Ortschaften and den

Zentren der Zonen entsprechend eines jeden räumlichen Prozesses gewählt.

Der obere Modellansatz ist angebracht für die Malariakartierung in Gebieten wo eine offensichtliche Teilung (z.B. in ökologische Zonen) besteht. Für Gebiete wo das nicht möglich ist wurde in Kapitel 5 ein weiteres nicht stationäres Modell entwickelt, was den Daten erlaubt die Anzahl und Gestalt der Teilstücke festzulegen und daher die verschiedenen räumlichen Prozesse zu bestimmen. Die Unterteilung der Studienfläche basierte auf zufälligen Voronoi-Diagrammen und die Modellparameter wurden mittels rückwärts verlaufenden Markov chain Monte Carlo Methoden (RJMCMC) bestimmt beruhend auf der Variablendimension des Parameterraumes.

In Kapitel 6 wurde die Realisierbarkeit der zuvor entwickelten mathematischen Malariatransmissionsmodelle untersucht, welche angepasst wurden an das Alter und die jahreszeitlichen Schwankungen, indem historische Malariastudien kartiert wurden. Das Transmissionsmodell wurde angepasst um im Alter der Teilnehmer schwankende Studiendaten von Mali in eine allgemein gültige Messgrösse für die Ausbreitungsintensität zu überführen. Ein Bayes'sches geostatistisches Modell für die Abschätzung der Ausbreitungsintensität wurde erstellt, indem als Kovariaten verschiedene umweltbezogene und klimatische Faktoren genutzt wurden. Bayes'sches Kriging wurde genutzt um gleichmässige Karten zur Ausbreitung zu erstellen, welche im Weiteren zu altersspezifischen Risikokarten umgewandelt wurden. Der Modellvergleich mit einigen Testorten zeigte, dass das Transmissionsmodell bessere Werte liefert als die direkte Modellierung von Prävalenzdaten. Dieser Ansatz wurde weiter getestet durch die Analyse der national repräsentativen Daten der Malaria Indikationsstudie (MIS) in Sambia. Obwohl die MIS Daten nicht die Selben Einschränkungen haben wie die historischen Daten, lag die Absicht dennoch darin die Karten zu vergleichenden, die 1) durch die direkte Modellierung der unveränderten Prävalenzdaten und 2) durch die Modellierung der Daten der Verbreitungsintensität aus dem Transmissionsmodell entstanden sind. Beide Karten sagen ähnliche Muster im Malariarisiko voraus, dennoch konnte die Karte basierend auf dem Transmissionsmodell die Endemizität ein wenig besser wiedergeben. Die Nutzung von Transmissionsmodellen für die Kartierung von Malaria erlaubt die Einbeziehung von jahreszeitlichen Schwankungen und die Altersabhängigkeit der Malariaprävalenz und es beinhaltet alle zur Verfügung stehenden historischen Daten die für die verschiedenen Altergruppen gesammelt wurden.

Acknowledgements

It is my great pleasure to acknowledge many people who contributed to this thesis.

First and foremost, this thesis would not have been possible without the great scientific support and the remarkable patience of my thesis advisor, Dr. Penelope Vounatsou. Her enthusiasm and passion for research had motivated and inspired me all these years. I owe her lots of gratitude for being an outstanding supervisor and a good friend. I could not have imagined having a better mentor for my PhD.

I also like to express my gratitude to Prof. Thomas Smith for his constructive comments and for his important support throughout this work. Thank you for being always available when I needed your advises. Special thanks to Prof. Christian Lengeler for many productive scientific discussions and to Prof. Don de Savigny for valuable comments and ideas. I would also like to thank Prof. Mitchell Weiss for a very good and warm working atmosphere in the department. I am especially grateful to Prof. Marcel Tanner for having welcomed me at the STI and for creating a stimulating environment for developing my research.

Many thanks to all the wonderful people in STI, who contributed to a fantastic working atmosphere. I appreciate the support of Magrit Slaoui, Christine Walliser, Eliane Ghilardi and Isabelle Bolliger who made my life as a student smooth and pleasant. I wish to thank the STI library team as well as to the IT support group for their assistance. Special thanks are addressed to Bianca Pluss for making sure I was getting enough coffee to keep me going, to Tippi Mak for the good chats, her scientific and friendly advices, to Nadine Riedel for translating the summary of this thesis in German and for always surprising me with new learned romanian words, to Claudia Sauerborn for her encouragements and for providing me with good books, to Amanda Ross for the late night chats in the last phase of our PhD's, to Niggi Maire for his help with LaTeX and not only and to Ellen Stamhuis for

the delicious waffles. A "grand merci" to Nafomon Sogoba for the lively discussion and for being a good friend. My warm thanks are addressed to colleagues and friends with whom I shared the office and good laughter: Daryl Somma, Wilson Sama, Shr-Jie Wang, Dan Anderegg, Collins Ahorlu, Honorati Masanja, Nadine Köler, Lena Fiebig, Mwifadni Mrisho, Ricarda Merkle, Josh Yukich, Andri Christen, Gonzalo Duran, Susan Rumisha, Amina Msengwa. Thanks to those people who provided me with so much whether they know it or not: Musa Mabaso, Marlies Craig, Michael Bretscher, Melissa Penny, Nakul Chitnis, Barbara Matthys, Stefan Dongus, Connie Pfeiffer, Manuel Hetzel, Tobias Erlanger, Giovanna Raso, Elisabetta Peduzzi, Simona Rondini and Monica Daigl.

A BIG thank you to Dora, Lavinia and Romeo who have made Basel a very special place over all these years. I wish to thank Nicoleta and Marco for the nice week-ends spend together and for their friendship. My sincerest thanks go to Lucas and a big hug to Anna and Nadia with whom we shared joyful moments.

The support of many friends back home has been indispensable and I would like particularly to acknowledge Anca, Mihaela, George, Maria, Gabriel, our godparents Mihaela and Marian Borcan, my sister-in-law Patricia and my brother-in-law Augustin for their constant encouragements throughout this entire journey. I would like to express my warmest thanks to my parents-in-law for their unconditional support at each turn of the road ("Va multumesc din suflet").

I am forever indebted to my parents for the sacrifices made to ensure that I had an excellent education, for their understanding and endless patience. ("Va voi ramane mereu indatorata"). My special gratitude is due to my brother for his affection and encouragements when they were most required.

Last but never least, I would like to express my deepest gratitude to Dominic for his endless love and for the incredible amount of patience he had with me in the last months. Thank you for accompanying me in this journey.

A lots of other people have contributed in different ways to this thesis. To all of you MANY THANKS.

This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Global malaria distribution

Malaria is the most important tropical disease, remaining widespread throughout the tropical and subtropical regions, including parts of Africa, Asia and Americas (Figure 1.1). It is a major cause of illness and death in large areas of the developing world, especially Africa. According to the World Malaria Report (WHO, 2005), at the end of 2004 there were 107 malaria endemic countries and 3.2 billion people were at risk of malaria. Malaria causes at least 300 million and possibly as many as 500 million cases of acute illness each year, which result in 1-3 million deaths (Breman et al., 2004). Ninety percent of deaths occur in sub-Sahara Africa. The large majority of these deaths are in children younger than five years of age, being estimated that every 30 seconds a child dies because of malaria. These rough estimates are not reliable because of the inadequate malaria case reporting in most endemic countries and lack of national wide malaria surveys. Accurate estimates of the burden of disease are required for planning, implementation and evaluation of malaria control programs. Hence, there is an urgent need for precise estimates of the number of people at risk of malaria to optimize the use of limited resources in high-risk areas.

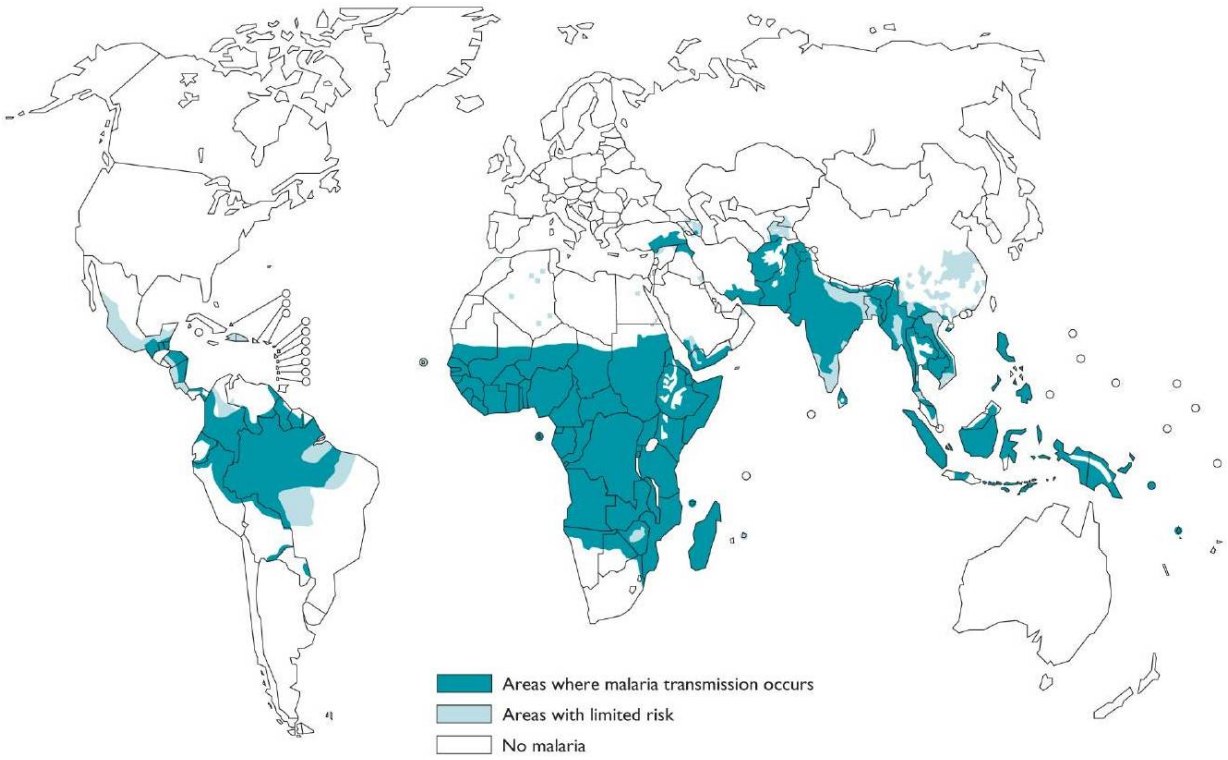

Figure 1.1: Geographic distribution of malaria.

1.2 Malaria disease and transmission

1.2.1 The malaria parasite

Malaria is an infectious disease caused by four parasitic protozoa of the genus *Plasmodium*, namely: *P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae*. The human malaria parasites require the presence of two hosts to complete their life cycle: the human, which represents the intermediate host and the female mosquito, which is the defitinive host. In mosquito, the parasite reproduces sexually by combining sex cells, whereas in human the parasite reproduces asexually, first in liver cells and then in red blood cells.

A human infection begins when an infected female mosquito takes a blood meal, passing sporozoites into the human's bloodstream. The sporozoites travel with the blood to the liver and enter the liver cells. There, the sporozoites mature into schizonts, which rupture and release merozoites into the blood stream. Merozoites invade the red blood cells, where they undergo the second asexual reproduction into human, forming again schizonts. When the schizonts mature, the cell rupture and merozoites burtst out, infecting other red blood cells. This repeating cycle depletes the body of oxygen and causes fever, triggering the onset of disease symptoms. In the red blood cells some merozoites develop into sex cells known as male and female gametocyte. When a female mosquito bites an infected human the gametocyte are ingested. In mosquito's stomach the gametocyte undergo sexual reproduction, forming a zygote. The zygote multiplies to form sporozoites, which make their way to the mosquito's salivary glands. Inoculation of the sporozoites into a new human host perpetuates the malaria life cycle.

1.2.2 The *Anopheles* vector

The distribution of the malaria parasite is largely determined by the distribution of the mosquito vectors which, in the case of malaria parasites of humans, are all of the genus *Anopheles*. 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 (Service and Townson, 2002). 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.

1.2.3 Malaria transmission

Measures of malaria endemicity and transmission

Measures of malaria transmission quantify malaria risk and endemicity levels and they are the basis of decision making in malaria control. These measures include parameters related to malaria transmission from mosquito to humans (i.e. entomological inoculation rates, force of infection, incidence rates, parasite prevalence) and parameters related to malaria vectors (i.e. mosquito survival, infection probability).

The most commonly used measure of malaria endemicity is the prevalence of human infections within a community. Information on malaria prevalence is collected through community-based surveys by computing the percentage of individuals found with a positive blood slide. In 1950s WHO classified malaria endemicity using both the parasite and the spleen rates (percent of children with enlarged spleen) as: hypoendemic, mesoendemic, hyperendemic and holoendemic. Malaria prevalence of the same population may vary in time, depending on the seasonality and stability of the disease.

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 1 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 bait 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 review of EIR estimates across Africa (Hay et

al, 2000) found the mean annual EIR of 121 infective bites, ranging from a minimum of 0 in Burkina Faso, The Gambia and Senegal to a maximum of 884 in Sierra Leone.

Incidence of malaria 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. Incidence data are usually collected in health facilities. In most settings in sub-Sahara Africa is not possible to perform laboratory confirmation of malaria diagnoses, therefore incidence of fever is used as a proxy for incidence of malaria.

Force of infection is the rate at which susceptible individuals become infected by malaria parasite. These data are an alternative to malaria incidence data which are difficult to be collected in communities where the prevalence of infection reached saturation. MacDonald was the first to propose an estimate of force of infection using infant parasite conversion rates for malaria.

The basic reproductive rate R_0 quantifies the transmission potential and is defined as the average number of successful offspring that a parasite is intrinsically capable of producing. Other two important parameters related to malaria vectors are the mosquito survival per gonotrophic cycle,P and the infection probability (sometimes called infectious reservoir) K, that is the probability that a mosquito becomes infected when it takes a feed. These quantities cannot be determined directly from field data, therefore transmission models are needed to estimate them.

Determinants of malaria transmission

Malaria transmission is affected by different factors like environmental conditions, poverty (socio-economic status), population movement (migration, urbanization), limited access to health services, poor quality of the public health services or water management methods (e.g. irrigation, dam constructions) that increase the mosquitoes population near human habitats. Climate is the main driver of malaria transmission with climate variability influencing the level of transmission intensity.

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.

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.

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.

Vegetation type 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 use changes may influence climatic conditions like temperature or evapotranspiration (Patz et al., 2005), which are main determinants of the abundance and longevity of mosquitoes. In addition, agriculture practices and human-made environmental alterations could influence the malaria vector population.

1.2.4 Malaria control interventions

In recent years several initiatives have been launched to tackle malaria in various parts of the world. In 1998 WHO initiated the Roll Back Malaria (RBM) Partnership with the goal of reducing malaria burden by at least 50% by the year 2010, applying evidence-based

intervations through strengthened health services. The Global Fund to Fight AIDS, Tuberculosis and Malaria (GFATM) was established in 2002, giving malaria-endemic countries access to additional external funding for malaria control. In 2005 President Bush launched the President's Malaria Initiative (PMI), a five-year programme which aims to reduce deaths due to malaria by 50% in 15 African countries.

Control measures are directed at each component involved in the malaria transmission cycle: the human host, the parasite and the mosquito vector. Complete cure of clinical malaria requires treatment with several drugs over several days and this creates problems of costs and compliance. Prophylaxis drugs have been of great benefit and widely used as a measure of malaria control, but they are no longer effective in many tropical areas because the parasite developed resistance to drugs.

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 and outdoor house insecticide spraying, the use of insecticide treated nets (ITN) and environmental measures such as management of water bodies and vegetation clearance. Applications of these techniques, alone or in combination, reduce human-mosquito contact, vector abundance and vector infectivity. ITN's are increasingly being promoted as a an efficient method for reducing the burden of malaria. In trials conducted between 1980 and 2000 ITN's were shown to reduce childhood malarial deaths in endemic areas in Africa by 17% and roughly halve the number of clinical malaria episodes (Lengeler, 2004). Although the use of ITN's provide significant individual effect, it still remains unclear what are the effects of ITN's on the wider community of bednet users and non-users.

Malaria control is a dynamic process which depends on the local epidemiological situation and of the facilities and resources available. Therefore, it is important that maps of malaria transmission are available for guiding control measures to high risk areas. These maps can also provide a baseline to evaluate the effectiveness of interventions programs.

1.3 Mapping malaria transmission

Although a lot of effort and resources have been put into the control of malaria, reliable estimates and mapping of malaria burden are not available. Maps of malaria distribution are valuable in increasing the effectiveness of decision-making. Maps are also useful to assess the effect of intervention programs by estimating the malaria distribution prior and post interventions. Recently, there is a renewed interest in mapping malaria (Snow et al., 2005; Hay et al., 2006) as well as different efforts in assembling existing malaria data (WHO, 2007; Malaria Atlas Project (MAP) (Guerra et al., 2007); MARA project newly funded by Bill and Melinda Gates foundation in 2007; Mekong malaria (Socheat et al, 2003)). Malaria is an environmental disease since its transmission depends upon the distribution and abundance of the mosquitoes, which are sensitive to climate. Hence, mapping malaria distribution is based on availability of malaria and environmental data as well as appropriate methods to analyze these data.

1.3.1 Types of data for malaria mapping

The main sources of data for mapping malaria are the following.

(i) historical malaria prevalence data. The "Mapping Malaria Risk in Africa" (MARA/ARMA, 1998) represents the most comprehensive database on malaria data in Africa. It contains malaria prevalence data collected over $10,000$ geographically positioned surveys from gray or published literature across the whole continent. The project was initiated over a decade ago to provide comprehensive, empirical and standardized maps of malaria distribution and endemicity in Africa. The database is currently being updated. A parallel project for assembling historical malaria data is being carried out by the MAP (Guerra et al., 2007). Maps from historical data may not reflect the current malaria situation at a given location, which could be influenced by control measures. Unfortunately, information on historical interventions is not generally available, therefore it is not possible to account for it. The limited number of surveys during the recent years requires inclusion of surveys from earlier years for mapping purposes. On the other hand, historical data are useful for looking at temporal changes of the malaria situation. The major drawback of these type of data is the heterogeneity in season and age since they are collected at non-standardized seasons and include overlapping age groups of the population. In addition, the data are sparse in time and space. These constraints make it difficult to consider seasonality and age adjustment

in malaria mapping (Gemperli et al., 2005).

(ii) parasite prevalence data derived from nationally representative surveys. These type of data are useful in countries with high transmission intensity and are collected form Demographic and Health surveys (DHS), Multiple Indicator Cluster Surveys (MICS) and Malaria Indicator Surveys (MIS). Zambia successfully conducted the first nationally representative household survey assessing coverage of malaria interventions and malaria-related burden among children under five years of age. The survey was conducted during May and June 2006 and was led by the Ministry of Health through the National Malaria Control Center (NMCC) in collaboration with many Zambia RBM partners. The MIS included information on intervention (IRS and ITNs) coverage, morbidity and background characteristics (i.e. household assets). The goal of the project was to provide baseline data against which to measure progress toward achieving its goals set forth in the National Malaria Strategic Plan for 2006-2010.

(iii) clinical malaria incidence data. These data are appropriate in areas with low malaria risk, like many countries in South-East Asia, since it is unlikely for people in the community to tolerate the parasite without being sick. Incidence data depend on precise estimates of the population at risk of malaria. Many countries in Africa do not have a reliable disease surveillance system, therefore routine malaria statistics can not be used (exception: South Africa, Zimbabwe, Botswana, Namibia). The major drawback of this type of data is that for countries with limited access to laboratory confirmation of cases - like most of the countries in sub-Sahara Africa - the new malaria cases refer to patients who are suspected to have malaria based on clinical signs and symptoms.

(iv) entomological data. They 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, hence prediction of these data at unsampled locations is difficult (Hay et al., 2000).

Recently, incidence and entomological data were both combined by Mabaso (2007) for mapping seasonality of malaria transmission in Africa. Seasonal dynamics of malaria

transmission are important for timing malaria control and preventive strategies, as well as for mapping malaria transmission via malaria transmission model (Gemperli et al., 2005; Gemperli et al., 2006).

1.3.2 Tools for mapping malaria

GIS and remote sensing

Remote sensing (RS) imagery is a powerful tool for determining the environmental predictors of malaria transmission. It is an important source that can provide such spatially rich information.

In recent years, significant progress has been made in the development of geographic information systems (GIS) and their applications in public health and spatial epidemiology. GIS are computerized systems capable of collecting, storing, handling, analyzing and displaying all forms of geographically referenced information. In GIS, information from different sources are represented as layers and linked in a spatial context. However, further research is needed on the relation between satellite-derived proxies on environmental conditions and ground climate data.

Since 1990s RS and GIS provide useful tools for mapping malariological indicators in Africa. Craig et al. (1999) produced a climatic suitability map of malaria transmission in sub-Sahara Africa and Snow et al. (1999) estimated the number of people at risk of malaria worldwide, by continent. In addition, integrated RS and GIS were used to produce maps of malaria vector distribution (Coetzee et al., 2000) and maps of vector breeding sites (Wood et al., 1992).

Malaria mapping is based on estimating the relation between malaria transmission and environmental/climatic factors and using this relation to predict malaria transmission at locations where the information is not available. Although the integrated GIS and remote sensing are valuable tools, they are not able to quantify the malaria-climate relation and to produce model-based predictions. Some GIS softwares have limited statistical capabilities, but these are inadequate for analyzing prevalence survey data.

Statistical modeling

Statistical models are useful for quantifying the relation between malaria risk and environmental factors and upon this relation predicting malaria risk at locations without observed malaria data. Malariological data are correlated in space since locations in close proximity have similar risk. Standard statistical methods assume independence of the observations and are not appropriate for analyzing spatially correlated data because they underestimate the standard error and thus the significance of the risk factors would be overestimated (Ver Hoef et al. 2001).

The type of spatial statistical methods used to analyze spatially correlated data depends on the nature of the geographical information. There are three kinds of spatial data: point-level (geostatistical), areal (lattice) and point patterns. Geostatistical data arise from observations collected at fixed locations over a continuous study region. Analysis of geostatistical data aims to identify environmental factors that determine the distribution of malaria in the presence of spatial correlation and kriging, that is spatial prediction at unobserved locations. Geographical dependence is considered as a function of the distance between locations. Areal data usually consist of counts or rates aggregated over a particular set of contiguous units. 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. Spatial proximity is defined by a neighboring structure. Point pattern data arise when the locations of particular events are not fixed, but random quantities. Questions of interest with these data center on whether events appear sporadically or they are clustered and which are the risk factors associated with such clusters.

Exploratory tools (variogram for geostatistical data, Moran's I and Geary's C for areal data and clustering statistics for point pattern data) describe the geographical pattern of the data and are available in most statistical packages. However, they are unable to filter the noise present in the data due to variable sample size between locations and produce smooth maps highlighting disease patterns.

Spatial models introduce at each data location (in the case of geostatistical data) or at each area (in the case of lattice data) an additional parameter on which the spatial correlation is incorporated. Hence these models are highly parametrized and fitting them could be challenging. Methods based on the maximum likelihood approaches are not appropriate

because they are not able to estimate simultaneously the malaria-climate relation and the spatial correlation. Bayesian hierarchical approaches avoid the computational problems in likelihood-based fitting by relying inference on Markov chain Monte Carlo (MCMC) simulation methods, hence are the best alternative in analyzing spatially correlated data.

For areal data, the mostly used prior distribution for random effects are simultaneously autoregressive (SAR) models (Whittle, 1954), conditional autoregressive (CAR) models (Clayton and Kaldor, 1987) and multivariate CAR models for multinomial response data (Vounatsou et al., 2000). In malaria epidemiology these approaches were employed to map malaria vector densities in a single village in Tanzania (Smith et al., 1995), malaria incidence rates in KwaZulu-Natal, South Africa (Kleinschmidt et al., 2001b) and in the state of Para, Brazil (Nobre et al., 2005) and to study malaria seasonality in Zimbabwe (Mabaso et al., 2005).

In the case of geostatistical data, the random effects model the underlying spatial process via a multivariate Normal distribution with the covariance matrix defined as a function of the distance between locations. Geostatistical models were introduced by Diggle et al. (1998) and have been employed to map malaria transmission in The Gambia (Diggle et al., 2002), Mali (Gemperli et al., 2005) and West-Africa (Gemperli et al., 2006).

Bayesian hierarchical models have become powerful methods in modeling spatial data due to development of simulation techniques like 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).

1.3.3 Malaria mapping in Africa - a review

The first global map of malaria endemicity was produced by Lysenko and Semashko almost 40 years ago (Lysenko and Semashko, 1968). The map combines data derived form historical documents and maps of several malariometric indices with expert opinion and simple climatic/geographical iso-lines. Historical maps at country level include the map of Namibia (de Meillon, 1951), Tanzania (Wilson, 1956), Kenya (Nelson, 1959) and Botswana (Chayabejata et al., 1975). The major drawback of the historical maps is that they made

limited use of empirical evidence and they did not capture the spatial and temporal heterogeneity of malaria transmission.

Hay et al. (1998) analyzed RS-derived climatic and malaria admission data recorded over 5 years in Kenya to produce a map of malaria seasonality. Thomson et al. (1999) used RS surrogate of climate to predict the number of children infected with *P.falciparum* according to different levels of bednet usage in The Gambia. Craig et al. (1999) used GIS techniques to spatially interpolated weather station data and based on the assumption that malaria transmission at continental level is limited by temperature and rainfall, defined climatic suitability for malaria transmission in sub-Sahara Africa. Using the malaria distribution model and a population distribution model (Deichmann, 1996), Snow et al. (1999) estimated the number of people at risk of malaria in sub-Sahara Africa. A spatial statistics approach was used by Kleinschimdt et al. (2000) and Kleinschimdt et al. (2001) to map malaria risk in Mali and West Africa, respectively, by fitting a standard regression model and applying classical kriging on the model residuals. Rogers et al. (2002) investigated the factors that influence the dynamics of malaria vector population and transmission and produce a map of EIR in Africa. Tanser el al. (2003) predicted potential effect of climate change on malaria transmission in Africa. Omumbo et al. (2005) defined two ecological zones in East Africa and modeled malaria risk using high-spatial resolution satellite imagery, as well as urbanization, water bodies and land use parameters.

The pioneers in using Bayesian statistics in spatial epidemiology of malaria were Diggle et al. (2002) who fitted a geostatistical model on malaria survey data from The Gambia without having produced a malaria risk map. Gemperli et al. (2005) and Gemperli et al. (2006) produced maps of malaria transmission in Mali and West Africa, respectively, making use of the Garki transmission model and Bayesian kriging. Gemperli (2003) was also the first to consider the non-stationarity feature of malaria and map malaria risk in Mali. 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. Using the Zambia National MIS data, Riedel et al. (unpublished) mapped malaria risk in Zambia.

1.3.4 Some methodological issues

There are a number of methodological problems related with malaria modeling, that is: i) modeling very large non-Gaussian geostatistical data; ii) analysis of non-stationary non-Gaussian geostatistical data; iii) modeling the non-linear effect of environmental/climatic factors on malaria risk and iv) adjusting for age and seasonality.

Fitting geostatistical models for non-Gaussian data requires repeated inversions of the covariance matrix of the spatial random effects which, for very large number of locations (N), is not feasible. In spatial data analyzes this computational challenge is referred to as the "large N problem". A number of strategies have been suggested for handling large spatial data sets (Gelfand et al., 2000; Rue and Tjelmeland, 2002; Stein et al., 2004; Gemperli and Vounatsou, 2006; Paciorek, 2007). Xia and Gelfand (2005) proposed an approach based on the assumption that a spatial random process can be approximated by a linear combination of $M << N$ random variables. This approach has the advantage of avoiding the inversion of the large NxN covariance matrix by reducing the problem to the inversion of a much smaller size matrix MxM.

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. There are a number of approaches for modeling non-stationarity in statistical literature (Haas, 1995; Higdon et al., 1998; Fuentes and Smith, 2002; Sampson and Guttorp, 1992). However, the most attractive method is the one developed by Kim et al. (2005) who modeled non-stationarity by partitioning the study area in random tiles, assuming an independent stationary process in each tile and independence between tiles. The only reference to non-stationarity in malaria mapping is by Gemperli (2003) who extended the work of Kim et al. (2005) for non-Gaussian data. The tessellation approach tackles another issue in Bayesian geostatistical modeling, that is computation of the inverse of the covariance matrix of the spatial process which appears in the prior distribution of the random effects. When the number of data locations is very large matrix inversion may not be feasible within time constraints.

The relation between malaria transmission and climate is complex and often non-linear. Nonparametric regression methods relax the assumption of linearity and the relation between outcome variable and the associated predictor variables is determined by the data,

not by a pre-specified model like in the parametric case. There are a lot of nonparametric modeling alternatives; here we mention local polynomial regression (Cleveland, 1979), kernel smoothing (Silverman, 1986), fractional polynomials (Royston and Altman, 1994) and spline smoothing. Splines are flexible models that take the form of piecewise polynomials joined at knots, where continuity constraints are imposed so that the function is smooth. Spline smoothing methods include regression splines (Eubank, 1988), B-splines (de Boor, 1978) and penalized splines (Eiler and Marx, 1996). The latter was implemented in the Bayesian framework by Crainiceanu et al. (2005), allowing simultaneous estimation of smooth functions and smoothing parameters. In malaria epidemiology field the most popular methods for modeling non-linearity are the use of the predictors in categories or functional transformations of the predictors (Kleinschmidt et al, 2001a; Gemperli et al., 2006).

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. Gemperli et al. (2005; 2006) demonstrated the feasibility of using malaria transmission models in malaria risk mapping by employing the Garki malaria transmission model (Dietz et al, 1974) to convert observed prevalence data into an estimated entomological measure of transmission intensity. They fitted a Bayesian geostatistical model on the estimates of EIR and employed Bayesian kriging to obtain a smooth map of EIR. The transmission model was applied again to convert the predicted EIR values into estimates of malaria prevalence for specific age groups of the population.

1.4 Objectives of the thesis

The overall objectives of this thesis were: 1) to develop Bayesian models for the analysis of Gaussian and non-Gaussian (binomial and negative binomial) geostatistical non-stationary data and 2) to validate and implement this methodology in the field of malaria epidemiology to produce maps of malaria risk and malaria transmission intensity and to assess the spatial effect of malaria control interventions on child mortality.

The specific methodological objectives were:

(i) development of geostatistical models for negative binomial data which allow applications to large data sets (Chapter 2);

(ii) development and validation of methods for non-stationary prevalence data appropriate for malaria mapping (Chapter 3 and Chapter 5);

(iii) modeling non-linear relation between malaria risk and environmental predictors (Chapter 4);

(iv) development and validation of models for mapping malaria transmission intensity (Chapter 6).

The above mentioned models were applied on data extracted from the Demographic Surveillance System (DSS), MARA and Malaria Indicator Surveys (MIS) databases to:

(1) evaluate the spatial effect of bednet use on child mortality in Kilombero Valley, Tanzania;

(2) identify environmental predictors of malaria transmission and produce smooth maps of malaria risk in Mali;

(3) produce smooth maps of malaria risk in West Africa based on the non-linear relation between climate and malaria risk;

(4) produce age and seasonality adjusted malaria risk maps from heterogeneous malaria survey data in Mali and Zambia.

Chapter 2

Spatial effects of mosquito bednets on child mortality

Gosoniu L.[1] ,Vounatsou P.[1], Tami A.[1,2], Nathan R.[3], Grundmann H.[4], Lengeler C.[1]

[1] Swiss Tropical Institute, Basel, Switzerland
[2] Royal Tropical Institute, Biomedical Research, Amsterdam, The Netherlands
[3] Ifakara Health Research and Development Center, Ifakara, Tanzania
[4] Centre for Infectious Diseases Epidemiology, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

This paper has been published in *BMC Public Health* 2008, 8:356

Summary

Background: Insecticide treated nets (ITN) have been proven to be an effective tool in reducing the burden of malaria. Few randomized clinical trials examined the spatial effect of ITNs on child mortality at a high coverage level, hence it is essential to better understand these effects in real-life situation with varying levels of coverage. We analyzed for the first time data from a large follow-up study in an area of high perennial malaria transmission in southern Tanzania to describe the spatial effects of bednets on all-cause child mortality.

Methods: The study was carried out between October 2001 and September 2003 in 25 villages in Kilombero Valley, southern Tanzania. Bayesian geostatistical models were fitted to assess the effect of different bednet density measures on child mortality adjusting for possible confounders.

Results: In the multivariate model addressing potential confounding, the only measure significantly associated with child mortality was the bed net density at household level; we failed to observe additional community effect benefit from bed net coverage in the community.

Conclusions: In this multiyear, 25 village assessment, despite substantial known inadequate insecticide-treatment for bed nets, the density of household bed net ownership was significantly associated with all cause child mortality reduction. The absence of community effect of bednets in our study area might be explained by (1) the small proportion of nets which are treated with insecticide, and (2) the relative homogeneity of coverage with nets in the area. To reduce malaria transmission for both users and non-users it is important to increase the ITNs and long-lasting nets coverage to at least the present untreated nets coverage.

2.1 Introduction

Plasmodium falciparum malaria is a leading infectious disease, accounting for approximately 300 to 500 million clinical cases each year and causing over one million deaths, mostly in African children younger than 5 years. Insecticide treated nets (ITN) have been proven to be an effective tool in reducing the burden of malaria (D'Alessandro et al., 1995; Binka et al., 1996; Nevill et al., 1996). Numerous trials all over the world have shown that such nets can reduce child mortality in endemic areas in Africa by 17% and roughly halve the number of clinical malaria episodes (Lengeler, 2004). These results were later confirmed under programme implementation (D'Alessandro et al., 1995; Schellenberg et al., 2001). It is well known that the use of ITNs provides significant individual protection, but direct and indirect effects on malaria transmission of treated and untreated nets on the wider community of bednet users and non-users are still little understood, despite some recent progresses (Killeen et al., 2007). Randomised trials in different malaria transmission regions examined the effect of ITNs on mortality of children without bednets. A study carried out in northern Ghana estimated that mortality risk in individuals without insecticide nets increased by 6.7% with every 100 m shift away from the nearest intervention compound (Binka et al., 1998). In western Kenya households without ITNs but within 300 m of ITN villages received nearly full protection (Hawley et al., 2003). These results conflict with those found from studies in The Gambia which concluded that protection against malaria seen in children using ITN is due to personal rather than community effect (Lindsay et al., 1993; Thomson et al., 1995, Quinones et al., 1998). A better understanding of these spatial effects in real-life situations is paramount for setting control targets, especially for understanding equity issues since these spatial effects mainly improve the situation of unprotected individuals, who are on average poorer. Moreover, the spatial effects of ITNs on non-bednet users in relation with the degree of density of bednets will indicate the type and level of bednet coverage that control programs need to achieve in order to maximize protection of non-bednet users. Here we present for the first time results for the spatial effects of ITNs in a "real-life" programme. One of the limitations of previous studies is that they used standard statistical methods which assume independence between observations. When these methods are applied to spatially correlated data, they underestimate the standard errors and thus overestimate the statistical significance of the covariates (Ver Hoef et al., 2001). In this paper we analyzed data from a large follow-up study in a highly malaria endemic area in southern Tanzania. Making use of a demographic surveillance system (DSS) we tracked child mortality prospectively and assessed

the relation between all-cause child mortality rates and the spatial effect of bednet density. To account for spatial clustering we fitted Bayesian geostatistical models with household-specific random effects. Models for geostatistical data introduce the spatial correlation in the covariance matrix of the household-specific random effects and model fit is based on Markov chain Monte Carlo methods (MCMC). MCMC estimation requires repeated inversions of the covariance matrix which, for large number of locations is computationally intensive and time consuming. To address this problem we propose a convolution model for the underlying spatial process which replaces large matrix inversion by the inversion of much smaller matrices.

2.2 Methods

2.2.1 Study area and population

The study was carried out from October 2001 to September 2003 in the 25 villages covered by a demographic surveillance system (DSS) in the Kilombero Valley, southern Tanzania. The DSS updates every 4 months demographic information on a population of about $73,000$ people living in $12,000$ dispersed households (Figure 2.1) in two districts - Kilombero and Ulanga (Armstrong Schellenberg et al., 2002). Most residents practice subsistence farming with rice and maize being the predominant crops. The climate is marked by a rainy season from November to May with annual rainfall ranging from 1200 to 1800 mm. Malaria is the foremost health problem, for both adults and children (Tanner et al., 1991). The prevailing malaria vectors in this region are *Anopheles gambiae* and *Anopheles funestus* with an estimated average entomological inoculation rate estimated of over 360 infective bites per person a year (Killeen et al., unpublished data). A large-scale social marketing programme of ITNs for malaria control has been running in this area since 1997 (Schellenberg et al., 2001).

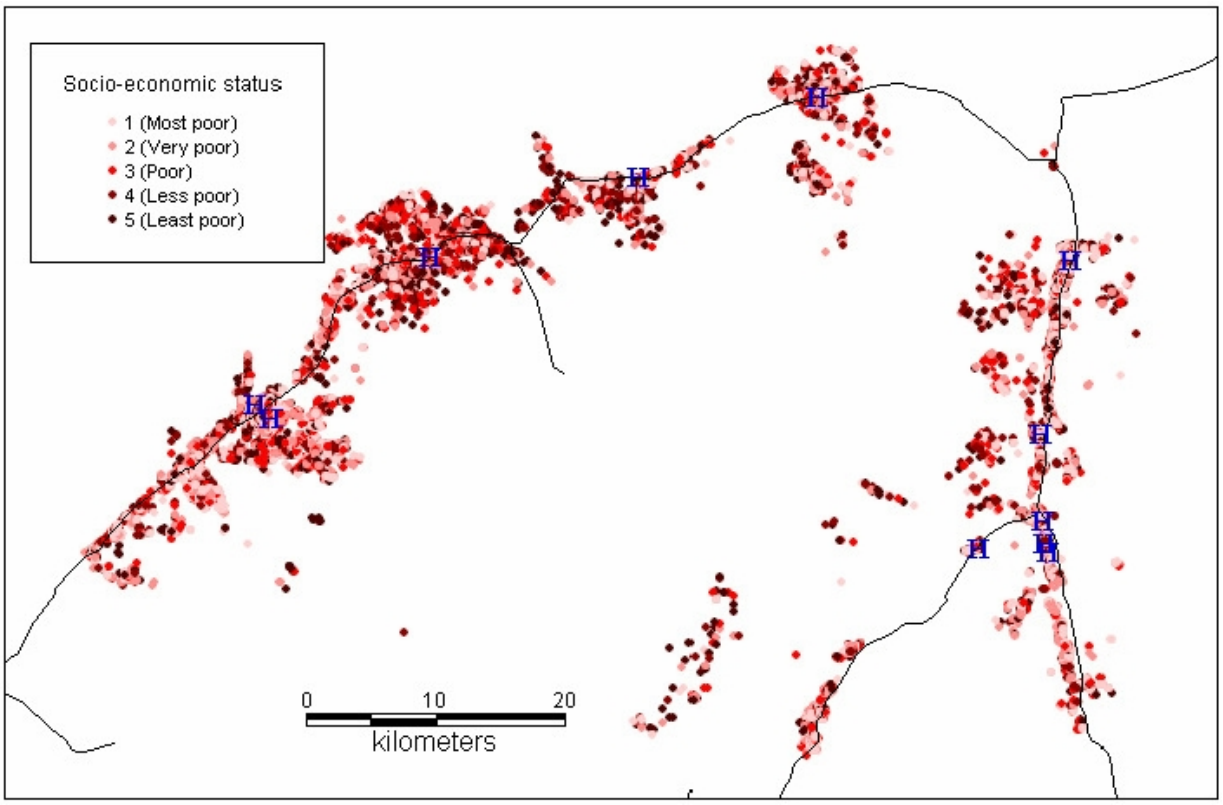

Figure 2.1: Distribution of the DSS households according to their socio-economic status and of the health facilities.

2.2.2 Data collection

Mortality data were obtained prospectively and continuously over a two-year period from the DSS, which allowed us to register age and sex data, births and migrations in and out the study area. Exact procedures are described in (Armstrong Schellenberg et al., 2002).

An additional survey was carried out in the DSS population in 2002 to collect socio-economic information. The survey questionnaire included a list of household assets (e.g. bednet), housing characteristics (e.g. type of roofing material) and type of energy and light. Although information on ITNs ownership was also collected, we did not use these data in our analysis since it was shown (Erlanger et al., 2004) that in this area two-thirds of the nets that were reported as having been re-treated within the last 12 months had insufficient insecticide to be effective.

Households and health facilities were geolocated using a hand-held Global Positioning

System (Garmin GPS 12, Garmin corp.) and Euclidean distances between houses and the health facilities were calculated.

2.2.3 Statistical analysis

Bednet density was defined as the number of bednets per person within a certain radius around each household. The following radii were chosen: 0 m (bednet coverage at household level), 50 m, 100 m, 150 m, 200 m, 300 m, 400 m, 500 m and 600 m.

A wealth index was calculated as a weighted sum of household assets. It has been shown that there is an inverse relationship between mortality and socio-economic status (Gwatkin, 2005); therefore the weights of the wealth index were obtained from the coefficients of a negative binomial model which estimated the effect of assets on all-age mortality. The weight of asset i was calculated as $w_i = \frac{b_i}{\sqrt{\sum_i b_i^2}}$, where b_i is the regression coefficient corresponding to asset i . The wealth index was divided into quintiles corresponding to poorest, very poor, poor, less poor and least poor groups of the population.

Negative binomial models were fitted to assess the effect of different bednet density measures on child mortality after adjusting for possible confounders: sex, wealth index and distance to the nearest health facility, using STATA v. 9.0 (Stata Corporation, College Station, TX, USA).

To estimate the effect of bednet density on the mortality of children without nets we performed a similar analysis. In particular, we defined bednet density as above, considering as index households the ones without any bednet. We then fitted the negative binomial models adjusted for the above mentioned confounders.

The household mortality data are correlated in space since common environmental risk factors, proximity to breeding sites and socio-economic exposures may influence the mortality outcome similarly in households within the same geographical area. The independence assumption of the standard negative binomial models may result in overestimation of the significance of the bednet coverage covariate. To address this problem Bayesian geostatistical negative binomial models were fitted with household-level random effects. Spatial correlation was modeled by assuming that the random effects are distributed according to a multivariate normal distribution with variance-covariance matrix related to an exponential correlation function between household locations, i.e. $\sigma^2 exp(-d_{ij}\rho)$, where d_{ij} is the

Euclidean distance between households i and j, σ^2 is the geographic variability known as the sill and ρ is the rate of correlation decay. The distribution of random effect defines the so called Gaussian spatial process. Model fit requires the inversion of a covariance matrix with the same size as the sample size. Due to the large number of observations in our dataset, the estimation of model parameters becomes unstable and unfeasible. To overcome this problem we propose a model based on a convolution representation that is, we approximate the spatial random process by a weighted sum of a small number of stationary spatial processes. The size of the covariance matrix that needs to be inverted is then much smaller, therefore the method is computationally efficient. We employed Markov chain Monte Carlo simulation to estimate the model parameters. Further details on this modeling approach are given in the appendix. The analysis was implemented using software written by the authors in FORTRAN 95 (Compaq Visual FORTRAN Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithm Group Ltd.).

2.3 Results

A total number of $11,134$ children from $7,403$ households had information available on both geolocation and socio-economic covariates.

The pooled data revealed an overall all-age crude mortality rate of 9.5 per 1000 person-years and an overall child mortality of 26.2 per 1000 person-years with no difference between the two districts ($P = 0.98$ and $P = 0.73$, respectively).

The insecticide treatment status of the nets was difficult to ascertain, therefore the results reported in this section refer to bednets only, whether treated or not. The mean bednet density in Kilombero Valley was 270 nets per 1000 inhabitants. $10,160$ households (85%) had at least one bednet and the mean number of bednets per household was 1.64.

Table 2.1 shows the overall child mortality rates together with district-specific child mortality rates by sex, socio-economic status, distance to the nearest health facility and bednet density at household level. Since there were no significant differences between child mortality rates in Kilombero and Ulanga Districts, all further analysis was done by pooling the data of the two districts. Males had a slightly lower mortality rate than females, but sex was not significantly associated with childhood mortality rates (Incidence-Rate Ratios (IRR) $= 0.90$, $P = 0.22$). Similarly, socio-economic status was not significantly associated

with child mortality ($P = 0.12$), but we could notice a trend for children from the relatively better off households to have a lower mortality rate than their poorer counterparts. No significant association was observed with distance to the nearest health facility, but children living ≥ 1 km away from the nearest health facility tended to have higher mortality rates than those living in close proximity.

Explanatory variables	Number of children(%)	Child mortality rate[a]			P-value
		Overall	Kilombero	Ulanga	
Sex					
Female	5669 (50.9)	27.6	29.8	25.0	0.81
Male	5465 (49.1)	24.7	27.1	21.6	0.80
Socio-economic status					
Poorest	2203 (19.8)	31.1	36.5	26.0	0.75
Very poor	2265 (20.3)	26.0	27.5	24.1	0.92
Poor	2281 (20.5)	25.7	29.5	20.6	0.79
Less poor	2239 (20.1)	21.3	21.1	21.6	0.99
Least poor	2146 (19.3)	27.1	28.9	24.5	0.90
Distance to nearest health facility					
< 1 km	2793 (25.1)	23.3	25.9	20.7	0.86
$1 - 4.9$ km	4666 (41.9)	27.2	29.4	24.1	0.82
≥ 5 km	3675 (33.0)	26.9	29.0	24.7	0.87
Bednet density at household level[b]					
0	1199 (10.8)	28.9	40.4	19.5	0.66
$0 - 0.2$	2531 (22.7)	27.9	28.9	26.8	0.95
$0.2 - 0.3$	3426 (30.8)	28.5	30.3	28.5	0.87
$0.3 - 0.5$	3026 (27.2)	22.3	22.4	22.3	0.99
> 0.5	952 (8.5)	22.6	28.9	13.9	0.79

[a] : Mortality rate per 1000 person years.
[b] : Number of bednets per person within a 0 m radius around each household.

Table 2.1: Overall and district-specific child mortality rates by sex, socio-economic status, distance to the nearest health facility and bednet density at household level

A simple bivariate analysis showed that bednet density at household level was significantly associated with reduced child mortality ($IRR = 0.50$, $P = 0.02$). There was a tendency for mortality rates to decrease for children living in households with at least 30% bednet density coverage.

The effect of various bednet density measures on child mortality after adjusting for possible confounders is shown in Table 2.2. Surprisingly, the only measure significantly associated with child mortality was the bednet density at household level (R_0) ($IRR = 0.53$, $P = 0.04$). We noted that the mean bednet density was similar for all radii, whereas the standard deviation tended to become smaller as the radius was increasing.

Bednet density	Mean (St.dev.)	% of households without bednets	IRR[a]	95% CI	LRT[b]	P-value[c]
R_0	0.25 (0.15)	0.00	0.53	(0.29,0.97)	4.37	0.04
R_{50}	0.18 (0.20)	0.09	0.64	(0.40,1.03)	3.48	0.06
R_{100}	0.24 (0.18)	12.68	1.13	(0.73,1.74)	0.27	0.61
R_{150}	0.25 (0.13)	13.84	1.18	(0.61,2.30)	0.24	0.62
R_{200}	0.26 (0.12)	14.44	1.69	(0.79,3.61)	1.79	0.18
R_{300}	0.26 (0.09)	15.04	2.51	(0.96,6.55)	3.46	0.06
R_{400}	0.27 (0.08)	15.04	2.10	(0.79,5.59)	2.05	0.15
R_{500}	0.27 (0.07)	15.18	2.40	(0.70,8.25)	1.89	0.17
R_{600}	0.27 (0.07)	15.23	2.89	(0.74,11.25)	2.32	0.13

[a] : IRR:Incidence-rate ratios.
[b] : LRT:Likelihood ratio test.
[c] : P-value based on likelihood ratio test (LRT).

Table 2.2: Summary of bednet density measures and estimates of the effect of bednet measures on child mortality, adjusted by sex, socio-economic status and distance to the nearest health facility. Results obtained by fitting negative binomial models.

The results of the bivariate and multivariate non-spatial negative binomial models are shown in Table 2.3. None of the explanatory variables were significantly associated with child mortality, except the fourth wealth quintile. After taking into account the spatial correlation present in the data, the effect of the covariates remained non-significant. However, the confidence intervals became wider, confirming the importance of taking into account spatial correlation when analyzing geographical data (Cressie, 1993). The parameters σ^2

and ρ shown in Table 2.3 measure the spatial variance and the rate of correlation decay (smoothing parameter), respectively. The estimates of the smoothing parameter ρ indicate a low spatial correlation in the child mortality rate data. In fact ρ was estimated to be 6.97 km, which in our exponential setting is translated to a minimum distance for which spatial correlation decrease to 0.05 of only around 0.43 km.

Indicator	Bivariate model		Multivariate model		Spatial model	
	IRR[a]	95% CI	IRR[a]	95% CI	IRR[a]	95% CI
Sex						
Female	1.0		1.0		1.0	
Male	0.90	(0.75,1.07)	0.89	(0.75,1.06)	0.88	(0.73,1.06)
Socio-economic status						
Most poor	1.0		1.0		1.0	
Very poor	0.83	(0.64,1.09)	0.84	(0.65,1.11)	0.87	(0.63,1.21)
Poor	0.82	(0.63,1.08)	0.84	(0.64,1.10)	0.82	(0.64,1.05)
Less poor	0.69	(0.52,0.91)	0.70	(0.53,0.93)	0.68	(0.51,0.94)
Least poor	0.87	(0.67,1.14)	0.90	(0.69,1.18)	0.90	(0.68,1.20)
Bednet density at household level						
0	1.0		1.0		1.0	
$0 - 0.2$	0.96	(0.71,1.32)	0.99	(0.73,1.36)	1.03	(0.83,1.29)
$0.2 - 0.3$	0.99	(0.73,1.33)	1.02	(0.76,1.39)	1.04	(0.81,1.39)
$0.3 - 0.5$	0.77	(0.57,1.06)	0.81	(0.59,1.11)	0.84	(0.56,1.13)
> 0.5	0.78	(0.52,1.17)	0.81	(0.54,1.23)	0.76	(0.54,1.24)
Distance to nearest health facility	0.80	(0.16,3.96)	0.60	(0.10,3.68)	0.23	(0.03,3.71)
Spatial parameters						
σ^2					0.75	(0.35,1.16)
Range $(3/\rho)^b$					0.43	(0.39,0.48)

[a] : IRR:Incidence-rate ratios.
[b] : Spatial correlation is significant ($> 5\%$) within this distance.

Table 2.3: Results of the association of sex, socio-economic status, bednet density at household level and distance to nearest health facility with child mortality, resulting from the bivariate and multivariate non-spatial models and spatial model.

Table 2.4 depicts the effect of different bednet density measures on the mortality of children without any bednet after adjusting for sex, socio-economic status and distance to the nearest facility. The results show no significant association between any bednet density measure and mortality of children without nets, indicating no detectable community effect.

Bednet density	Incidence risk ratio				P-value[a]
	No bednet	$0 - 0.2$	$0.2 - 0.3$	> 0.3	
R_{50}	1.0	0.88 (0.42,1.83)	0.70 (0.31,1.60)	0.89 (0.42,1.71)	0.94
R_{100}	1.0	0.64 (0.29,1.42)	1.04 (0.51,2.10)	1.15 (0.57,2.34)	0.52
R_{150}	1.0	0.82 (0.34,1.98)	0.91 (0.38,2.17)	2.06 (0.91,4.64)	0.08
R_{200}	1.0	0.74 (0.30,1.79)	0.68 (0.28,1.63)	1.35 (0.57,3.20)	0.30
R_{300}	1.0	1.18 (0.33,4.16)	1.64 (0.48,5.62)	1.40 (0.38,5.08)	0.71
R_{400}	1.0	1.40 (0.31,6.28)	1.90 (0.44,8.24)	1.49 (0.32,7.00)	0.81
R_{500}	1.0	1.97 (0.26,15.15)	2.31 (0.31,17.38)	1.80 (0.22,14.54)	0.73
R_{600}	1.0	1.22 (0.16,9.33)	1.63 (0.22,12.03)	1.14 (0.14,9.06)	0.81

[a] : P-value based on likelihood ratio test (LRT).

Table 2.4: Estimated effect of bednet measures on mortality of children without nets, adjusted by sex, socio-economic status and distance to the nearest health facility, obtained by fitting negative binomial models.

Pearsons correlation coefficient between bednet density and bednet usage was 0.83, indicating a strong correlation between the two measures. Hence, the results regarding the bednet density could be extended to bednet usage.

2.4 Discussion

We examined the effect of a variety of factors on child mortality in an area of high perennial malaria transmission in southern Tanzania and identified that the density of household bed net ownership was the only factor significantly associated with child mortality reduction. The spatial effects of bednets on all-cause child mortality in an area of high perennial malaria transmission in southern Tanzania have been presented here. The effect of different bednet density measures was estimated after adjusting for possible confounders like sex, socio-economic status and distance to the nearest health facility. We concentrated on all-cause child mortality because in rural Africa it is difficult to assess malaria-specific

mortality. Most deaths occur at home and verbal autopsy is the only tool available to determine the cause of mortality. It has been shown (Snow et al., 1992; Todd et al., 1994) that this is an inaccurate method to detect malaria, having a low sensitivity and specificity.

Our results indicated a surprising lack of community effect of bednets on childhood mortality. This conclusion is based on the fact that only the bednet density at household level had a significant protective effect on child mortality. When net density within $\geq$ 50m was considered, the risk of child mortality increased slightly but the relation was not significant. Our findings contrast with previous studies in Africa, which demonstrated a strong community-wide effect of ITNs on child mortality (Binka et al., 1998, Hawley et al., 2003). However, our study differed from the studies mentioned above in a number of ways.

Firstly, the epidemiological studies that demonstrated the mass effect of ITNs on child mortality were all designed as community-trial interventions, ensuring a uniformly high coverage of treated nets in the intervention group, with a control group almost not using any sort of nets. This creates a strong gradient of ITN at the margins use, which allows a good measure of spatial effects. By contrast, net usage, treated or not, was uniformly high in our study area, with the result that any sort of spatial effects would be more difficult to detect unless there would be heterogeneity in coverage, which was not the case.

Secondly, we were not able to distinguish between treated and untreated nets in the field because there is no reliable testing method to do this at present. Armstrong-Schellenberg et al. (2002b) and Erlanger et al. (2004) showed that in our study area use of insecticide re-treatment is relatively low, with only 32% of the nets having enough insecticide to ensure an entomological impact. Since untreated nets are less effective than treated ones (Lengeler, 2004; Maxwell et al., 1999; D'Alessandro et al., 1995b), this had certainly an impact on the analysis by reducing differences between users and non-users.

Lastly, as specific data on bednet use was not available for the whole sample, we created a different measure of the impact of bednets: the "bednet density" defined as the ratio between the number of bednets owned and the number of people living in a specific area. Previous studies in this region showed that on average 2 people sleep under a bednet with an overall bednet use of about 75% (Killeen et al., unpublished data).

Most analyzes of bednets effect on different malaria-related outcomes so far have been based on the assumption of independence between observations. However, household mortality data are spatially correlated due to common exposures. When the spatial correlation

present in the data is ignored, the statistical significance of the covariates is overestimated. We could control for that by using a Bayesian geostatistical approach to assess the child mortality-bednet density relation. Bayesian computation implemented via MCMC enabled simultaneous estimation of all model parameters together with their standard errors, a feature that is not available in the maximum likelihood based framework.

Despite these limitations, our results are consistent with the analysis of ITNs protective efficacy against malaria transmission in Kilombero Valley (Killeen et al., 2006), which predicted little community-level protection for the individuals not using ITNs. The most likely explanations for this were the small proportion of re-treated nets and the insufficient concentration of insecticide present in the bednets, leading to diversion of mosquitoes. A recently developed model for the transmission of malaria using data collected in Tanzania (Killeen et al., 2007) predicted that modest bednet coverage (35% - 65%) of the entire population, rather than just high-risk groups (pregnant women and young children) is needed to achieve community-wide protection similar to, or greater than, individual protection. Hence, there is clearly a strong case for improving the status of insecticide treatment through the introduction of long-lasting insecticidal nets (LLINs) which are now becoming increasingly available (Guillet et al., 2001) and for the wide-use of ITNs and LLINs by the whole population. We expect that achieving a high coverage with LLINs will result in further substantial reductions of malaria transmission and hence malaria-related mortality and morbidity for both users and non-users.

Acknowledgments

The authors would like to acknowledge the residents of the Kilombero Valley for their commitment during the study and the Ifakara DSS field and office workers who carried out the surveys and compiled the database. We would like to thank Prof. Tom Smith and Prof. Dr. Don de Savigny for stimulating discussions. Many thanks also to Dr. Hassan Mshinda, Dr. Honorati Masanja, Oskar Mukasa and Dr. Salim Abdulla for their overall support. This manuscript has been published with kind permission of Dr Andrew Kitua, Director of the National Medical Research Institute, Tanzania. This study was funded by the Swiss National Science Foundation (Grant number 3252B0-102136/1). Ethical review and approval for this study was provided by the Medical Research Coordination Committee of NIMR (Ref. No NIMR/HQ/R.8a/VOL.X/12, dated 28/4/1998).

2.5 Appendix

Let Y_{il} be the mortality outcome of child l at site s_i $i = 1, \ldots, n$ taking value 1 if the child is dead and 0 otherwise. We assume that Y_{il} arises from a negative binomial distribution, that is $Y_{il} \sim NegBin(p_{il}, r)$, where p_{il} is the probability that child l at location s_i is dead and r is the parameter that quantifies the amount of extra Poisson variation. To account for spatial variation in the data, location-specific random effects were integrated in the negative binomial model. The probability p_{il} is modeled as $p_{il} = \frac{r}{r+z_{il}}$, with $log(z_{il}) = log(pyrs_{il}) + X_{il}^T\boldsymbol{\beta} + \phi_i$, where X_{il} is the vector of associated covariates, $\boldsymbol{\beta}$ are the regression coefficients and ϕ_i's are the spatial random effects. $pyrs_{il}$ represents the number of person-years corresponding to child l at location s_i and $log(pyrs_{il})$ is considered as covariate with regression coefficient fixed to 1 and is referred as offset.

The standard approach to model the spatial dependence is to assume that the covariance of ϕ_i's at every two locations s_i and s_j decreases with their distance d_{ij}, that is $\Sigma_{ij} = \sigma^2 f(d_{ij}; \rho)$ with $f(d_{ij}; \rho) = exp(-d_{ij}\rho)$, where $\rho > 0$ is a smoothing parameter that controls the rate of correlation decay with increasing distance and σ^2 quantifies the amount of spatially structured variation. Estimation of the location-specific random effects and of the spatial parameters requires repeated inversions of the covariance matrix Σ. Due to the large number of locations in our dataset $(7,403)$, matrix inversion is computationally intensive and is not feasible within practical time constrains. To overcome this issue we develop a convolution model for the underlying spatial process. In particular, we choose a small number of locations t_k, $k = 1, \ldots, K$ over the study region, assume a stationary spatial process ω_k over these locations and we model the spatial random effect ϕ_i at each data location s_i as a weighted sum of the fixed location stationary processes. That is, $\phi_i = \sum_{k=1}^{K} a(i,k)\omega_k$, where the weights $a(i,k)$ are decreasing functions of the distance between data location s_i and the fixed location t_k and $\boldsymbol{\omega_k} \sim N(\mathbf{0}, \Sigma_k)$, with $(\Sigma_k)_{hl} = \sigma^2 exp(-d_{hl}\rho)$, where d_{hl} is the distance between the fixed locations t_h and t_l. This approach avoids the inversion of the large covariance matrix nxn , reducing the problem to the inversion of a much smaller size matrix KxK. For this specific analysis we have chosen $K = 200$.

For the correlation function chosen, the minimum distance for which spatial correlation between locations is below 5% is $3/\rho$ (range). The above specification of spatial correlation is isotropic, assuming that correlation is the same in all directions.

Following a Bayesian model specification, we adopt prior distributions for the model parameters as follows: non-informative uniform prior distributions for the regression coefficients $\boldsymbol{\beta}$, inverse gamma prior distribution for σ^2 and gamma prior distribution for the decay parameter ρ and the over-dispersion parameter r.

We estimate the model parameters using Markov chain Monte Carlo simulation. In particular we implemented Gibbs sampler (Gelfand and Smith, 1990), which requires simulating from the full conditional distributions of all parameters iteratively until convergence. The full conditional distribution of σ^2 is an inverse gamma distribution and it is straightforward to simulate from. The conditional posterior distribution of $\boldsymbol{\beta}$, ρ, and r do not have known forms. We simulate from these distributions using the Metropolis algorithm with a Normal proposal distribution having the mean equal to the parameter estimate from the previous Gibbs iteration and the variance equal to a fixed number, iteratively adapted to optimize the acceptance rates. We have run a five-chain sampler with a burn-in of $10,000$ iterations and we assessed the convergence by inspection of ergodic averages of selected model parameters after $200,000$ iterations.

Chapter 3

Bayesian modeling of geostatistical malaria risk data

Gosoniu L.[1] ,Vounatsou P.[1], Sogoba N.[2] and Smith T.[1]

[1] Swiss Tropical Institute, Basel, Switzerland
[2] Malaria Research and Training Center, Universite du Mali, Bamako, Mali

This paper has been published in *Geospatial Health* 1, 2006, 127-139

Summary

Bayesian geostatistical models applied to malaria risk data quantify the environment-disease relations, identify significant environmental predictors of malaria transmission and provide model-based predictions of malaria risk together with their precision. These models are often based on the stationarity assumption which implies that spatial correlation is a function of distance between locations and independent of location. We relax this assumption and analyze malaria survey data in Mali using a Bayesian non-stationary model. Model fit and predictions are based on Markov chain Monte Carlo simulation methods. Model validation compares the predictive ability of the non-stationary model with the stationary analogue. Results indicate that the stationarity assumption is important because it influences the significance of environmental factors and the corresponding malaria risk maps.

Keywords: Bayesian inference; malaria risk; Markov chain Monte Carlo; non-stationarity; kriging

3.1 Introduction

Malaria is the most prevalent human parasitic disease. Although reliable estimates are not available, rough calculations suggest that globally, 250 million new cases occur each year resulting in more than one million deaths (Bruce-Chwatt, 1952; Greenwood, 1990; WHO, 2004). Around 90% of these deaths happen in sub-saharan Africa, mostly in children less than 5 years old. The malaria parasite is transmitted from human to human via the bite of infected female *Anopheles* mosquitoes. Transmission depends on the distribution and abundance of the mosquitoes which are sensitive to environmental factors mainly temperature, rainfall and humidity. By determining the relations between the disease and the environment, the burden of malaria can be estimated at places where data on transmission are not available and high risk areas can be identified. Reliable maps of malaria transmission can guide intervention strategies and thus optimize the use of limited human and financial resources to areas of most need. In addition, early warning systems can be developed to predict epidemics from environmental changes.

Remote sensing is a useful source of satellite-derived environmental data. Geographic Information Systems (GIS) has emerged over the last 15 years as a powerful tool for linking and displaying information from many different sources such as environmental and disease data, in a spatial context. Integrated GIS and remote sensing have been applied to map malaria risk in Africa (Snow et al., 1996; Craig et al., 1999; Thomson et al., 1999; Hay et al., 2000; Kleinschmidt et al., 2001; Rogers et al., 2002). However, the mapping capabilities of existing GIS software are rather limited as they are unable to quantify the relation between environmental factors and malaria risk and to produce model-based predictions. GIS is also used in early warning systems for malaria epidemics (Abeku et al., 2004; Grover et al., 2005; Thomson et al., 2006), however the thresholds for environmental factors have been based on expert opinion rather than observed data.

Statistical modeling gives mathematical descriptions of the environment-disease relations, identifies significant environmental predictors of malaria transmission and provides predictions of malaria risk based on the above relations together with their precision. The standard statistical models assume independence of observations. However, malaria infectious cases cluster due to underlying common environments. When spatially correlated data are analyzed this independence assumption leads to overestimation of the statistical significance of the covariates (Cressie, 1993). Spatial models incorporate the spatial correlation according to the way the geographical information is available. For areal data

(typically counts or rates aggregated over a particular set of contiguous units) the spatial correlation is defined by a neighborhood structure. For geostatistical data (collected at fixed locations over a continuous study region) the spatial correlation is usually considered as a function of the distance between locations.

Linear regression is applied for modeling geostatistical continuous data which are normally distributed (Gaussian). The spatial correlation is introduced in the residuals (error terms) of the model. The parameters can not be estimated simultaneously, thus iterative methods are used. The generalized least squares approach (GLS) estimates the regression coefficients conditional on the spatial correlation parameters. The correlation parameters can be estimated conditional on the regression coefficients empirically from the residuals or using maximum likelihood based approaches (Zimmerman and Zimmerman, 1991).

In this paper we present models for geostatistical prevalence data derived from malaria surveys carried out at a number of fixed locations. For this type of data and in general for non-Gaussian geographical data, spatial models introduce at each location an error term (random effect) and incorporate spatial correlation on these parameters. Estimation can use generalized linear mixed models (GLMM). However, this is difficult to apply for spatial problems with large number of locations (Gemperli and Vounatsou, 2004). In addition, estimation of standard errors depends on asymptotic results, which in the case of geostatistical models, do not give unique estimates (Tubila, 1975).

Bayesian geostatistical models implemented via Monte Carlo methods avoid asymptotic inference and the computational problems encountered in likelihood-based fitting. They were introduced for the analysis of geostatistical data by Diggle et al. (1998) and have been employed in modeling the spatial distribution of parasitic diseases (Diggle et al., 2002; Gemperli et al., 2004; Raso et al., 2004; Abdulla et al., 2005; Gemperli et al., 2005; Raso et al., 2005; Clements et al., 2006; Gemperli et al., 2006; Raso et al., 2006). Most health applications of Bayesian geostatistical models have relied on an assumption of stationarity, which implies that the spatial correlation is a function of the distance between locations and independent of locations themselves. This assumption is questionable when malariological indices are modeled since local characteristics related to human activities, land use, environment and vector ecology influence spatial correlation differently at the different locations.

In this paper we present and compare Bayesian stationary and non-stationary models for mapping malaria risk data in Mali. Using model validation we assess the assumption of

stationarity and show the impact it can have on inference when non-stationary data are analyzed. In Section 3.2 we describe the malaria data which motivated this work and the environmental predictors we extracted from remote sensing and GIS databases. Section 3.3 introduces the stationary and non-stationary Bayesian geostatistical models as well as the model validation approaches. The results are presented in Section 3.4 and the paper ends with final remarks and suggestions for future work given in Section 3.5.

3.2 Data

3.2.1 Malaria data

The malaria data were extracted from the "Mapping Malaria Risk in Africa" (MARA/ ARMA,1998) database. This is the most comprehensive database on malariological indices initiated to provide a malaria risk atlas by collecting published and unpublished data from over $10,000$ surveys across Africa. We analyzed malaria prevalence data from surveys carried out in children between 1 and 10 years old at 86 sites in Mali (Figure 3.1) between 1977 and 1995, including a total of $43,492$ children.

3.2.2 Climatic and environmental data

The environmental data and the databases from which they were extracted are given in Table 3.1. Preliminary non-spatial analysis indicated that the following factors and their transformation should be included in the analysis: Normalized Difference Vegetation Index (NDVI), NDVI squared, length of malaria season, amount of rainfall, maximum temperature, squared maximum temperature, minimum temperature, squared minimum temperature, distance to the nearest water body and squared distance to the nearest water body.

Factor	Resolution	Source
NDVI	$8km^2$	NASA AVHRR Land data sets
Temperature	$5km^2$	Hutchinson et al. 1996
Rainfall	$5km^2$	Hutchinson et al. 1996
Water bodies	$1km^2$	World Resources Institute 1995
Season length	$5km^2$	Gemperli et al. 2006

Table 3.1: Spatial databases used in the analysis.

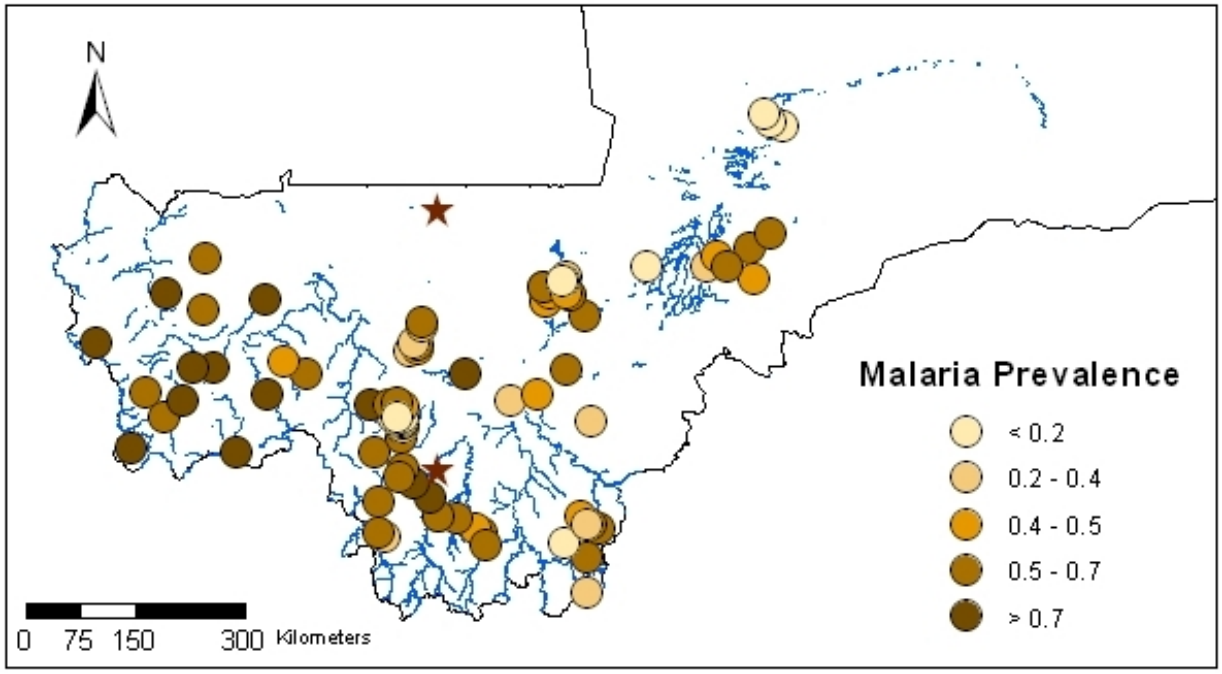

Figure 3.1: Sampling locations with dot shading indicating the observed malaria prevalence. The stars indicate the centroids of two fixed tiles used to account for non-stationarity.

The NDVI values were extracted from satellite information conducted by the NOAAA/ NASA Pathfinder AVHRR Land Project (Agbu and James, 1994). NDVI is shown to be highly correlated with other measures of vegetation (Justice et al., 1985) and used as a proxy of vegetation and soil wetness. Index values can vary from -1 to 1 with higher values $(0.3-0.6)$ indicating the presence of green vegetation, and negative values indicating water. We used the logarithm of the yearly mean NDVI over the malaria season. The temperature and rainfall data were obtained from the "Topographic and Climate Data Base for Africa (1920-1980)" Version 1.1 by Hutchinson et al. (1996). We used the yearly averages over the months suitable for transmission according to the map of Gemperli et al. (2006). The distance to the nearest water source was calculated based on permanent rivers and lakes extracted from "African Data Sampler" (WRI, 1995). The length of malaria season was defined using the seasonality model of (Gemperli et al., 2006). The covariates were standardized prior to the analysis.

3.3 Bayesian geostatistical models

3.3.1 Model formulation

The malaria data are derived from surveys carried out at the various locations. These are typical binomial data and modeled via logistic regression. Let N_i be the number of children tested at location s_i, $i = 1, \ldots, n$, Y_i be the number of those found with malaria parasites in a blood sample and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ be the vector of p associated environmental predictors observed at location s_i. We assume that Y_i arises from a Binomial distribution, that is $Y_i \sim Bin(N_i, p_i)$ with parameter p_i measuring malaria risk at location s_i and model the relation between the malaria risk and environmental covariates $\boldsymbol{X}_i$ via the logistic regression $logit(p_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are the regression coefficients. This model assumes independence between the surveys. However, the geographical location introduces correlation since the malaria risk at nearby locations is influenced by similar environmental factors and therefore it is expected that the closer the locations the similar the way malaria risk varies. To account for spatial variation in the data we introduce an error term (random effect) ϕ_i at each location s_i, that is $logit(p_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \phi_i$ and model the spatial correlation on the ϕ_i parameters, that is the ϕ_i's are not independent but they derive from a distribution which models the correlation or equivalently the covariance between every pair of random effects. We adopt the multivariate Normal distribution for the ϕ_i's since they represent error terms and therefore they are defined on a continuous scale, that is $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T \sim N(\mathbf{0}, \Sigma)$. Σ is a matrix with elements Σ_{ij} quantifying the covariance $Cov(\phi_i, \phi_j)$ between every pair (ϕ_i, ϕ_j) at locations s_i and s_j respectively. The distribution of random effect $\boldsymbol{\phi}$ defines the so called Gaussian spatial process.

Stationary model

Assuming stationarity, spatial correlation is considered to be a function of distance only and irrespective of location. Under this assumption, we take $\Sigma_{ij} = \sigma^2 corr(d_{ij}; \rho)$, where $corr$ is a parametric correlation function of the distance d_{ij} between locations s_i and s_j. Several correlation functions have been suggested by Ecker and Gelfand (1997). In this application, we choose an exponential correlation function $corr(d_{ij}; \rho) = exp(-d_{ij}\rho)$, where $\rho > 0$ measures the rate of decrease of correlation with distance and it is known as the range parameter of the spatial process. For the correlation function chosen, the minimum distance for which the correlation becomes less than 5 % is $3/\rho$. σ^2 measures within location

variation and it is known as the sill of spatial process. The above specification of spatial correlation is isotropic, assuming that correlation is the same in all directions.

Non-stationary model

The assumption of stationarity is not always justified, especially over large geographical areas. Differences in agro-ecological zones, health systems and socio-economic indicators may change geographical correlation differently at various locations. In recent years, non-stationary specifications are based on piecewise Gaussian processes (Kim et al., 2002, Gemperli et al., 2003) kernel convolution methods (Higdon et al., 1999; Fuentes et al. 2002) and normalized distance-weighted sums of stationary processes (Banerjee et al., 2004). In Raso et al. (2005) we extended the Banerjee et al. (2004) model for non-Gaussian prevalence data to map hookworm risk in the region of Man in Cote d'Ivoire. In this paper, we use the same approach to analyze the Mali malaria prevalence data.

The study area is partitioned into K subregions, a stationary spatial process $\boldsymbol{\omega_k}$ is assumed in each subregion $k = 1, \ldots, K$ that is $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kn})^T \sim N(\mathbf{0}, \Sigma_k)$ and the spatial random effect ϕ_i at each location s_i is modeled as a weighted sum of the subregion-specific stationary processes, that is $\phi_i = \sum_{k=1}^{K} a_{ik}\omega_{ki}$, where a_{ik} are decreasing functions of the distance between location s_i and the centroid of the subregion k. This is equivalent to say that $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T \sim N(\mathbf{0}, \sum_{k=1}^{K} A_k \Sigma_k A_k)$, where $A_k = diag\{a_{1k}, a_{2k}, \ldots, a_{nk}\}$ is a matrix which has the elements $a_{1k}, a_{2k}, \ldots, a_{nk}$ on the main diagonal and 0 outside the main diagonal. The Σ_k are specified using exponential correlation functions as in the case of the stationary model described in the previous section, that is $(\Sigma_k)_{ij} = \sigma_k^2 exp(-d_{ij}\rho_k)$. Note that the spatial parameters σ_k and ρ_k are specific for each subregion k.

Three non-stationary models were fitted with $K = 2, 3, 4$. Due to relatively small number of locations included in our data we have not investigated models with larger number of tiles to avoid estimating spatial parameters from tiles with few locations and thus over-parametrising the models. The sub-regions were obtained by overlaying a rectangular grid over the study area. We first divide the rectangle in half north-to-south and then, to obtain four sub-regions, each of these rectangles is partitioned in half west-to-east. For $K = 3$ we divide the north part of our study area in two rectangles and consider the south area as one sub-region. The centroids of two fixed tiles are shown in Figure 3.1.

3.3.2 Bayesian specification and implementation

The Bayesian approach to inference allows parameter estimation using information coming from the data via the likelihood function as well as information coming from other sources prior seen the data (i.e. previous studies, subjective judgments) which is formalized via prior distributions. Bayes theorem combines the likelihood function and the prior distribution defining a new quantity, known as posterior distribution which forms the basis of Bayesian inference. Parameters are considered as random and their estimation results not only in a single value, but in the probabilities of their possible values which are given by their probability distribution, known as marginal posterior distribution.

To complete the Bayesian model formulation of the geostatistical models mentioned above we need to specify prior distributions for their parameters. For the regression coefficients we adopt a non-informative uniform prior distributions with bounds $-\infty$ and ∞ which reflects lack of prior knowledge other than that the regression coefficients can take any positive or negative value. For the spatial parameters σ^2, ρ, σ_k^2 and ρ_k we adopt inverse gamma and gamma prior distributions respectively with parameters chosen to have mean equal to 1 and very large variance.

We estimate the parameters of the model using Markov chain Monte Carlo (MCMC) simulation and in particular Gibbs sampling (Gelfand and Smith, 1990). Starting with some initial values about the parameters, the algorithm iteratively updates the parameters by simulating from their full conditional distributions, that is the posterior distribution of each parameter conditional on the remaining parameters. The full conditional distributions of σ^2 and σ_k^2, $k = 1, \ldots, K$ are inverse gamma distributions and simulation from them is straightforward. The rest of the parameters do not have full conditional distributions of known forms. We simulate from the non-standard distributions by employing a random walk Metropolis algorithm (Tierney, 1994), having a Normal proposal density with mean equal to the estimate of the corresponding parameter from the previous Gibbs iteration and variance equal to a fixed number, iteratively adapted to optimize the acceptance rates. We run five chains with a burn-in of $5,000$ iterations. Convergence was assessed by inspection of ergodic averages of selected model parameters.

The analysis was implemented in Fortan 95 (Compaq Visual Fortran Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.).

3.3.3 Prediction model

Bayesian kriging (Diggle et al., 1998) is used to predict the malaria risk at locations where malaria data are not available. This approach treats the malaria risk at a new location as random and calculates its predictive posterior distribution, which provides not only a single estimate of the risk but a whole range of likely values together with their probabilities to be the true values at a specific location. This makes it possible to estimate the prediction error, a substantial advantage over the classical kriging methods. We estimated the predictive posterior distributions at new locations via simulation. Predictions were made for $28,000$ pixels, covering the whole area of south Mali. Further details are given in the Appendix of this chapter.

3.3.4 Model validation

In total we fitted 4 models (a stationary and three non-stationary). Model fit was carried out on a randomly selected subset of our data (training set) including 69 locations. The remaining dataset of 20 locations was used for validation (testing set).

The goodness-of-fit of each model was assessed using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). This quantity considers the fit of the data but penalizes models that are very complex.

The predictive ability of the models was assessed using a Bayesian "p-value" analogue calculated from the predictive posterior distribution. In particular, for each one of the test locations we calculated the area of the predictive posterior distribution which is more extreme than the observed data. The model predicts the observed data well for a specific location when the observed data is close to the median of the predictive posterior distribution and therefore the "p-value" close to 0.5. A box plot is used to summarize the "p-values" calculated from the 20 test locations under a particular model. The box plot displays the minimum, the 25th, 50th, 75th quantile as well as the maximum of the distribution of the 20 "p-values". We consider as best the model with median "p-value" closer to 0.5. The "p-value" is calculated using simulation-based inference by $\frac{1}{1000}\sum_{j=1}^{1000}\min(I(p_i^{rep(j)} > p_i^{obs}), I(p_i^{rep(j)} < p_i^{obs}))$, where p_i^{obs} is the observed prevalence at test site s_i and $\boldsymbol{p}_i^{rep} = (p_i^{rep(1)}, \ldots, p_i^{rep(1000)})$ are 1000 replicated data from the predictive distribution at test location s_i.

A χ^2-based measure was also calculated as an alternative way of comparing the predictive ability of the models. For every test location s_i, we calculated the statistic $\chi_i^2 = \frac{(Y_i^{obs} - \hat{Y}_i)^2}{\hat{Y}_i}$ where Y_i^{obs} are the observed count at test location s_i and $\hat{Y}_i$ is the median of the predictive posterior distribution at s_i. For each model, we obtained the distribution as well as the sum T_{χ^2} of the χ_i^2 values over the 20 test points. The best model was the one with the lowest median and T_{χ^2}, estimating predicted counts which are closer to the observed ones.

In addition to the above approaches, for each model we calculated 5 credible intervals (the equivalent of confidence intervals in the Bayesian framework) with probability coverage equal to 5%, 25%, 50%, 75% and 95% respectively of the posterior predictive distribution at the test locations. The model which gave better predictions was the one with the highest percentage of locations within the interval of smallest coverage.

3.4 Results

The pooled data have shown an overall malaria prevalence of 44.0%(19, 156 children). The median malaria prevalence estimated at village level was 51.3%, ranging from 5.3% to 95.5%.

The univariate non-spatial analysis showed that the following environmental indicators and their transformations were associated with malaria prevalence: NDVI, length of malaria season, rainfall, maximum temperature, minimum temperature and distance to the nearest water body. The relation with NDVI was best described by the logarithmic transformation and the relation with minimum, maximum temperature and distance to water by polynomial terms of order 2. The results of the bivariate non-spatial logistic regression are summarized in Table 3.3. All covariates significant at a 15% significance level were included in the spatial analysis.

Figure 3.2 compares the predictive ability of the stationary and 3 non-stationary (with 2, 3 and 4 tiles respectively) multiple logistic regression models using the Bayesian "p-value" approach. Each box plot summarize the distribution of the 20 "p-values" calculated from the predictive posterior distribution of the 20 test locations. The median of this distribution for the non-stationary model with two tiles is the closest to 0.5, suggesting that this is the best model. The same conclusion was drawn by comparing the models using the chi-squared measure. Figure 3.3 shows that the non-stationary models with two and three

tiles have similar medians of the distribution of χ^2-values over the 20 test locations, but the non-stationary model with two tiles had the lowest T_{χ^2} value, indicating the smallest deviations between the observations and model predictions.

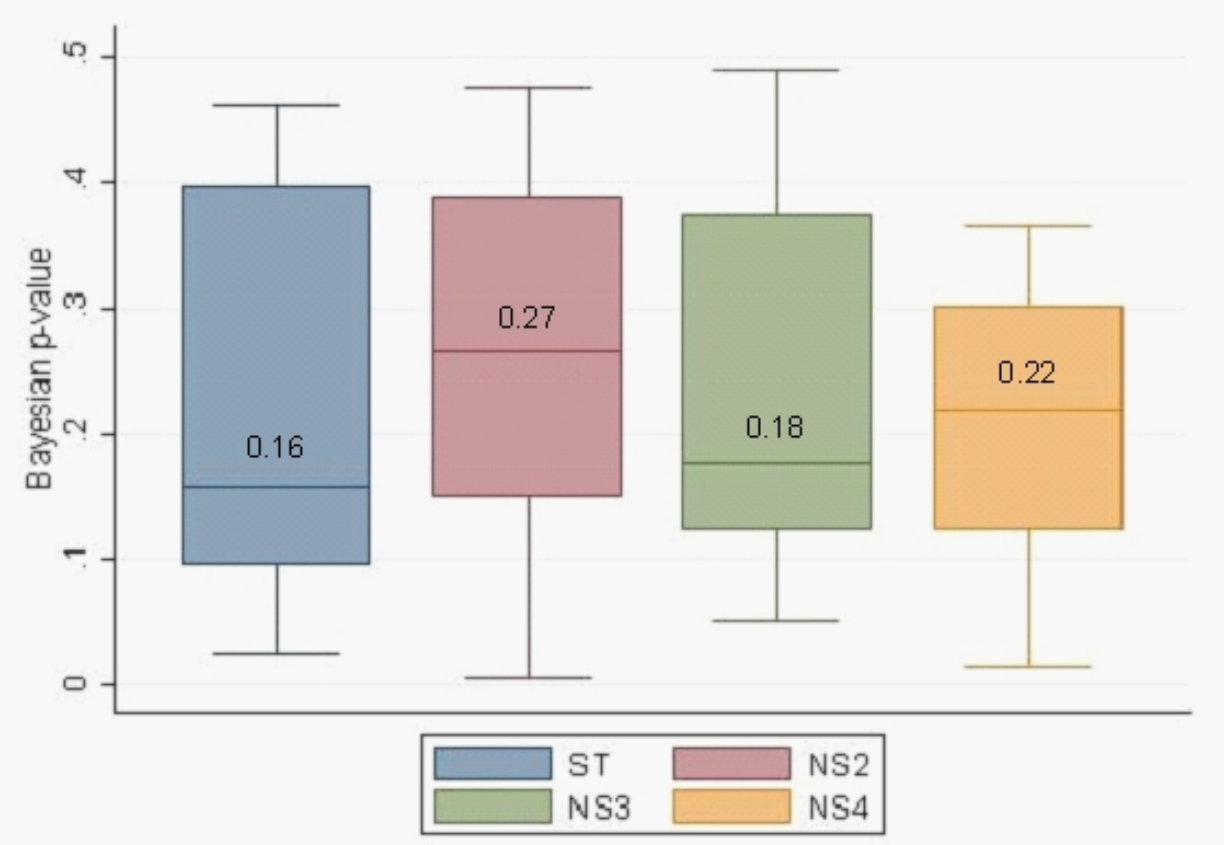

Figure 3.2: The distribution of Bayesian p-values for the stationary model (ST), and the non-stationary with 2 (NS2), 3 (NS3), and 4 (NS4) tiles.

In Table 3.2 are presented the percentages of test locations with malaria prevalence which falls in each of the 5 credible intervals of the posterior predictive distribution. We observe that the non-stationary model with two fixed tiles includes 10% of the test locations in the narrowest interval of 5% probability content. This is the highest percentage in comparison to the remaining fitted models. Also in the 95% credible interval the non-stationary model with two fixed tiles has the highest percentage of observed prevalences at test locations, namely 80% in comparison with 75% reported by the other three models.

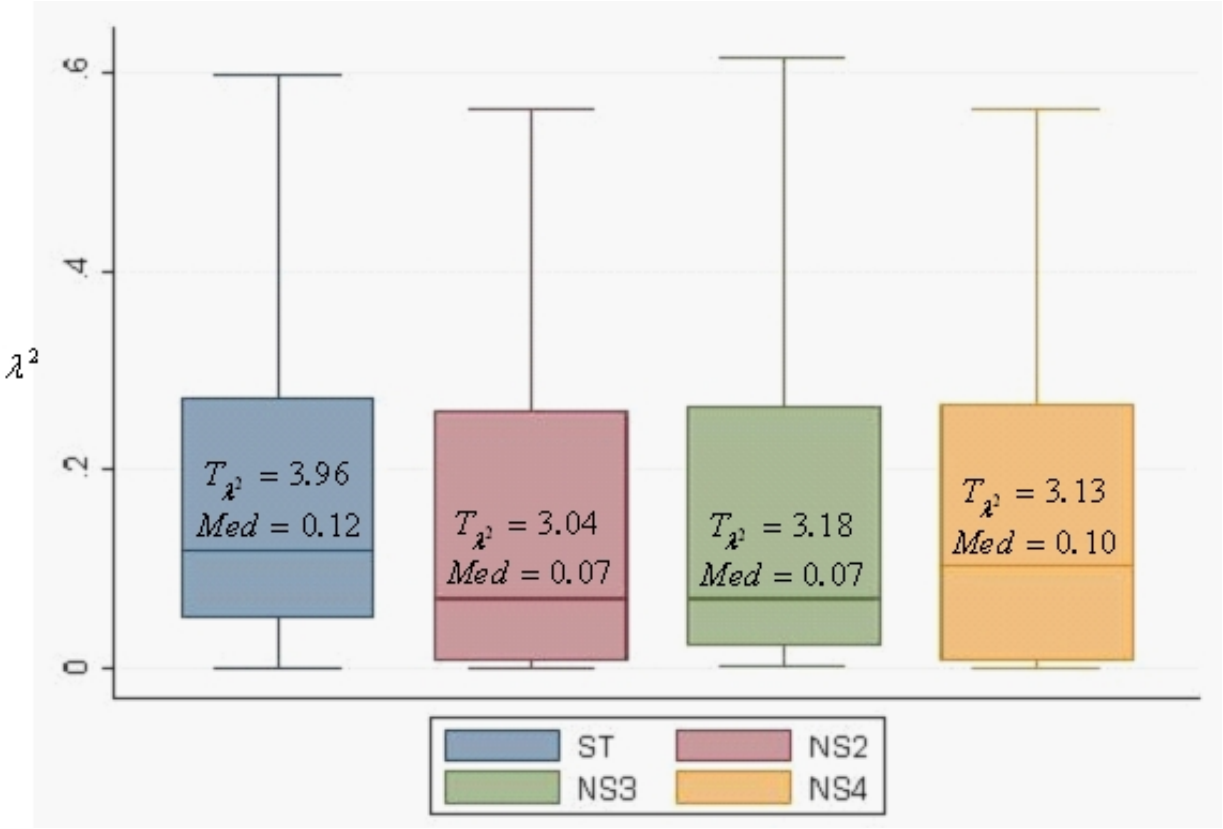

Figure 3.3: The distribution and the sum T_{χ^2} of the χ^2-values over the 20 test points.

Credible Interval	Bayesian geostatistical model			
	Stationary	NS-2 tiles	NS-3 tiles	NS-4 tiles
5%	5%	10%	0%	5%
25%	15%	25%	15%	25%
50%	30%	55%	50%	35%
75%	55%	60%	65%	55%
95%	75%	80%	75%	75%

Table 3.2: Percentage of test locations with malaria prevalence falling in the 5%, 25%, 50%, 75% and 95% credible intervals of the posterior predictive distribution.

Table 3.3 depicts the results of the stationary and the best fitting non-stationary model with two tiles. The stationary model suggested that the following environmental factors are associated with malaria risk: NDVI (in logarithmic scale), maximum, minimum temperature and distance to the nearest water body (in polynomial forms of order 2) and rainfall.

In the non-stationary model the rainfall as well as the second order polynomial of the distance to water were not any more related with the malaria risk. As we were expected, the higher the value of the NDVI (indicating the presence of green vegetation) the higher the malaria risk. A negative relation with maximum temperature showed that the lower the maximum temperature the higher the malaria risk. Also, malaria risk increases with an increase in the minimum temperature. Surprisingly, the models estimated a positive relation with the distance to water, implying that the risk increases with the distance from permanent water bodies.

The stationary model calculates a posterior median for ρ equal to 2.63 (95 % credible interval: $1.11, 6.09$) which, in our exponential setting indicates that the minimum distance for which the spatial correlation is smaller than 5 % is equal to $3/\rho = 1.14$ km (95 % credible interval: $0.49, 2.71$). The best fitting 2-tile non-stationary model confirms that spatial correlation changes as we move from the North to the South part of the country. In particular the minimum distance with negligible correlation is 0.86 km (95 % credible interval: $0.40, 2.19$) in the North and 8.90 km (95 % credible interval: $1.79, 26.88$) in the South part. It is interesting to see that although the models differ in their predictive ability (Figure 3.2 and Figure 3.3), the goodness of fit DIC measure does not favor any of the models, showing that it is not able to assess which model has the best predictions.

The smooth maps of malaria prevalence in sub-Sahara Mali obtained from the stationary and non-stationary spatial model with two tiles are shown in Figures 3.4 and 3.5. Both maps predicted high malaria prevalence in the region of Kayes (South-West Mali), with the exception of the district of Kayes and in the region of Segou (East-Center Mali). Low prevalence was predicted in the regions of Gao, Tombactou and Kidal (North Mali) and in the district of Kati (Center Mali). Differences between the stationary and non-stationary models appear in the districts of Ansongo, Gourma Rharous, Douentza and western district of Tombactou region (Goundam). Figures 3.6 and 3.7 depict the prediction error from the stationary and non-stationary models respectively. The error is higher in the North Mali where the observed data were very sparse. The prediction error obtained from the non-stationary model was lower, ranging from 0.36 to 5.7 in comparison to that obtained from the stationary one which varied from 0.70 to 8.11.

Variable	Bivariate non-spatial model		Stationary spatial model		Non-stationary (2 tiles) spatial model	
	Median	95% CI[a]	Median	95% CI[a]	Median	95% CI[a]
Intercept			0.13	(-0.25, 0.50)	0.21	(-0.20, 0.63)
Log(NDVI)	0.26	(0.24, 0.28)	0.97	(0.43, 1.49)	0.85	(0.28, 1.40)
Log(NDVI)^2	-0.11	(-0.12, -0.10)	0.20	(-0.15, 0.56)	0.13	(-0.25, 0.47)
Season Length	0.24	(0.22, 0.26)	-0.37	(-0.90, 0.15)	-0.27	(-0.85, 0.30)
Rainfall	0.23	(0.21,0.25)	-0.78	(-1.24, -0.30)	-0.60	(-1.13, 0.01)
Maximum Temperature	-0.40	(-0.42, -0.37)	-1.26	(-1.90, -0.62)	-1.02	(-1.73, -0.18)
Maximum Temperature ^{2}	-0.13	(-0.14, -0.12)	0.07	(-0.21, 0.32)	0.05	(-0.21, 0.32)
Minimum Temperature	-0.05	(-0.07, -0.03)	0.94	(0.37, 1.52)	0.90	(0.28, 1.48)
Minimum Temperature ^{2}	-0.22	(-0.23, -0.21)	-0.36	(-0.72, 0.01)	-0.30	(-0.69, 0.09)
Distance to water	0.4	(0.38, 0.42)	0.48	(0.11, 0.87)	0.42	(0.03, 0.81)
Distance to water ^{2}	0.10	(0.08, 0.12)	-0.17	(-0.33, -0.002)	-0.15	(-0.31, 0.01)
σ^{2b}			0.81	(0.58, 1.17)	0.88	(0.56, 1.45)
					0.65	(0.31, 1.44)
ρ^b			2.63	(1.11, 6.09)	0.34	(0.11, 1.68)
					3.49	(1.37, 7.51)
DIC			507.47		507.50	

[a] : Credible intervals (or posterior intervals).
[b] : In the case of non-stationary spatial model with 5 fixed tiles we get a set of spatial parameters for each tile.

Table 3.3: Posterior estimates for model parameters.

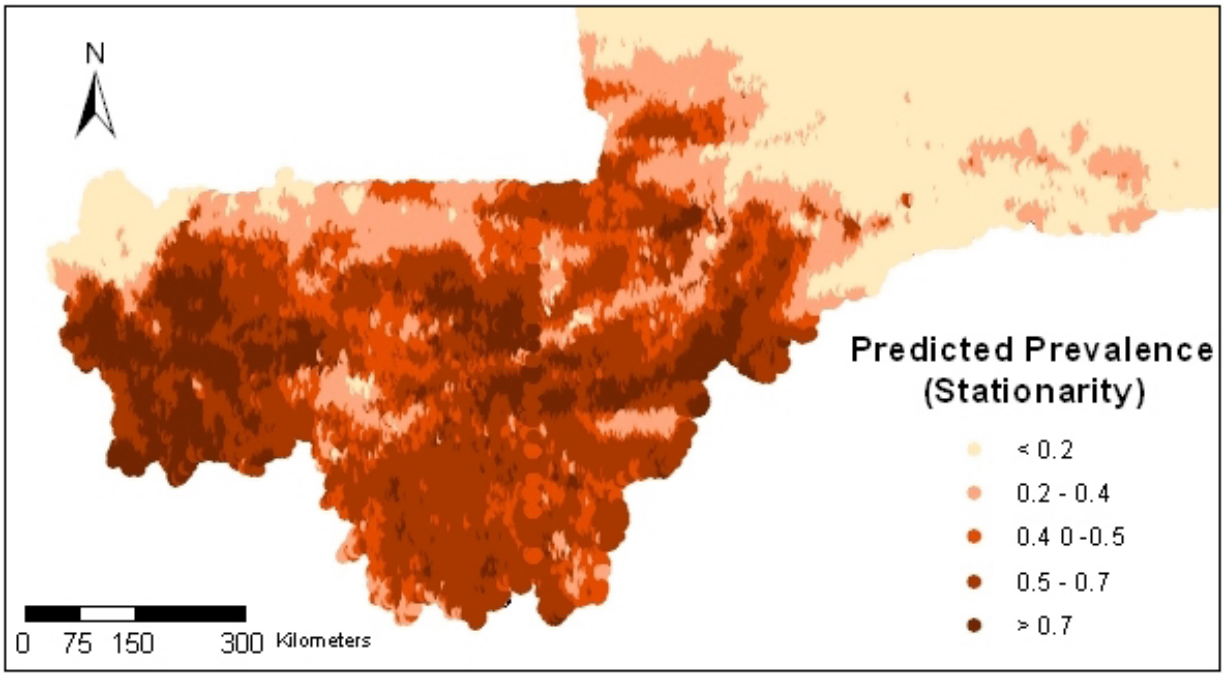

Figure 3.4: Map of predicted malaria risk for southern Mali using the stationary model.

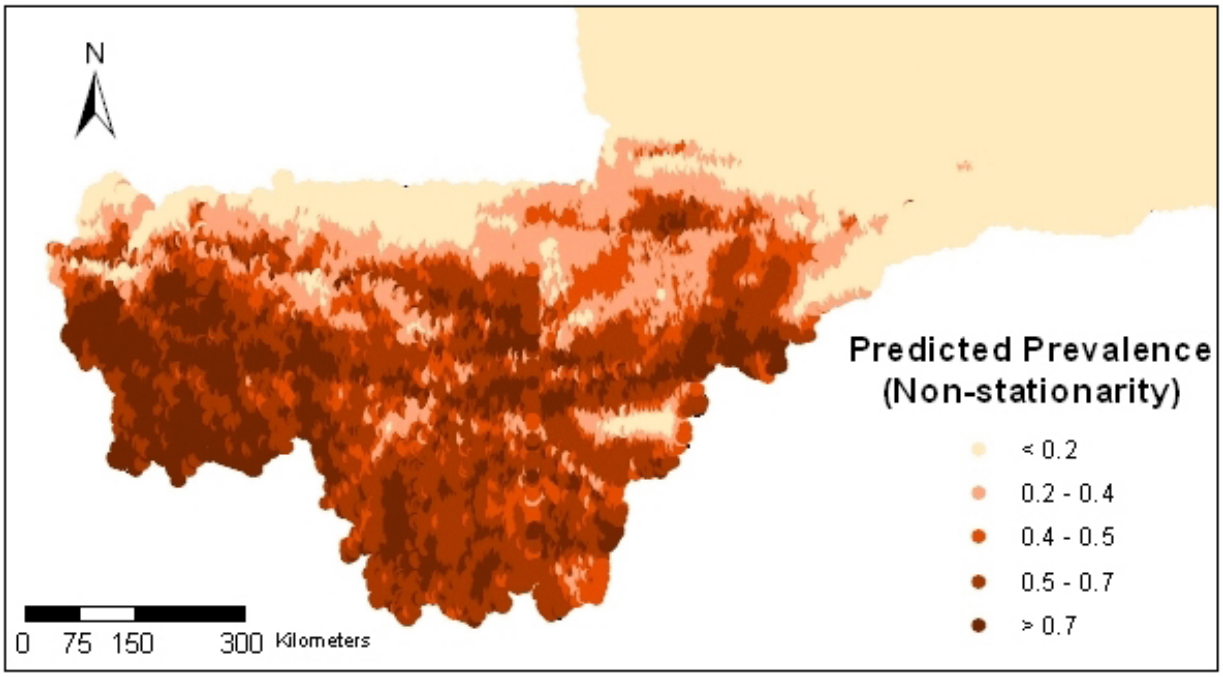

Figure 3.5: Map of predicted malaria risk for southern Mali using the non-stationary model with 2 fixed tiles.

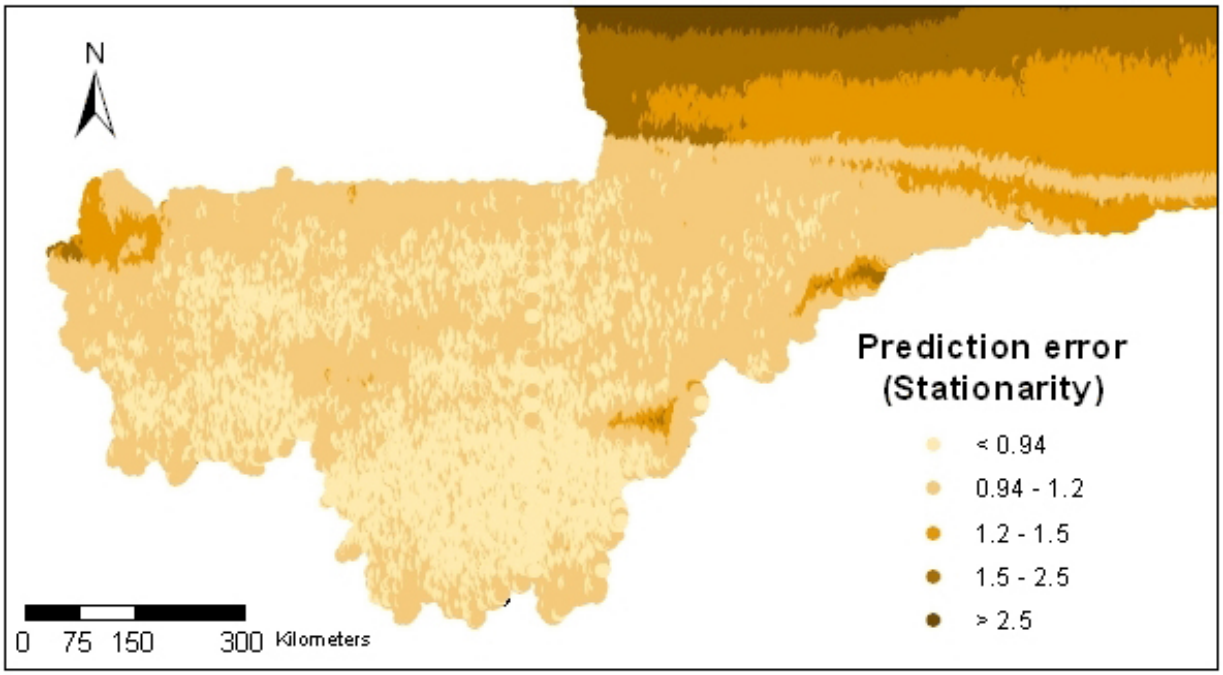

Figure 3.6: Map of prediction error for southern Mali using the stationary model.

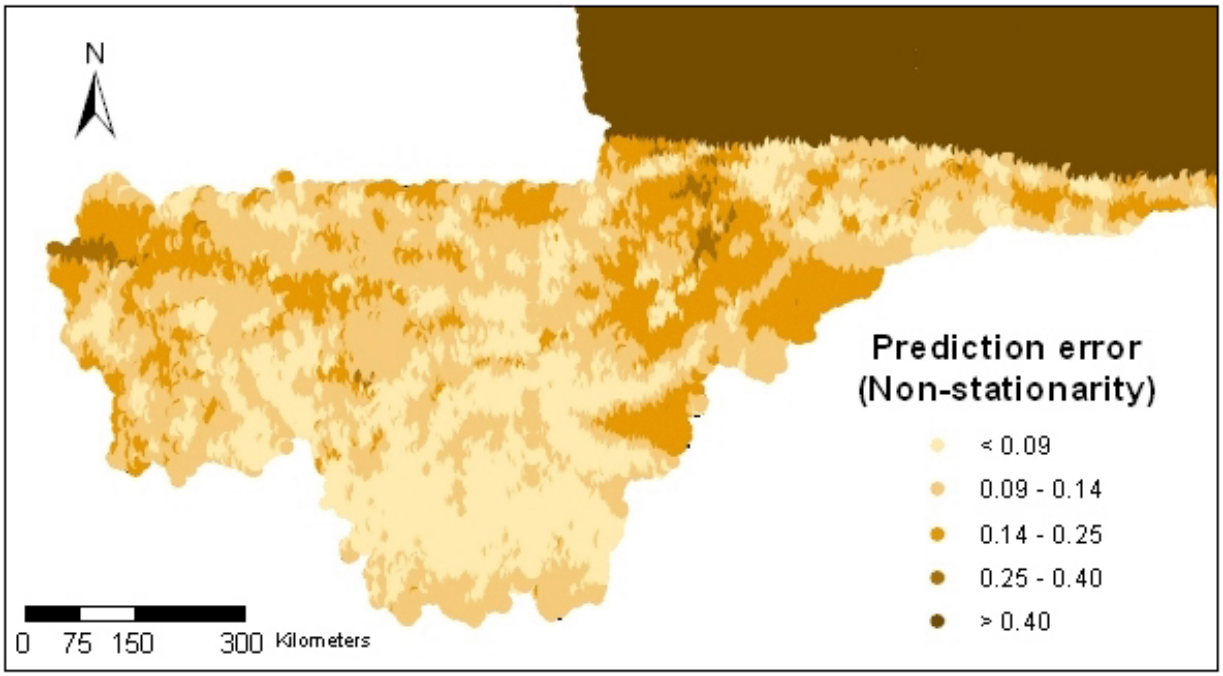

Figure 3.7: Map of prediction error for southern Mali using the non-stationary model with 2 fixed tiles.

3.5 Discussion

Accurate maps of malaria risk are important tools in malaria control as they can guide interventions and assess their effectiveness. These maps rely on predictions of risk at locations without observed prevalence data. Malaria is an environmental disease and environmental factors are good predictors of transmission, but the relation between environmental factors, mosquito abundance and malaria prevalence is not linear. This relation can be established only by means of adequate spatial statistical models which can be used for improving predictions of malaria transmission not only in space (for risk mapping) but also in time (for developing early warning systems for malaria epidemics). In this study we present Bayesian geostatistical approaches to assess the malaria-environmental relation for the purpose of malaria risk mapping.

The Bayesian stationary and non-stationary models we presented for analyzing the malaria survey data in Mali showed that the statistical modeling approach plays an important role in inference. It influences not only the estimation of parameters related with the spatial structure of the data but also the significance of the malaria risk predictors, the resulting malaria risk maps and the associated predicted errors. Model validation should routinely accompany any model fitting exercise. For the purpose of validation, we recommend to carry out the model fitting on the 80% of the data locations and compare the predictive ability of the models on the remaining locations. For the purpose of mapping, we suggest, once the best model is selected, to apply it to the whole dataset so that the final maps are based on as much data as possible.

Non-stationarity is an important feature of malaria data which is often ignored. Gemperli (2003) developed a non-stationary model for analyzing malaria risk data which divides the study region in random tiles, assuming a separate correlation structure within region but independence between tiles. The number and configuration of tiles are random parameters estimated by the data. The non-stationary modeling approach we adopt here relies on a partition of the study region into fixed tiles. We assume a separate correlation structure within tile as well as correlation between tiles. This modeling approach is more appropriate when modeling malaria data over large areas covering different ecological zones which define the fixed partition. An extension of the model will allow different covariate effects in each zone. We are currently working on such an approach and implementing it in analyzing MARA malaria risk data from West and Central Africa. A further extension of the methodology presented here is to assume random rather than fixed partition of region

in tiles. This methodology could be applied in mapping malaria data over large areas with no clear way of finding a fixed partition (i.e no clearly defined ecological zones).

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. Both the stationary and especially the non-stationary models have a large number of parameters. Bayesian computation implemented via MCMC enables simultaneously estimation of all model parameters together with their standard errors. In addition, Bayesian kriging allows model-based predictions (together with the prediction error) taking into account the non-stationary feature of the data. This is not possible in a maximum likelihood based framework.

The significant positive association between our data and the distance to water was unexpected. Possible explanation could be because the majority of the main cities (most populated areas) in Mali are located along the river Niger. During the dry season the receding of the river create numerous water pools which serve as vector breeding habitats. The time lag between the rainfall and vector abundance and between vector abundance and the occurrence of the disease may have also played an important role.

Earlier analyzes of the MARA data in Mali (Kleinschmidt et al., 2000; Gemperli, 2003) differ in the way the spatial structure is incorporated in the model as well as in the way the covariate effects were modeled. Kleinschmidt et al. (2000) determined the relation between malaria prevalence and environmental predictors by fitting an ordinary logistic regression by maximum likelihood method without taking into account spatial correlation. The prediction map was improved by kriging the residuals and adding them to the map on a logit scale. The main weaknesses of this analysis are firstly that estimation of environmental effects did not take into account the spatial correlation and thus the significance of the covariates may have been underestimated; and secondly the kriging assumes normality, which usually does not hold for the residuals of the logistic regression. Gemperli (2003) re-analyzed the data using the Bayesian non-stationary model with random tiles mentioned above. Both previous analyzes found a negative relation between malaria risk and distance to water, while Gemperli (2003) suggested also a positive relation with rainfall. Neither analyzes assessed non-linear covariate effects. The different analyzes reported different covariate effects and produced different maps of prevalence from essentially the same database. Neither performed model validation on test data.

The predicted prevalence map from the non-stationary model with 2 tiles is in a better agreement with the eco-geographical descriptive epidemiology of malaria in Mali (Doumbo

et al., 1989) than the maps obtained from the other models. The two maps of predicted malaria prevalence obtained from the stationary and the non-stationary model with two tiles were shown to different malaria epidemiologists in Mali. They all agreed that that the non-stationary model predicts better the epidemiological situation of malaria in Mali. However, they found that the prevalence in the western part of the country (Kayes region) is over-estimated in comparison with the southern region of Mali (Sikasso). Also previous mapping approaches (Kleinscmidt et al., 2000; Gemperli, 2003) suggested high malaria prevalence in the western region of Mali. The relatively high predictive standard deviation observed in the North-Western (region of Kayes) and the desert fringes (Tombouctou, Gao and Kidal regions) of the country is probably because of the very few number of data points in these areas rather than the statistical approach. Only one survey has been carried out in the northern regions since 1988.

Further analyzes which include recent data, particularly in areas where very few number of data points were observed such as in the north part are needed because environmental changes in the last decades are likely to have influenced malaria transmission dynamics in Mali.

Acknowledgments

The authors would like to acknowledge the MARA / ARMA collaboration for making the malaria prevalence data available. We are thankful to Dr. Sekou F. Traour, Dr. Seydou Doumbia, Dr. Madama Bouar and Dr. Mahamoudou Tour for their useful comments on the predicted risk maps. This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

3.6 Appendix

Once the spatial parameters are estimated and the environmental covariates $\boldsymbol{X}_0$ at unsampled locations are known, we can predict the malaria risk at new sites $\boldsymbol{s}_0 = (s_{01}, s_{02}, \ldots, s_{0l})^T$ from the predictive distribution

$$P(\boldsymbol{Y}_0|\boldsymbol{Y}, \boldsymbol{N}) = \int P(\boldsymbol{Y}_0|\boldsymbol{\beta}, \boldsymbol{\phi}_0) P(\boldsymbol{\phi}_0|\boldsymbol{\phi}, \sigma^2, \rho) P(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \rho|\boldsymbol{Y}, \boldsymbol{N})\, d\boldsymbol{\beta}\, d\boldsymbol{\phi}_0\, d\boldsymbol{\phi}\, d\sigma^2\, d\rho,$$

where $\boldsymbol{Y}_0 = (Y_{01}, Y_{02}, \ldots, Y_{0l})^T$ are the predicted number of cases at locations $\boldsymbol{s}_0$, $P(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \rho|\boldsymbol{Y}, \boldsymbol{N})$ is the posterior distribution and $\boldsymbol{\phi}_0$ is the vector of random effects at new site $\boldsymbol{s}_0$. The distribution of $\boldsymbol{\phi}_0$ at unsampled locations given $\boldsymbol{\phi}$ at observed locations is normal

$$P(\boldsymbol{\phi}_0|\boldsymbol{\phi}, \sigma^2, \rho) = N(\Sigma_{01}\Sigma_{11}^{-1}\boldsymbol{\phi}, \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{01}^T)$$

with $\Sigma_{11} = E(\boldsymbol{\phi}\boldsymbol{\phi}^T)$ the covariance matrix built by including only the sampled locations $s_1, s_2, \ldots, s_n$, $\Sigma_{00} = E(\boldsymbol{\phi}_0\boldsymbol{\phi}_0^T)$ the covariance matrix formed by taking only the new locations $s_{01}, s_{02}, \ldots, s_{0l}$ and $\Sigma_{01} = E(\boldsymbol{\phi}_0\boldsymbol{\phi}^T)$ describing covariances between unsampled and sampled locations. For the non-stationary models, $\boldsymbol{\phi}_0 = \sum_{k=1}^{K} a_{0k}\omega_{k0}$, where a_{0k} are decreasing functions of the distance between new locations $\boldsymbol{s}_0$ and the centroid of the subregion k.

Conditional on ϕ_{0i} and $\boldsymbol{\beta}$, Y_{0i} are independent Bernoulli variates $Y_{0i} \sim Ber(p_{0i})$ with malaria prevalence at unsampled site s_{0i} given by $logit(p_{0i}) = X_{0i}^t\boldsymbol{\beta} + \phi_{0i}$. For the test locations the predicted number of cases Y_{ti} arise from a Binomial distribution $Y_{ti} \sim Bin(N_{ti}, p_{ti})$, where N_{ti} is the number of tested children and p_{ti} is the predicted prevalence at test site s_{ti}. The predictive distribution is numerically approximated by the average

$$\tfrac{1}{r}\sum_{q=1}^{r}\left[\prod_{i=1}^{l} P(Y_{0i}^{(q)}|\boldsymbol{\beta}^{(q)}, \phi_{0i}^{(q)}\right] P(\boldsymbol{\phi}_0^{(q)}|\boldsymbol{\phi}^{(q)}, \sigma^{2(q)}, \rho^{(q)}),$$

where $(\boldsymbol{\beta}^{(q)}, \boldsymbol{\phi}^{(q)}, \sigma^{2(q)}, \rho^{(q)})$ are samples drawn from the posterior $P(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \rho|\boldsymbol{Y}, \boldsymbol{N})$.

Chapter 4

Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model

Gosoniu L.[1] ,Vounatsou P.[1], Sogoba N.[2], Maire N.[1], Smith T.[1]

[1] Swiss Tropical Institute, Basel, Switzerland
[2] Malaria Research and Training Center, Universite du Mali, Bamako, Mali

This paper has been accepted for publication in *Computational Statistics and Data Analysis*, Special Issue on SPATIAL STATISTICS

Summary

Malaria transmission is highly influenced by environmental and climatic conditions but their effects are often not linear. The climate-malaria relation is unlikely to be the same over large areas covered by different agro-ecological zones. Similarly, spatial correlation in malaria transmission arisen mainly due to spatially structured covariates (environmental and human made factors), could vary across the agro-ecological zones, introducing non-stationarity. Malaria prevalence data from West Africa extracted from the "Mapping Malaria Risk in Africa" database were analyzed to produce regional parasitaemia risk maps. A non-stationary geostatistical model was developed assuming that the underlying spatial process is a mixture of separate stationary processes within each zone. Non-linearity in the environmental effects was modeled by separate P-splines in each agro-ecological zone. The model allows smoothing at the borders between the zones. The P-splines approach has better predictive ability than categorizing the covariates as an alternative of modeling non-linearity. Model fit and prediction was handled within a Bayesian framework, using Markov chain Monte Carlo (MCMC) simulations.

Keywords: Bayesian inference; geostatistics; malaria risk; Markov chain Monte Carlo; non-stationarity; P-splines; kriging.

4.1 Introduction

Plasmodium falciparum malaria is a major cause of mortality and morbidity in sub-Sahara Africa. Malaria is a vector-borne disease and is transmitted from human to human by female mosquitoes of the genus *Anopheles*. It is an environmental disease since its transmission depends on the distribution and abundance of the mosquitoes, which are sensitive to factors like temperature, rainfall and humidity. Climatic and environmental factors play an important role in changes of the malaria distribution and endemicity, influencing the survival and the development rate of both the parasite and the mosquito vector. The relation between climatic/environmental factors and malaria transmission allows the prediction of malaria risk at locations without observed data and therefore estimation of the geographical distribution of the disease. Accurate maps of malaria transmission and malaria risk are needed to estimate the burden of disease, to improve control and intervention strategies and to optimize the use of limited resources in high-risk areas.

Reliable maps of malaria distribution are based on the availability of the disease data and on appropriate methods for analyzes. The "Mapping Malaria Risk in Africa" (MARA/ARMA, 1998) represents the most comprehensive database on malaria data in Africa and contains malaria prevalence data collected over $10,000$ surveys from all available sources across the whole continent. Using the MARA database a number of malaria risk maps have been produced at both country and regional level. Initially, maps were developed using spatially interpolated weather station data that were used to define climatic suitability for malaria transmission (Craig et al., 1999). The mapping work continued with the application of spatial statistical methods used to quantify the relations between environmental factors and malaria risk and based on these relations to produce model-based predictions (Kleinschmidt et al., 2000; Kleinschmidt et al., 2001; Gemperli et al., 2005; Gemperli et al., 2006).

Different assumptions and estimation approaches may result in different malaria risk maps. Spatial models introduce at each data location an additional parameter on which the spatial correlation is incorporated. Fitting these models is challenging because of the large number of parameters. The first modeling efforts were based on maximum likelihood approaches and could not estimate the environmental effects and the spatial correlation simultaneously (Kleinschmidt et al., 2000). The Bayesian geostatistical models introduced by Diggle et al. (1998) avoid the computational problems encountered in maximum likelihood-based fitting. Most geostatistical modeling of malaria has been based on the assumption of stationarity (Gemperli et al., 2005; Gemperli et al., 2006), which implies that the covariance

between any two points depends only on the distance between them. This assumption is not justifiable when malariological indices are modeled, especially over large areas, since local features related to human activities, land use, environment and vector ecology may affect geographical correlation differently at various locations (Gemperli, 2003). Gosoniu et al. (2006) analyzed malaria prevalence data in Mali under both assumptions of stationarity and non-stationarity and performed model comparison which revealed that the non-stationary model captured better the spatial correlation present in malaria data. Previous mapping efforts over large areas assumed that the relation climate-malaria remains the same over the entire area (Gemperli et al., 2006). However, the effect of environmental/climatic factors on malaria risk depends on local ecological factors.

Modeling non-stationary spatial processes has recently received much attention (Banerjee et al., 2004) due to computational advances in geostatistical model fit. Kernel convolution was introduced by Higdon et al. (1998) who convolve spatially evolving kernels and Fuentes et al. (2002) who used a Gaussian process as a function of locations to obtain a non-stationary covariance structure. Kim et al. (2005) used piecewise Gaussian processes to model a non-stationary process by partitioning the study region in random tiles, assuming an independent Gaussian stationary process within each tile and independence between tiles. This approach has the computational advantage of inverting several covariance matrices of smaller dimensions instead of the global large covariance matrix since the covariance matrix is reduced to a block-diagonal form. Gemperli (2003) extended the work of Kim et al. (2005) and model non-Gaussian malaria prevalence data using random tessellations. This was the first time the non-stationarity issue was addressed in the mapping malaria risk field. Although the assumption of independence between tiles facilitates the matrix inversion, it ignores the spatial correlation between neighboring points located at the edges and which belong to different tiles. To address the between-tile independence problem Gosoniu et al. (2006) defined the spatial process as a mixture of tile-specific stationary spatial processes and demonstrated this methodology for a fixed space partitioning, modeling malaria risk data in Mali. This approach is more appropriate when modeling malaria data over large areas with fixed partitions defined by different ecological zones.

Another statistical issue which occurs in model-based malaria prevalence mapping is the assumption of linearity between environmental factors and malaria risk which may not always hold. The most popular method adopted to model the non-linearity is categorizing the covariates. The results are easy to understand, although it is unreasonable to

conclude that the risk is increasing (or decreasing) abruptly as a category cut point is crossed. There is considerable literature on nonparametric regression modeling, which allows the shape of the relationship between outcome and covariates to be determined by the data, whereas in the parametric framework the model determines this relationship. Non-parametric modeling alternatives include kernel smoothing (Silverman, 1985; Hardle, 1990), local polynomial regression (Cleveland, 1979), fractional polynomials (Royston and Altman, 1994) and spline smoothing. Spline approaches comprise regression splines (Eubank, 1988), B-splines (de Boor, 1978) and penalized splines (P-splines). The latter approach was first introduced by O'Sullivan (1986), but was popularized by Eilers and Marx (1996). Bayesian approaches to P-splines allow simultaneous estimation of smooth functions and smoothing parameters.

In this paper we develop non-stationary models to produce a smooth malaria risk map of West Africa, modeling a non-linear relation between climate factors and malaria risk separately in each agro-ecological zone. Non-linear environmental effects were modeled via categorizing the covariates and by P-splines. Modeling validation is applied to identify the best approach to capture non-linearity. To model the non-stationary spatial process we extended the work of Gosoniu et al. (2006) considering as fixed tiles the four agro-ecological zones in West Africa. The computing time in the implementation of the geostatistical modeling was reduced using a volunteer computer platform (www.malariacontrol.net) which is based on Berkeley Open Infrastructure for Network Computing (http://boinc.berkeley.edu/). The article is structured as follows. In Section 4.2 we describe the malaria data together with the environmental and climatic data used in the analysis. The description of both the nonparametric approach and Bayesian geostatistical non-stationary model as well as the model validation approaches are provided in Section 4.3. The results are presented in Section 4.4. Concluding remarks and suggestions for future work are given in Section 4.5.

4.2 Data

Malaria prevalence data were extracted from the MARA/ARMA database (MARA/ARMA, 1998). Only included surveys conducted after 1950 and on children between 1 and 10 years old were included in the analysis. The final data set we analyzed was collected at 265 distinct locations over 374 surveys (Figure 4.1), including $56,672$ observed individuals.

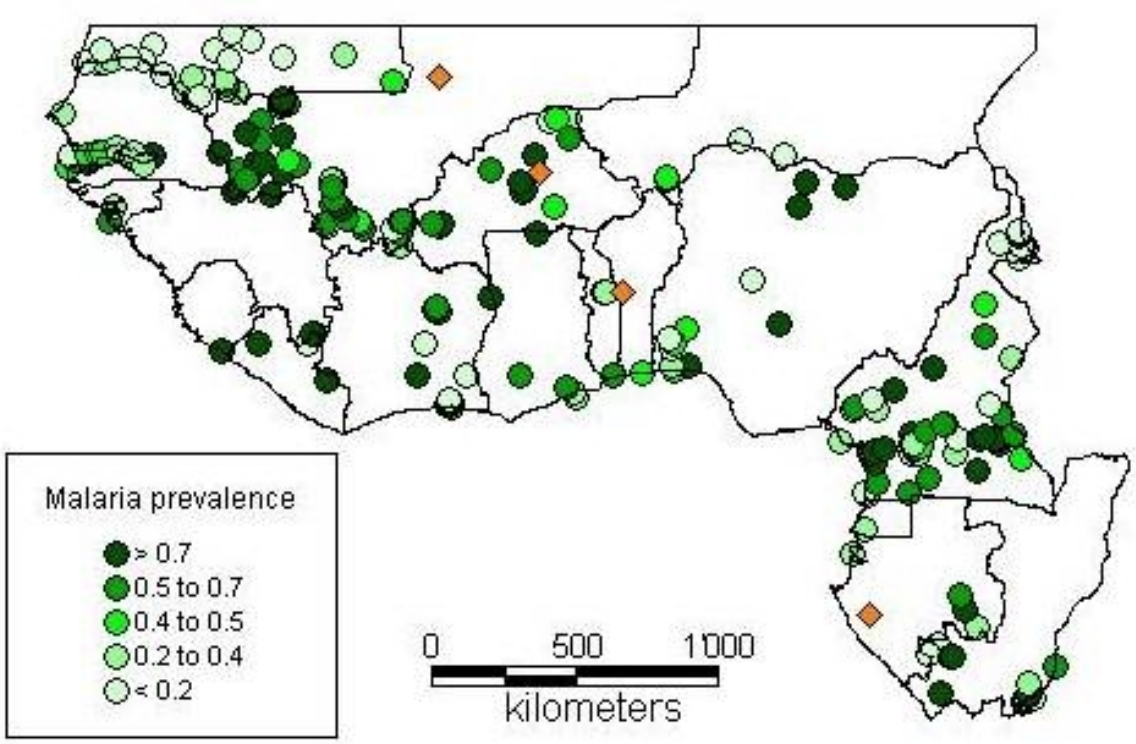

Figure 4.1: Sampling locations of the MARA surveys included in the analyzes in West Africa with dot shading indicating the observed malaria prevalence. The diamonds indicate the centroids of the four fixed tiles (AEZ) used to account for non-stationarity.

The environmental and climatic predictors used in this analysis are similar to the ones used by Gemperli et al. (2006) for mapping malaria transmission in West Africa: minimum and maximum temperature, amount of rainfall, the normalized difference vegetation index (NDVI), the soil water storage (SWS) index, the length of the malaria transmission season, the distance to the nearest water body and the land use.

Data on temperature and rainfall were extracted from the "Topographic and climate database for Africa" (Hutchinson et al., 1996). The data were obtained from weather stations between 1920 and 1980 and were extrapolated by fitting thin plate spline functions of latitude, longitude and elevation to values of monthly mean daily minimum temperature, daily maximum temperature and rainfall in the whole continent. Monthly averages over the years with available data were calculated.

NDVI is used to indicate green vegetation cover and is calculated from the red and near infra-red reflectance observed by the AVHRR (Advanced Very High Resolution Radiometer) sensor on NOAA meteorological satellites (Agbu and James 1994) at a spatial resolution of 8km^2. Index values can range from -1 to 1 with high values $(0.3 - 0.6)$ indicating

high levels of healthy (green) vegetation cover, whereas values near zero and negative values are associated with non-vegetated features such as barren surfaces and water. Monthly NDVI values were obtained by averaging the maximum monthly index values over the period $1985 - 1995$.

SWS estimates the amount of water that is stored in the soil within the plant's root zone. Monthly estimates of the SWS index were obtained at 5km^2 resolution using the procedure given by Droogers et al. (2001).

The length of malaria season was defined using the seasonality model of Gemperli et al. (2006). They defined a region and month as suitable for malaria transmission when rainfall, temperature and NDVI values are higher than pre-specified cut-offs .

Using Idrisi software (Clark Labs, Clark University) calculation for the distance to the nearest water source was based on permanent rivers and lakes extracted from "African Data Sampler" (WRI, 1995).

Land use classification was based on land use/ land cover database maintained by the United States Geological Survey and the NASA's Distributed Active Archive Center. We used the 24-category classification scheme described by Anderson et al. (1979) and regrouped them into six broad categories.

The region of West Africa was divided in four agro-ecological zones on the basis of the period when the water is available for vegetative production on well-drained soils, according to the procedure described in FAO (1978). The four agro-ecological zones are defined as follows: Sahel (< 90 days), Sudan Savanna ($90 - 165$ days), Guinea Savanna ($165 - 270$ days) and Equatorial Forest (> 270 days).

4.3 Spatial modeling

The malaria prevalence data were treated as binomial data and modeled via the logistic regression. Let Y_i be the number of observed malaria cases out of N_i children tested at location i, $i = 1, \ldots, n$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ be the vector of p associated environmental predictors observed at location i. We assume that Y_i are binomially distributed, that is $Y_i \sim Bin(N_i, p_i)$ with parameter p_i measuring malaria prevalence at location i.

4.3.1 Non-linearity of covariates effect

The relation between environmental factors and malaria risk may not be linear. It is difficult to foresee the form of the relationship due to considerable residual variability in the data. Parametric models are not appropriate, even if one assumes a non-linear relation captured by a low-order polynomial, since the true relationship may not have a simple polynomial form. In this analysis we focus on nonparametric regression models using P-splines because they can be easily implemented in the Bayesian framework (Crainiceanu et al., 2005).

The logit of p_i is modeled non-parametrically $logit(p_i) = \sum_{j=1}^{p} f_j(X_{ij}) + \phi_i$, where ϕ_i's are error terms, modeling between-location variation and $f(\cdot)$ is an unknown but a smooth function of the environmental predictors. That is:

$$f_j(X_{ij}) = \sum_{k=1}^{K} u_{kj}|X_{ij} - s_{kj}|^3,$$

where $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{Kj})^T$ is the vector of regression coefficients, $s_{1j} < s_{2j} < \ldots < s_{Kj}$ are fixed knots and $|X_{ij} - s_{kj}|^3$ is a truncated 3-rd order polynomial spline basis, all corresponding to covariate X_j. In a simple regression spline approach the unknown regression coefficients are estimated using standard maximum likelihood algorithms for linear models. The problem in using this equation is the choice of the number and position of the knots. Choosing a small number of knots would lead to a smooth function that is not flexible enough to capture the variability in the data, while a large number of knots may result in over-parametrization and over-fitting. To overcome these difficulties, one can use a penalty on the spline coefficients $\boldsymbol{u}$ to achieve a smooth fit, that is we can penalize $\boldsymbol{u}_j$ by the quadratic form $\lambda \boldsymbol{u}_j^T D \boldsymbol{u}_j$, where λ is the smoothing parameter and D is a penalty matrix chosen according to the data. The penalty controls the degree of smoothness. The P-spline approach states that minimization of:

$$\sum_{i=1}^{n} \{logit(p_i) - \sum_{j=1}^{p} f_j(x_{ij})\}^2 + \tfrac{1}{\lambda} \boldsymbol{u}_j^T D \boldsymbol{u}_j$$

leads to a smooth vector of coefficients $\boldsymbol{u}$. Eilers and Marx (1996) defined the penalty function to be based on differences between neighboring spline coefficients.

The final model can be written as:

$$logit(\mathbf{p}) = Z\mathbf{b} + \boldsymbol{\phi}, \text{ with } Cov \begin{pmatrix} \mathbf{b} \\ \boldsymbol{\phi} \end{pmatrix} = \begin{pmatrix} \sigma_b^2 \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \sigma_\phi^2 \mathbf{I}_n \end{pmatrix},$$

where $\mathbf{b} = Q^{1/2}\boldsymbol{u}$ and $Z = Z_K Q^{-1/2}$ with Z_K the matrix having the ith row $Z_{Ki} = \{|x_i - s_1|^3, \ldots |x_i - s_K|^3\}$ and Q the matrix having the element $(Q)_{ij} = |s_i - s_j|^3$ (Crainiceanu et al., 2005).

In this paper we consider the knots to be based on sample quantiles of the covariates, but one could take the knots to be equally spaced.

Since the environmental factors which influence malaria transmission are different in each agro-ecological zone, a separate nonparametric model was derived for each of the four zones.

The non-linear relation between malaria risk and environmental factors was also modeled categorizing the risk factors. Scatter plots were used to choose the cutoffs points of the categories.

4.3.2 Spatial correlation and non-stationarity

The parasitaemia risk at neighboring locations is influenced by similar environmental factors and therefore it is expected that the malaria risk varies similarly at locations within the neighborhood. We model spatial heterogeneity via a geostatistical model by introducing a random effect ϕ_i at each location i, that is $logit(p_i) = Z_i b_i + \phi_i$. The spatial correlation is modeled on these parameters as a function of the distance between locations. The design matrix Z is the one obtained in the previous section.

Most geostatistical models assume stationarity of the spatial process, that is the spatial correlation structure does not vary across the study area. This assumption is questionable, especially when modeling malaria indices over large geographical areas. Differences in agro-ecological zones, health systems and socio-economic indicators may change geographical correlation differently at various locations.

Following the approach described in Gosoniu et al. (2006) the space was partitioned in fixed tiles corresponding to the agro-ecological zones in West Africa. Thus the study area is divided into 4 subregions and a stationary Gaussian process $\boldsymbol{\omega}_k$ is assumed in each subregion $k = 1, \ldots, 4$, that is $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kn})^T \sim N(\mathbf{0}, \Sigma_k)$, with $(\Sigma_k)_{ij} = \sigma_k^2 corr(d_{ij}; \rho_k)$, where $corr$ is a parametric correlation function of the distance d_{ij} between locations i and j. In most epidemiological applications, the exponential correlation function $corr(d_{ij}; \rho_k) = exp(-d_{ij}/\rho_k)$ is used, where $\rho_k > 0$ measures the rate of decrease of correlation with

distance and it is known as the range parameter of the spatial process. Parameter σ_k^2 measures within location variation and it is known as the sill of spatial process. Note that the spatial parameters σ_k^2 and ρ_k are specific for each agro-ecological zone k.

The spatial random effect ϕ_i at location i were modeled as a weighted sum of the tile-specific stationary processes, that is $\phi_i = \sum_{k=1}^{K} a_{ik}\omega_{ki}$, where a_{ik} are decreasing functions of the distance between location i and the centroid of the subregion k. Then $\boldsymbol{\phi}$ is a non-stationary spatial process $\boldsymbol{\phi} \sim N(\mathbf{0}, \sum_{k=1}^{K} A_k \Sigma_k A_k)$, where A_k is a diagonal matrix with $(A_k)_{ii} = a_{ik}$.

The Bayesian formulation of the model is completed by specifying prior distributions for the model parameters $\boldsymbol{\beta}$, σ_k^2 and ρ_k (see Section 3.5).

The model described above have a large number of parameters. Bayesian computation implemented via Markov chain Monte Carlo (MCMC) simulation methods enables simultaneously estimation of all model parameters together with their standard errors. More details are given in Section 4.3.5.

4.3.3 Prediction

Bayesian kriging (Diggle et al., 1998) was employed to predict malaria risk at unsampled locations. This approach has the advantage over the classical kriging that calculates the predictive distribution of parasitaemia risk at new location and therefore, makes it possible to estimate of prediction error.

Estimates of the malaria risk at any unsampled location $\boldsymbol{s}_0 = (s_{01}, s_{02}, \ldots, s_{0m})^T$ were obtained by the predictive distribution

$$P(\boldsymbol{Y}_0|\boldsymbol{Y}, \boldsymbol{N}) = \int P(\boldsymbol{Y}_0|\boldsymbol{\beta}, \boldsymbol{\phi}_0) P(\boldsymbol{\omega}_{k0}|\boldsymbol{\omega}_k, \sigma_k^2, \rho_k) P(\boldsymbol{\beta}, \boldsymbol{\omega}_k, \sigma_k^2, \rho_k|Y, N)\, d\boldsymbol{\beta}\, d\boldsymbol{\phi}_0\, d\boldsymbol{\omega}_k, d\sigma_k^2\, d\rho_k,$$

where $\boldsymbol{Y}_0 = (Y_{01}, Y_{02}, \ldots, Y_{0m})^T$ are the predicted number of children found with malaria parasites in a blood sample at location $\boldsymbol{s}_0$, $P(\boldsymbol{\beta}, \boldsymbol{\omega}_k, \sigma_k^2, \rho_k|\boldsymbol{Y}, \boldsymbol{N})$ is the posterior distribution and $\boldsymbol{\phi}_0$ is the vector of random effects at new sites $\boldsymbol{s}_0$. Following the non-stationary model, $\boldsymbol{\phi}_0 = \sum_{k=1}^{4} a_{0k}\omega_{k0}$, where a_{0k} are decreasing functions of the distance between new locations $\boldsymbol{s}_0$ and the centroid of the subregion k.

The distribution of $\boldsymbol{\omega}_{k0}$ at unsampled locations given $\boldsymbol{\omega}_k$ at observed locations is normal

$$P(\boldsymbol{\omega}_{k0}|\boldsymbol{\omega}_k, \sigma_k^2, \rho_k) = N((\Sigma_{01})_k(\Sigma_{11})_k^{-1}\boldsymbol{\omega}_k, (\Sigma_{00})_k - (\Sigma_{01})_k(\Sigma_{11})_k^{-1}(\Sigma_{01})_k^T),$$

where $(\Sigma_{11})_k$ is the covariance matrix created using the sampled locations $s_1, s_2, \ldots, s_n$, $(\Sigma_{00})_k$ is the covariance matrix built by taking only the unsampled locations $s_{01}, s_{02}, \ldots, s_{0m}$ and $(\Sigma_{01})_k$ depicts the covariance between sampled and unsampled locations.

The predicted malaria prevalence at new location s_{0i} is given by $logit(p_{0i}) = \boldsymbol{X}_{0i}^T\boldsymbol{\beta} + \phi_{0i}$, where $\boldsymbol{X}_{0i}$ are the environmental covariates corresponding to the unsampled location s_{0i}.

4.3.4 Model validation

We considered two non-stationary Bayesian geostatistical models which address the non-linearity of the covariates effect by using P-splines and categorized risk factors respectively. The goodness-of-fit of each model was evaluated using the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002). To assess the best way of modeling non-linearity, model fit was carried out on a randomly selected subset of 239 locations (training set). The remaining 26 locations, comprising a simple random sample, were used for validation as testing points (holdout set).

Following Gosoniu et al. (2006) the predictive ability of the two models was assessed by using 1) a Bayesian "p-value" analogue, 2) the probability coverage of the shortest credible interval and 3) the Kullback-Leibler difference between observed and predicted prevalences. For each test location we calculate the area of the predictive posterior distribution which is more extreme than the observed data and we assert that the model which predicts better is the one with the "p-value" closer to 0.5.

Another approach to assess the accuracy of the prediction for the two models is to calculate different coverages credible intervals of the posterior predictive distribution and to compare the percentages of test locations with observed malaria prevalence falling in these intervals. We calculated 12 credible intervals of the posterior predictive distribution at the test locations with probability coverage equal to 5 %, 10 %, 20 %, 25 %, 30 %, 40 %, 50 %, 60 %,70 %, 80 %, 90 % and 95 %, respectively.

Finally, the Kullback-Leibler divergence from the observed prevalence to the predictive posterior distribution was calculated as $KL(j) = \sum_{i=1}^{26} p_i^{obs} * log(\frac{p_i^{obs}}{p_i^{rep(j)}}), j = 1, \ldots, 1000$, where p_i^{obs} is the observed prevalence at test site s_i and $\boldsymbol{p}_i^{rep} = (p_i^{rep(1)}, \ldots, p_i^{rep(1000)})$ are 1000 replicated data from the predictive distribution at test location s_i.

The results of the model validation are presented in Section 4.4.

4.3.5 Implementation details

Although the choice of the number of knots is not essential for penalized splines (Ruppert, 2002), a minimum number of knots are needed to capture the spatial variability in the data. We fixed the number of knots to 5 and we saw that increasing this number has no significant change to the fit given by P-splines due to the penalty parameter. The P-spline regression was performed in R version 2.4.0.

For the environmental variables with monthly values assigned we calculated summary statistics over the months suitable for malaria transmission and assess which measure leads to a better model fit in each agro-ecological zone. The calculated summary statistics were maximum and average for the minimum temperature, maximum temperature, NDVI and SWS and maximum, average and total for the rainfall. Table 4.1 depicts the measures used for each environmental predictor and each agro-ecological zone.

Agro-ecological zone	Environmental predictors				
	Min. Temp.	Max. Temp.	Rainfall	NDVI	SWS
Sahel	Average	Average	Maximum	Maximum	Average
Sudan Savanna	Average	Average	Average	Maximum	Average
Guinea Savanna	Average	Average	Total	Average	Average
Equatorial Forest	Average	Average	Total	Maximum	Maximum

Table 4.1: Measures of each environmental predictor which lead to a better model fit in each agro-ecological zone.

The prior distributions used to complete the Bayesian formulation of the model were as following. For the predictor's coefficients β_k we adopt non-informative Normal distribution $\beta_k \sim N(0, 10^2)$. We adopt an inverse Gamma prior for the variance parameters $\sigma_k^2 \sim IG(a_1, b_1)$ and a Gamma prior distribution for the range parameters $\rho_k \sim G(a_2, b_2)$ with the hyperparameters of these distributions chosen to have mean equal to 1 and variance equal to 100. The model parameters were estimated by implementing the Gibbs sampler (Gelfand and Smith, 1990) with five parallel chains, which requires simulating from the conditional posterior distributions of all parameters, iteratively until convergence. The full

conditional distributions of σ_k^2 are inverse Gamma distributions and it is straightforward to simulate from. The conditional posterior distributions of β_k, $\boldsymbol{\omega}_k$ and ρ_k do not have known forms. We simulate from these distributions using the Metropolis algorithm with a Normal proposal distribution having the mean equal to the parameter estimate from the previous Gibbs iteration and the variance equal to a fixed number, iteratively adapted to optimize the acceptance rates. We have run a five chain sampler of $200,000$ iterations with a burn-in of $10,000$ iterations and we assessed the convergence by examining the ergodic averages of selected parameters. The analysis was implemented in Fortan 95 (Compaq Visual Fortran Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.).

4.4 Results

First, the results of the model validation are described because inference and mapping are based on the model with the best predictive ability. Figure 4.2 shows the distribution of the "p-values" of test locations and the distribution of the Kulback-Leibler difference measure estimated by the P-spline model and the model with categorized covariates. The median "p-value" of the former model is closer to 0.5, suggesting that this is the best model. The P-spline model has also the smallest Kulback-Leibler value, supporting the results of the previous validation approach.

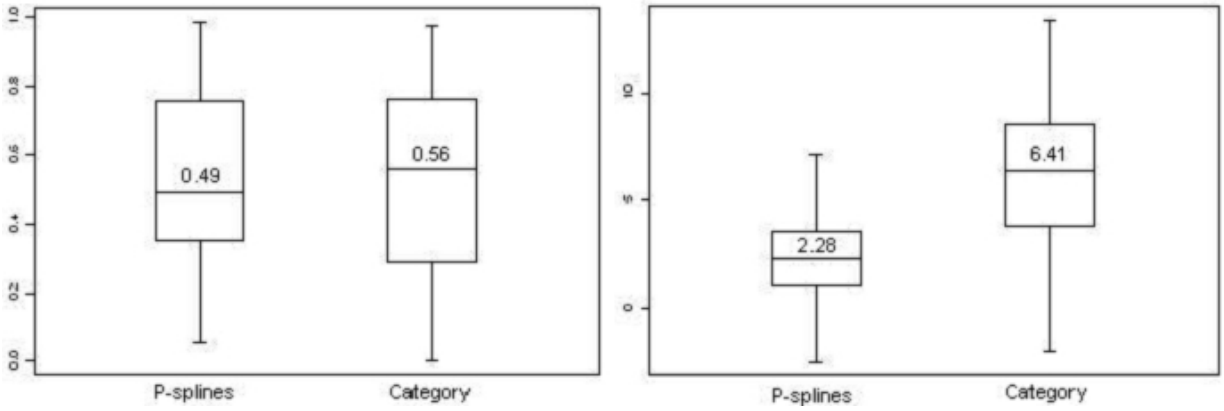

Figure 4.2: The distribution of Bayesian p-values (left) and Kulback-Leibler difference measure (right) for the two non-stationary Bayesian geostatistical models. The box plots display the minimum, the 25^{th}, 50^{th}, 75^{th} and the maximum of the distribution.

Figure 4.3 presents the percentages of test locations with malaria prevalence falling into credible intervals of coverage ranging from 5% to 95% for both geostatistical models. For the 95% credible interval the P-spline model included the highest precentage of test locations (96.15% versus 84.62%). Consistently, the P-sline model includes the highest percentage of observed locations in all coverage intervals.

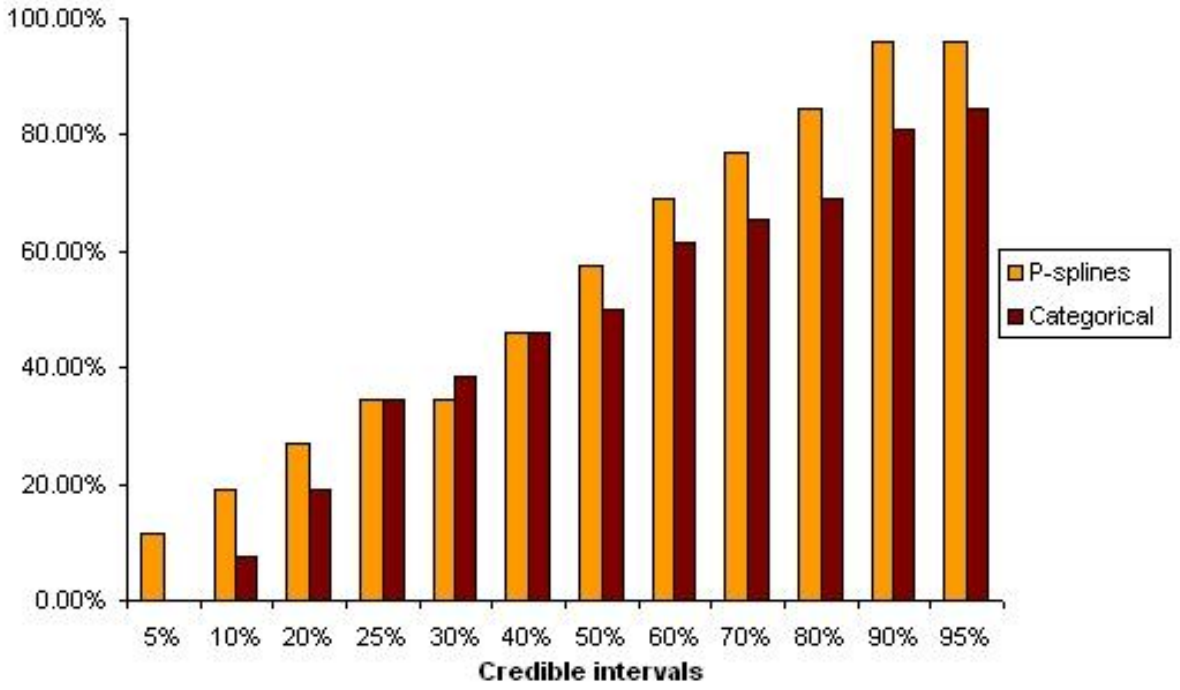

Figure 4.3: Percentage of test locations with malaria prevalence falling in the 5%, 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 95% credible intervals of the posterior predictive distribution.

Although the models differ in their predictive ability, the goodness of fit DIC measure does not favor any of the models. In particular, the model with categorized covariates had a DIC of 1636.1, while the P-spline model had a DIC of 1634.2.

Based on the results of the model validation, we use the P-spline non-stationary model for estimating the relation climate-malaria and produce a smooth map of malaria risk.

Figures 4.4 through 4.7 show the non-linear effect of the environmental factors in all four agro-ecological zones. The plots depict the posterior means and the 95% credible intervals. The effect of each covariate changed from one agro-ecological zone to another, suggesting that it was important to consider a separate nonparametric model in each zone.

Figure 4.4 shows that in the Sahel zone the only covariates with significant effect on malaria risk were distance to the nearest water body and season length since the credible intervals of the posterior means do not include zero. The effect of distance to water on parasitaemia risk is more or less constant. We detect a constant trend for the length of transmission season up to 3 months and then a decreasing trend up to 5 months.

The impact of the environmental variables in the Sudan Savanna zone are presented in Figure 4.5. The only credible interval that did not include zero was the one corresponding to the posterior mean of the minimum temperature, therefore we conclude that this was the only variable with a significant effect on the malaria risk in Sudan Savanna area. We notice that minimum temperature had a constant effect between 19°C and 21°C and then it showed an increasing risk effect from 21°C to 25°C.

In Guinea Savanna the environmental factors significant associated with the parasitaemia risk were: distance to the nearest water body, maximum temperature, rainfall and length of the malaria transmission season (Figure 4.6). The distance to the nearest water body had a slightly increasing effect up to 10 km, indicating high malaria prevalence in areas within 10 km away from a water source, which in this study was considered permanent river or lake. The risk of malaria decreased in regions 10 km further away from a water body. The maximum temperature had a decreasing effect in Guinea Savanna between 28°C and 32°C. As expected, the effect of rainfall on malaria prevalence was not linear. We observe an increase in parasitaemia prevalence when rainfall was between 800 mm and 1300 mm. The risk of malaria decreased when the amount of rainfall exceeded 1300 mm since excessive rainfall may flush out the eggs or larvae out of the pools impeding the development of mosquito eggs or larvae. Malaria prevalence slightly decreased for length of malaria transmission season between 6 and 8 months and increased when the malaria transmission season exceeds 9 months.

Figure 4.7 shows the non-linear effect of the predictors on malaria prevalence in Equatorial Forest. The only variable with a significant risk effect was rainfall, which showed an increased effect on malaria risk. All the other environmental factors had zero included in the credible intervals of the posterior means, thus their influence is minimal.

We notice that for some variables the credible intervals tend to widen at the tails of the splines. This could be explained by the fact that only few data locations had extreme values for those variables.

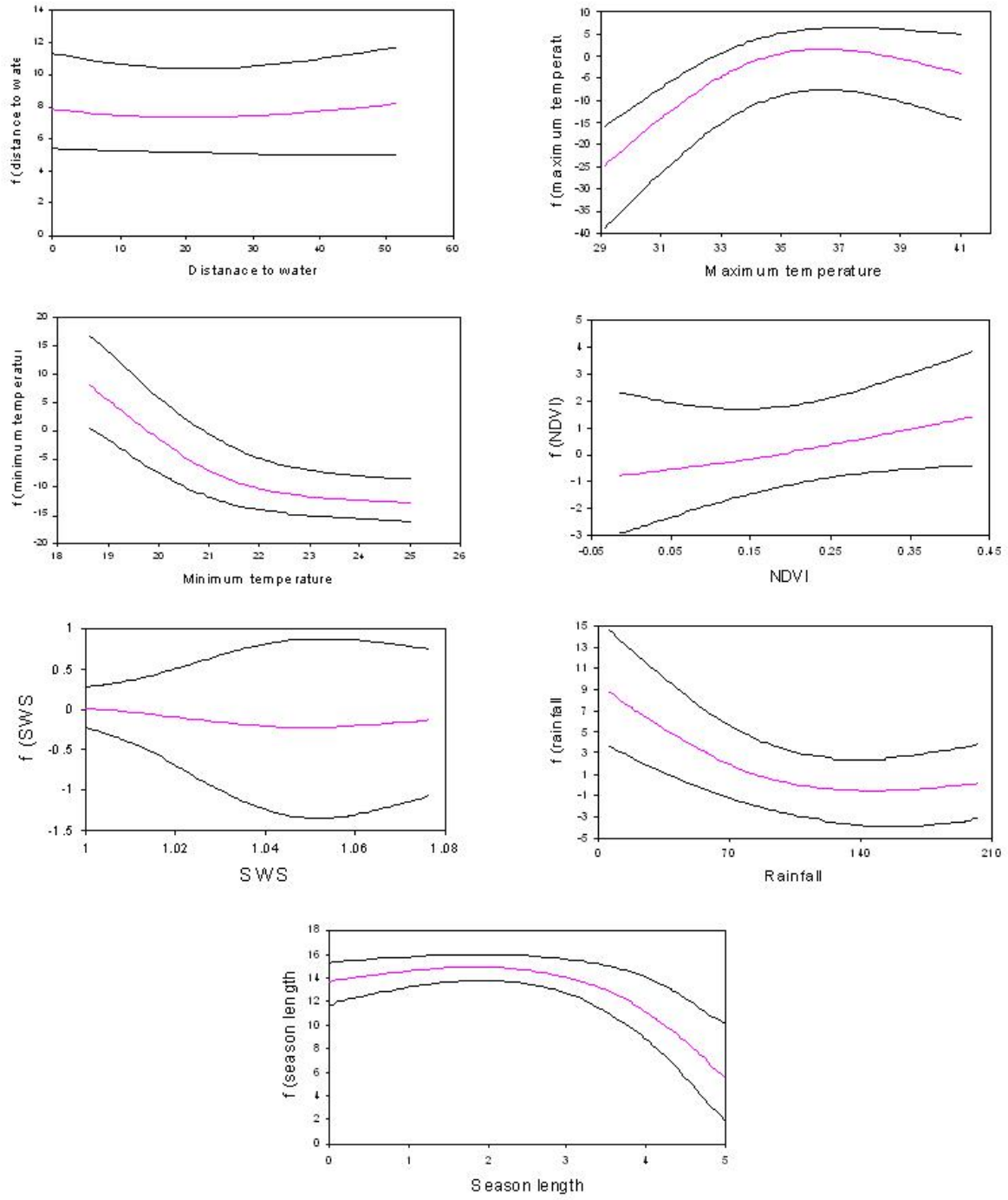

Figure 4.4: Estimated non-linear effect (P-spline) of environmental factors on malaria risk in West Africa, Sahel. The posterior mean (pink) and the 95% credible interval are shown.

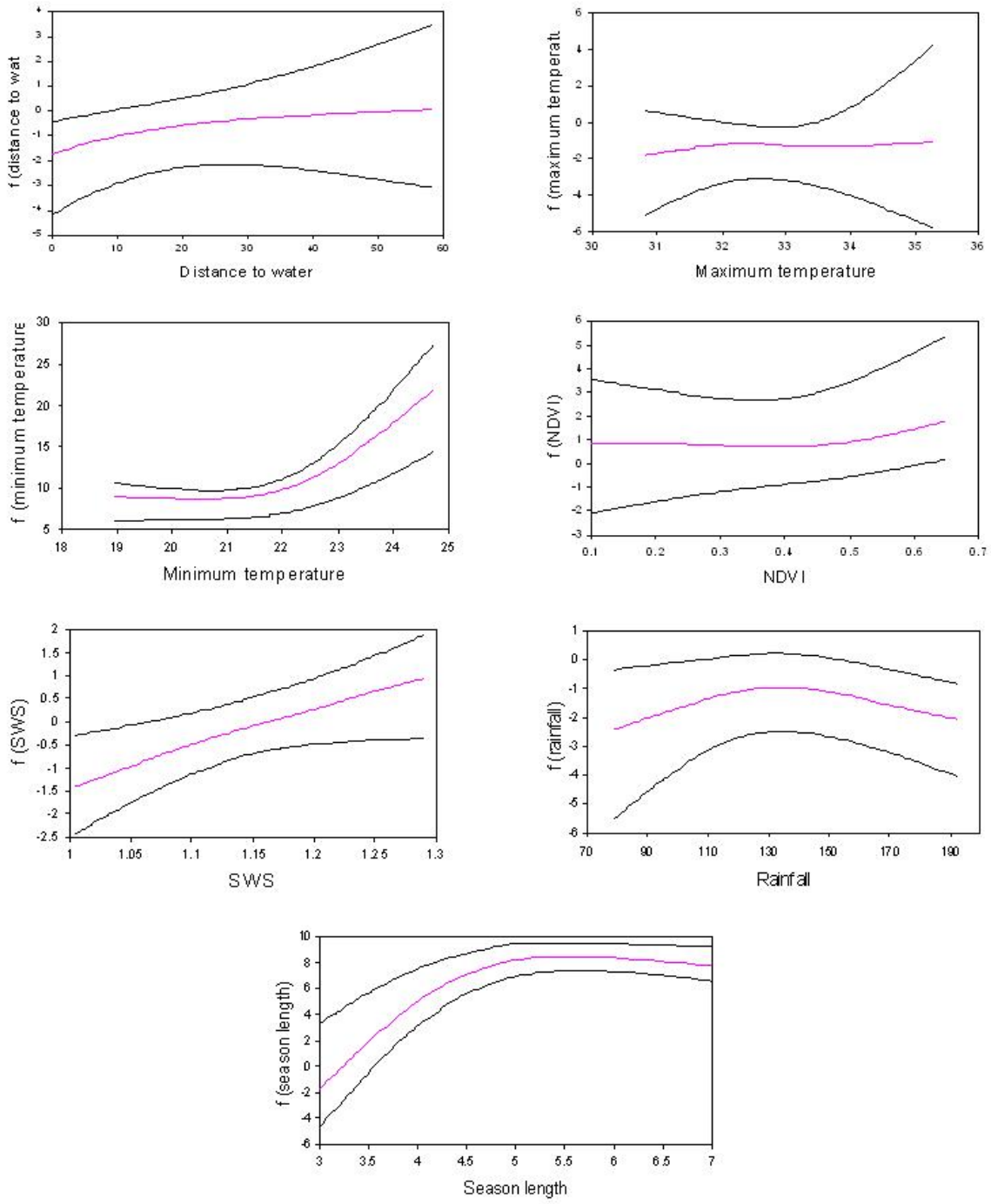

Figure 4.5: Estimated non-linear effect (P-spline) of environmental factors on malaria risk in West Africa, Sudan Savanna. The posterior mean (pink) and the 95% credible interval are shown.

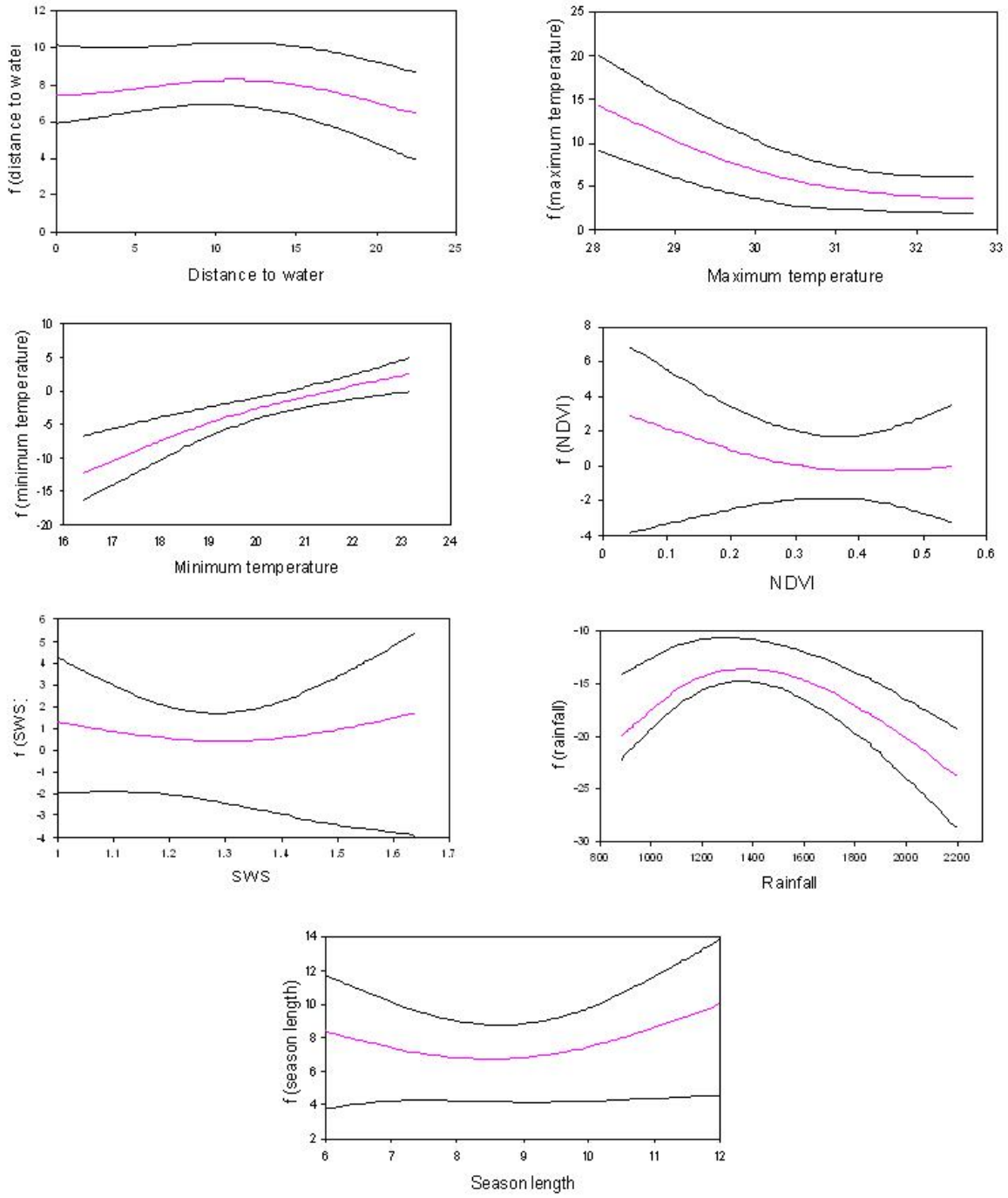

Figure 4.6: Estimated non-linear effect (P-spline) of environmental factors on malaria risk in West Africa, Guinea Savanna. The posterior mean (pink) and the 95% credible interval are shown.

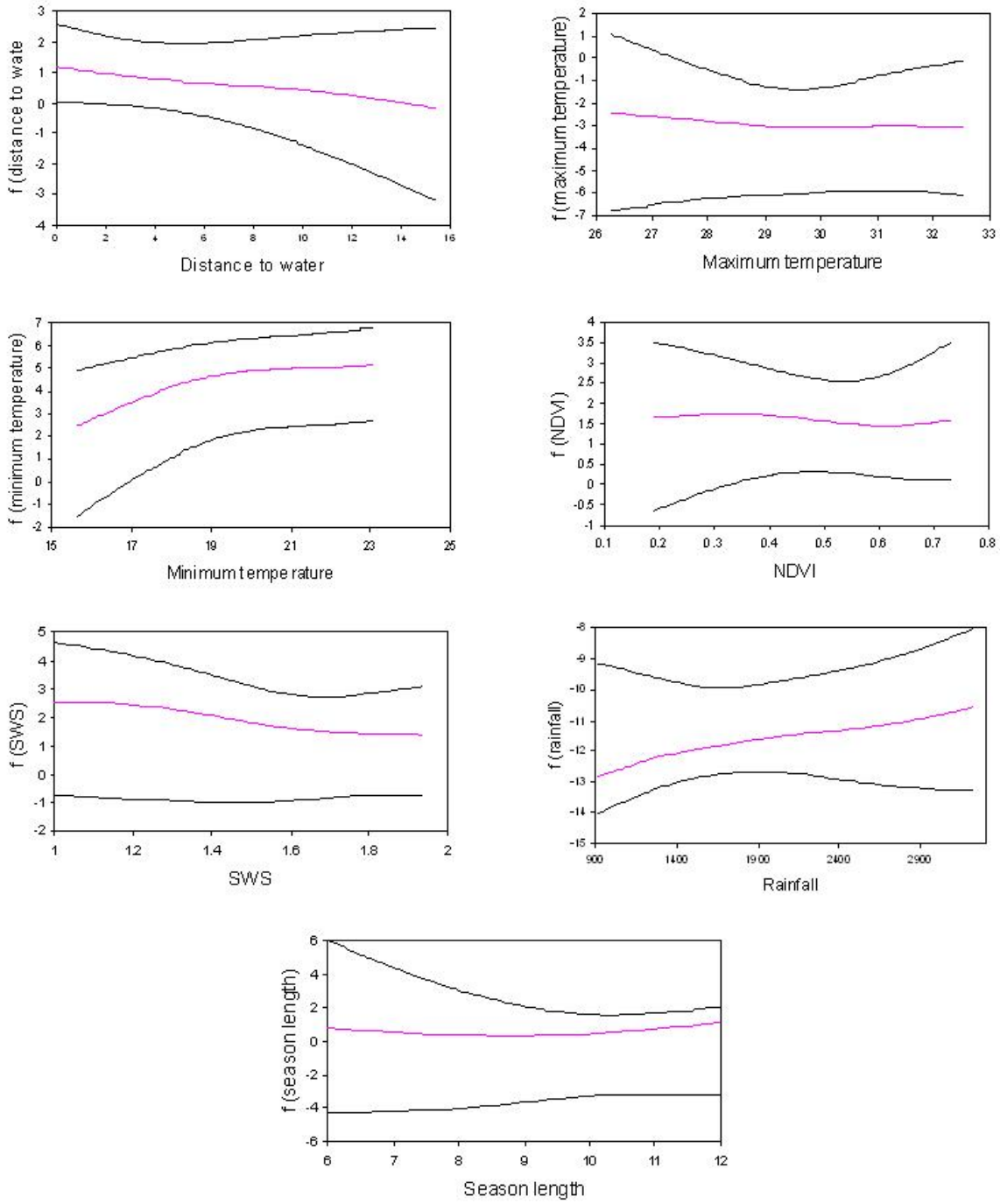

Figure 4.7: Estimated non-linear effect (P-spline) of environmental factors on malaria risk in West Africa, Equatorial Forest. The posterior mean (pink) and the 95% credible interval are shown.

In Table 4.2 we present the posterior estimates (odds ratio and credible intervals) corresponding to the land use regression coefficients in each agro-ecological zone obtained from the Bayesian geostatistical non-stationary model. In Sahel and Guinea Savanna the land use was not significantly associated with the malaria risk when spatial correlation was taken into account. The odds of malaria for locations situated around cropland were significantly higher than those in urban area in Sudan Savanna (OR=4.13, 95 % credible interval: $1.25, 19.79$). In Equatorial Forest malaria odds were significantly higher in Grass/Shrub land/Savanna (OR=4.06, 95 % credible interval: $1.12, 19.63$) and Forest (OR=3.53, 95 % credible interval: $1.03, 13.78$) compared with the urban areas.

Agro-ecological zone	Land use	OR	95% CI[a]
Sahel	Cropland	1.00	
	Grass/Shrub land/Savanna	0.37	(0.02, 3.13)
	Water bodies	0.06	(0.00, 1.71)
	Urban area	0.17	(0.02, 2.16)
Sudan Savanna	Urban area	1.00	
	Cropland	4.13	(1.25, 19.79)
	Grass/Shrub land/Savanna	3.50	(0.79, 17.76)
	Water bodies	2.05	(0.26, 14.22)
	Wetland	1.86	(0.24, 19.77)
Guinea Savanna	Urban area	1.00	
	Grass/Shrub land/Savanna	0.87	(0.05, 11.52)
	Forest	0.22	(0.01, 2.24)
	Water bodies	0.61	(0.00, 15.12)
Equatorial Forest	Urban area	1.00	
	Cropland	2.25	(0.13, 78.49)
	Grass/Shrub land/Savanna	4.06	(1.12, 19.63)
	Forest	3.53	(1.03, 13.78)
	Water bodies	2.48	(0.36, 17.01)
	Wetland	1.13	(0.14, 8.87)

[a] : Credible intervals (or posterior intervals).

Table 4.2: Posterior estimates for land use coefficients. The land use is the only categorical variable in the P-spline non-stationary model.

Posterior estimates of the spatial parameters (spatial variance and decay parameter) in the four zones are shown in Table 4.3. In Sahel the posterior median of ρ was equal to 6.23 (95 % credible interval: $1.15, 22.31$), which in the current exponential setting corresponds to a minimum distance for which the spatial correlation becomes negligible of 0.48 km (95 % credible interval: $0.13, 2.61$). In the other 3 zones the decay parameters were similar and indicated ranges of 4.23 km (95 % credible interval: $0.61, 28.64$) in Sudan Savanna, 4.58 km (95 % credible interval: $0.56, 98.84$) in Guinea Savanna and 4.86 km (95 % credible interval: $0.54, 44.78$) in Equatorial Forest. We conclude that the spatial correlation is weak in the Sahel zone and strong in the other agro-ecological zones. The spatial variance varied from 0.83 (95 % credible interval: $0.34, 1.67$) in the Sahel to 3.33 (95 % credible interval: $1.48, 7.64$) in Guinea Savanna.

Agro-ecological zone	Spatial parameter	Median	95% CI[a]
Sahel	σ^2	0.83	(0.34, 1.67)
	ρ^b	6.23	(1.15, 22.31)
Sudan Savanna	σ^2	1.31	(0.55, 2.98)
	ρ^b	0.70	(0.10, 4.88)
Guinea Savanna	σ^2	3.33	(1.48, 7.64)
	ρ^b	0.66	(0.03, 5.39)
Equatorial Forest	σ^2	1.78	(1.10, 3.02)
	ρ^b	0.62	(0.07, 5.53)

[a] : Credible intervals (or posterior intervals).
[b] : Based on ρ we calculate the range parameter $3/\rho$ (in km).

Table 4.3: Posterior estimates of spatial parameters.

The map of predicted malaria prevalence for sub-Sahara West Africa is shown in Figure 4.8. We find a relatively high malaria risk with only few exceptions. High levels of prevalence were predicted in the center-east of Sénégal, the north of Ghana, the south of Togo, some areas in the west and east of Nigeria and north-west of the Democratic Republic of Congo. In addition, the areas along the Atlantic Ocean were estimated to have high malaria risk. Low levels of malaria prevalence were observed in the north and west of Sénégal, Guinea Bissau, Guinea, north-west of Cote d'Ivoire, the south of Ghana, the north and north-east of Nigeria, the north of Cameroon, a small region in center Gabon and Republic of Congo. The prediction error from the Bayesian geostatistical non-stationary model is depicted in Figure 4.9.

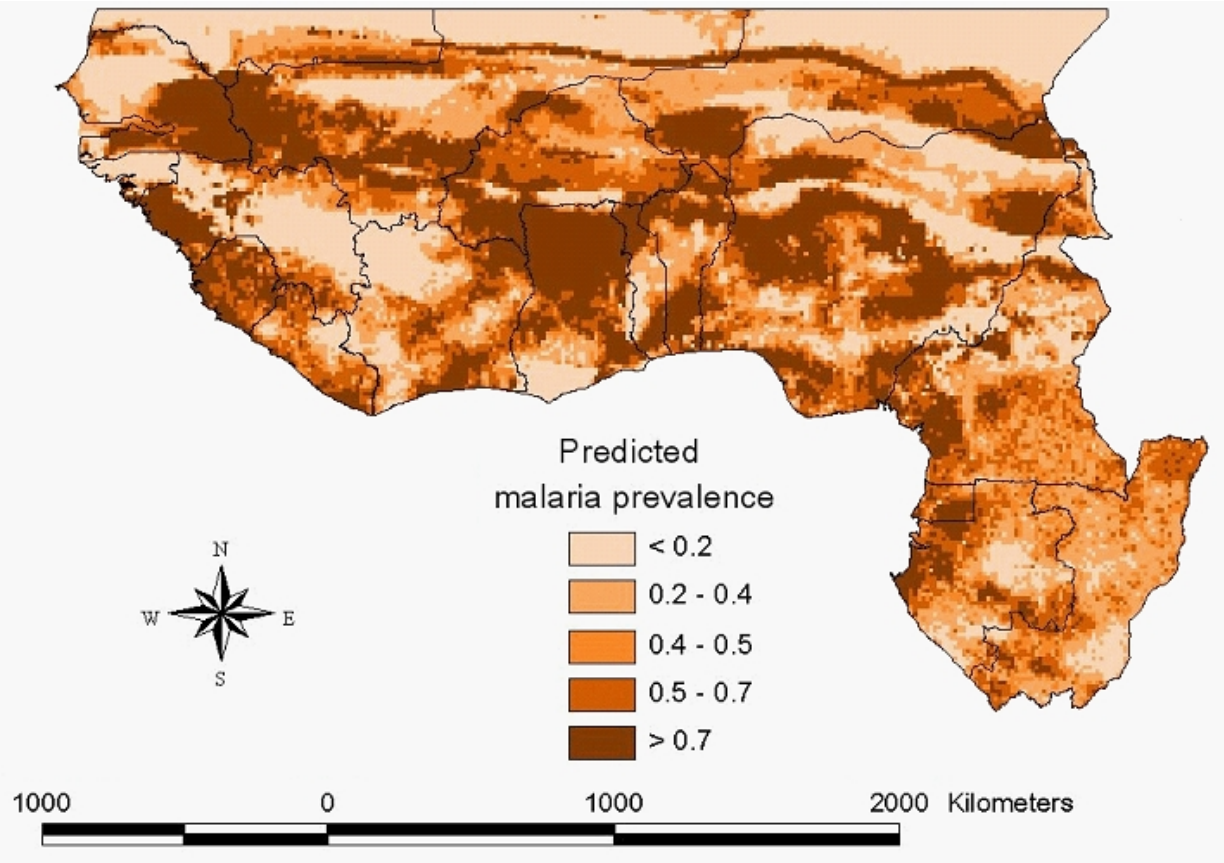

Figure 4.8: Map of predicted malaria prevalence in children $1-10$ years for West Africa.

4.5 Discussion

Malaria is an environmental disease and its endemicity depends on the density and infectivity of anopheline vectors. These are highly influenced by meteorological variables, but the relation is often not linear. In this analysis we modeled this relation using Bayesian P-splines. The comparison between Bayesian P-splines and the widely used method which considers the covariates as categorical variables (Section 4.3.4) shows that the former model fits better the relation between malaria risk and environmental factors. It is the first time the non-parametric spline approach is used to obtain a smooth map of malaria prevalence. Previous mapping efforts in West Africa assumed the same climate-malaria relation across the agro-ecological zones and modeled the non-linearity by using polynomial functions (Kleinschmidt et al., 2001) and by comparing different functional forms of the predictors (Gemperli et al., 2006). However, the polynomial functions and the pre-specified functional forms may not be able to capture the true underlying relations.

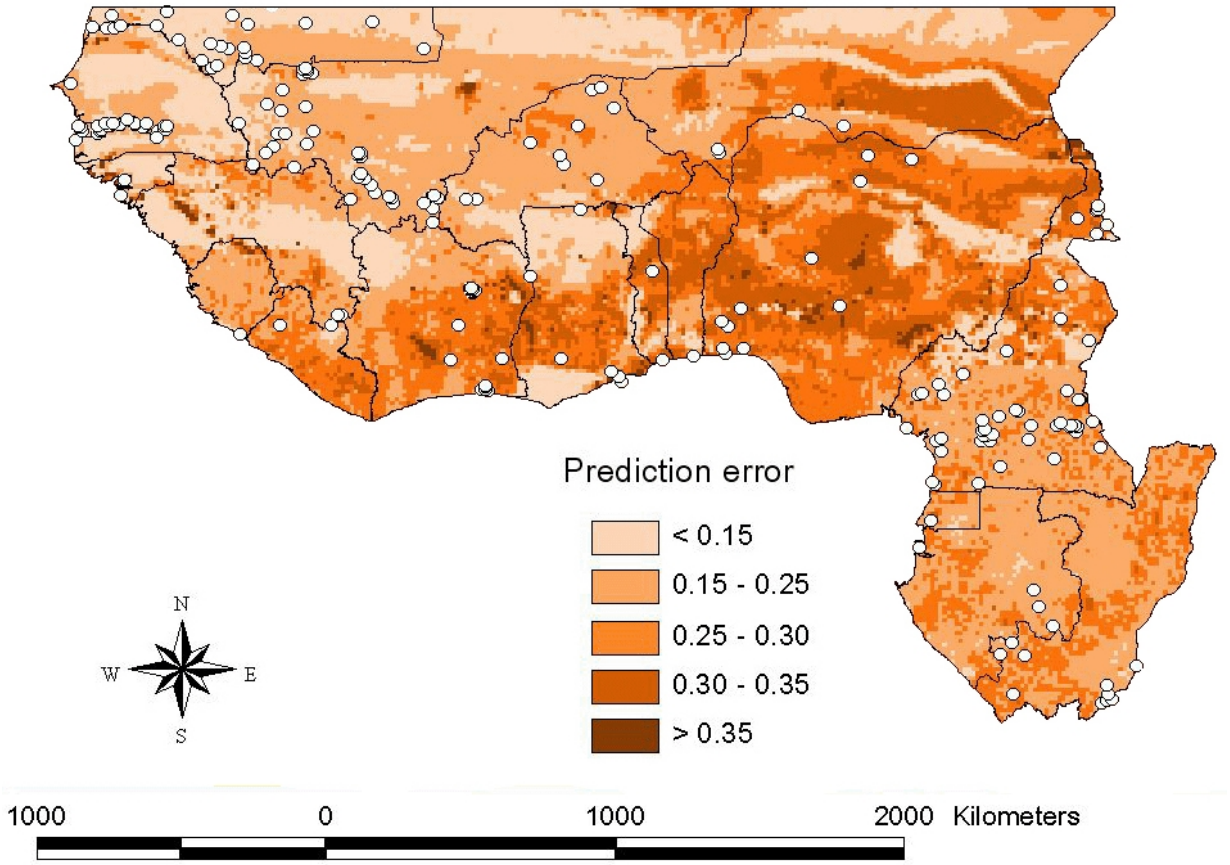

Figure 4.9: Map of prediction error.

We assume that the environmental factors have a different influence on malaria transmission in the four agro-ecological zones and we consider a separate nonparametric model in each zone. Kleinschmidt et al. (2001) addressed this issue by a non-Bayesian spatial model. However, the resulting map showed discontinuities at the borders between the zones, which were further smoothed. This problem was prevented in our case since we used a mixture of spatial processes over the whole area.

In the Sahel, the temperature is high and the amount of rainfall is very low. Thus mosquito populations increase rapidly at the onset of the rain, because of short vector developmental cycles (Picq et al., 1992). Consequently, the intensity of transmission may depend on the number of rainy months. This supports the significant association between the length of transmission season and malaria prevalence found in the Sahel.

The temperature affects the transmission cycle in many different ways, but the effects on

the duration of the sporogonic cycle of the parasite and vector survival are particularly important. In Sudan Savanna the minimum temperature had a significant effect on the parasitaemia risk, with values above 21°C indicating an increased risk of malaria. Our findings correspond to the ones of Craig et al. (1999) who showed that temperatures above 22°C are suitable for stable malaria transmission.

The Guinea Savanna represents the optimum climate envelope for malaria transmission because it offers favorable conditions for *An. gambiae s.l.* and *An. funestus* mosquitoes. Usually, the limiting factor of transmission in this zone is the heavy rainfall which flushes out many larvae and pupae out of the pools or decreases the temperatures. This was also reflected by our results.

Malaria data observed over large areas have non-stationary characteristics. Ignoring this feature could lead to a misspecification of the spatial correlation and therefore to wrong estimates of the standard error of both the covariates and the prediction. In our recent work (Gosoniu et al., 2006) we addressed non-stationarity in mapping malaria risk in Mali by dividing the study area into fixed number of tiles, assuming a separate stationary spatial process in each tile and correlation between tiles. In this paper the fixed tiles correspond to the four different agro-ecological zones in West Africa. This method is appropriate for malaria mapping over large areas with a clear way of defining the partitioning. We are currently extending this model by assuming random space partitioning to accommodate mapping over areas with no clear way of finding a fixed tessellation.

Comparing our malaria risk map for children $1-10$ years old in West Africa with those of Kleinschmidt et al. (2001) and Gemperli et al. (2006) we observe similar pattern of malaria prevalence. We all estimated high prevalence at the border between Mali and Sénégal, the north part of Ghana, the south area of Togo and the west part of Cameroon. Low levels of prevalence were estimated by all three maps in the north-west of Sénégal, Guinea Bissau and the north of Cameroon. We notice several differences in the three maps: in Guinea we estimated a low prevalence (< 0.2), Gemperli et al. (2006) estimated a prevalence between 0.3 and 0.4, while Kleinschmidt et al. (2001) showed a prevalence between 0.3 and 0.7. In Sierra Leone both our map and the one of Kleinschmidt et al. (2001) show a high prevalence, whereas Gemperli et al. (2006) predicted a lower level of malaria risk. Our map and Kleinschmidt et al. (2001) map estimate a low prevalence in south of Ghana, while the map of Gemperli et al. (2006) shows higher risk.

Gemperli et al. (2006) used the malaria transmission Garki model (Molineaux and Gramiccia, 1980) to produce age and seasonality adjusted maps from the MARA survey data which are heterogeneous in age and seasonality across locations. In this study we discarded the surveys carried out on population outside the range of 1 to 10 years old. Currently we are employing the method developed in this paper and newly developed stochastic transmission model (Smith et al., 2006) to produce age-adjusted maps in West and Central Africa, making use of all the available MARA data.

Although we used a mixture of spatial processes over the area of interest, the map of predicted malaria risk show discontinuities between the agro-ecological zones on the right hand side. This limitation could be explained by the position of the fixed centers of the tiles. We have chosen the coordinates of the fixed centers as the average of the coordinates of the data points belonging to the specific tile. However, the distribution of the survey locations was higher in the east side of the study area, therefore the centers of the tiles were shifted towards that part of the region. The weights used in the calculation of location-specific random effects were based on the distance between the fixed centers of the tiles and the specific points, hence for the points in the west side of the map the weights were very small and unable to smooth the tile-specific spatial processes at the borders. In this situation (and usually in the case when the tiles have more a rectangular rather than square or circle shape) one may consider more than one fixed points as "centers" per tile and calculate the weight for each location from the nearest "center" of the tile. We are currently exploring this approach by re-analyzing malaria data in West and Central Africa.

Acknowledgments

The authors would like to thank the MARA/ARMA collaboration for making the malaria data available. We are grateful to the volunteer computer platform malariacontrol.net and to all the people around the world who made available their computers to help us predict malaria distribution in a feasible period of time. This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

Chapter 5

Non-stationary partition modeling of geostatistical data for malaria risk mapping

Gosoniu L.[1] ,Vounatsou P.[1]

[1] Swiss Tropical Institute, Basel, Switzerland

This paper has been submitted to *Journal of Applied Statistics*.

Summary

The most common assumption in geostatistical modeling of malaria is stationarity, that is spatial correlation is a function of the distance between locations and independent of the locations themselves. However, local factors (environmental or human related activities) may influence geographical dependence in malaria transmission differently at different locations, introducing non-stationarity. Ignoring this characteristic in malaria spatial modeling may lead to inaccurate estimates of the standard errors for both the covariate effects and the predictions. In this paper a model based on random Voronoi tessellation that takes into account non-stationarity was developed. In particular, the spatial domain was partitioned into sub-regions (tiles), a stationary spatial process was assumed within each tile and between-tile correlation was taken into account. The number and configuration of the sub-regions are treated as random parameters in the model and inference is made using reversible jump Markov chain Monte Carlo (RJMCMC) simulation. This methodology was applied to analyze malaria survey data from Mali and to produce a country-level smooth map of malaria risk.

Keywords: Bayesian inference; geostatistics; kriging; MARA; malaria risk; prevalence data; non-stationarity; reversible jump Markov chain Monte Carlo; Voronoi tessellation.

5.1 Introduction

Malaria is one of the most common infectious diseases and a major international public health problem. According to the World Health Organization, every year between 300 and 500 million people are infected with malaria. Most cases occur in sub-Sahara Africa (with approximately 2 million deaths each year), but many infections persist there in an asymptomatic state. Accurate estimates of the burden of malaria are needed for evidence-based planning of malaria control. In endemic areas malaria burden is best measured by the age-specific prevalence of infection, but for most African countries no comprehensive surveys have been carried out and it is needed to use ad hoc collections of local surveys. The survey data are correlated in space, therefore geostatistical methods are the most appropriate for obtaining smooth maps of malaria risk and estimates of number of people at risk of malaria. Malaria is an environmental disease because its transmission depends on the distribution and abundance of mosquitoes, which are sensitive to environmental and climatic factors. Estimating the environment-parasitaemia relation we can predict malaria transmission at locations where data are not available .

In geostatistics it is commonly assumed that the spatial dependence between two points is a function of only the separation distance and independent of the absolute locations. However, malaria data observed over large areas have non-stationary characteristics because spatial correlation may be influenced differently at various locations due to local characteristics like environmental factors, intervention measures, mosquito ecology, human activities (e.g. irrigation, dam construction), health services etc. Ignoring this feature could lead to wrong estimates of the standard error for both covariates coefficients and predictions (Gosoniu et al., 2006).

In recent years, much attention has been concentrated toward the development of models that allow for non-stationary spatial covariance structure. Obled and Creutin (1986) proposed the empirical orthogonal function (EOF) approach based on an eigenfunction expansion of the covariance function and is very popular in the spatial analysis of environmental data. Nychka et al. (2002) replaced the basis of the orthogonal functions with wavelet basis functions. Although these methods are very flexible, they are not easily interpreted from a geostatistical point of view. The spatial deformation method introduced by Sampson and Guttorp (1992) assumes that the spatial process is stationary and isotropic only after some nonlinear transformation of the original sampling space. A Bayesian framework for the deformation method was proposed independently, by Damian et al. (2001) and

Schmidt and O'Hagan (2003). Haas (1990, 1995) modeled non-stationary spatial processes using a moving window approach which assumes stationarity only within the window constructed around the locations where kriging is performed. Higdon et al. (1998) assumed that any stationary Gaussian process can be expressed as the convolution of a Gaussian white noise process with a kernel function (i.e. bivariate Gaussian density function). They account for non-stationarity by allowing the kernel to vary smoothly over space. Kim et al. (2005) modeled non-stationary Gaussian data by partitioning the region of interest into sub-regions via a Voronoi tessellation, assuming stationarity within each tile but across sub-regions the data are assumed to be independent. The shape and the number of sub-regions are estimated by the data. Fuentes (2001) represents a non-stationary process as a weighted average of locally stationary processes defined within disjoint sub-regions of the area of interest. The weights of the local processes are based on kernel functions. The shape of the sub-regions is known a priori and the number of sub-regions is chosen using an AIC or a BIC criterion. Banerjee et al. (2004) replaced the kernels by decreasing functions of the distance between the data points and the centroids of the sub-regions.

In the field of disease mapping the non-stationarity issue was previously addressed by Gosoniu et al. (2006) who extended the work of Banerjee et al. (2004) to model non-Gaussian malaria prevalence data in Mali. Raso et al. (2005) applied the model developed by Gosoniu et al. (2006) to map hookworm infection prevalence in western Cote d'Ivoire.

In this paper we further extend our previous work (Gosoniu et al., 2006) by considering random rather than fixed tiles to map malaria prevalence in Mali. In particular, we partition the spatial domain into sub-regions using a Voronoi tiling, assume stationary spatial covariance structure within sub-regions and correlation between them. The shape and the number of tiles are treated as unknown variables and inference is made using reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). A description of the malaria data which motivated this work and the environmental variables used as predictors is given in section 5.2. Section 5.3 describes the Bayesian partition model and details of the implementation by RJMCMC are given in Section 5.4. We analyzed the malaria prevalence data in Section 5.5. The paper ends with final remarks and conclusions in Section 5.6.

5.2 Motivating example: mapping malaria risk in Mali

The data which motivated this work are malaria prevalence data extracted from the "Mapping Malaria Risk in Africa" (MARA/ARMA,1998) database. MARA is the most comprehensive database of malaria prevalence data in Africa containing over 10.000 distinct age-specific prevalence values extracted from published and unpublished sources across the whole continent. We selected prevalence data from malaria surveys carried out between 1977 and 1995 at 89 locations in Mali on children between 1 and 10 years old (Figure 5.1). The children were considered malaria positive if *Plasmodium falciparum* was found in the blood smears collected. The size of the surveys varied from 14 to 4835. There were 43.492 sampled children and 44% were found malaria positives. Environmental data used as malaria predictors such as rainfall, maximum and minimum temperature, vegetation index, were extracted from remote sensing. Details on sources and resolutions of the environmental data are given in Gosoniu et al. (2006).

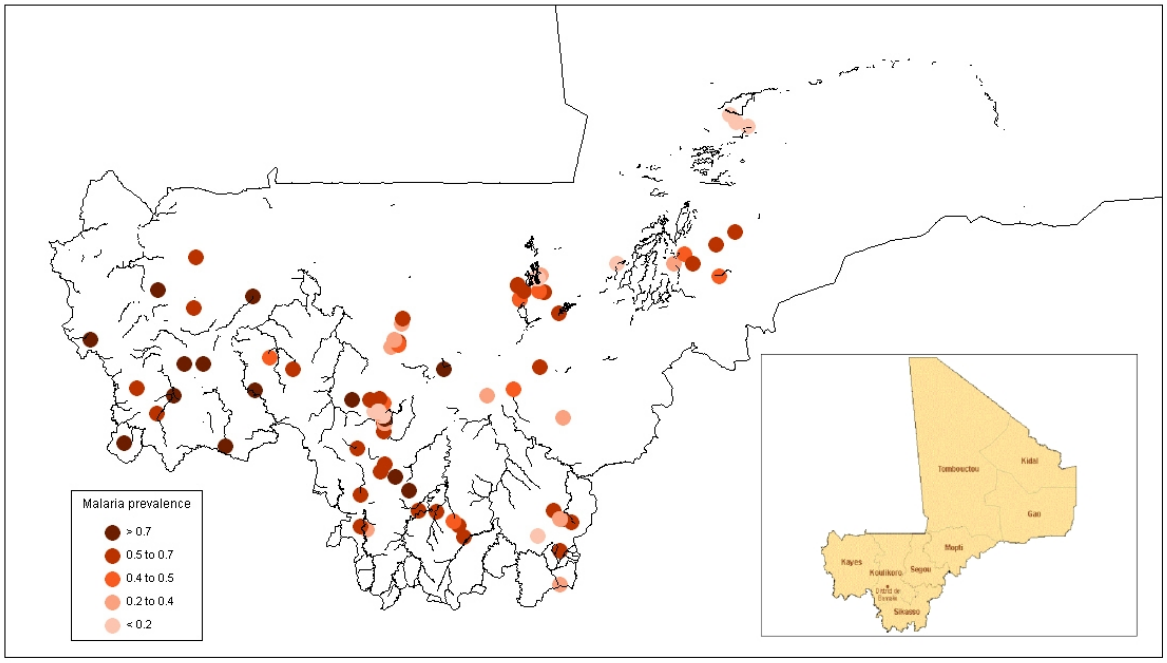

Figure 5.1: Sampling locations in sub-Saharan Mali with dot shading indicating the observed malaria prevalence.

5.3 Modeling non-stationarity via dependent spatial processes

At each location $s_i \in A \subset R^2$, $i = 1, \ldots, n$ the data are available in the form of pairs, giving the number of children tested N_i and the number of those found positive to *P.falciparum* parasitaemia Y_i. These are typical binomial data and modeled via logistic regression. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be a collection of p environmental explanatory variables available at each location s_i. We model the relation between the environmental covariates $\boldsymbol{X}_i$ and the malaria risk p_i at location s_i via the logistic regression, that is $logit(p_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of regression coefficients. Parasitaemia risk at close geographical proximity is influenced by similar factors therefore we can not assume independence of observations. To account for the spatial variation present in the data we introduce at each location s_i a random effect ϕ_i, that is $logit(p_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \phi_i$ and model the spatial correlation on these parameters. Most geostatistical models assume that $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T$ models a latent stationary spatial process with spatial correlation structure that does not vary across the study area. This assumption is questionable, especially when modeling malaria indices over large geographical areas. Differences in agro-ecological zones, health systems and socio-economic indicators may change geographical correlation differently at various locations, so it cannot be expressed as a simple homogeneous function of the distance between the locations.

We relax the assumption of stationarity by defining a spatial partitioning via Voronoi tessellation. In particular, we divide the area of interest A into several subregions (tiles) $T_k, k = 1, \ldots, K$, such that $A = \bigcup_{k=1}^{K} T_k$ and $T_i \cap T_j = \oslash, \forall i \neq j$. Denoting by $\boldsymbol{c} = (c_1, \ldots, c_K)^T$ the centroids of the Voronoi tiles $T = (T_1, \ldots, T_K)^T$, we define the tile T_k as $T_k = \{s_i \in A | d(s_i, c_k) < d(s_i, c_l), \forall l \neq k\}$, where d is a distance measure (usually Euclidean distance). The number and centroids of the tiles are treated as unknown parameters of the model.

Spatial partitioning via Voronoi tiles was proposed by Kim et al. (2005) to model non-stationary spatial processes. However, the authors assumed a separate stationary process in each tile and independence of the data across the tiles.

Although the assumption of independence between tiles facilitates the matrix inversion by converting the spatial covariance matrix into block diagonal form, it ignores the spatial correlation between neighboring points located in different tiles. In this paper we extend

the random tessellation model to address the between-tile correlation by defining the spatial process as a mixture of tile-specific stationary spatial processes.

In each tile k we assume a stationary spatial process $\boldsymbol{\omega}_k$, that is $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kn})^T \sim N(\mathbf{0}, \Sigma_k)$. Here $(\Sigma_k)_{ij} = \sigma_k^2 corr(d_{ij}; \rho_k)$, where σ_k^2 corresponds to the sill of the spatial process and *corr* is a parametric correlation function of the distance d_{ij} between locations s_i and s_j belonging to the area A. For the current example we choose the exponential correlation function $corr(d_{ij}; \rho_k) = exp(-d_{ij}\rho_k)$, where ρ_k captures the rate of correlation decay with distance and is known as the range parameter. We then model the spatial random effect ϕ_i at location s_i as a weighted sum of the tile-specific stationary processes, that is $\phi_i = \sum_{k=1}^{K} a_{ik}\omega_{ki}$, where a_{ik} are decreasing functions of the distance between location s_i and the centroid of the subregion k. Then $\boldsymbol{\phi}$ is a non-stationary spatial process $\boldsymbol{\phi} \sim N(\mathbf{0}, \sum_{k=1}^{K} A_k \Sigma_k A_k)$, where A_k is a diagonal matrix with $(A_k)_{ii} = a_{ik}$. The number K and the centroids $\boldsymbol{c} = (c_1, \ldots, c_K)^T$ of the tiles are unknown parameters of the geostatistical model.

5.4 Implementation details

5.4.1 Model fit

We assume that the prior distribution for the number of tiles K has the form $Pr(K) \propto (1-\alpha)^K$, with the parameter $\alpha \in [0, 1)$ pre-defined. Knorr-Held and Rasser (2000) recommend small values for α to specify a non-informative prior. Given a number of tiles k, for the vector of tile centers $c_k = (c_1, \ldots, c_k)^T$ we assume the prior distribution $Pr(c_k/k) = \frac{(n-k)!}{n!}$.

For the regression coefficients $\boldsymbol{\beta}$ we adopt a non-informative uniform prior distribution $Pr(\beta) = U(-\infty, \infty)$. For the spatial parameters σ_k^2 and ρ_k we choose inverse gamma and gamma prior distributions respectively, that is $Pr(\sigma_k^2) = IG(a_1, b_1)$ and $Pr(\rho_k) = G(a_2, b_2)$. The hyperparameters a_1, b_1 and a_2, b_2 are chosen so that the prior distributions are proper but non-informative so that the inference is driven by the data.

The unknown parameters are estimated using the Metropolis-Hastings algorithm (Hastings, 1970). These methods are useful for sampling from the posterior distribution when the dimension of the parameter vector is fixed. In our application the number of tiles is not fixed, therefore at each iteration the number of parameters is unknown. To estimate the

parameters of our model, reversible jump MCMC (RJMCMC) was used. The RJMCMC algorithm introduced by Green (1995) is an extension of the Metropolis-Hastings algorithm which allows simulation from posterior distribution on spaces of varying dimensions. The method is called *reversible jump* due to the ability of the Markov Chain to jump between parameter spaces of different dimensions. At each iteration t we consider four possible steps, that is: STAY, BIRTH, DEATH and MOVE. Each of these steps are randomly chosen with probabilities Q_s, Q_b, Q_d, Q_m, such that $Q_s + Q_b + Q_d + Q_m = 1$.

In the STAY step the number of tiles k and the centroids c_k remain at the current value and we estimate the remaining model parameters using the Gibbs sampling. The only conjugate distributions are the full conditional distributions of σ_k^2 which are inverse gamma and simulation from them is straightforward. The remaining parameters can not be sampled directly from the full conditionals, hence we employ a random walk Metropolis algorithm (Tierney, 1994) having a Normal proposal density with mean equal to the estimate of the corresponding parameter from the previous Gibbs iteration and variance equal to a fixed number, iteratively adapted during the burn-in period to optimize the acceptance rates.

In the BIRTH step the number of tiles increases by one, adding a new center c_{K+1}, uniformly selected form the $n - K$ points which are not already centers. In this case the dimension of the vector parameter changes from $p + n + 2K$ to $p + n + 2K + 2 + n$ by introducing 2 additional spatial parameters σ_{K+1}^2 and ρ_{K+1} and a new stationary spatial process $\boldsymbol{\omega}_{K+1} = (\omega_{(K+1)1}, \ldots, \omega_{(K+1)n})^T \sim N(\mathbf{0}, \Sigma_{K+1})$, where $(\Sigma_{K+1})_{ij} = \sigma_{K+1}^2 exp(-d_{ij}\rho_{K+1})$, corresponding to the new tile. To match the dimensions of the parameter spaces between successive iterations with variable number of parameters we introduce the parameters $\boldsymbol{u}^{(t-1)} = (u_{\sigma^2}^{(t-1)}, u_{\rho}^{(t-1)}, \boldsymbol{u}_{\boldsymbol{\omega}}^{(t-1)})^T$. $u_{\sigma^2}^{(t-1)}$ and $u_{\rho}^{(t-1)}$ are generated from a gamma distribution and $\boldsymbol{u}_{\boldsymbol{\omega}}^{(t-1)}$ from a multivariate Normal distribution, independently from the value of parameters at previous iteration. We then define $\sigma_{K+1}^2 = f(u_{\sigma^2}^{(t-1)})$, $\rho_{K+1} = f(u_{\rho}^{(t-1)})$ and $\boldsymbol{\omega}_{K+1} = f(\boldsymbol{u}_{\boldsymbol{\omega}}^{(t-1)})$, with f the identity function $\mathbf{1}$. Then the location-specific random effects at a birth step will be $\phi_i^{new} = \sum_{k=1}^{K+1} a_{ik}\omega_{ki}, i = 1, \ldots, n$. The birth step is accepted with probability $\alpha_b = min(1, L_b A_b P_b \left|J_b\right|)$, where

$$L_b = \frac{L(\boldsymbol{Y},\boldsymbol{N},\boldsymbol{\beta},\boldsymbol{\phi}^{new})}{L(\boldsymbol{Y},\boldsymbol{N},\boldsymbol{\beta},\boldsymbol{\phi})}$$

is the likelihood ratio,

$$A_b = \frac{Pr(\boldsymbol{\omega}_{K+1}^{(t+1)})Pr(\sigma_{K+1}^{2(t+1)})Pr(\rho_{K+1}^{(t+1)})}{1} \frac{Pr(K+1)}{Pr(K)} \frac{1}{(n-K)}$$

is the prior ratio and

$$P_b = \frac{Q_d}{Q_b} \frac{(n-K)}{\Gamma(a_{\sigma^2}, b_{\sigma^2})\Gamma(a_\rho, b_\rho) N(\mathbf{0}, \Sigma_{K+1})}.$$

is the proposal ratio. $a_{\sigma^2}, b_{\sigma^2}$ and a_ρ, b_ρ are the parameters of the proposal gamma distribution for the spatial parameters σ^2 and ρ, respectively. The determinant of the Jacobian resulting from the potential change of dimension of the parameter vector $|J_b| = 1$ since we draw the new spatial parameters independent of the current parameters.

In the DEATH step the number of tiles decreases by deleting a center, uniformly chosen from the K existing ones. The spatial parameters corresponding to the deleted tile are σ_d^2 and ρ_d. We also delete the stationary spatial process $\boldsymbol{\omega}_d = (\omega_{(d1}, \ldots, \omega_{dn})^T \sim N(\mathbf{0}, \Sigma_d)$, where $(\Sigma_d)_{ij} = \sigma_d^2 exp(-d_{ij}\rho_d)$. Similarly to the birth step, the death step is accepted with probability $\alpha_d = min(1, L_d A_d P_d |J_d|)$. The likelihood ratio is

$$L_d = \frac{L(\boldsymbol{Y}, \boldsymbol{N}, \boldsymbol{\beta}, \boldsymbol{\phi}^{new})}{L(\boldsymbol{Y}, \boldsymbol{N}, \boldsymbol{\beta}, \boldsymbol{\phi})}$$

where $\phi_i^{new} = \sum_{k=1}^{K-1} a_{ik}\omega_{ki}, i = 1, \ldots, n$. The determinant of the Jacobian $|J_d|$ is equal to 1. The prior ratio and the proposal ratio have the same form as the corresponding birth step, except that the ratio is inverted, that is:

$$A_d = \frac{1}{Pr(\boldsymbol{\omega}_d^{(t)}) Pr(\sigma_d^{2(t)}) Pr(\rho_d^{(t)})} \frac{Pr(K-1)}{Pr(K)} \frac{(n-K+1)}{1}$$

and

$$P_d = \frac{Q_b}{Q_d} \frac{\Gamma(a_{\sigma^2}, b_{\sigma^2})\Gamma(a_\rho, b_\rho) N(\mathbf{0}, \Sigma_d)}{(n-K+1)}$$

In the MOVE step the number of tiles K remains at the current value and we uniformly select a center from the current ones and propose a new location for it by uniformly sampling a data point from the $n - K$ available points. The tessellation structure is modified but the values of the parameters are not altered. The MOVE step is accepted with probability $\alpha_m = min(1, L_m A_m P_m |J_m|)$, with proposal ratio $P_m = 1$ and the determinant of the Jacobian $|J_m| = 1$. The likelihood ratio is defined by:

$$L_d = \frac{L(\boldsymbol{Y}, \boldsymbol{N}, \boldsymbol{\beta}, \boldsymbol{\phi^{new}})}{L(\boldsymbol{Y}, \boldsymbol{N}, \boldsymbol{\beta}, \boldsymbol{\phi})}$$

with $\phi_i^{new} = \sum_{k=1}^{K} a_{ik}\omega_{ki}, i = 1, \ldots, n$. In this case $\boldsymbol{a}_k$ and $\boldsymbol{\omega}_k$ are the weights and the tile-specific random effects corresponding to the new tessellation. The prior ratio is $A_m = 1$.

5.4.2 Prediction

We predict malaria risk at a set of unsampled locations $\boldsymbol{s}^{(0)} = (s_1^{(0)}, s_2^{(0)}, \ldots, s_l^{(0)})^T$ by using Bayesian kriging. In particular, we obtain estimates of the number of cases $\boldsymbol{Y}^{(0)} = (Y_1^{(0)}, Y_2^{(0)}, \ldots, Y_l^{(0)})^T$ at locations $\boldsymbol{s}^{(0)}$ from the predictive distribution

$$P(\boldsymbol{Y}^{(0)}|\boldsymbol{Y}, \boldsymbol{N}) =$$
$$\int P(\boldsymbol{Y}^{(0)}|\boldsymbol{\beta}, \boldsymbol{\omega}_k^{(0)}) P(\boldsymbol{\omega}_k^{(0)}|\boldsymbol{\omega}_k, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, K, \boldsymbol{c}_k) P(\boldsymbol{\beta}, \boldsymbol{\omega}_k, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, K, \boldsymbol{c}_k|\boldsymbol{Y}, \boldsymbol{N})\, d\boldsymbol{\beta}\, d\boldsymbol{\omega}_k^{(0)}\, d\boldsymbol{\omega}_k\, d\boldsymbol{\sigma}^2\, d\boldsymbol{\rho}\, dK\, d\boldsymbol{c}_k,$$

where $P(\boldsymbol{\beta}, \boldsymbol{\omega}_k, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, K, \boldsymbol{c}_k|\boldsymbol{Y}, \boldsymbol{N})$ is the posterior distribution and $\boldsymbol{\omega}_k^{(0)} = (\omega_{k1}^{(0)}, \ldots, \omega_{kl}^{(0)})^T$ is the vector of tile-specific random effects at new site $\boldsymbol{s}^{(0)}$. The distribution of $\omega_{ki}^{(0)}$ given $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kn})^T$ at observed locations is Normal

$$P(\omega_{ki}^{(0)}|\boldsymbol{\omega}_k, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, K, \boldsymbol{c}_k) = N(\Sigma_k^{01}(\Sigma_k^{11})^{-1}\boldsymbol{\omega}_k, \Sigma_k^{00} - \Sigma_k^{01}(\Sigma_k^{11})^{-1}(\Sigma_k^{01})^T),$$

with $\Sigma_k^{11} = E(\boldsymbol{\omega}_k\boldsymbol{\omega}_k^T)$ the covariance matrix built by including only the sampled locations $s_1, s_2, \ldots, s_n$, $\Sigma_k^{00} = E(\boldsymbol{\omega}_k^{(0)}\boldsymbol{\omega}_k^{(0)T})$ the covariance matrix formed by taking only the new locations $s_1^{(0)}, s_2^{(0)}, \ldots, s_l^{(0)}$ and $\Sigma_k^{01} = E(\boldsymbol{\omega}_k^{(0)}\boldsymbol{\omega}_k^T)$ describing covariances between unsampled and sampled locations. The location-specific random effect at new site $s_i^{(0)}$ is calculated as $\phi_i^{(0)} = \sum_{k=1}^K a_{ik}^{(0)}\omega_{ki}^{(0)}$, where $a_{ik}^{(0)}$ are decreasing functions of the distance between new location $s_i^{(0)}$ and the centroid of tile k. Conditional on $\phi_i^{(0)}$ and $\boldsymbol{\beta}$, $Y_i^{(0)}$ are independent Bernoulli variates $p(Y_i^{(0)}|\boldsymbol{\beta}, \phi_i^{(0)}) \sim Ber(p_i^{(0)})$ with $logit(p_i^{(0)}) = X_i^{(0)T}\boldsymbol{\beta} + \phi_i^{(0)}$, where $\boldsymbol{X}^{(0)}$ are environmental covariates at new locations $\boldsymbol{s}^{(0)}$.

5.5 Analysis of the malaria prevalence data

The performance of the spatial logistic model described in Section 5.3 is illustrated on malaria prevalence data presented in Section 5.2. The relation between the malaria risk and the environmental factors is not linear (Chapter 4). A preliminary non-spatial analysis was run to find the best combination and transformation of the environmental factors based on the AIC. The factors and their transformations included in the analysis were: Normalized Difference Vegetation Index (NDVI), NDVI squared, length of malaria season, amount of rainfall, maximum temperature, squared maximum temperature, minimum temperature, squared minimum temperature, distance to the nearest water body and squared distance to the nearest water body.

We used Metropolis-Hastings algorithm to sample from the full conditional distributions for

all model parameters. The starting values for the regression coefficients $\boldsymbol{\beta}$ were set equal to the estimates of the non-spatial logistic regression. The location-specific random effects ϕ_i, $i = 1, \ldots, n$ were all initialized with 0. The initial values for the spatial variances σ_k^2 were fixed to 0.1 and for the decay parameters ρ_k to 1.0, $k = 1, \ldots, K$. The probabilities Q_s, Q_b, Q_d, Q_m of the RJMCMC were set to 0.4, 0.2, 0.2 and 0.2 respectively. The parameter α of the prior distribution of K was fixed to 0.1 to obtain a non-informative prior.

We ran a single chain of $120,000$ iterations with a burn-in of $5,000$ iterations. Convergence was assessed by inspection of ergodic averages of selected model parameters. After convergence we collected a sample of size $1,000$ from the posterior distribution by taking every 10th value from the chain to avoid autocorrelation in the sample.

Table 5.1 shows the posterior estimates of the effects of environmental predictors. The environmental covariates significantly related to malaria risk were rainfall, maximum temperature and distance from the water. We found a negative association between malaria prevalence and the average amount of rainfall during the malaria transmission season. A similar negative association was observed between the average of maximum monthly temperature during the transmission season and the malaria risk. The coefficient corresponding to the distance from nearest water body indicates high malaria risk in areas away from water sources. A similar result was found by Gemperli (2003) in the analysis of MARA data from Mali.

The posterior frequency for the number of partitions is summarized in Figure 5.2. The most frequent tessellation favors two tiles, suggesting two spatial processes. Figure 5.3 displays the lower and upper quartiles and the median of the posterior distribution for the spatial covariance parameters σ^2 and ρ. The parameter ρ varies from 0.0154 to 0.2781, which in our exponential setting indicates that the minimum distance for which the spatial correlation is less than 5% varies between 10.79 and 194.81 kilometers.

Variable	Median	95% CI[a]
Intercept	0.65	(0.09, 1.3)
Log(NDVI)	0.63	(-0.47, 1.04)
Log(NDVI)^2	-0.19	(-0.53, 0.98)
Seson Length	0.51	(-0.1, 1.24)
Rainfall	-1.48	(-2.48, -0.004)
Maximum Temperature	-1.31	(-2.38, 0.73)
Maximum Temperature ^{2}	-0.03	(-2.27, -0.02)
Minimum Temperature	1.80	(-2.18, 2.71)
Minimum Temperature ^{2}	0.05	(-0.03, 0.1)
Distance to water	1.08	(0.46, 2.51)
Distance to water ^{2}	-0.29	(-0.56, 0.09)

[a] : Credible intervals (or posterior intervals).

Table 5.1: Posterior estimates for environmental covariate effects.

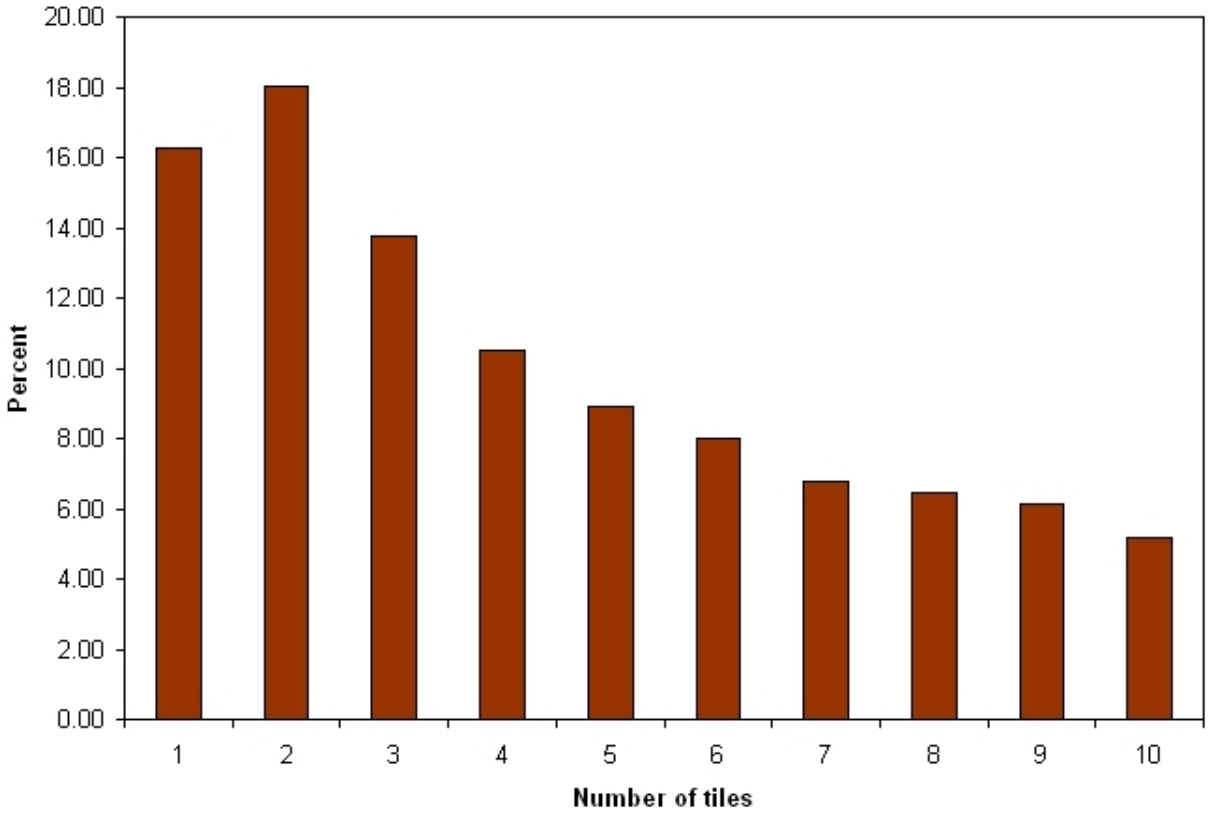

Figure 5.2: Posterior distribution of the number of tiles.

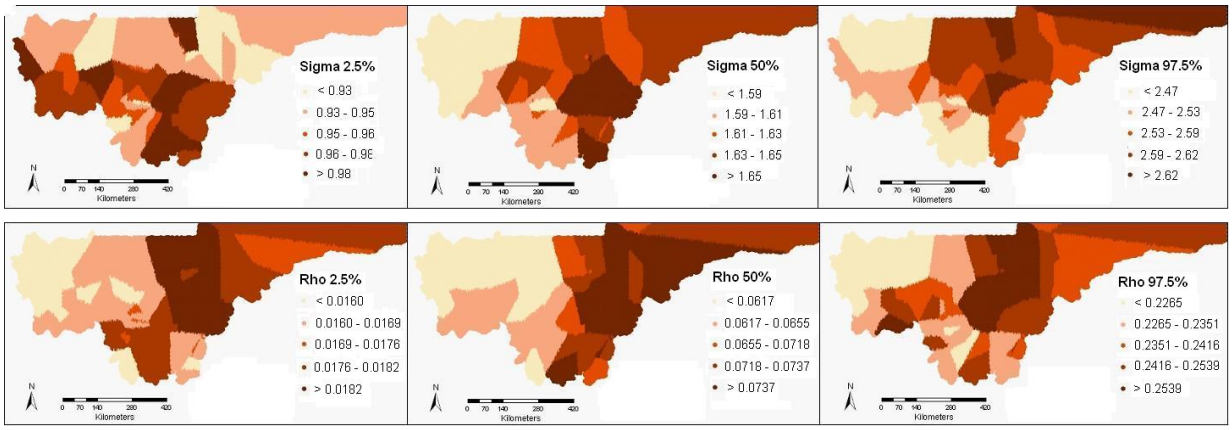

Figure 5.3: Percentiles (2.5th, 50th and 97.5th) of the posterior distribution for σ^2 and ρ.

Predictions of malaria risk at $60,000$ unsampled locations covering the whole area of sub-Sahara Mali were carried out using Bayesian kriging. The map of malaria prevalence is shown in Figure 5.4. High malaria prevalence was predicted in the west part of the country (Kayes region), a small area in the center of Mali (Koulikoro and Segou regions) and in the south-east part of Mali. Low levels of malaria risk were predicted in the north part of the country, the region at the border with Mauritania and the center of Koulikoro region. The prediction error is depicted in Figure 5.5. We observe that predictions have lower variances in areas around the data locations and the prediction error is higher in regions remote from the sampling locations.

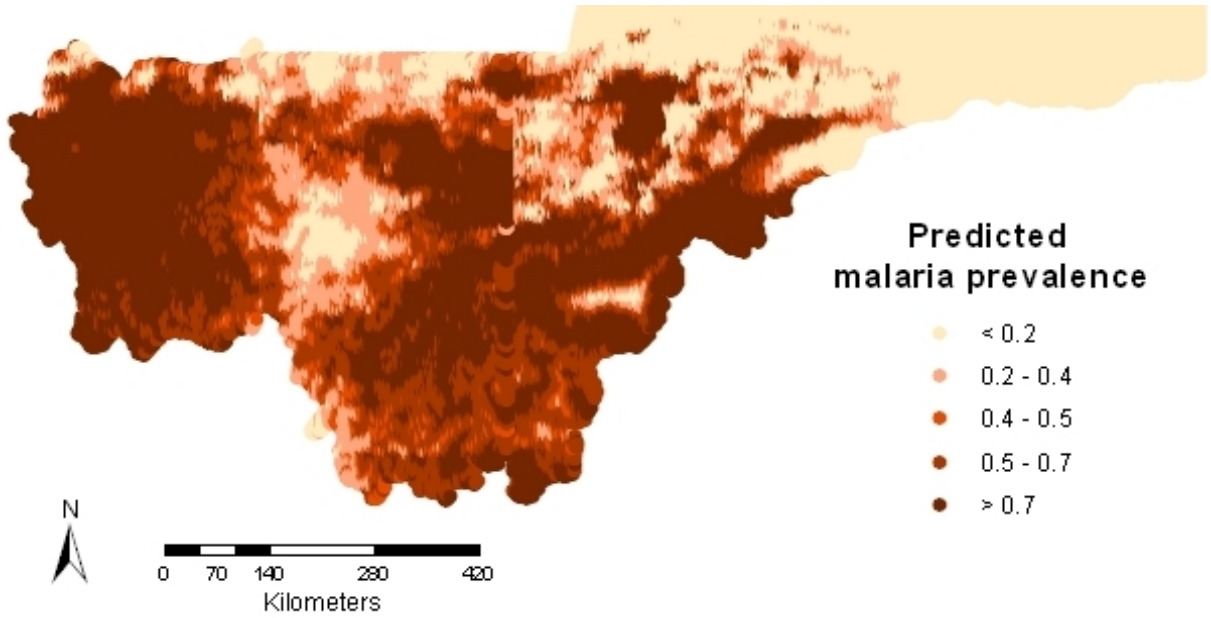

Figure 5.4: Map of predicted malaria prevalence for sub-Sahara Mali, estimated from the median of the posterior predictive distribution.

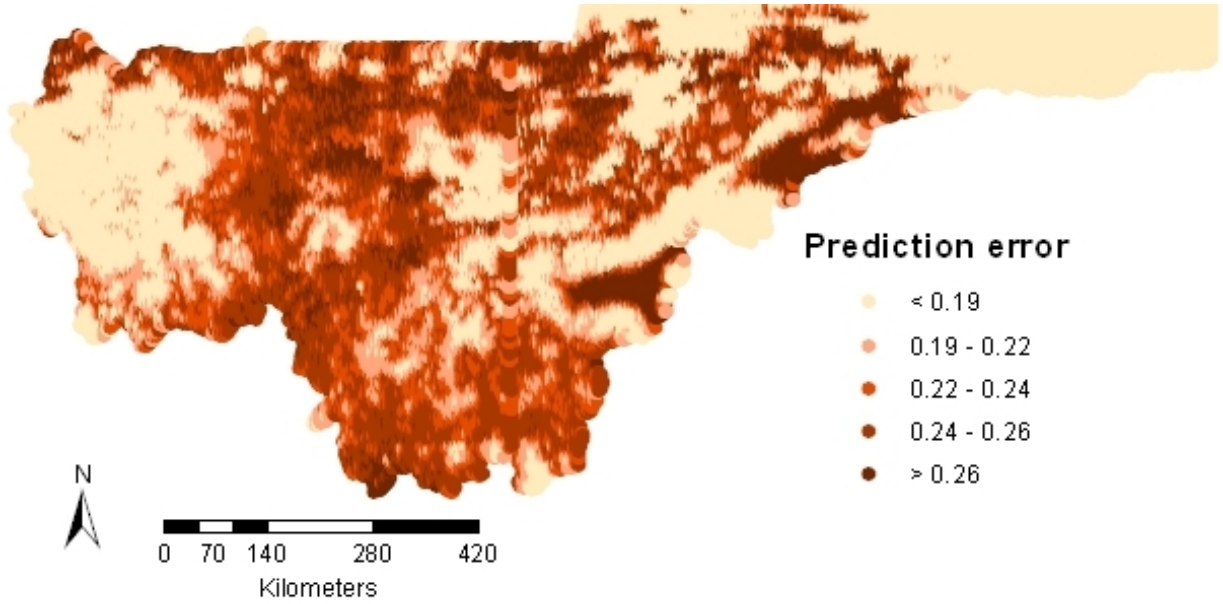

Figure 5.5: Map of prediction error for sub-Sahara Mali, estimated from the standard deviation of the posterior predictive distribution.

5.6 Discussion

In this paper we extended our earlier work (Gosoniu et al., 2006) which considered the non-stationary feature of malaria by dividing the area of interest in fixed number of regions, assumed a separate stationary spatial process in each region and correlation between the regions. That approach is more appropriate when the data analyzed cover a large area with a fixed space partitioning such as ecological zones defined by precipitation, evaporation and availability of water. Non-stationarity may also be explained by factors other than the environmental ones (socio-economic status, human activities etc.) which may influence the spatial correlation differently over the study area. However, in practical situations a fixed partitioning may not be obvious. In the present approach we let the statistical model capture the different spatial processes by allowing the number and the configuration of the regions to be random. The model parameter space has variable dimensions, depending on the tessellation, therefore we used RJMCMC for model fit. A random tessellation model has been employed also by Gemperli (2003) for mapping malaria in Mali. However, the authors assumed independence of the data across the regions. This assumption implies that neighboring points located in different tiles are not correlated. Our contribution was to develop a random tessellation model which takes into account the between tile correlation.

In Gosoniu et al. (2006) the model validation results between a stationary and a fixed tile-based non-stationary model were reported, showing that malaria mapping is sensitive to these model assumptions. Ignoring the non-stationarity characteristic may lead to unreliable estimates of the risk factors' effects on malaria and to inaccurate prediction estimates of malaria prevalence. In addition, the spatial correlation gives an indication of the importance of geographically structured factors as well as of the unmeasured local factors, such as human behavior.

The malaria map based on the non-stationary model developed in this paper shows overall, similar risk patterns as the previous two maps that allowed for non-stationary spatial covariance structure (Gemperli, 2003; Gosoniu et al., 2006), as well as significant differences. In particular, all three approaches estimated high malaria risk (> 0.7) in the region of Kayes and low prevalence (< 0.2) in the region of Tombouctou. However, our map indicated larger areas with high parasitaemia risk, as well as certain areas with smaller risk compared with the other two approaches. In particular, in the south of Koulikoro region and the west of Mopti region we predicted higher parasitaemia risk that the other two methods. Similarly, at the border with Burkina Faso we predicted malaria prevalence

higher than 0.7 in our case, while Gemperli (2003) and Gosoniu et al. (2006) estimated a prevalence of $0.4 - 0.7$. Our map shows lower level of malaria prevalence at the border with Mauritania and in the region of Gao compared with the map based on the approach developed by Gemperli (2003). The malaria risk map produced by the current model was validated by expert opinion who suggested that the estimates obtained reflect better the malaria situation on the ground. The maps indicate that the assumptions involved in modeling non-stationarity influence the resulting maps.

Previous approaches that modeled non-stationarity (Gemperli, 2003; Kim et al., 2005) assumed independence between neighboring points located in different tiles. We addressed this issue by considering correlation between the random tiles. Although our model methodologically improves the previous modeling approaches, it has larger number of parameters because at each location it includes as many random effects as the number of tiles. Further research on simulated data is required to show if the more complex model is able to capture better the spatial processes over the study region than the simpler model, with fewer parameters but stronger assumptions.

Fitting geostatistical models for non-Gaussian data involves repeated inversion of the spatial covariance matrices and, for large number of locations the operation becomes computationally intensive. The previous methods that assumed tile-independence have the advantage that facilitates the matrix inversion by converting the spatial covariance matrix into block diagonal form. Our model requires inversion of the spatial covariance matrix as many times as the number of tiles. To overcome this computational challenge one could approximate the tile-specific spatial processes by sparse Gaussian processes (Seeger et al., 2003; Snelson and Ghahramani, 2006), estimating the spatial processes from a subset of $m << n$ locations within each tile. In this way the problem would be reduced to inversion of much smaller matrices mxm, instead of the original nxn matrices.

We are currently further developing this approach by assuming that the relation between the environmental factors and the malaria risk is tile specific.

Acknowledgments

The authors would like to thank the MARA collaboration for making the malaria data available. This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

Chapter 6

Mapping malaria using mathematical transmission models

Gosoniu L.[1] ,Vounatsou P.[1], Riedel N.[1], Sogoba N.[2], Miller J.[3],Steketee R.[3], Smith T.[1]

[1] Swiss Tropical Institute, Basel, Switzerland
[2] Malaria Research and Training Center, Universite du Mali, Bamako, Mali
[3] The Malaria Control and Evaluation Partnership in Africa, Zambia

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

Summary

Historical malaria field survey data used for mapping the geographical distribution of the disease have a number of limitations that make their analyzes challenging. The surveys are carried out at different seasons with non-standardized and overlapping age groups of the population. To overcome these problems, we propose the use of a newly developed mathematical malaria transmission model to translate the heterogeneous age prevalence data into a common measure of transmission intensity like entomological inoculation rate. This approach was applied on malaria data extracted from the Mapping Malaria Risk in Africa (MARA) database for Mali. A Bayesian geostatistical model was fitted to the estimates of transmission intensity and using Bayesian kriging we produced a smooth map of annual entomological inoculation rates in Mali. This map was converted to age specific malaria risk maps using again the transmission model. Model validation revealed that the geostatistical model based on the estimates derived from the transmission model had a better predictive ability compared to the one modeling the raw prevalence data. To further assess the malaria risk maps based on the transmission model, data from the nationwide Malaria Indicator Survey (MIS) in Zambia were analyzed. Maps based on both the transmission model and the raw prevalences were compared. Results showed that the map based on the transmission model predicts similar patterns of malaria risk with the one obtained by analyzing directly the MIS data. The proposed approach in malaria mapping has the advantages that i) age and seasonality adjusted malaria risk maps can be obtained from prevalence data compiled from different sources ii) all survey data can be used in mapping despite the age-heterogeneity and iii) maps of different malaria transmission measures can be produced from survey data.

Keywords: Bayesian kriging, entomological inoculation rate, malaria mapping, Markov chain Monte Carlo, parasite prevalence.

6.1 Introduction

Mapping the geographical patterns of malaria risk and estimates of population at risk are important for planning, implementation and evaluation of malaria control programs. They provide important information for optimizing the resource allocation for malaria control in high risk areas. The national and global estimates of burden of disease are imprecise because of the inadequate malaria case reporting in most endemic countries as well as the lack of national wide malaria surveys. Consequently there is a renewed effort in mapping malaria risk (Hay et al., 2006; WHO, 2007) with the aim of producing baseline maps of malaria transmission.

Most empirical maps of malaria in sub-Sahara Africa are based on field survey data on prevalence of infection. Todate the main sources for these data are published and unpublished reports. The compiled databases have a number of problems which complicate their analyzes. First, the surveys are carried out in different seasons,hence it is difficult to account for seasonality in modeling malaria transmission. Second, the population covered by different surveys has non-standardized and overlapping age groups, therefore adjusting for age could be challenging.

The most complete database on malariometric data across Africa is the "Mapping Malaria Risk in Africa" (MARA/ARMA, 1998). It contains over $10,000$ distinct age-specific prevalence values since the early 1960s. The advantage of analyzing historical data, in particular data extracted from the MARA database is that they provide a wealth of information in assessing temporal changes of malaria. However, the derived malaria maps may not indicate the actual situation of malaria at a specific location, which could be affected by control interventions or human activities (e.g. irrigation, dams). Most of the analysis done so far on MARA data (Kleinschmidt et al., 2001; Gemperli et al, 2003; Gosoniu et al., 2006) were concentrated on a specific age group, discarding surveys with overlapping age groups of the population. In this way much of the data are unused, resulting in unreliable malaria transmission estimates in areas with sparse data.

A new source of malaria data is Malaria Indicator Surveys (MIS), developed by Roll Back Malaria (RBM) in 2004 for monitoring coverage of malaria prevention and treatment. The MIS package is a stand-alone survey tool that can be used to collect malaria-related data in countries that are lacking such data for malaria program management. The Zambia National MIS is the first nationally representative household survey assessing coverage of malaria interventions and malaria-related burden in children under 5 years of age during

May-June 2006. Similar surveys have been conducted in Angola, Mozambique, Ethiopia and Senegal.

To overcome the problems related to the MARA database, Gemperli et al. (2005; 2006) employed the Garki malaria transmission model (Dietz et al, 1974) to convert observed prevalence data into an estimated age-independent entomological measure of transmission intensity, which was further used for mapping purposes. This work demonstrated the feasibility of using malaria transmission models in malaria risk mapping. However, the Garki model was developed on field data from the savanna zone of Nigeria, therefore it cannot be generalized to other regions in Africa with different environmental conditions and levels of malaria endemicity. Recently, the modeling group of the Swiss Tropical Institute developed a new malaria transmission model (Smith et al., 2006) which overcomes a number of limitations of the Garki model. This is an individual-based stochastic model which simulates age-specific malaria epidemiological outcomes (i.e. parasitaemia, morbidity, mortality) at a given location, conditional on the seasonal pattern in the entomological inoculation rate (EIR). In addition, a seasonality model has been developed (Mabaso, 2007) to estimate from climatic factors monthly EIR due to *Anopheles gambiae s.l.* at any location in Africa.

We employed this mathematical model to map malaria survey data from Mali, adjusting for age and seasonality. In particular, we translated the observed MARA prevalence data into estimates of EIR. Using environmental and climatic data from remote sensing (RS) as predictors we fitted a Bayesian geostatistical model on the estimates of EIR. We further employed Bayesian kriging to obtain a smooth map of EIR for Mali. Applying again the mathematical model we converted the predicted EIR values into estimates of malaria prevalence for children less than 5 years old. We assessed the predictive ability of the geostatistical intensity model in malaria mapping and compare it with the model that analyzed directly the raw prevalence data.

To further assess the malaria risk maps based on the transmission model, we analyzed the Zambia MIS data by employing the transmission model as well as modeling directly the prevalence data and compare the resulting maps. The Zambia MIS data were previously analyzed by Riedel et al. (unpublished) by fitting Bayesian geostatistical models. Bayesian P-splines (Eilers and Marx, 1996) were employed within a geostatistical model to take into account the non-linear relation between parasitaemia risk and environmental factors in the analyzes of Mali and Zambia data.

6.2 Materials and methods

6.2.1 Malaria data

Malaria prevalence data for Mali were extracted from the MARA/ARMA database. We analyzed data from 497 surveys carried out at 115 distinct locations between 1962 and 2001. The surveys covered different age groups of the population (Table 6.1). During these surveys, 104,689 individuals were examined and a proportion of 46.67% were found with *P. falciparum* parasites in the blood sample.

Age groups (years)	Nb. of surveys	Age groups (years)	Nb. of surveys
0-9	13	6-99	1
0-12	6	14-32	1
0-15	3	15-99	1
0-44	9	0-5 & 6-10	4
0-99	1	0-9 & 0-20	3
1-1	1	2-4 & 5-9	17
1-9	2	2-9 & 10-99	22
1-12	2	5-9 & 10-14	30
1-15	1	8-14 & 15-19	8
1-70	1	0-1 & 2-4 & 5-9	9
2-9	47	1-2 & 3-5 & 6-10	30
2-15	3	1-15 & 5-9 & 6-14	7
5-9	6	2-4 & 5-9 & 10-14	30
5-15	1	2-4 & 5-9 & 10-15	33
6-14	3	1-1 & 2-4 & 5-9 & 10-14	180
6-18	4	0-4 & 0-12 & 2-9 & 5-9 & 10-18	14
6-20	1	1-4 & 2-4 & 5-9 & 10-14 & 15-19 & 15-24	8

Table 6.1: Age groups of the population included in the MARA surveys in Mali between years 1962-2001.

Data on malaria prevalence in Zambia are available from the Zambia National Malaria Indicator Survey conducted during May-June 2006. The sample frame of the survey was the list of 17,000 Standard Enumeration Areas (SEAs) from the 2000 Population Census. The survey was carried out on a sample of 120 SEAs with 25 households in every SEA, including 2,364 children under 5 years of age. Malaria prevalence data were collected by finger prick blood sampling in children under five. The coordinates of the households were recorded using personal digital assistants (PDAs). The data were aggregated at SEA level and we calculated the coordinates of SEAs as the mean of the households coordinates belonging to each SEA. After eliminating the children with incomplete information, the final dataset analyzed here was collected at 109 distinct SEAs and comprised 1,324 children.

6.2.2 Environmental data

In the analyzes of malaria data in Mali we used the following environmental and climatic variables: normalized difference vegetation index (NDVI), soil water storage index (SWS), rainfall, temperature, distance to the nearest water body and land use. The predictors used in the analyzes of MIS data from Zambia were: NDVI, rainfall, temperature, potential evapotranspiration (PET) and land use. The climatic and environmental data used in this analysis were obtained from different sources and at different resolutions for both Mali and Zambia (Table 6.2 and Table 6.3).

The land cover data downloaded for Mali included 24 categories. We regrouped them in the following 3 classes: urban/ dry/ barren/ sparsely vegetated (land use 1), crop/ grassland/ mosaic(land use 2) and wet/ irrigated crop land/ savanna (land use 3). We created a two kilometers buffer area around each data point and calculated the relative frequencies of the land use classes in each buffer. Every land use class is considered as a separate variable with values between 0 and 1. In Zambia we obtained the land cover in 16 categories which we regrouped into 5 broad classes, namely: wet area, forest, urban, shrub land and other. Similarly to Mali, we created for every land use class a variable corresponding to the relative frequency of the pixels of the different land use categories inside a two kilometers buffer.

Factor	Resolution	Source
Normalized Difference Vegetation Index (NDVI)	$8km^2$	NASA AVHRR Land data sets
Soil Water Storage Index (SWS)	$5km^2$	Droogers et al., 2001
Rainfall	$5km^2$	Hutchinson et al., 1996
Temperature	$5km^2$	Hutchinson et al., 1996
Water bodies	$1km^2$	World Resources Institute, 1995
Land use	$1km^2$	USGS-NASA

Table 6.2: Spatial databases used in the spatial analysis in Mali.

Factor	Resolution	Source
Normalized Difference Vegetation Index (NDVI)	$250m^2$	MODIS
Rainfall	$8km^2$	ADDS
Temperature	$1km^2$	MODIS
Evapotranspiration (PET)	$8km^2$	ADDS
Land use	$1km^2$	MODIS

Table 6.3: Spatial databases used in the spatial analysis in Zambia.

6.2.3 Statistical analysis

Malaria transmission models

The transmission model of Smith et al. (2006) simulates malaria epidemiological outcomes (i.e. parasitaemia, morbidity, mortality) at a specific location, conditional on the seasonal pattern in the EIR. Treating EIR as a known quantity, but allowing for temporal variations and the influence of age host, infections are introduced into a population of simulated humans via a stochastic process dependent on the EIR and then sample the subsequent parasite densities using 5-day time steps. Multiple field datasets across Africa have been used to optimize parameter estimates. A detailed description of the model is given in Smith et al. (2006) and Smith et al. (unpublished).

Geostatistical models

Fitting malaria transmission intensity data

A Bayesian geostatistical model was fitted on the Box Cox transformed EIR values Y_i estimated from the transmission model at each location i, $i = 1, \ldots, n$. We assume that Y_i are normally distributed, $Y_i \sim N(\mu_i, \tau^2)$ and introduce the environmental covariates on the mean structure μ_i. To overcome the non-linearity in the relation between environmental predictors and malaria transmission we employed a non-parametric regression model using P-splines. Details on the implementation of this approach are given in Chapter 4. In brief, μ_i is modeled as $\mu_i = \sum_{j=1}^{p} f_j(X_{ij}) + \phi_i$, where $f(\cdot)$ is a smooth function of the environmental predictors, that is $f_j(X_{ij}) = \sum_{k=1}^{K} u_{kj}|X_{ij} - s_{kj}|^3$, where $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{Kj})^T$ is the vector of regression coefficients, $s_{1j} < s_{2j} < \ldots < s_{Kj}$ are fixed knots and $|X_{ij} - s_{kj}|^3$ is a truncated 3-rd order polynomial spline basis, all related to covariate X_j. We consider the knots to be based on sample quantiles of the covariates.

The parameters ϕ_i are location-specific random effects modeling the spatial correlation by assuming that they derive from a multivariate normal distribution $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T \sim MVN(\mathbf{0}, \Sigma)$ with the variance-covariance function related to an exponential correlation function between locations. In particular, $\Sigma_{ij} = \sigma^2 exp(-d_{ij}\rho)$, where d_{ij} is the Euclidean distance between locations i and j, σ^2 captures within locations spatial variation and is known as the sill and ρ is called the decay parameter and measures the rate of decrease of correlation with distance. In this exponential setting, the decay parameter is translated into the range parameter, that is the minimum distance for which the spatial correlation is less than 5%, by calculationg $3/\rho$. Markov chain Monte Carlo (MCMC) simulation techniques were employed to estimate the model parameters. We used Bayesian kriging to predict Box Cox transformed EIR values at locations where malaria data are not available and produce smooth maps of EIR in Mali and Zambia.

We applied the relations estimated by the malaria transmission model and transformed back the predicted EIR values to age related malaria prevalence. We produced maps of malaria risk for children under 5 years old and $1 - 10$ years old for Mali and only for children under 5 in Zambia. The software used in this analysis was written by the authors in Fortran 95 (Compaq Visual Fortran Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.). Further details on the Bayesian modeling approach are given in the Appendix of this chapter.

Fitting malaria prevalence data

We also analyzed directly the malaria survey data from Zambia, by fitting a Bayesian geostatistical model, using as predictors the environmental factors mentioned in Section 6.2.2. Let N_i be the number of children screened at location i and Y_i the number of children found positive to *P. falciparum* parasitaemia. We assume that Y_i arises from a Binomial distribution, that is $Y_i \sim Bin(N_i, p_i)$ with parameter p_i measuring malaria risk at location i. Similarly to the previous section we assume a non-linear environment-malaria relation and model it using P-splines via the logistic regression, that is $logit(p_i) = \sum_{j=1}^{p} f_j(X_{ij}) + \phi_i$. The spatial dependency is modeled on the covariance matrix of the location-specific random effects. We assumed that the covariance $Cov(\phi_i, \phi_j)$ between every pair (ϕ_i, ϕ_j) decreases with their distance d_{ij} and, like in the previous geostatistical model, we choose an exponential correlation function. More details of this approach are given in the Appendix of this chapter.

Model validation

Using the malaria dataset from Mali, we assessed the predictive abilities of both the model that analyzed directly the raw prevalence data as well as the one that models transmission intensity data derived from the mathematical model. Surveys carried out in children between $1-10$ years old were selected to model directly the prevalence data. Model fit was implemented in 85% of the survey locations and validation was performed in the remaining ones. The geostatistical model based on the estimates derived from the transmission model was fitted on the entire MARA dataset, except for the test locations used for validation.

Following the approaches developed in Gosoniu et al. (2006), the predictive ability of the two models was assessed using Bayesian "p-values" and the probability coverage of the shortest credible interval. In particular, for each test location we calculated the area of the predictive posterior distribution which is more extreme than the observed data. The model predicts well the observed data when the median of the posterior predictive distribution is close to the observed data. We assert that the model with the best predictive ability is the one with the "p-value" closer to 0.5. In addition we calculated 12 credible intervals of the predictive posterior distribution at the test locations with probability coverage equal to 5 %, 10 %, 20 %, 25 %, 30 %, 40 %, 50 %, 60 %,70 %, 80 %, 90 % and 95 %, respectively and compared the proportions of test locations with observed malaria prevalence falling in the above intervals.

6.3 Results

The Mali MARA dataset contained 330 surveys conducted in children between $1-10$ years old at 87 distinct locations. The geostatistical model based on the raw prevalence data was fitted on randomly selected 317 surveys at 74 distinct locations (training set). Surveys at the remaining 13 locations were used for validation (test points). The geostatistical model based on the transmission intensity data was fitted on 476 surveys over 102 locations, that is on the entire MARA data set except the 13 test locations.

Each box-plot in Figure 6.1 summarizes the distribution of the 13 Bayesian "p-values" calculated from the predictive posterior distribution of the test locations for the model that directly analyzed the prevalence data (left) and for the approach that modeled the transmission intensity data derived from the mathematical malaria transmission model (right). The median of this distribution for the latter model in closer to 0.5, suggesting that this is the best model.

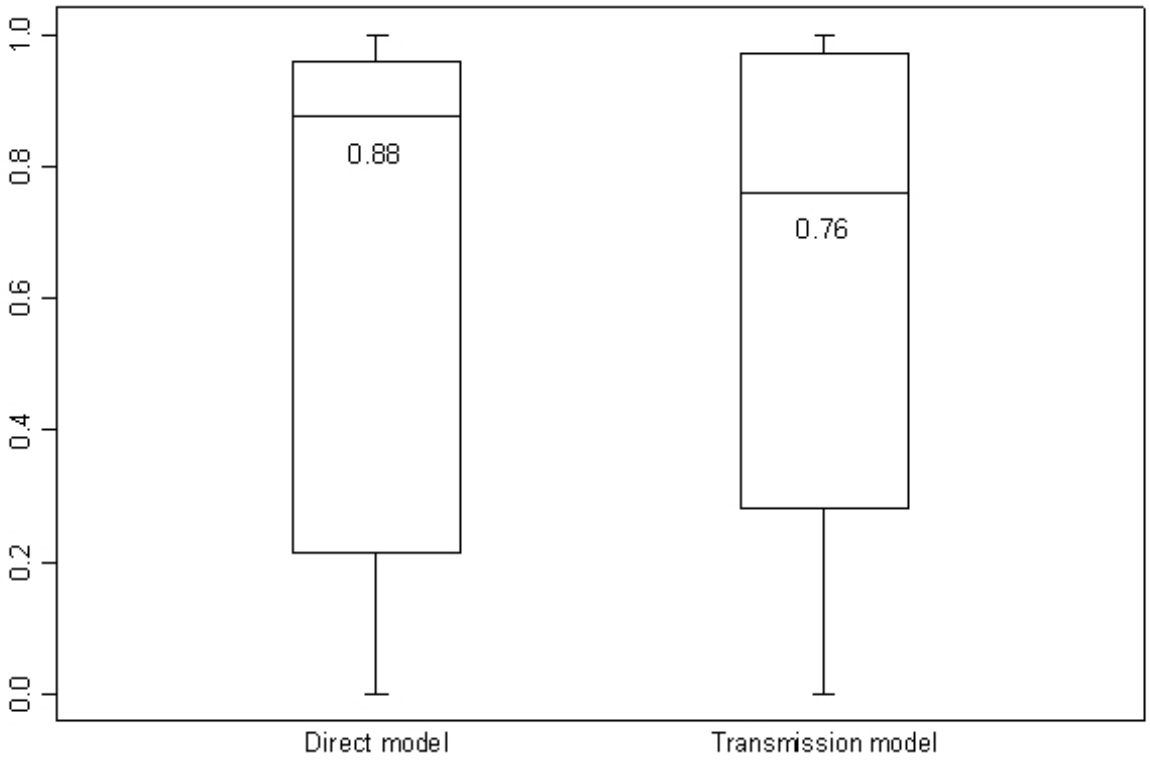

Figure 6.1: The distribution of Bayesian p-values. The box plots display the minimum, the 25^{th}, 50^{th}, 75^{th} and the maximum of the distribution.

Figure 6.2 shows the percentages of test locations with malaria prevalence falling in each of the 12 credible intervals of the posterior predictive distribution for both geostatistical models. We observe that the wider credible interval (95%) includes the same percentage of test locations (69.23%) for both models, but for the credible intervals ranging from 50% to 90% the approach based on the transmission model include, consistently, the highest percentage of observed locations.

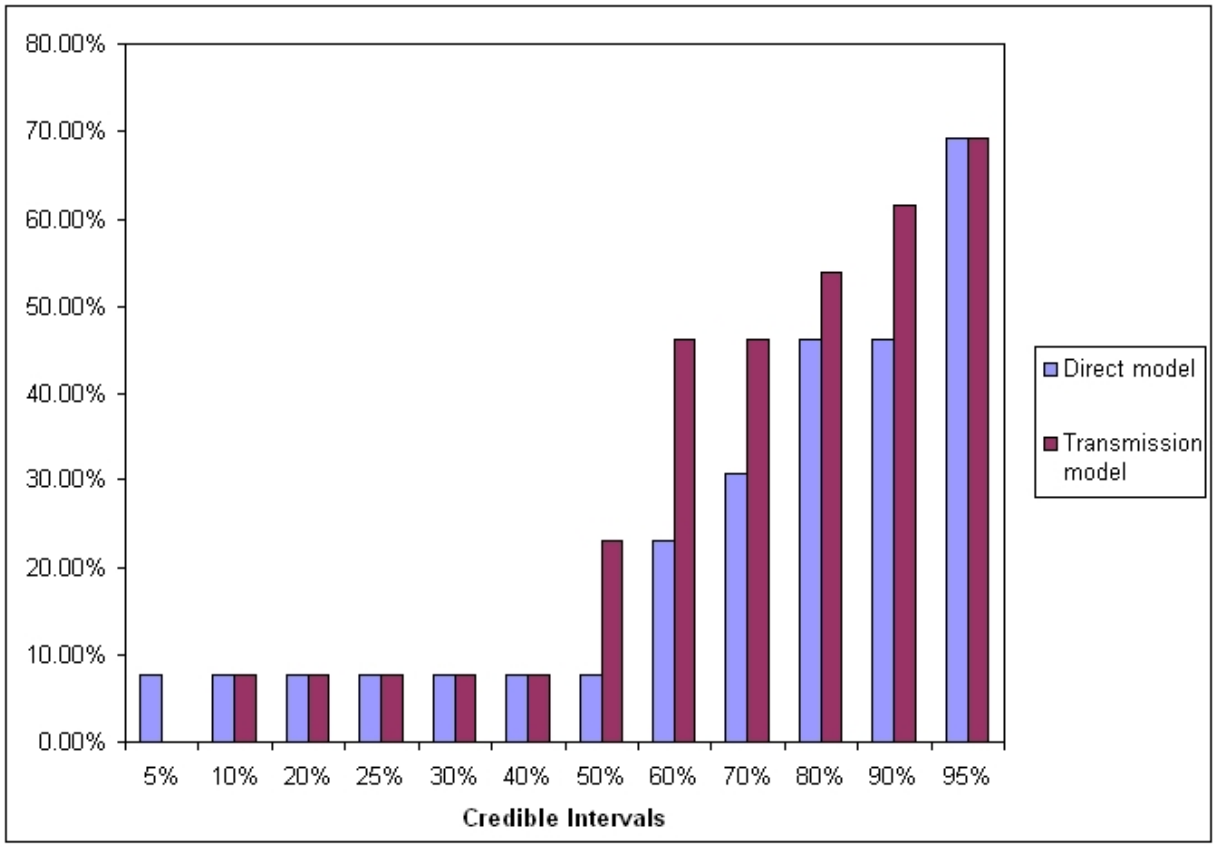

Figure 6.2: Percentage of test locations with malaria prevalence falling in the 5%, 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 95% credible intervals of the posterior predictive distribution.

The nonlinear effects of the environmental factors on malaria transmission in Mali and Zambia are shown in Figure 6.3 and Figure 6.4, respectively. The graphs show the posterior median and the 95% credible intervals. In Mali, the credible intervals of the posterior medians for all the variables contain 0, therefore none environmental factor was significantly associated with annual EIR. In Zambia, the only credible interval that did not include zero was the one corresponding to the posterior median of the evapotranspiration,

therefore we conclude that this was the only variable accounting for significant spatial variation in malaria transmission in this dataset. We notice a slightly decreasing effect of evapotranspiration on annual EIR.

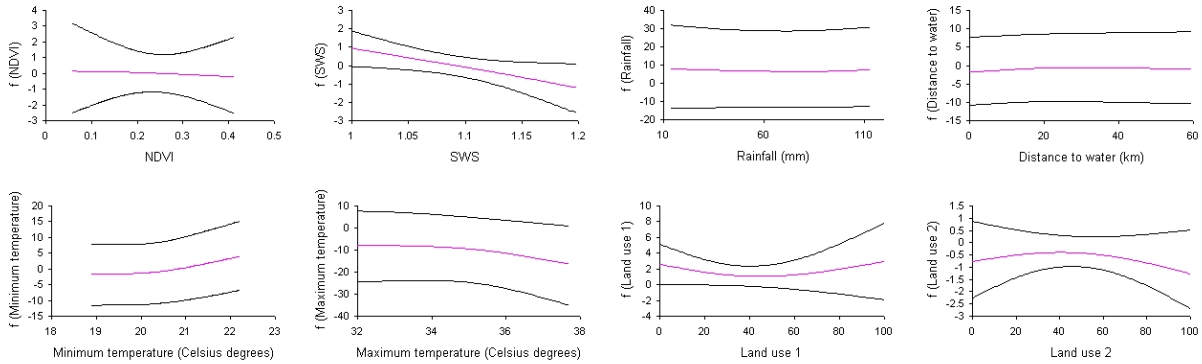

Figure 6.3: Estimated effect (P-spline) of environmental factors on EIR in Mali. The posterior median (pink) and the 95% credible interval are shown.

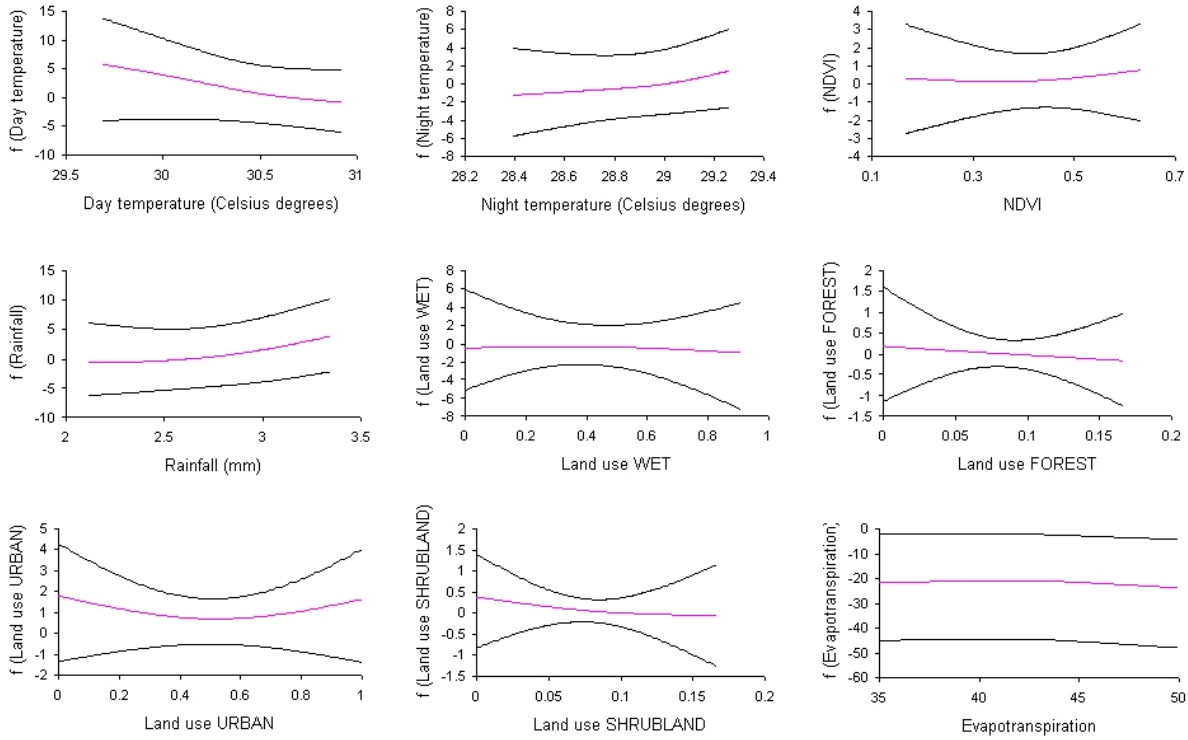

Figure 6.4: Estimated effect (P-spline) of environmental factors on EIR in Zambia. The posterior median (pink) and the 95% credible interval are shown.

Table 6.4 shows the posterior estimates of the spatial parameters: decay parameter, spatial variance and residual non-spatial variance. In Mali, the posterior median of the decay parameter ρ was equal to 2.60 km (95% credible interval: $0.73, 8.01$), which translates to a minimum distance where correlation drops below 5% of around 1.15 km (95% credible interval: $0.37, 4.14$). The posterior distribution of ρ had the median equal to 0.11 km (95% credible interval: $0.04, 0.59$) in Zambia, corresponding to a range of 28.52 km (95% credible interval: $5.10, 68.45$). Estimates of the range parameters suggest a weak spatial correlation in Mali and a strong spatial correlation in Zambia. In Mali, the spatial variation was slightly smaller ($\sigma^2 = 1.74$) than the residual non-spatial variation ($\tau^2 = 2.62$). The opposite situation was in Zambia, where the spatial variation was very high ($\sigma^2 = 7.23$) compared to the non-spatial variation ($\tau^2 = 0.11$).

Country	Spatial parameter	Median	95% CI[a]
Mali	ρ	2.60	(0.73, 8.01)
	σ^2	1.74	(0.98, 2.92)
	τ^2	2.62	(2.27, 3.03)
Zambia	ρ	0.11	(0.04, 0.59)
	σ^2	7.23	(2.84, 10.90)
	τ^2	0.11	(0.01, 3.95)

[a] : Credible intervals (or posterior intervals).

Table 6.4: Posterior estimates of spatial parameters.

The smooth map of predicted annual EIR in sub-Saharan Mali is shown in Figure 6.5. It depicts high malaria transmission in the region of Kayes (west of Mali), in the central part of Gao region and in the east of Mopti region (at the border with Burkina Faso). Low transmission of malaria is predicted at the border with Mauritania, east side of region Mopti, south of Gao region and some parts in the region of Sikasso (south of Mali). The corresponding prediction error is shown in Figure 6.6. The error is very low near the sampling locations and it increases toward the north of Mali, where malaria surveys are very sparse.

We employed the malaria transmission model and converted the predicted EIR values to malaria prevalence for children under 5 years old and for children 1 to 10 years old. The two malaria risk maps for Mali are shown in Figure 6.7 and Figure 6.8. Both maps depict

a pattern similar with the predicted annual EIR map, with malaria risk for children 1 to 10 years old uniformly higher than for children under 5 years old.

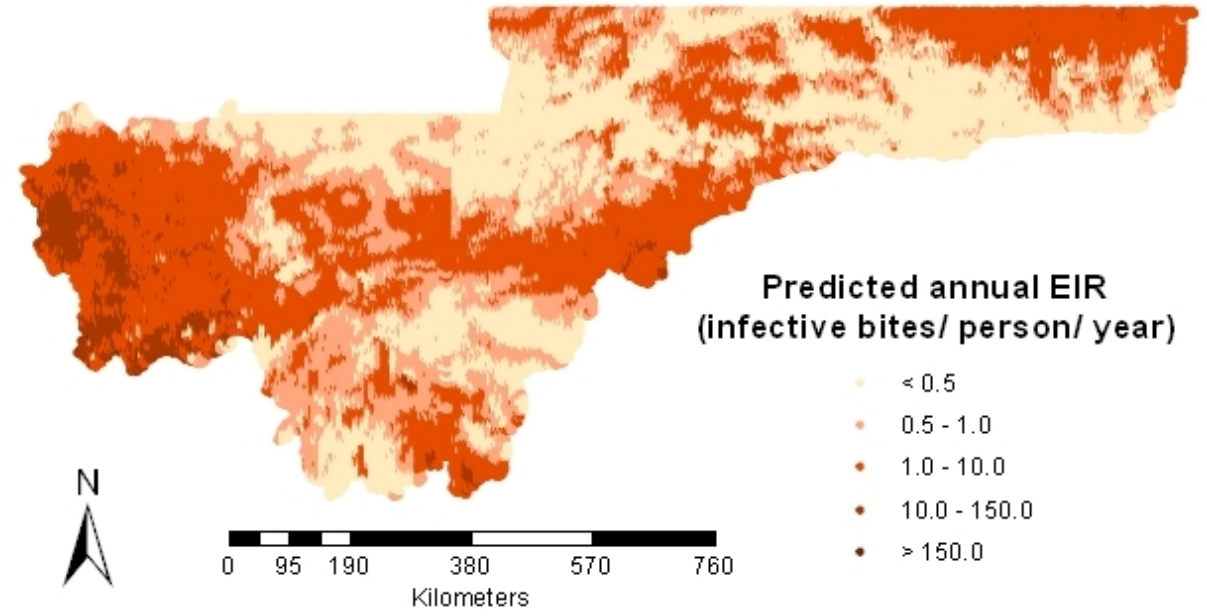

Figure 6.5: Predicted annual entomological inoculation rate (EIR) in Mali estimated from the median of the posterior predictive distribution.

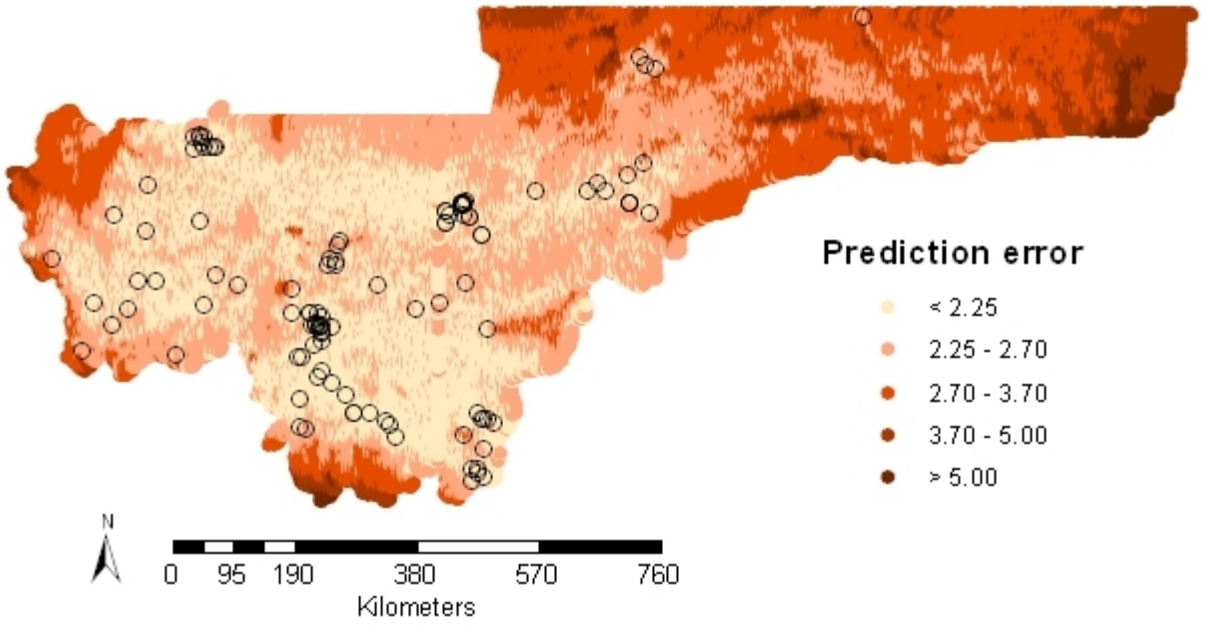

Figure 6.6: Prediction error of annual entomological inoculation rate (EIR) in Mali estimated from the standard deviation of the posterior predictive distribution. Sampling locations are indicated by circles.

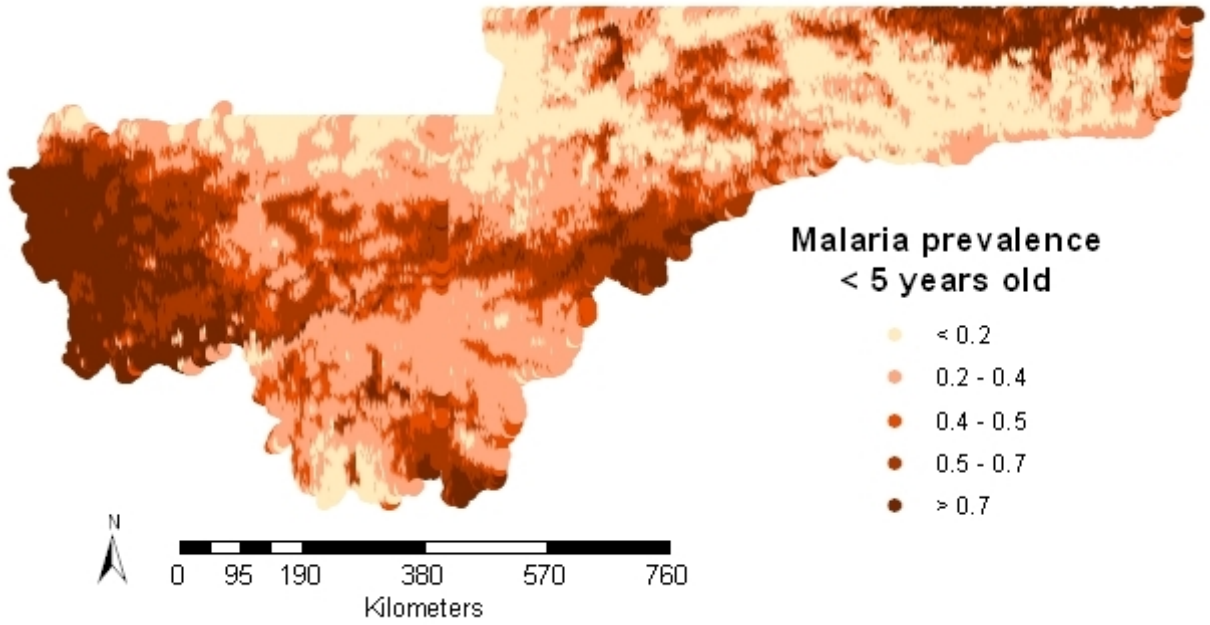

Figure 6.7: Predicted malaria prevalence for children under 5 years old in Mali estimated from the median of the posterior predictive distribution.

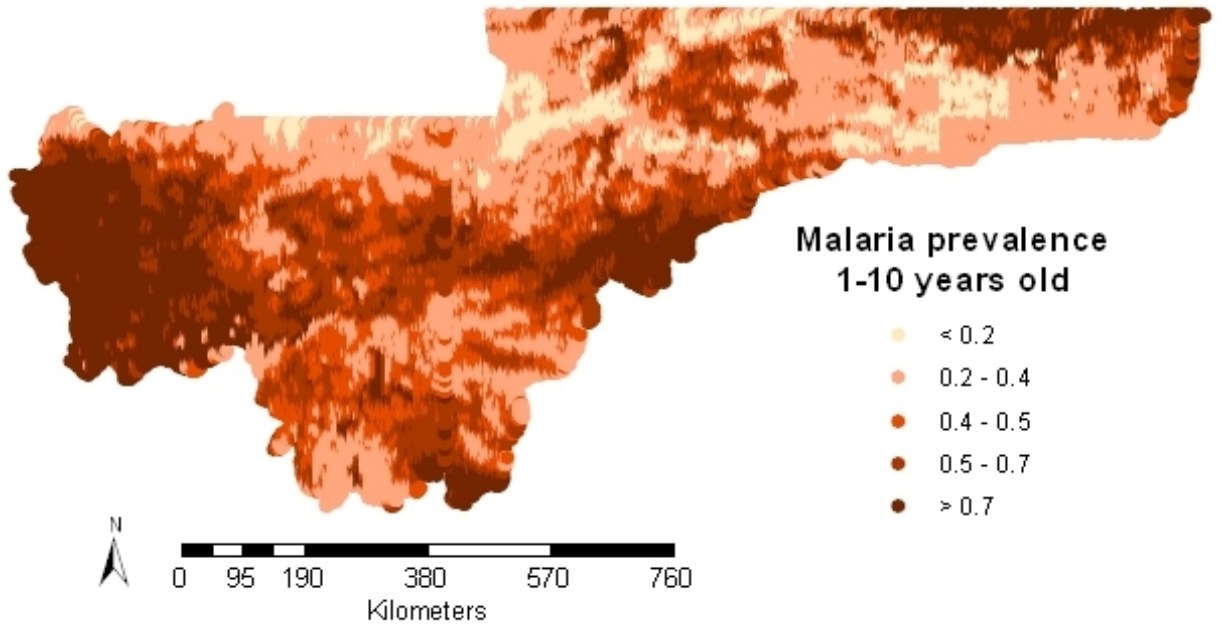

Figure 6.8: Predicted malaria prevalence for children between 1 and 10 years old in Mali estimated from the median of the posterior predictive distribution.

Figure 6.9 depicts the predicted annual EIR in Zambia. Overall, the model predicts low level of malaria transmission. The highest level of transmission is observed in the eastern part of the country and some regions in the northern part. The lowest malaria transmission was predicted in the north-west of Zambia and a small area in the south of the country. Figure 6.10 presents the prediction error of Zambia model. We observe higher error in the north-west and southern part of the country, at locations remote from the survey locations and low prediction error in areas around the sampling locations.

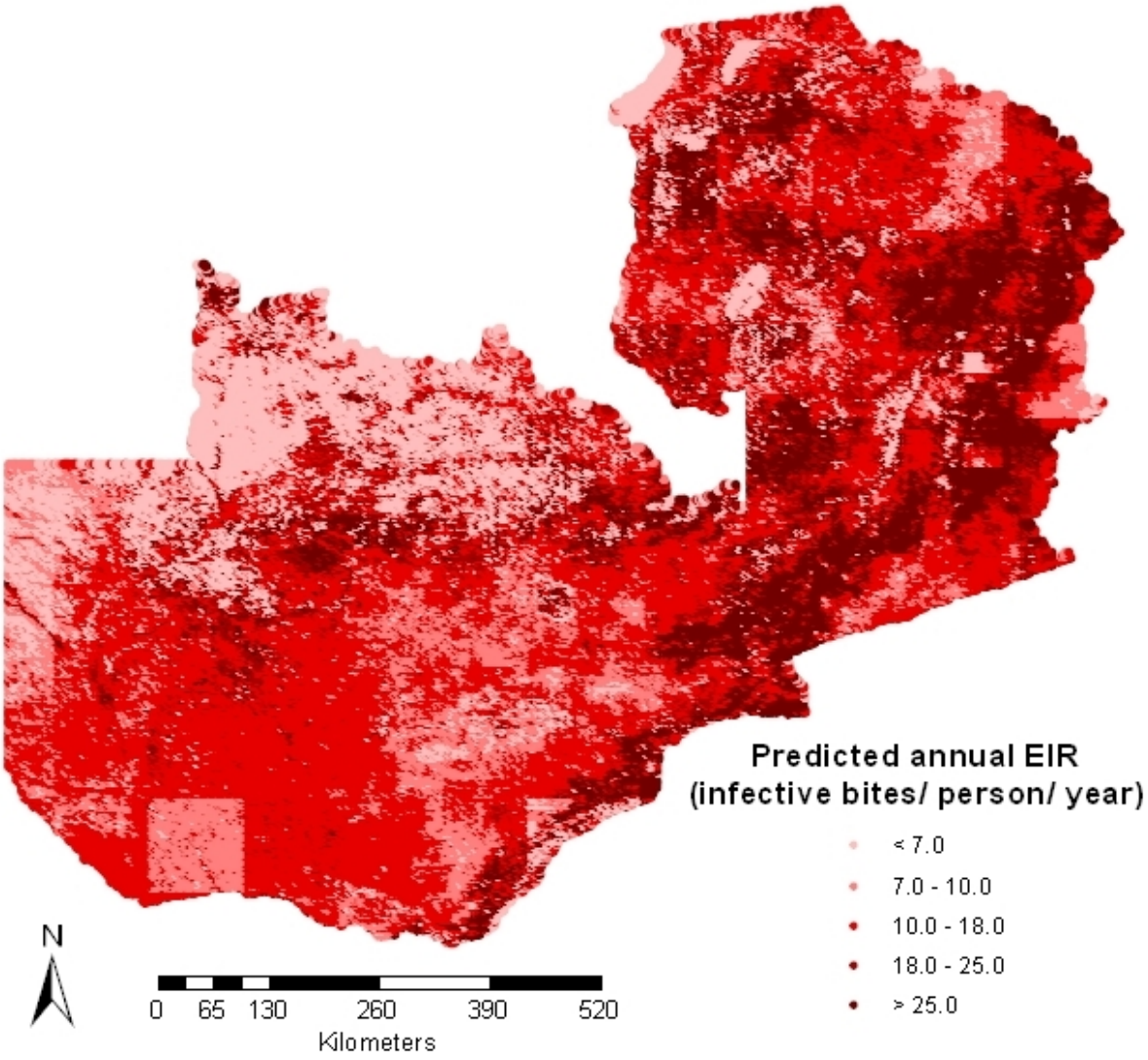

Figure 6.9: Predicted annual entomological inoculation rate (EIR) in Zambia estimated from the median of the posterior predictive distribution.

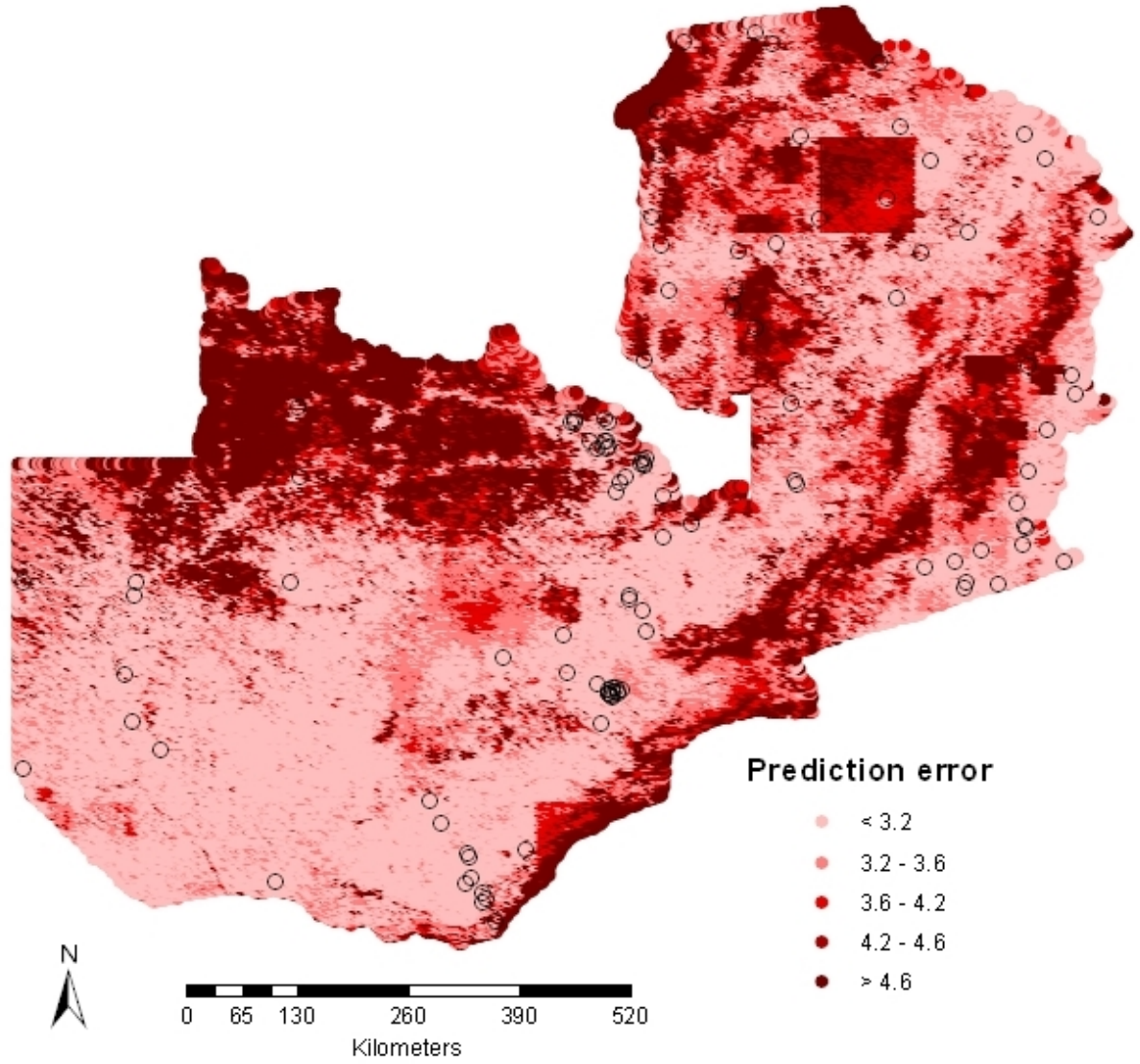

Figure 6.10: Prediction error of annual entomological inoculation rate (EIR) in Zambia estimated from the standard deviation of the posterior predictive distribution. Sampling locations are indicated by circles.

A smooth map of malaria risk for children under 5 years in Zambia is shown in Figure 6.11. Similar to the map of transmission intensity, high malaria risk is predicted in the eastern part and some regions in the north of Zambia, as well as at the border with Zimbabwe. Figure 6.12 depicts the smooth map of parasitaemia risk for children under 5 years old in Zambia obtained by fitting the logistic geostatistical model on the malaria survey data. This map shows a similar pattern of malaria risk to the one obtained by employing

the transmission model. However, the former map predicts lower level of parasitaemia prevalence (< 0.2) compared with the latter one (0.2-0.4). We calculated the Kendall's correlation coefficient to measure the degree of correspondence between parasitaemia risk estimated by the mathematical model and by directly analyzing the prevalence data. We obtained Kendall's tau equal to 0.22 and p-value < 0.001, indicating that the two measures are statistically significant associated.

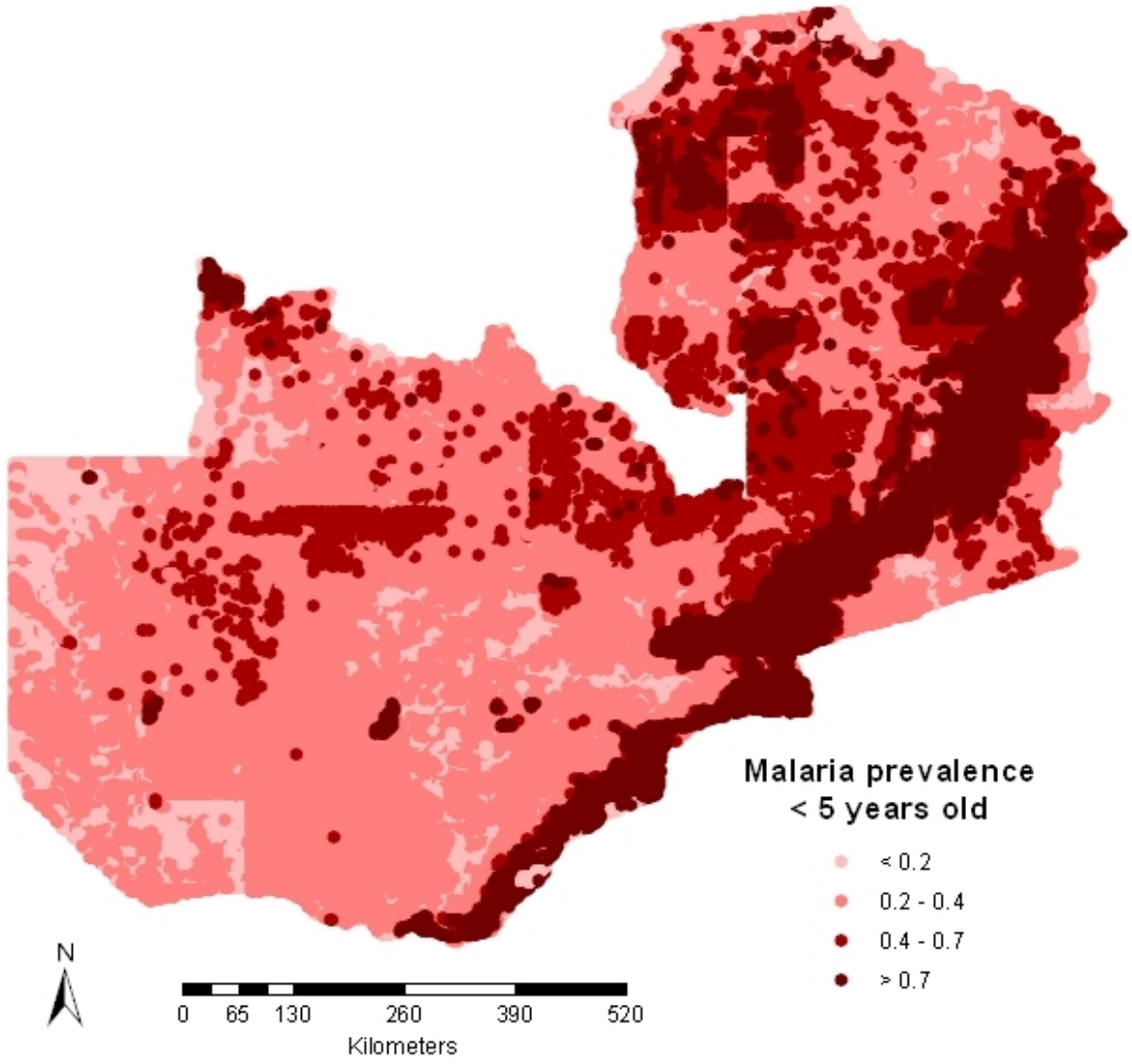

Figure 6.11: Predicted malaria prevalence for children under 5 years old in Zambia estimated from the median of the posterior predictive distribution.

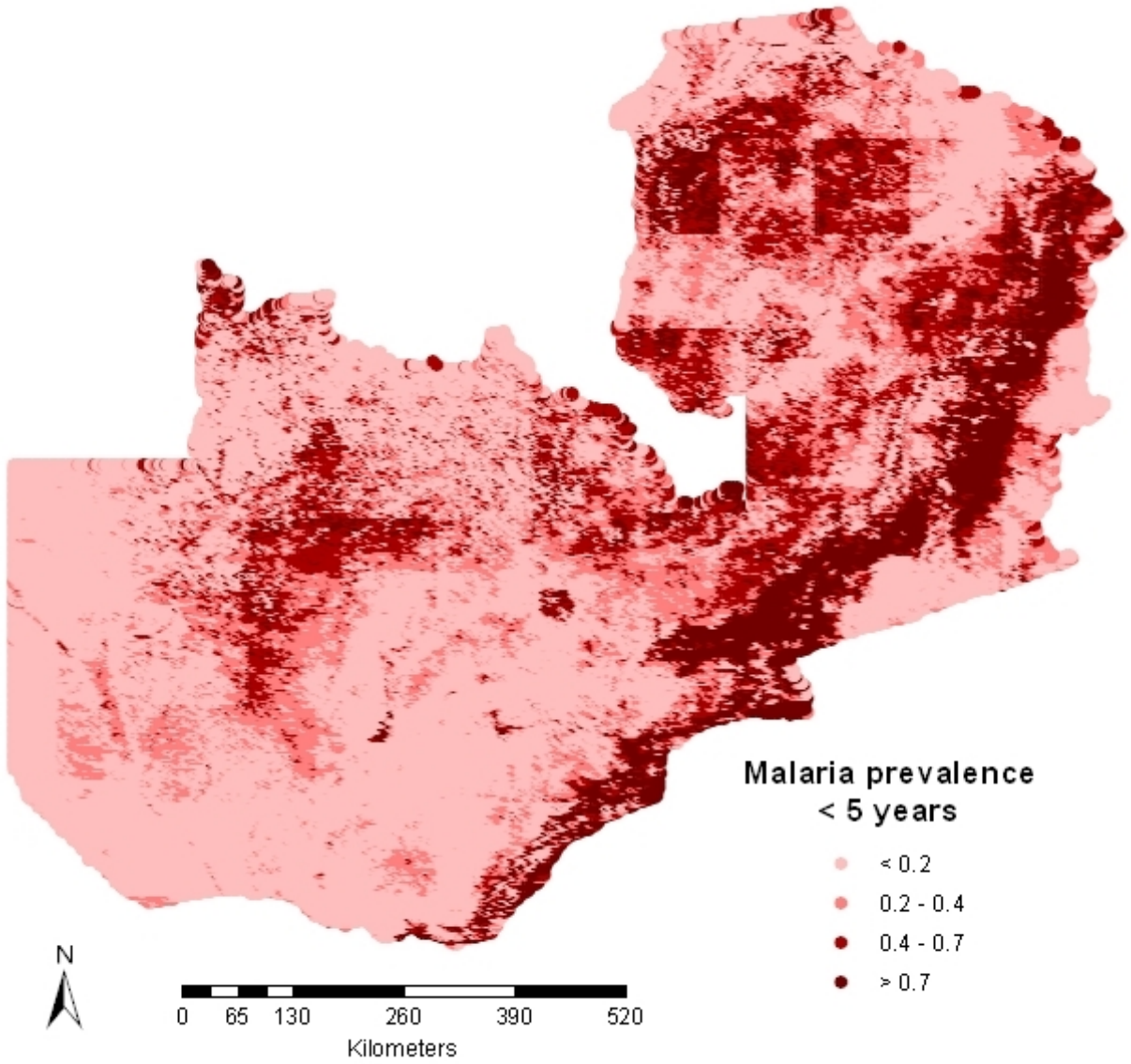

Figure 6.12: Predicted malaria prevalence for children under 5 years old in Zambia estimated from the median of the posterior predictive distribution obtained by directly analyzing the prevalence data.

6.4 Discussion

In this paper, a newly developed malaria transmission model was employed to produce age-adjusted malaria risk maps from survey data extracted from different sources. The age-heterogeneous prevalence data were converted into an age independent measure of transmission, the so-called entomological inoculation rate (EIR). Smooth maps of EIR were converted to age-specific malaria risk maps, using the relation between prevalence and transmission intensity described by the mathematical transmission model for each age group.

This approach was applied to analyze MARA survey data from Mali. Model validation revealed that the geostatistical intensity model has a better predictive ability in mapping the MARA survey data than the geostatistical prevalence model. To further assess the malaria risk maps based on the transmission model, data from the nationwide MIS in Zambia were analyzed. Maps based on both the transmission model and the raw prevalences were compared. Results showed that the map based on the transmission model predicts similar patterns of malaria risk with the one obtained by analyzing directly the MIS data. The advantage of using the transmission model in malaria mapping is that it makes possible age and seasonality adjustment as well as the inclusion of all available malaria historical data collected at different age groups. In addition, maps of different malaria transmission measures can be produced from survey data.

The Bayesian P-spline regression approach models the non-linear relation between environmental/climatic factors and malaria transmission in a flexible way. In Mali, none of the environmental variables used in the analyzes were statistically significant associated with annual EIR. In Zambia the model estimates a slightly negative association between evapotranspiration and malaria transmission. The spatial correlation present in the data was modeled in a Bayesian framework, using MCMC simulation techniques which enables simultaneously estimation of all model parameters together with their standard errors. We assessed the precision of the smooth maps of annual EIR by quantifying the prediction error using Bayesian kriging.

Gemperli et al. (2005) made use of the Garki malaria transmission model and produced maps of malaria transmission intensity and malaria risk for Mali. However, the Garki model was developed using field data from the savanna zone of Nigeria, therefore it should not be generalized to other regions in Africa with different environmental conditions and malaria endemicity. In addition, they ignored the seasonal patterns of malaria transmission,

whereas we used the seasonality model developed by Mabaso et al. (2007) and adjust for season at which data were collected. Comparing the map of annual EIR estimated by our model with the map obtained by Gemperli et al. (2005) we observe that overall we predicted lower level of EIR. This difference could be explained by the fact that Gemperli et al. (2005) did not adjust for seasonality in transmission. The malaria risk maps for Mali (for the two age groups) are similar with the maps obtained from previous work (Gosoniu et al., 2006).

The maps produced for Zambia indicate low level of malaria transmission intensity. This could be related to the reduced amount of rainfall experienced by Zambia ($2.0 - 3.5$ mm) and to the recent malaria control intervention implemented in the country. The only recent available data on EIR in Zambia are given by Kent et al. (2007) who conducted a study in two villages in the Southern Province of Zambia during November 2004-May 2005 and November 2005-May 2006 to investigate the seasonal intensity of malaria transmission in this region. During the first period malaria transmission was undetectable because of severe drought experienced by Zambia, but in the next season EIR was estimated at 1.6 and 18.3 infective bites per person per transmission season in the two villages, respectively.

The modeling approach used in this study was based on the assumption of stationarity, that is the spatial correlation was considered a function of only the distance between locations and irrespective of locations themselves. Malaria has non-stationary feature since local characteristics (environment, land use, vector ecology) influence the spatial correlation differently at various locations. Gosoniu et al. (2006) showed the importance of statistical modeling approach when analyzing malaria data. The methodology presented here could be extended to take into account the non-stationary characteristic of malaria by using the Bayesian partition modeling approach for geostatistical data developed in Chapter 5.

The P-spline approach which was used to model the non-linear effect of environmental conditions could be further developed by allowing the model to choose the number and the position of the knots instead of the fixed knots based on sample quantiles of the covariates used in this study. These models would have a variable number of parameters, therefore fitting would require the use of Reversible Jump MCMC (RJMCMC) (Green et al., 1995).

Acknowledgments

This work was supported by the Swiss National Foundation grant Nr.3252B0-102136/1.

6.5 Appendix

Geostatistical model of transmission intensity

Let Y_i be the Box Cox transformation of the annual EIR at location i, $i = 1, \ldots, n$. We assume that Y_i are normally distributed, $Y_i \sim N(\mu_i, \tau^2)$ with the mean structure μ_i a function of the environmental factors and τ^2 the unexplained non-spatial variance. The non-linear effect of the environmental predictors is modeled using Bayesian P-splines, that is $\mu_i = \sum_{j=1}^{p} f_j(X_{ij}) + \phi_i$, where $f(\cdot)$ is a smooth function of the environmental variables. In particular,

$$f_j(X_{ij}) = \sum_{k=1}^{K} u_{kj}|X_{ij} - s_{kj}|^3,$$

where $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{Kj})^T$ is the vector of regression coefficients, $s_{1j} < s_{2j} < \ldots < s_{Kj}$ are fixed knots and $|X_{ij} - s_{kj}|^3$ is a truncated 3-rd order polynomial spline basis, all corresponding to covariate X_j. The knots were chosen to be based on sample quantiles of the covariates.

Spatial correlation was modeled by assuming that the random effects ϕ_i introduced at each location i are distributed according to a multivariate normal distribution $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T \sim MVN(\mathbf{0}, \Sigma)$, where Σ_{ij} is a parametric function of the distance d_{ij} between locations i and j. A commonly used parametrization is $\Sigma_{ij} = \sigma^2 corr(d_{ij}; \rho)$, where σ^2 is the spatial variance and $corr(d_{ij}; \rho)$ is a valid correlation function. For the current application we chose the exponential correlation function $corr(d_{ij}; \rho) = exp(-d_{ij}\rho)$, where the parameter ρ captures the scale of correlation decay with distance. Ecker and Gelfand (1997) proposed several other parametric correlation forms, such as Gaussian, Cauchy and spherical.

Following the Bayesian modeling specification, we need to adopt prior distributions for all the parameters in the model. We adopt non-informative normal prior for the regression coefficients $p(\boldsymbol{u}) \sim N(0, 100)$ and the following priors for σ^2, τ^2 and ρ: $p(\sigma^2) \sim$ Inverse Gamma(a_1, b_1), $p(\tau^2) \sim$ Inverse Gamma(a_2, b_2) and $p(\rho) \sim$ Gamma(a_3, b_3) with the hyper-parameters a_1, b_1, a_2, b_2 and a_3, b_3 chosen to obtain a prior mean equal to 1 and a variance equal to 100. Bayesian inference is based on the joint posterior distribution

$$p(\boldsymbol{u}, \boldsymbol{\phi}, \sigma^2, \tau^2, \rho|\boldsymbol{Y}) = L(\boldsymbol{u}, \boldsymbol{\phi}; \boldsymbol{Y})p(\boldsymbol{u})p(\boldsymbol{\phi}|\sigma^2, \rho)p(\sigma^2)p(\tau^2)p(\rho),$$

where $L(\boldsymbol{u}, \boldsymbol{\phi}; \boldsymbol{Y})$ is the likelihood function and $p(\boldsymbol{\phi}|\sigma^2, \rho)$ is the distribution of the random effects, $p(\boldsymbol{\phi}|\sigma^2, \rho) \sim MVN(\mathbf{0}, \Sigma)$.

The model parameters were estimated using Markov chain Monte Carlo (MCMC) simulation and in particular Gibbs sampling. The conditional distribution of the regression coefficients is conjugate normal from which we can easily draw samples. The remaining parameters do not have conditional distributions of standard form and we employed a random walk Metropolis algorithm for sampling. We adopt a Gaussian proposal for the random effects $\boldsymbol{\phi}$ and a Gamma proposal for σ^2, τ^2 and ρ. The mean of the proposal distribution was the parameter estimate from the previous iteration and the variance was adaptively adjusted to achieve an acceptance rate of about 0.4.

Prediction of the Box Cox transformations of annual EIR at locations were malaria data are not available was carried out using Bayesian kriging. Let $\boldsymbol{Y}^0 = (Y_1^0, \ldots, Y_m^0)^T$ be the values to predict at the new m locations. Given the random effects $\boldsymbol{\phi}$ at the observed locations, we predict the random effects $\boldsymbol{\phi}^0$ at the new locations by sampling from the normal distribution

$$P(\boldsymbol{\phi}^0|\boldsymbol{\phi}, \sigma^2, \rho) = N(\Sigma_{01}\Sigma_{11}^{-1}\boldsymbol{\phi}, \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{01}^T),$$

where $\Sigma_{11} = E(\boldsymbol{\phi}\boldsymbol{\phi}^T)$, $\Sigma_{00} = E(\boldsymbol{\phi}^0\boldsymbol{\phi}^{0T})$ and $\Sigma_{01} = E(\boldsymbol{\phi}^0\boldsymbol{\phi}^T)$.
The predictive distribution is

$$P(\boldsymbol{Y}^0|\boldsymbol{Y}) = \int P(\boldsymbol{Y}^0|\boldsymbol{u}, \boldsymbol{\phi}^0)P(\boldsymbol{\phi}^0|\boldsymbol{\phi}, \sigma^2, \rho)P(\boldsymbol{u}, \boldsymbol{\phi}, \sigma^2, \tau^2, \rho|\boldsymbol{Y})\, d\boldsymbol{u}\, \boldsymbol{\phi}^0\, \boldsymbol{\phi}\, d\sigma^2\, d\tau^2\, d\rho,$$

where $P(\boldsymbol{u}, \boldsymbol{\phi}, \sigma^2, \tau^2, \rho|\boldsymbol{Y})$ is the posterior distribution. Conditional on $\boldsymbol{u}$ and ϕ_i^0, Y_i^0 is drawn from the normal distribution $N(\sum_{j=1}^p f_j(X_{ij}^0) + \phi_i^0, \tau^2)$, where $\boldsymbol{X}_i^0 = (X_1^0, \ldots, X_p^0)^T$ are the environmental covariates at the new location. Bayesian spatial prediction is performed by consecutive drawing samples from the posterior distribution, the distribution of the new random effects and the normal distribution of the predicted outcome.

Geostatistical model of prevalence

Let Y_i be the number of observed malaria cases out of N_i children examined at location i, $i = 1, \ldots, n$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ be the vector of p associated environmental predictors observed at location i. We assume that Y_i are binomially distributed, that is $Y_i \sim Bin(N_i, p_i)$ with parameter p_i measuring malaria risk at location i. The non-linear effect of environmental conditions is modeled non-parametrically on the logit transformation of p_i, that is $logit(p_i) = \sum_{j=1}^p f_j(X_{ij}) + \phi_i$. The smooth function $f(\cdot)$ is defined as in the intensity geostatistical model.

The spatial correlation is modeled on the covariance matrix of the location-specific random

effects, that is $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T \sim MVN(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \sigma^2 exp(-d_{ij}\rho)$. The spatial covariance parameters are described in the previous section, as well as their prior distributions. Model fit is handled via MCMC simulations. The conditional distribution of the spatial variance σ^2 is conjugate Gamma and we use Gibbs sampler to estimate the parameter. For the remaining parameters we employed Metropolis algorithm for sampling, since their conditional distributions have non-standard forms.

We obtain estimates of the malaria risk at any unsampled location by the predictive distribution

$$P(\boldsymbol{Y}^0|\boldsymbol{Y}, \boldsymbol{N}) = \int P(\boldsymbol{Y}^0|\boldsymbol{\beta}, \boldsymbol{\phi}^0) P(\boldsymbol{\phi}^0|\boldsymbol{\phi}, \sigma^2, \rho) P(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \rho|\boldsymbol{Y}, \boldsymbol{N})\, d\boldsymbol{\beta}\, d\boldsymbol{\phi}^0\, d\boldsymbol{\phi}\, d\sigma^2\, d\rho,$$

where $\boldsymbol{Y}^0 = (Y_1^0, \ldots, Y_m^0)^T$ are the predicted number of cases at new locations,
$P(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \rho|\boldsymbol{Y}, \boldsymbol{N})$ is the posterior distribution and $\boldsymbol{\phi}^0 = (\phi_1^0, \ldots, \phi_m^0)^T$ is the vector of random effects at new sites. The distribution of $\boldsymbol{\phi}^0$ at new locations given $\boldsymbol{\phi}$ at observed locations is normal

$$P(\boldsymbol{\phi}^0|\boldsymbol{\phi}, \sigma^2, \rho) = N(\Sigma_{01}\Sigma_{11}^{-1}\boldsymbol{\phi}, \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{01}^T),$$

with $\Sigma_{11} = E(\boldsymbol{\phi}\boldsymbol{\phi}^T)$, $\Sigma_{00} = E(\boldsymbol{\phi}^0\boldsymbol{\phi}^{0T})$ and $\Sigma_{01} = E(\boldsymbol{\phi}^0\boldsymbol{\phi}^T)$.

Conditional on ϕ_i^0 and $\boldsymbol{\beta}$, Y_i^0 are independent Bernoulli variates $Y_i^0 \sim Ber(p_i^0)$ with malaria prevalence at unsampled site given by $logit(p_i^0) = X_i^0\boldsymbol{\beta} + \phi_i^0$. We predicted malaria prevalence at the same unsampled locations where the Box Cox transformation of the annual EIR was predicted.

Chapter 7

General discussion and conclusions

The epidemiological questions that motivated the work in this thesis were the following: (1) assessing the spatial effect of bednets use on child mortality; (2) producing different maps of malaria risk at country and regional level and (3) producing malaria transmission intensity maps and malaria risk maps adjusted for age and seasonality. These applications led to the development of novel statistical methods for: (1) modeling large, negative binomial, geostatistical data; (2) modeling binomial, geostatistical data under the assumption of non-stationarity; (3) geostatistical models validation and (4) modeling the non-linear effects of covariates on the outcome variable.

Each chapter in the thesis provides a detailed discussion on the findings. Here is presented a summary of the main contributions, the implication of our results in malaria control and some ideas for potential future research.

From methodological point of view, the developed novel statistical methods have several original contributions in: (i) facilitating estimation of large non-Gaussian geostatistical models; (ii) estimation and prediction of non-stationary non-Gaussian spatial processes; (iii) comparing the predictive ability of geostatistical models and (iv) estimation of non-linear relations between covariates and outcome variables.

In Chapter 2 the spatial effect of bednets use on child mortality was assessed by analyzing data extracted from the Demographic Surveillance System (DSS) set up in Kilombero Valley, Tanzania. The spatial correlation in mortality data was considered as a function of the distance between locations and the model was fitted in the Bayesian framework,

using Markov chain Monte Carlo (MCMC) simulations. The DSS data are collected at very large number of households, therefore fitting these geostatistical models is computational intensive because it requires repeated inversions of the variance-covariance matrix. To overcome this challenge a convolution model for the underlying spatial process was developed. In particular, the spatial process was estimated by a subset of locations and then the location-specific random effects were approximated by a weighted sum of the subset of location-specific random effects with the weights inversely proportional to the separation distance. This approach has the advantage that it reduces the size of the matrix to be inverted. There are other methods that overcome very large matrix computations involved in geostatistical model fit, such as the ones suggested by Rue and Tjelmeland (2002), Stein et al. (2004), Lee et al. (2005), Xia and Gelfand (2005). An approach similar to the one developed in this thesis would be the approximation of the spatial process by the projected latent variables algorithm (PLV) developed by Seeger (2003). Further research may include validation and comparison of different approaches to accelerate large matrix inversion on simulated datasets.

The main epidemiological result coming from Chapter 2 is that in an area of high perennial malaria transmission in southern Tanzania, a community effect of bednets use on all-cause child mortality was not evident. This result is explained by the small proportion of insecticide treated nets (ITN), as well as by the homogeneity of bednets coverage in the study area. To achieve a significant reduction of malaria transmission, hence of child mortality, the coverage of ITNs and long-lasting insecticidal nets (LLINs) should reach at least the level of present bednets coverage. In fact, three initiatives namely the Roll Back Malaria Partnership, the United Nations Millennium Development Goals, and the US President's Malaria Initiative have set a target of at least 80% use of ITNs by the people most vulnerable to malaria (young children and pregnant women) by the year 2010. Killeen et al. (2007) have shown that a coverage of $35\% - 65\%$ of ITN use by the entire population would have a similar degree of community-wide protection as the personal protection. Hence, the wide-scale of ITN use by the whole population should be promoted, keeping still as a priority the use of ITNs by the most vulnerable people.

In Chapter 3 maps of malaria risk in Mali were produced. Maps of malaria distribution are important tools for increasing the effectiveness of malaria control programs because they provide useful information on which regions are at high risk of malaria and optimize the allocation of resources to areas of most need. In addition, the malaria transmission maps

can be used to assess the effectiveness of intervention programs. Several maps of malaria risk have been produced for Mali (Kleinschmidt et al., 2001; Gemperli, 2003; Gemperli et al., 2006; Sogoba, 2007). Although these maps predicted superficially similar patterns of malaria transmission, there are significant differences between them because 1) they were based on different malaria data and environmental predictors and 2) they were produced using different statistical methods. Spatial correlation in malaria data is likely to be influenced differently at various locations due to local characteristics like environmental factors, intervention measures, mosquito ecology, health services etc. Therefore, non-stationarity is an important feature of malaria that can not be ignored when analyzing parasitaemia data. Maps of malaria risk in Mali under both assumptions of stationarity and non-stationarity were produced using the same dataset. Non-stationarity was modeled by partitioning the study area into fixed tiles, assuming a stationary spatial process in each subregion and correlation between the tiles. Comparison of the predictive ability of the two models indicated that the geostatistical model that allowed for non-stationary spatial covariance structure predicts better the malaria risk. In addition, the stationarity assumption influenced the significance of the parasitaemia risk predictors, as well as the estimation of the spatial parameters. We conclude that malaria mapping is sensitive to the assumptions about the spatial process. Non-stationarity is an important feature that should be taken into account in malaria mapping.

The approach that allows modeling non-stationarity, proposed in Chapter 3, is more appropriate when modeling malaria data over large areas with fixed partitions outlined, for example, by various ecological zones. This methodology was employed in Chapter 4 to obtain a malaria risk map in West Africa, considering as fixed tiles the four agro-ecological zones that partition the region. In addition, the non-linear relation between parasitaemia risk and environmental factors was modeled via Bayesian P-splines, separately in each agro-ecological zone. Previous malaria maps for West Africa and West and Central Africa were produced by Kleinschmidt et al. (2001) and Gemperli et al. (2006), respectively, both under the assumption of stationarity. Kleinschmidt et al. (2001) modeled the non-linear effects of environmental factors on malaria risk using fractional polynomials and considered a separate statistical model in each agro-ecological zone. The resulting map showed discontinuities at the edges of the zones, which were further smoothed. However, the modeling approach of Kleinschmidt et al. (2001) did not allow estimation of the prediction error. The discontinuities at the borders between the zones were avoided in our case because the spatial correlation was modeled by a mixture of spatial processes over the entire region,

with the mixing proportions chosen to be exponential functions of the distance between locations and the centers of the tiles corresponding to each of the spatial processes.

This work, as well as the previous work of Kleinschmidt et al. (2001) was based on prevalence data collected on specific age groups of the population (less than 10 years old), while Gemperli et al. (2006) made use of all available data by employing the Garki transmission model to adjust for age. Gemperli et al. (2006) modeled the non-linear environment-malaria relation by including interaction terms and using non-linear functional forms of the predictors (i.e. logarithm, polynomials). However, they assumed the same relation over the entire region of West and Central Africa. The model developed in Chapter 4 addressed the drawbacks of the previous efforts in mapping malaria in West Africa, by relaxing the assumptions of stationarity, linearity and the common effect of environmental conditions on malaria transmission over the different ecological zone. All three maps predict similar patterns of malaria transmission, but there are also significant discrepancies. The regions with main differences between the three malaria maps coincide with the areas with very few or no survey data collected. To improve the current maps, future efforts should be concentrated on the collection of new data, as well as on the development of novel statistical models for analyzing the data.

In Chapter 4 the model validation methodology developed in Chapter 3 was further employed to compare the Bayesian P-spline approach with the method which considers the covariates as categorical variables. The results indicated that the former model fits better the non-linear environment-malaria relation. The knots used to define the splines were fixed, based on sample quantiles of the independent variables. From methodological point of view, this approach could be extended by allowing the number and the position of the knots to be random, chosen by the model.

In Chapter 5 a non-stationary model of malaria risk was fitted by assuming a random rather than a fixed tile configuration, giving a smooth map of parasitaemia risk in Mali. In fact, the non-stationarity characteristic of malaria data may be explained not only by the variation in environmental factors, but also by the variation in other factors such as socio-economic status, malaria control interventions, health care services and human activities which may influence the spatial correlation differently in various part of the study area. While the climatic and environmental variables may define a fixed partitioning, this may not be obvious for the other factors, therefore the data should be allowed to

decide on the number of tiles and thus to identify the different spatial processes. Non-stationarity was modeled by considering the number and configuration of the tiles to be random parameters in the model. The parameter space has variable dimensions, depending on the tessellation, hence model fit was handled via a Reversible Jump MCMC sampler. A random tessellation approach has been developed also by Gemperli (2003) who accounts for non-stationarity when analyzing malaria prevalence in Mali. However, the author used a different set of environmental predictors from the ones used in our study and assumed independence between the tiles. Although this assumption facilitates the inversion of the variance-covariance matrix, it is not reasonable to assume that neighboring points located in different tiles are not correlated. In Chapter 5 this assumption was relaxed and a random partitioning model which takes into account the correlation between tiles was developed. The model introduces at each locations as many random effects as the number of tiles and therefore it requires the inversion of the spatial covariance matrix as many times as the number of tiles. Future research may consider the approximation of tile-specific spatial processes by sparse Gaussian processes (Seeger, 2003), as discussed earlier, estimating the spatial processes from a subset of locations in each subregion. Inversion of the large, original matrices would be replaced by the inversion of much smaller size matrices.

Models that account for non-stationarity allow mapping of spatial covariance parameters. Estimates of these parameters are useful in improving the estimation of prediction error, which helps quantifying the precision of the map. Low values of the range parameter indicate that the spatial correlation decreases rapidly over short distances, suggesting that the parasitaemia risk may be influenced by local factors such as human behavior, rather than factors that vary over large areas.

In this thesis, the malaria maps of Mali and West Africa (Chapter 3 - Chapter 5) were based on the analyzes of data extracted form the MARA database. This consists of published and unpublished surveys which took place at different locations, including non-standardized age groups of the study population. In addition, the season when the data were collected varied between locations, making difficult seasonality adjustment of parasitaemia prevalence. Due to age dependence of malaria prevalence, the maps were based on a subset of the MARA database, which included malaria surveys carried out on a population with age ranging from 1 to 10 years old. Gemperli et al. (2005) and Gemperli et al. (2006) have prove the feasibility of using mathematical models of malaria transmission in converting heterogeneous malaria prevalence into a measure of transmission intensity, by employing the Garki

model. However, this model has been developed using field data from the savanna zone of Nigeria and may not be reliable for other regions in Africa, with different environmental conditions and levels of malaria endemicity. In Chapter 6 a newly developed malaria transmission model (Smith et al., 2006) was employed and maps of malaria transmission and age-specific malaria risk maps for Mali were produced. The model is based on the seasonal pattern of malaria by employing the seasonality model of Mabaso (2007). Model validation revealed that the geostatistical model based on the estimates derived from the transmission model had a better predictive ability compared to the one modeling the raw prevalence data.

MARA is the most extensive database of historical malaria survey data in Africa. The data collected since 1900's contained in the database make MARA the perfect dataset for the analyzes of temporal changes in malaria transmission. However, the maps based on these data may not reflect the current situation of malaria because this could be influenced by intervention measures that are not documented and therefore it is not possible to adjust for them. Another major shortcoming of the MARA prevalence data is that they are collected mainly in high endemic areas and therefore introduce selection bias. The necessity of standard surveys and up to date, better quality malaria data led to the development of Malaria Indicator Surveys (MIS), carried out in few African countries. The MIS are nationally representative household surveys, running in a specific period of time. The data corresponds to the current malaria situation and avoid sampling biases. In Chapter 6 parasitaemia data from the Zambia National MIS were analyzed and smooth maps of malaria risk were obtained. To validate the use of mathematical model in malaria mapping the parasitaemia risk map obtained by employing the transmission model was compared with the risk map obtained directly by analyzing the prevalence data. Both maps predicted similar patterns of malaria risk.

Throughout this thesis the importance of Bayesian geostatistical models to identify the relation between environmental conditions and parasitaemia risk and to produce model-based maps of malaria transmission was shown. These maps are essential for implementation of malaria control programs because they offer valuable information on the areas which are at high risk of malaria. Together with demographic data, the malaria distribution maps may produce accurate estimates of the burden of disease and therefore optimize the allocation of human and financial resources for malaria control. In addition, the malaria maps provide a baseline against which the effectiveness of malaria intervention programs can be

assessed.

The country and regional malaria maps produced in this thesis can have an important role in planning the intervention programs in these regions as long as they reach the key people in malaria control programs, departments of health and research institutions in African countries. Next steps toward the dissemination of this products include publication of this work, distribution of poster sized malaria maps as well as organization of workshops in collaboration with local partners.

Recently, there is a renew interest in malaria mapping (WHO, 2007), therefore new efforts are put in collection of malaria data. Despite the above mentioned disadvantages of the historical malaria survey data, we are currently updating MARA because it is one of the only database to date which contains survey data across all Africa. The newly established MIS are currently running in very few selected countries. We are planning to implement the methodology developed in this thesis on the updated MARA database and together with local collaborators to produce regional maps for East and South Africa and a continental malaria map. The work in this thesis has shown that modeling assumptions make a difference in malaria mapping and therefore it is important not only to collect data but also to develop appropriate statistical models to obtain more accurate maps. This thesis, following the work of Gemperli (2003), is an important step toward this direction.

Based on the results in this thesis the following suggestions can be made:

- Non-stationarity is an important characteristic of malaria that should be taken into account when mapping the disease. In particular, the geostatistical non-stationary model based on the fixed partitioning developed in Chapter 3 should be employed when modeling malaria over regions with an apparent fixed partitioning. If the factors that determine non-stationarity do not define obvious fixed subregions within the study area, it is recommended the approach based on random Voronoi tessellations developed in Chapter 5.

- The Bayesian P-splines approach is recommended for modeling non-linear effects of environmental/climatic factors on malaria transmission.

- The mathematical transmission model employed in Chapter 6 can be incorporated in malaria mapping to produce age and seasonality adjusted risk maps derived from compiled survey data.

Bibliography

Bibliography

[1] Abdulla S, Gemperli A, Mukasa O, Armstrong Schellenberg Jr, Lengeler C, Vounatsou P, Smith T, 2005. Spatial effects of the social marketing of insecticide-treated nets on malaria morbidity. Tropical Medicine and International Health 10, 11-18

[2] Abeku TA, Hay SI, Ochola S, Langi P, Beard B, De Vlas SJ, Cox J, 2004. Malaria epidemic early warning and detection in African highlands. Trends in Parasitology 20, 400-405.

[3] Agbu PA, James ME, 1994. NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual, Goddard Distributed Active Archive Center. NASA Goddard Space Flight Center, Greenbelt

[4] Anderson JR, Hardy EE, Roach JT, Witmer RE, 1979. A land use and land cover classification system for use with remote sensor data. US Geological Survey Professional Paper, 964.

[5] Armstrong Schellenberg JRM, Mukasa O, Abdulla S, Marchant T, Lengeler C, 2002. The Ifakara Demographic Surveillance System. INDEPTH Monograph Series: Demographic Surveillance Systems for Assessing Populations and their Health in Developing Countries. Volume 1: Population, Health and Survival in INDEPTH Sites. Ottowa, IyDRC/CRDI.

[6] Armstrong Schellenberg JRM, Nathan R, Abdulla S, Mukasa O, Marchant TJ et al., 2002b. Risk factors for child mortality in rural Tanzania. Tropical Medicine and International Health 6, 506-511.

[7] Banerjee S, Gelfand AE, Knight JR, Sirmans CF, 2004. Spatial modeling of house prices using normalized distance-weighted sum of stationary processes. Journal of Business and Economic Statistics 22, 206-213.

[8] Binka FN, Kubaje A, Abjuik M, Williams LA, Lengeler C, 1996. Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. Tropical Medicine and International Health 1, 147-154.

[9] Binka FN, Indome F, Smith T, 1998. Impact of spatial distribution of permethrin-impregnated bed nets on child mortality in tutal northern Ghana. American Journal of Tropical Medicine and Hygiene 59, 80-85.

[10] Breman JG, Alilio MS, Mills A, 2004. Conquering the intolerable burden of malaria: what's new, what's needed: a summary. American Journal of Tropical Medicine and Hygiene 71(S2), 1-15.

[11] Bruce-Chwatt LJ, 1952. Malaria in African infants and children in Southern Nigeria. Annals of Tropical Medicine and Parasitology 46, 173-200.

[12] Clayton DG, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

[13] Clements ACA, Lwambo NJS, Blair L, Nyandindi U, Kaatano G, Kinung'hi S,Webster JP, Fenwick A, Brooker S, 2006. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Tropical Medicine and International Health 11, 490-503.

[14] Cleveland WS, 1979. Robust locally weighted regression and smoothing scatter plots. Journal of the American Statistical Association 74, 829-836.

[15] Coetzee M, Craig M, Le Sueur D, 2000. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. Parasitology Today 16, 74-77.

[16] Craig MH, Snow RW, Le Sueur D, 1999. A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitology Today 15, 105-111.

[17] Crainiceanu CM, Ruppert D, Wand MP, 2005. Bayesian Analysis for Penalized Spline Regression Using WinBUGS. Journal of Statistical Software 14(14), 1-24.

[18] Cressie NAC, 1993. Statistics for Spatial Data. New York, Wiley.

[19] D'Alessandro U, Olaleye BO, McGuire W, Langerock P, Bennett S, 1995. Mortality and morbidity from malaria in Gambian children after introduction of an impreganted bednet programme. Lancet 345, 479-483.

[20] D'Alessandro U, Olaleye BO, McGuire W, Thomson MC, Langerock P, 1995b. A comparison of the efficacy of insecticide-treated and untreated bed nets in preventing malaria in Gambian children. Transactions of the Royal Society of Tropical Medicine and Hygiene 89 (6), 596-598.

[21] Damian D, Sampson PD, Guttorp P, 2001. Bayesian estimation of semi-parametric nonstationary spatial covariance structure. Econometrics 12, 161-178.

[22] de Boor C, 1978. A practical guide to splines. Springer, Berlin.

[23] Deichmann U, 1996. African population database. Digital database and documentation. National Center for Geographic Information and Analysis, Santa Barbara, USA (http://grid2.cr.usgs.gov/globalpop/africa/).

[24] Dietz K, Molineaux L, Thomas A, 1974. A malaria model tested in the African savannah. Bulletin of the World Health Organization 50, 347-357.

[25] Diggle PJ, Tawn JA, 1998. Model-based geostatistics. Applied Statistics 47, 299-350.

[26] Diggle PJ, Moyeed R, Rowlinson B, Thomson M, 2002. Childhood malaria in the Gambia: a case-study in model-based geostatistics. Applied Statistics 51, 493-506.

[27] Doumbo O, Outtara NI, Koita O, Maharaux A, Toure YT, Traoure SF, Quilici M, 1989. Approche eco-geographique du paludisme en milieu urbain: ville de Bamako au Mali. Ecologie Humaine 8, 3-15.

[28] Droogers P, Seckler D, Makin I, 2001. Estimating the potential of rainfed agriculture. International Water Management Institute, Working Paper 20, available at www.iwmi.cgiar.org/pubs/working/Index.htm

[29] Ecker MD, Gelfand AE, 1997. Bayesian Variogram Modeling for an Isotropic Spatial Process. Journal of Agricultural, Biological and Environmental Statistics 4, 347-368.

[30] Eilers PHC, Marx BD, 1996. Flexible smoothing with B-splines and penalties. Statistical Scienc 11(2), 89-121.

[31] Erlanger TE, Enayati AA, Hemingway J, Mshinda H, Tami A et al., 2004. Field issues related to effectiveness of insecticide-treated nets in Tanzania. Medical and Veterinary Entomology 18, 153-160.

[32] Eubank RL, 1988. Spline smoothing and nonparametric regression. Decker, New York.

[33] FAO, 1978. Report on the agro-ecological zones project, Vol1, Methodology and results for Africa. World Soil Resources Report 40, 32-41.

[34] Fuentes M, 2001. A new high frequency kriging approach for nonstationary environmental processes. Envirometrics 12, 469-483

[35] Fuentes M, Smith RL, 2002. A new class of nonstationary spatial models. Technical Report, Statistics Department, North Carolina State University.

[36] Gelfand AE, Smith AFM, 1990. Sampling-based approach to calculating marginal densities. Journal of the American Statistical Society 85, 398-409.

[37] Gemperli A, 2003. Development of Spatial Statistical Methods for Modelling Point-Referecnced Spatial Data in Malaria Epidemiology. Doctoral Dissertation, Swiss Tropical Institute, University of Basel.

[38] Gemperli A ,Vounatsou P, 2004a. Fitting generalized linear mixed models for point-referenced data. Journal of Modern Applied Statistical Methods 2, 497-511.

[39] Gemperli A, Vounatsou P, Kleinschmidt I, Bagayoko M, Lengeler C, Smith T, 2004b. Spatial patterns of infant mortality in Mali; the effect of malaria endemicity. American Journal of Epidemiology 159, 64-72.

[40] Gemperli A, Vounatsou P, Sogoba N, Smith N, 2005. Malaria mapping using transmission models: application to survey data from Mali. American journal of Epidemiology 163, 289-297.

[41] Gemperli A, Sogoba N, Fondjo E, Mabaso M, Bagayoko M, Brïet O, Anderegg D, Liebe J, Smith T, Vounatsou P, 2006. Mapping Malaria Transmission in West- and Central Africa. Tropical Medicine and International Health 11(7), 1032-1046.

[42] Gemperli A, Vounatsou P, 2006. Strategies for fitting large, geostatistical data in MCMC simulation. Communications in Statistics - Simulation and Computation 35, 331-345.

[43] Gosoniu L, Vounatsou P, Sogoba N, Smith T, 2006. Bayesian modelling of geostatistical malaria risk data. Geospatial Health 1, 127-139.

[44] Green PJ, 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711-732

[45] Greenwood BM, 1990. Populations at risk. Parasitology Today 6, 188.

[46] Grover-Kopec E, Kawano M, Klaver RW, Blumenthal B, Ceccato P, Connor SJ, 2005. An online operational rainfall-monitoring resource for epidemic malaria early warning systems in Africa. Malaria Journal 21, 4-6.

[47] Guerra CA, Hay SI, Lucioparedes LS, Gikandi P, Tatem AJ, Noor AM, Snow RW, 2007. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. Malaria Journal 6: 17.

[48] Guillet P, Alnwick D, Cham MK, Neira M, Zaim M, Heymann D, Mukelabai K, 2001. Long-lasting treated mosquito nets: a breakthrough in malaria prevention. Bulletin World Health Organization 79 (10), p.0-0. ISSN 0042-9686.

[49] Gwatkin DR, 2005. Assessing economic inequalities in health: Contribution of the INDEPTH health equity project In: INDEPTH Network, eds. Measuring health equity in small areas: findings from demographic surveillance systems. Aldershot: Ashgate.

[50] Haas TC, 1990. Kriging and automated variogram modeling within a moving window. Atmospheric Environment 24A, 1759-1769.

[51] Haas TC, 1995. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. Journal of the American Statistical Association 90, 1189-1199.

[52] Hardle W, 1990. Applied Nonparametric regression. Cambridge Univ. Press.

[53] Hastings WK, 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97-109.

[54] Hawley WA, Phillips-Howard PA, ter Kuile FO, Terlouw DJ, Vulule JM et al., 2003. Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. American Journal of Tropical Medicine and Hygiene 68, supplement 4, 121-127.

[55] Hay SI, Snow RW, Rogers DJ, 1998. Predicting malaria seasons in Kenya using multi-temporal meteorological satellite sensor data. Transactions of Royal Society of Tropical Medicine and Hygiene 92, 12-20.

[56] Hay SI, Rogers DJ, Toomer JF, Snow R, 2000a. Annual *Plasmodium falciparum* entomological inoculation rates (EIR) across Africa: literature survey, internet access and review. Transactions of Royal Society of Tropical Medicine and Hygiene 94, 113-127.

[57] Hay SI, Omumbo JA, Craig MH, Snow RW, 2000b. Earth observation,geographic infortmation system and *Plasmodium falcipyrum* malaria in sub-Saharan Africa. Advances in Parasitology 47, 173-215.

[58] Hay SI, Snow RW, 2006. The Malaria Atlas Project: developing global maps of malaria risk. PLoS Medicine 3(12): e473.

[59] Higdon D, Swall J, Kern J, 1999. Non-stationary spatial modeling. Bayesian Statistics 6, 761-768.

[60] Hutchinson MF, Nix HA, McMahon JP, Ord KD, 1996. Africa - A topographic and climate database (CD-ROM). The Australian National University Canberra, ACT 0200, Australia.

[61] Justice CO, Townshend JRG, Holben BN, Tucker CJ, 1985. Analysis of the phenology of global vegetation using meteorological satellite data. International Journal of Remote Sensing 6, 1271-1318.

[62] Kent RJ, Thuma PE, Mharakurwa S, Norris DE, 2007. Seasonality, blood feeding behavior and transmission of *Plasmodium falciparum* by *Anopheles arabiensis* after an extended drought in Southern Zambia. The American Society of Tropical Medicine and Hygiene 76 (2), 267-274.

[63] Killeen GF, Kihonda J, Lyimo E, Oketch FR, Kotas ME et al., 2006. Quantifying behavioural interactions between humans and mosquitoes: Evaluating the protective efficacy of insecticidal nets against malaria transmission in rural Tanzania. BMC Infectious Diseases 6, 161.

[64] Killeen GF, Smith TA, Ferguson HM, Mshinda H, Abdulla S et al., 2007. Preventing Childhood Malaria in Africa by Protecting Adults from Mosquitoes with Insecticide-Treated Nets. PloS Medicine 4(7), e229, doi:10.1371/journal.pmed.0040229.

[65] Kim H-Y, Mallik B, Holmes C, 2005. Analyzing nonstationary spatial data using piecewise Gaussian processes. Journal of the American Statistical Association 100, 653-668.

[66] Kleinschmidt I, Bagayoko M, Clarke GPY, Craig M, Le Sueur D, 2000. A spatial statistical approach to malaria mapping. International Journal of Epidemilogy 29, 355-361.

[67] Kleinschmidt I, Omumbo JA, Briet O, van de Giesen N, Sogoba N, Mensah NK, Windmeijer P, Moussa M, Teuscher T, 2001a. An empirical malaria distribution map for West Africa. Tropical Medicine and International Health 6, 779-786.

[68] Kleinschmidt I, Sharp BL, Clarke GP, Curtis B, Fraser C, 2001b. Use of generalized linear mixed models in the spatial analysis of small area incidence rates in KwaZulu Natal, South Africa. American journal of Epidemiology 153, 1213-1221.

[69] Knorr-Held L, Rasser G, 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56, 13-21.

[70] Lee HKM, Higdon DM, Calder CA, Holloman C, 2005. Efficient models for correlated data via convolutions of intrinsic processes. Statistical Modeling 5, 53-74.

[71] Lengeler C, 2004. Insecticide-treated bed nets and curtains for preventing malaria (Cochrane Review). The Cochrane Database of Systematic Reviews, Issue 2.

[72] Lindsay SW, Alonso PL, Armstrong JRM, Hemingway J, Adiamah JH et al., 1993. A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, West Africa.7.Impact of permethrin-impregnated bed nets on malaria vectors. Transactions of the Royal Society of Tropical Medicine and Hygiene 87, supplement 2, 45-51.

[73] Mabaso MLH, Craig M, Vounatsou P, Smith T, 2005. Towards empirical description of malaria seasonality in southern Africa: the example of Zimbabwe. Tropical Medicine and International Health 10, 909-918.

[74] Mabaso M, 2007. Temporal variations in malaria risk in Africa. Doctoral Dissertation, Swiss Tropical Institute, University of Basel.

[75] MARA/ARMA, 1998. Towards an Atlas of Malaria Risk in Africa. First technical report of the MARA/ARMA collaboration (www.mara.org.za) South Africa

[76] Mardia CV, Goodall CR, 1992. Spatial-temporal analysis of multivariate environmental monitoring data. In: Multivariate Environmental Statistics 6. Bose NK, Patil GP, Rao CR. North Holland, Newyork, 347-385.

[77] Maxwell CA, Mimaba J, Njunwa KJ, Greenwood BM, Curtis CF, 1999. Comparison of bednets impregnated with different pyrethroids for their impact on mosquitoes and on re-infection with malaria after clearence of pre-existing infections with chlorproguanil-dapsone. Transactions of the Royal Society of Tropical Medicine and Hygiene 93, 4-11.

[78] Molineaux L, Gramiccia G, 1980. The Garki Project. Geneva, World Health Organization.

[79] Nevill CG, Some ES, Mung'ala VO, Mutemi V, New L et al., 1996. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. Tropical Medicine and International Health 1, 139-146.

[80] Nobre A, Schmidt A, Lopes H, 2005. Spatio-temporal models for mapping the incidence of malaria in Par. Environmetrics 16, 291-304.

[81] Nychka D, Wickle CK, Royle JA, 2002. Multi resolution models for nonstationary spatial covariance functions. Statistical Modelling 2, 315-331.

[82] Omumbo JA, Hay SI, Snow RW, Tatem RJ, Rogers DJ, 2005. Modelling malaria risk in East Africa at high-spatial resolution. Tropical Medicine and International Health 10 (6), 557-566.

[83] Obled C, Creutin JD, 1986. Some developments in the use of empirical orthogonal functions for mapping meteorological fields. Journal of Applied Meteorology 25, 1189-1204.

[84] OSullivan F, 1986 A statistical perspective on ill-posed inverse problems. Statistical Science 1, 502518.

[85] Paciorek CJ, 2007. Computational techniques for spatial logistic regression with large datasets.. Computational Statistics and Data Analysis 51, 3631-3653.

[86] Patz JA, Campbell-Lendrum D, Holloway T, Foley JA, 2005. Impact of regional climate change on human health. Nature 438, 310-317.

[87] Picq JJ, Delemont J, Nosny Y, 1992. Zones bioclimatiques tropicales et pathologie. Medecine D'Afrique Noire 39 (3), 157-161.

[88] Quinones ML, Lines J, Thomson MC, Jawara M, Greenwood BM et al., 1998. Permethrin-treated bed nets do not have a mass-killing effect on village populations of

Anopheles gambiae s.l. in The Gambia. Transactions of the Royal Society of Tropical Medicine and Hygiene 92, 373-378.

[89] Raso G, N'Goran EK, Toty A, Luginbuhl A, Adjoua CA, Tian-Bi NT, Bogoch II, Vounatsou P, Tanner M, Utzinger J, 2004. Efficacy and side effects of praziquantel against Schistosoma mansoni in a community of western Cote d'Ivoire. Transactions of the Royal Society of Tropical Medicine and Hygiene 98, 18-27.

[90] Raso G, Vounatsou P, Gosoniu L, Tanner M, N'Goran EK, Utzinger J, 2005. Risk factors and spatial patterns of hookworm infection among school children in a rural area of Western Cote d'Ivoire. International Journal for Parasitology 36, 201-210.

[91] Raso G, Vounatsou P, Singer BH, N'goran EK, Tanner M, Utzinger J., 2006. An integrated approach for risk profiling and spatial prediction of Schistosoma mansoni-hookworm coinfection. Proceedingof the National Academy of Sciences of the U S A 103, 6934-6939

[92] Rogers DJ, Randolph SE, Snow RW, Hay SI, 2002. Satellite imagery in the study and forecast of malaria. Nature 415, 710-715.

[93] Rogers DJ, Randolph SE, Snow RW, Hay SI, 2002. Updating historical maps of malaria transmission intensity in East Africa using remote sensing. Photogrammetric Engineering and Remote Sensing 68, 161-166.

[94] Royston P, Altman DG, 1994. Regression using fractinal polynomials of continuous covariates: parsimonious parametric modeling (with discussion). Applied Statistics 43, 429-467.

[95] Rue H, Tjelmeland H, 2002. Fitting Gaussian Markov random fields to Gaussian fields. Scandinavian Journal of Statistics 29, 3149.

[96] Ruppert D, 2002. Selecting the number of knots for Penalized Splines. Journal of Computational and Graphical Statistics 11, 735-757.

[97] Ruppert D, Wand MP Carroll R, 2003. Semiparametric regression. Cambridge Univ. Press, Cambridge, U.K.

[98] Sampson PD, Guttorp P, 1992. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 87, 108-119.

[99] Schellenberg JR, Abdulla S, Nathan R, Mukasa O, Marchant T et al., 2001. Effect of largescale social marketing of insecticide-treated nets on child survival in rural Tanzania. Lancet 357, 1241-1247.

[100] Schmidt A, O'Hagan A, 2003. Bayesian inference for nonstationary spatial covariance structures via spatial deformations. Journal of Royal Statistical Society, Series B 65, 743-758.

[101] Seeger M,Williams CKI, Lawrence ND, 2003. Fast forward selection to speed up sparse Gaussian process regression. In Bishop CM and Frey BJ, Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003.

[102] Service MW, Townson H, 2002. The *Anopheles* vector. In Warrell DA and Gilles HM. Essential malariology, Fourth Edition. Arnold: London

[103] Silverman BW, 1986. Density estimation for statistics and data analysis. Chapman and Hall, London

[104] Smith T, Charlwood JD, Takken W, Tanner M, Spiegelhalter DJ, 1995. Mapping the density of malaria vectors within a single village in Tanzania. Acta Tropica 59, 1-18.

[105] Smith T, Killeen GF, Maire N, Ross A, Molineaux L, Tediosi F, Hutton G, Utzinger J, Dietz K, Tanner M, 2006. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *Plasmodium falciparum* malaria: Overview . American Journal of Tropical Medicine and Hygiene 75 (Suppl 2), 1-10.

[106] Snelson E and Ghahramani Z, 2006. Sparse Gaussian Processes using Pseudo-inputs. Neural Information Processing Systems 18.

[107] Snow RW, Armstrong JRM, Forster D, Winstanley MT, Marsh VM et al., 1992. Childhood deaths in Africa: uses and limitations of verbal autopsies. Lancet 340, 351-355.

[108] Snow RW, Marsh K, Le Sueur D, 1996. The need for maps of transmission intensity to guide malaria control in Africa. Parasitology Today 12, 455-456.

[109] Snow RW, Craig MH, Deichmann U, le Sueur D, 1999. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature 434(7030), 214-217.

[110] Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI, 2005. A preliminary continental risk map for malaria mortality among African children. Parasitology Today 15(3), 99-104.

[111] Socheat D, Denis MB, Fandeur T, Zhang Z, Yang H, Xu J, Zhou X, Phompida S et al., 2003. Mekong malaria. II. Update of malaria, multi-drug resistance and economic development in the Mekong region of Southeast Asia. Southeast Asian Journal of Tropical Medicine and Public Health 34, 1-102.

[112] Sogoba N, Vounatsou P, Bagayoko MM, Doumbia S, Dolo G, Gosoniu L, Traore SF, Toure YT, Smith T, 2007. The spatial distribution of Anopheles gambiae sensu stricto and *An. arabiensis* (Diptera: Culicidae) in Mali. Geospatial Health 2, 213-222.

[113] Spiegelhalter DJ, Best N, Carlin BP, van der Linde A, 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64, 583-639.

[114] Tanner M, De Savigny D, Mayombana C, Hatz C, Burnier E et al., 1991. Morbidity and mortality at Kilombero 1982-88. In: Feachem RG, Jamison DT, editors. Disease and mortality in Sub-Saharan Africa. Oxford University Press, Oxford. pp. 286-305.

[115] Tanser FC, Sharp B, le Sueur D, 2003. Potential effect of climate change on malaria transmission in Africa. Lancet 362, 1792-1798.

[116] Thomson MC, Adiamah JH, Connor SJ, Jawara M, Bennett S et al., 1995. Entomological evaluation of the Gambias National Impregnated Bednet Programme. Annals of Tropical Medicine And Parasitology 89, 229-242.

[117] Thomson MC, Connor SJ, D'Alessandro U, Rowlingson B, Diggle P, Cresswell M, Greenwood B, 1999. Predicting malaria infections in Gambian children from satellite data and bed net surveys: the importance of spatial correlation in the interpretation of results. The American Society of Tropical Medicine and Hygiene 61(1), 2-8.

[118] Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, Phindela T, Morse AP, Palmer TN, 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439, 576-579.

[119] Tierney L, 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22, 1701-1762.

[120] Todd JE, De Francisco A, ODempsey TJ, Greenwood BM, 1994. The limitations of verbal autopsy in a malaria-endemic region. Annals of Tropical Paediatrics 14(1), 31-36.

[121] Tubilla A, 1975. Error convergence rates for estimates of multidimensional integrals of random functions. Technical Report 72, Department of Statistics, Stanford University, Stanford, CA.

[122] Ver Hoef JM, Cressie N, Fisher RN, Case TJ, 2001. Uncertainty and spatial linear models for ecological data. In: Hunsaker C, Goodchild M, Friedl M, Case T, editors. Spatial uncertainty for ecology: implications for remote sensing and GIS applications. Springer-Verlag, New-York.

[123] Vounatsou P, Smith T, Gelfand AE, 2000. Modeling of Multinomial data with latent structure: application to geographical mapping of human gene and haplotype frequencies. Biostatistics 1, 177-189.

[124] Whittle P, 1954. On stationary process in the plane. Biometrika 41, 439-449.

[125] World Health Organization, 2004. Making health research work for people - Progress 2003-2004. http://www.who.int/tdr/publications/publications/pdf/pr17/malaria.pdf

[126] World Health Organization, 2007. Meeting on malaria risk mapping and burden estimation. 21-22 March, Geneva, Switzerland.

[127] World Resources Institute, 1995. African Data Sampler (CD-ROM) Edition I.

[128] Xia G, Gelfand AE, 2005. Stationary process approximation for the analysis of large spatial datasets. Technical Report 24, Institute of Statistics and Decision Science, Duke University.

[129] Zimmerman DL, Zimmerman MB, 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometrics 33, 77-91.

Curriculum vitae

Laura Gosoniu
Date and place of birth: 27th May 1979 in Bucharest, Romania
Nationality: Romanian

Education

2004–2007	PhD studies in Epidemiology at the Swiss Tropical Institute, Basel on "Development of Bayesian geostatistical models for mapping malaria transmission in Africa" under the supervision of PD. Dr. P. Vounatsou
2001–2003	Master of Science (MSc) in Applied Statistics and Probabilities at Faculty of Mathematics, University of Bucharest. Thesis on "ARIMA Series" under the supervision of Prof. Dr. M. Dumitrescu
1997–2001	Bachelor of Science (BSc) in Applied Statistics at Faculty of Mathematics, University of Bucharest. Thesis on "Multivariate Normal distribution" under the supervision of Prof. Dr. M. Dumitrescu

Professional Activities and Teaching

2001–2003	Statistician in the National Institute for Statistics, Bucharest
2003	Statistics Lecturer in the College of Statistics, University of Bucharest
2006–2007	Statistics Lecturer at the European Course in Tropical Epidemiology

Publications

Gosoniu L, Vounatsou P, Sogoba N, Maire N, Smith T. Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. Computational Statistics and Data Analysis, Special Issue on SPATIAL STATISTICS, in press.

Gosoniu L, Vounatsou P, Tami A, Nathan R, Grundmann H, Lengeler C, 2008. Spatial effects of mosquito bednets on child mortality. BMC Public Health 8:356.

Sogoba N, Vounatsou P, Bagayoko M, Doumbia S, Dolo G, **Gosoniu L**, Traore S, Smith T, Toure Y, 2008. Spatial distribution of the chromosomal forms of anopheles gambiae in Mali. Malaria Journal 7:205.

Sogoba N, Vounatsou P, Bagayoko MM, Doumbia S, Dolo G, **Gosoniu L**, Traore SF, Toure YT, Smith T, 2007. The spatial distribution of *Anopheles gambiae sensu stricto* and *An. arabiensis* (Diptera: Culicidae) in Mali. Geospatial Health 1(2), 213-222

Matthys B, Tschannen AB, Tian-Bi NT, Comoe H, Diabate S, Traore M, Vounatsou P, Raso G, **Gosoniu L**, Tanner M, Cisse G, N'Goran EK, Utzinger J, 2007. Risk factors for Schistosoma mansoni and hookworm in urban farming communities in western Cte d'Ivoire. Tropical Medicine and International Health 12, 709-723.

Gosoniu L, Vounatsou P, Sogoba N, Smith T, 2006. Bayesian modelling of geostatistical malaria risk data. Geospatial Health 1, 127-139

Matthys B, Vounatsou P, Raso G, Tschannen AB, Becket EG, **Gosoniu L**, Ciss G, Tanner M, N'Goran EK, Utzinger J, 2006. Urban farming and malaria risk factors in a medium- sized town in Cote d'Ivoire. American Journal of Tropical Medicine and Hygiene 75, 1223-1231

Raso G, Vounatsou P, **Gosoniu L**, Tanner M, N'Goran EK, Utzinger J, 2005. Risk factors and spatial patterns of hookworm infection among school children in a rural area of Western Cote d'Ivoire. International Journal for Parasitology 36, 201-210