Development of Spatial Statistical Methods for Modelling Point-Referenced Spatial Data in Malaria Epidemiology

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Armin Gemperli
aus Hildisrieden (LU)

Basel, September 2003

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät der Universität Basel auf Antrag von Prof. Dr. M. Tanner, Prof. Dr. H. Becher, Dr. P. Vounatsou und Prof. Dr. T.A. Smith.

Basel, den 23. September 2003

Prof. Dr. Marcel Tanner
Dekan der Philosophisch-Naturwissenschaftlichen Fakultät

To my loving wife Daniela

Contents

List of Figures

List of Tables

Abbreviations

ACM	Association for Computing Machinery
ADS	African Data Sampler
AEZ	Agro-Ecological Zone
AICC	Bias Corrected Akaike Information Criterion
AVHRR	Advanced Very High Resolution Radiometer
BOD	Burden of Disease
CALGO	Collected Algorithms (of the ACM)
CAR	Conditional Autoregressive Regression
CCD	Cold Cloud Duration
CGM	Conjugate Gradient Method
CI	Confidence Interval
CIMMYT	International Maize and Wheat Improvement Center
CLAVR	Clouds from AVHRR
CPS	Contraceptive Prevalence Surveys
CPU	Central Processing Unit
CRES	Center for Resource and Environmental Studies
CV	Coefficient of Variation
DAAC	Distributed Active Archive Center
DEM	Digital Elevation Model
DHS	Demographic and Health Survey
DIC	Deviance Information Criterion
EDC	EROS Data Center
EIR	Entomological Inoculation Rate
ELISA	Enzyme-Linked Immunosorbent Assay
EM	Expectation-Maximization
EROS	Earth Resources Observation System
ESA	European Space Administration
ESHAW	Eco-System and Health Analysis Workshop
FAO	Food and Agriculture Organization of the United Nations
GAC	Global Area Coverage
GCM	Global Climate Models
GEE	Generalized Estimating Equations
GIS	Geographical Information System
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GLS	Generalized Least Squares
GPS	Global Positioning System
GPS	Gibbs-Poole-Stockmeyer
HDF	Hierarchical Data Format
HR	Hazard Ratio

IEEE	Institute of Electrical and Electronics Engineers
IFVO	Instantaneous Field of View
IG	Inverse Gamma
IMR	Infant Mortality Rate
IMSL	International Mathematical and Statistical Libraries
IRD	Institute for Resource Development
ISD	Importance Sampling Density
LST	Land Surface Temperature
MARA/ARMA	Mapping Malaria in Africa/Atlas du Risque de la Malaria en Afrique
MCMC	Markov Chain Monte Carlo
MEASURE	Monitoring and Evaluation to Assess and use Results
NAG	Numerical Algorithms Group
NASA	National Aeronautics and Space Administration
NCDC	National Climatic Data Center
NDAAC	NASA Distributed Active Archive Center
NDVI	Normalized Difference Vegetation Index
NESDIS	National Environmental Satellite Data and Information Service
NOAA	National Oceanic and Atmospheric Administration
OR	Odds Ratio
PHN	Population, Health and Nutrition
PQL	Penalized Quasi Likelihood
QC	Quality Control
QMD	Quotient Minimum Degree
RBC	Red Blood Cell
REML	Restricted (or Residual) Maximum Likelihood
RIS8	8-bit Raster Image
RJMCMC	Reversible Jump Markov Chain Monte Carlo
SAR	Simultaneously Autoregressive Regression
SDS	Scientific Data Sets
SDSD	Satellite Data Service Division
SIAM	Society for Industrial and Applied Mathematics
SIR	Sampling-Importance-Resampling
SPA	Service Provision Assessments
STI	Swiss Tropical Institute
STI	Sexually Transmitted Infection
SWS	Soil Water Storage
TAMSAT	Tropical Applications of Meteorology using Satellite
USAID	U.S. Agency for International Development
USGS	United States Geological Survey
WFS	World Fertility Survey
WHO	World Health Organization
WRI	World Resource Institute

Acknowledgements

The present thesis was undertaken under the joint supervision of Dr. Penelope Vounatsou and Prof. Dr. Tom Smith. During my research, I could profit tremendously from the scientific knowledge and experience Dr. Vounatsou was willing to share as my main technical supervisor. Only thanks to this close cooperation I was able to find the results presented in this thesis. Drs. Smith and Vounatsou jointly guided me through my time as a doctoral student by lending patiently a helping hand. My sincerest thanks are addressed to them for their personal and scientific contributions.

I would like to thank Prof. Dr. Marcel Tanner, Director of the STI, for establishing the framework and infrastructure for my research at the Institute's level and Prof. Dr. Mitchell Weiss at the Department's level. Without these human and material resources this work would not have been possible.

A special thank you goes to Prof. Dr. Heiko Becher from Ruprecht-Karls Universität Heidelberg, who was willing to act as a co-referee in the role of an external expert.

Prof. Dr. Alan Gelfand from Duke University, Durham, whom I thank for helpful, stimulating and entertaining discussions concerning computational and spatial statistics. His critical and amicable thoughts and suggestions helped to improve this thesis and make it fit into up-to-date statistical research.

My warmest thanks are addressed to Christine Walliser, Cornelia Naumann and Eliane Ghilardi for professional administrative support throughout the study. Thanks are also expressed to senior scientists, staff and my fellow students at STI who all helped in one way or another: Dr. Salim Abdulla, Sohini Banerjee, Marlies Craig, Tobias Erlanger, Rainer Fretz-Männel, Dr. Sébastien Gagneux, Gaby Gehler-Mariacher, Reto Hagmann, Felix Heckendorn, Dr. Abraham Hodgson, Jennifer Jenkins, Olivia Keiser, Dr. Immo Kleinschmidt, Dr. Frank Krönke, Irene Küpfer, PD. Dr. Christian Lengeler, Dr. Tanya Marchant, Dr. Hassan Mishinda, Musawenkosi Mabaso, Dr. Ivo Müller, Dr. Victor Mwanakasale, Lucy Ochola, Dr. Seth Owusu-Agyei, Dr. Shubhangi Parkar, Sama Wilson, Grégoire Yapi-Yapi, Guojing Yang, PD Dr. Jakob Zinsstag, Tu Zuwu and last but not least Daniel Anderegg.

For the excellent maintenance of computing resources my thanks go to Dr. Urs Hodel, Simon Roelly und Martin Baumann. They were always able to find good solutions to persistent problems without even being asked for.

My many thanks are addressed to the STI library team of Heidi Immler, which currently are Mehtap Tosun, Manuel Minder and Annina Isler. As the every-year record holder in number of orders, I gave them an intense and busy time. I specially thank Nils Hug for providing a lot of information and tricks on how to fetch any desired document.

I am bound in gratitude to Klaus Schwinn from former Systor AG who indirectly laid the financial base to make this thesis possible and prioritized my personal wishes in a very unworldly way.

Finally my deepest thanks go to my family, parents, sister, brother in law, niece, nephew and my godson Nino Joel. And to the Fasciati family in Val Bregaglia. My love, thanks and admiration cannot be rightfully expressed to you, Daniela. I owe you so much.

Summary

Plasmodium falciparum malaria is the world's most important parasitic disease and a major cause of morbidity and mortality in Africa. However figures for the burden of malaria morbidity and mortality are very uncertain, since reliable maps of the distribution of malaria transmission and the numbers of affected individuals are not available for most of the African continent. Accurate statistics on the geographical distribution of different endemicities of malaria, on the populations at risk, and on the implications of given levels of endemicity for morbidity and mortality are important for effective malaria control programs. These estimates can be obtained using appropriate statistical models which relate infection, morbidity, and mortality rates to risk factors, measured at individual level, but also to factors that vary gradually over geographical locations.

Statistical models which incorporate geographical or individual heterogeneity are complex and highly parameterized. Limitations in statistical computation have until recently made the implementation of these models impractical for non-normal response data, sampled at large numbers of geographical locations. Modern developments in Markov chain Monte Carlo (MCMC) inference have greatly advanced spatial modelling, however many methodological and theoretical problems still remain. For data collected over a fixed number of locations (point-referenced or geostatistical data) such as malaria morbidity and mortality data used in this study, spatial correlation is best specified by parameterizing the variance-covariance matrix of the outcome of interest in relation to the spatial configuration of the locations (variogram modelling). This has been considered infeasible for a large number of locations because of the repeated inversion of the variance-covariance matrix involved in the likelihood. In addition the spatial correlation in malariological data could be dependent not only on the distance between locations but on the locations themselves. Variogram models need to be further developed to take into account the above property which is known as non-stationarity.

This thesis reports research with the objectives of: a) developing Bayesian hierarchical models for the analysis of point-referenced malaria prevalence, malaria transmission and mortality data via variogram modelling for a large number of locations taking into account non-stationarity and misalignment, while present in the data; b) producing country specific and continent-wide maps of malaria transmission and malaria prevalence in Africa, augmented by the use of climatic and environmental data; c) assessing the magnitude of the effects of malaria endemicity on infant and child mortality after adjusting of socio-economic factors and geographical patterns.

A comparison of the MCMC and the Sampling-Importance-Resampling approach for Bayesian fitting of variogram models showed that the latter was no easier to implement, did not improve estimation accuracy and did not lead to computationally more efficient estimation. Different approaches were proposed to overcome the inversion of large covariance matrices. Numerical algorithms especially suited within the MCMC framework were implemented to convert large covariance matrices to sparse ones and to accelerate inversion. A tesselation-based model was developed which partition the space into random Voronoi

tiles. The model assumes a separate spatial process in each tile and independence between tiles. Model fit was implemented via reversible jump MCMC which takes into account the varying number of parameters arised due to random number of tiles. This approach facilitates inversion by converting the covariance matrix to block diagonal form. In addition, this model is well suited for non-stationary data. An accelerated failure time model was developed for spatially misaligned data to assess malaria endemicity in relation to child mortality. The misalignment arised because the data were extracted from databases which were collected at a different set of locations.

The newly developed statistical methodology was implemented to produce smooth maps of malaria transmission in Mali and West- and Central Africa, using malaria survey data from the Mapping Malaria Risk in Africa (MARA) database. The surveys were carried out at arbitrary locations and include non-standardized and overlapping age groups. To achieve comparability between different surveys, the Garki transmission model was applied to convert the heterogeneous age prevalence data to a common scale of a transmission intensity measure. A Bayesian variogram model was fitted to the transmission intensity estimates. The model adjusted for environmental predictors which were extracted from remote sensing. Bayesian kriging was used to obtain smooth maps of the transmission intensity, which were converted to age-specific maps of malaria risk. The West- and Central African map was based on a seasonality model we developed for the whole of Africa. Expert opinion suggests that the resulting maps improve previous mapping efforts. Additional surveys are needed to increase the precision of the predictions in zones were there are large disagreement with previous maps and data are sparse.

The survival model for misaligned data was implemented to produce a smooth mortality map in Mali and assess the relation between malaria endemicity and child and infant mortality by linking the MARA database with the Demographic and Health Survey (DHS) database. The model was adjusted for socio-economic factors and spatial dependence. The analysis confirmed that mothers education, birth order and preceding birth interval, sex of infant, residence and mothers age at birth have a strong impact on infant and child mortality risk, but no statistically significant effect of *P. falciparum* prevalence could be demonstrated. This may reflect unmeasured local factors, for instance variations in health provisions or availability of water supply in the dry Sahel region, which could have a stronger influence than malaria risk on mortality patterns.

Zusammenfassung

Plasmodium falciparum Malaria ist die weltweit bedeutendste parasitäre Krankheit und Hauptursache der hohen Sterberate in Afrika. Aktuelle Schätzungen malariabedingter Krankheits und Sterbehäufigkeit in Afrika sind allerdings ungenau, weil verlässliche Karten, welche die geographische Verteilung der Krankheit und der davon Betroffenen aufzeigen, nicht vorhanden sind. Damit Projekte zur Eindämmung von Malaria effizient durchgeführt werden können, ist es jedoch notwendig über eine genaue Statistik der Anzahl betroffener Menschen, sowie der Auswirkung von lokalem Malariavorkommen auf das Sterblichkeits- und Krankheitsrisiko, zu verfügen. Geeignete Schätzverfahren setzen Infektions-, Sterblichkeits- und Krankheitsrate in Beziehung zu Risikofaktoren. Bei diesen Faktoren kann es sich entweder um lokale Umweltfaktoren handeln, oder aber um Merkmale, die individuell für jede untersuchte Person gelten.

Statistische Modelle welche geographische oder individuelle Einflussfaktoren berücksichtigen sind komplex und wurden in der Malariaforschung bisher kaum eingesetzt. Dies gilt insbesondere für die Analyse nicht-normalverteilter, grossräumig erhobener Daten. Erst die moderne Errungenschaft der Markov chain Monte Carlo (MCMC) Methode vermochte die Schätzung für solche Daten signifikant verbessern, obwohl auch damit noch immer methodologische Probleme verbunden sind. Für Stichproben die an bestimmten, genau definierten Orten erhoben wurden (geostatistische Daten), wird die räumliche Abhängigkeit bevorzugt mit einer speziell parametrisierten Kovarianzmatrize modelliert (Variogrammodellierung). Diese Modellierung ist jedoch nicht mehr möglich, falls die Stichprobe an sehr vielen verschiedenen Orten erhoben wurde, weil dann die Grösse dieser Kovarianzmatrize eine numerische Analyse verunmöglicht. Bei der MCMC Methode muss die Kovarianzmatrize wiederholt invertiert werden. Dies ist bei grossen Matrizen zeitintensiv und kann zu einer nicht vernachlässigbaren Kummulation von numerischen Fehlern führen. Hinzu kommt, dass die räumliche Abhängigkeit von Malariadaten nicht bloss von der Distanz zwischen zwei Stichproben abhängt, sondern möglicherweise auch von deren absoluter Lage (nicht-stationäre Daten), was neuartige statistische Verfahren benötigt.

Die Forschung in Zusammenhang mit dieser Doktorarbeit hatte folgende Ziele: a) Entwicklung von bayesschen hierarchischen Methoden um geostatistische Malaria-Häufigkeits, Übertragungs- und Sterblichkeitsdaten mittels Variogrammodellierung zu analysieren, wobei auf das Problem der Nicht-Stationarität und die grosse Anzahl der Stichprobenorte eingegangen wird; b) Erstellen von Karten für den Afrikanischen Kontinent um die Häufigkeit und Übertragungsraten von Malaria, unter Berücksichtigung von Klima- und Umweltfaktoren, darzustellen; c) Schätzung der Wirkung die ein bestimmtes Malariarisiko auf die Säuglings- und Kindersterblichkeit ausübt, unter Berücksichtigung sozioökonomischer und räumlicher Aspekte.

Ein Vergleich von MCMC mit der Sampling-Importance-Resampling Methode für bayessches Schätzen von Variogrammen zeigte, dass die zweite Methode weder einfacher anzuwenden war, noch zu besseren Schätzern führte. Zudem war die Berechnung mit

dieser Methode nicht effizienter. Verschiedene Verfahren wurden vorgeschlagen um die Inversion grosser Kovarianzmatrizen zu erleichtern. Dies beinhaltete numerische Algorithmen um grosse Kovarianzmatrizen zu dünn besetzten Matrizen zu transformieren, was sich in Zusammenhang mit der MCMC Methode besonders gut eignet. Ein Partitionierungsverfahren, das den Raum in Voronoi Kacheln zerlegt, wurde entwickelt. Dabei wurde ein separater räumlicher Prozess für jede Kachel gebildet und Unabhängigkeit zwischen den Kacheln postuliert. Dieses Modell wurde mittels Reversible Jump MCMC (RJMCMC) geschätzt. Da die Kovarianzmatrize im Partitionierungsverfahren block-diagonale Struktur besitzt, wird die Matrizeninversion erleichtert. Diese Methode eignet sich zudem um nichtstationäre, räumliche Daten zu analysieren. Des Weiteren wurde ein Überlebensmodell entwickelt für die Analyse räumlicher, nicht-ausgerichteter Datensätze, um den Effekt, den das Malariarisiko auf die Kindersterblichkeit ausübt, abzuschätzen. Die Nicht-Ausrichtung der Daten rührt daher, dass die beiden Datensätze, von welchen die Mortalitätsrate, respektive das Malariarisiko extrahiert wurden, an verschiedenen Orten erhoben wurden.

Die neu entwickelten Methoden wurden angewendet um Karten der Übertragungsrate von Malaria für Mali sowie West- und Zentralafrika zu erstellen. Die zugrunde liegenden Daten stammen aus der "Mapping Malaria Risk in Africa" (MARA) Datenbank, einer Sammlung von beliebigen Erhebungen an unterschiedlichen Orten und nichtstandardisierten, überlappenden Altersgruppierungen. Um die verschiedenen Erhebungen vergleichen zu können wurde das Garki Modell angewendet, das alterspezifische Häufigkeitsdaten in ein einheitliches Malaria Übertragungsmass konvertiert. Ein bayessches Variogrammodel wurde für die errechneten Übertragungsraten geschätzt, wobei Umweltfaktoren aus Fernerkundungsdaten berücksichtigt wurden. Bayessches Kriging wurde angewandt um Karten der Übertragungsintensität von Malaria herzustellen. Diese wurden schliesslich zu alterspezifischen Häufigkeits-Karten transformiert. Die hergestellten Karten für West- und Zentral Afrika basieren auf einem eigens entwickelten Saisonalitätsmodell. Expertenmeinungen zeigen, dass diese Schätzungen bestehende Karten verbessern. Allerdings werden weitere Erhebungen nötig sein um die Genauigkeit in jenen Gebieten zu erhöhen, wo grössere Abweichungen im Vergleich zu früheren Karten bestehen, oder wo wenig Stichproben erhoben wurden und deswegen wenig Datenmaterial vorhanden ist.

Das Überlebensmodell mit nicht-ausgerichteten Daten wurde verwendet um eine Sterblichkeitskarte für Mali zu produzieren und um die Säuglings- und Kindersterblichkeit in Abhängigkeit des Malariarisikos zu modellieren. Dafür wurden die MARA Datenbank und die "Demographic and Health Survey" (DHS) Datenbank kombiniert. Das Modell berücksichtigte sozio-ökonomische Faktoren und räumliche Abhängigkeiten. Die Analyse bestätigte, dass die Schulbildung der Mutter, die Geburtenfolge, die Länge des vorhergehenden Geburtintervalls, das Geschlecht des Kindes, der Wohnort sowie das Alter der Mutter bei der Geburt des Kindes einen statistisch signifikanten Einfluss auf die Säuglings- und Kindersterblichkeit haben. Jedoch konnte kein Zusammenhang zwischen der Sterblichkeit und dem Auftreten von *P. falciparum* festgestellt werden. Es ist denkbar, dass in den untersuchten Gebieten unberücksichtigte Faktoren, wie die Gesundheitsversorgung oder die Verfügbarkeit von Wasser, einen stärkeren Einfluss auf das Sterblichkeitsrisiko ausüben, als Malaria.

CHAPTER 1

Introduction: Biology and epidemiology of malaria

1.1 The malaria parasite in the human

Malaria is a vector born disease caused by protozoan parasites of the genus *Plasmodium*. There are four malaria parasite species in humans, namely *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale*. Parasites are transmitted from person to person by female mosquitoes of the genus *Anopheles*. Different species appear in different regions. The transmission can be seasonal, depending on the dynamics of the vector population.

The life cycle of the parasite is depicted in figure 1.1. It starts with the inoculation of the parasite into the human blood by the bite of a female *Anopheles* mosquito. Within half an hour, the sporozoites reach the liver and invade the liver cells. Within the liver cells, the trophozoites start their intracellular asexual division. At the completion of this phase, thousands of erythrocytic merozoites are released from each liver cell. The time taken for the completion of the tissue phase is variable, depending on the infecting species; (5–6 days for *P. falciparum*). The merozoites invade the red blood cell (RBC), and then develop through the stages of rings, trophozoites, early- and mature schizonts; each mature schizont consists of thousands of erythrocytic merozoites. These merozoites are released by the lysis of the RBC and immediately invade uninfected red cells.

This whole cycle of invasion - multiplication - release - invasion takes about 48 hours in *P. falciparum* infections. The contents of the infected cell that are released with the lysis of the RBC stimulate the Tumor Necrosis Factor and other cytokines, which results in the characteristic clinical manifestations of the disease. A small proportion of the merozoites undergo transformation into gametocytes. Mature gametocytes appear in the peripheral blood after a period of 8–11 days of the primary attack in *P. falciparum*, they rise in number until three weeks and decline thereafter, but circulate for several weeks. The gametocytes enter the mosquito when it bites an infected individual.

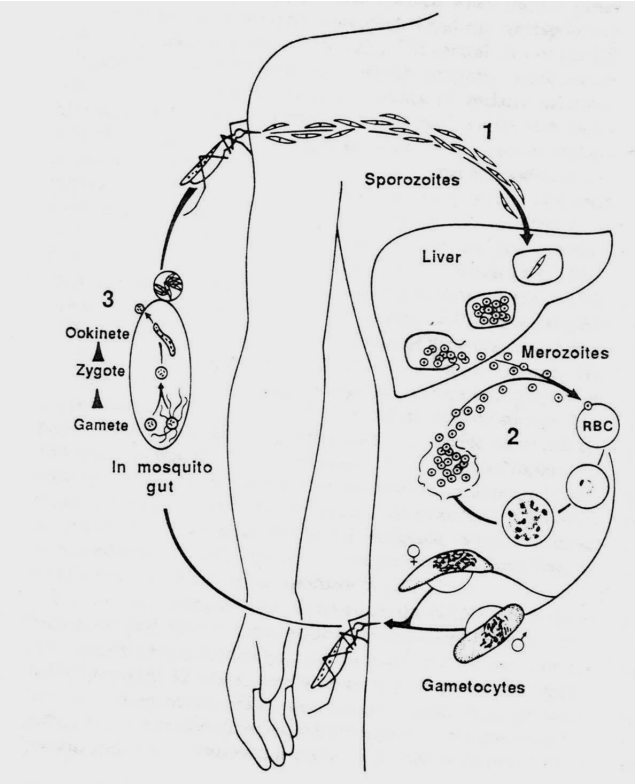


Figure 1.1: The life cycle of malaria.

1.2 The malaria parasite in the vector

Human malaria is transmitted by mosquitoes of the genus *Anopheles*. Out of the 360 species there are about 45 with the ability to transmit malaria of humans. *Anopheles* live worldwide, but the transmission of malaria occurs predominantly in tropical and sub-tropical zones (figure 1.2). Free of *Anopheles*, always means free of malaria, but not vice-versa.

When, after the blood meal, the malaria parasite enters the mosquito, the gametocytes continue their development (Sporogony). The male and female gametes fuse and form into a zygote. This transforms into an ookinete which penetrates the gut wall and becomes an oocyst. The oocyst divides asexually into numerous sporozoites which reach the salivary gland of the mosquito, where they can be transmitted when the mosquito next takes a blood-meal. The sporogony in the mosquito takes about 10–20 days dependent on air temperature and thereafter the mosquito remains infective for 1–2 months, if it survives. There is no sporogony at a temperature below 15°C.

Only the female mosquito takes a blood meal (male *Anopheles* feed on nectar) which is necessary for the development of eggs. Two to three days after the blood meal, which is taken during the night or at dawn, the female anopheline lays around hundred eggs. During her life of several weeks, she can therefore produce more than 1,000 eggs. The eggs

Figure 1.2: Countries with endemic malaria transmission (WHO, 2000).

are always laid on water surface, with preference for swamps or shallow water. They may also breed in water containers or tree holes. The oval eggs are one millimeter long and require about two weeks to develop into adult mosquitoes. They fly only short distances of a few kilometers. Their preferred location is close to human houses.

There are behavioral differences between mosquito species, which are important for the study of the geographical distribution of the vector. The most important *Anopheles* species in Africa are members of the *A. gambiae* complex and *A. funestus*. Five species of the *A. gambiae* complex are vectors of malaria and two of them (*A. gambiae s.s.* and *A. arabiensis*) are the most widely distributed throughout sub-Saharan Africa. *A. arabiensis* predominates in drier and *A. gambiae s.s.* in more humid areas. Their preferred breeding sites are sunlit temporary pools or rice fields. *A. arabiensis* feeds on humans and animals while *A. gambiae s.s.* feeds on humans predominantly, prefers indoor locations for biting and resting, and has a higher vectorial capacity than other species. Two salt water species of the *A. gambiae* complex (*A. melas* and *A. merus*) are found in West- and East Africa, respectively where *A. merus* feeds mainly on animals and *A. melas* bites humans or animals. Another major vector of malaria in many parts of tropical and sub-tropical Africa is *A. funestus* of the *A. funestus* group. It feeds mainly on humans and rests and bites indoors. It breeds in semi-permanent and permanent water with vegetation and swamps and is associated with all-year malaria transmission.

1.2.1 Vector ecology

The short fly range and the preferred locations for hosting and breeding are responsible for large local differences in the geographical distribution of the anopheline. The effect the environment has on the malaria vector is further determined by rainfall and temperature which affect mosquito survival and the duration of the parasite life cycle in the vector.

Temperature

Temperature influences the survival of the parasite during its life-cycle in the *Anopheles* vector. All species have the shortest development cycle around 27–31°C which lasts from 8 to 15–21 days depending on species. The lower the temperature, the longer the cycle. Below 19°C for *P. falciparum*, the parasites are unlikely to complete their cycle and hence to further propagate the disease. Temperature also modifies the vectorial capacity of the *Anopheles*. Optimal temperature values, ranging from 22°C to 30°C, lengthen the life-span of the mosquitoes and increase the frequency of blood meals taken by the females, to up to one meal every 48 hours. Higher temperatures also shorten the aquatic life cycle of the mosquitoes from 20 to 7 days and reduce the time between emergence and oviposition, as well as the time between successive ovipositions.

Temperature affects also the vector. In tropical climate the *Anopheles* eggs hatch within 2–3 days of laying, whereas for colder temperatures it can require 2–3 weeks. At minimum temperatures near the freezing point, African vector populations are effectively obliterated and at very high temperatures of above 40°C, the *Anopheles* die (Craig et al., 1999). As a consequence of all the temperature requirements malaria transmission becomes less frequent at high altitudes. Near the equator there are no *Anopheles* above 2,500 meters altitude and in the other regions there are none above 1,500 meters altitude.

Rainfall and humidity

Rainfall and humidity impact to a great extend the living conditions of the *Anopheles* (Thomson et al., 1996). Temporal ponds, created by increasing rainfall, are responsible for ideal vector breeding conditions. However rainfall can also destroy existing breeding places: Heavy rain can change breeding pools into streams, impede the development of mosquito eggs or larvae, or simply flush the eggs or larvae out of the pools (Ribeiro et al., 1996; Craig et al., 1999). Conversely exceptional drought conditions can turn streams into pools. The appearance of such opportunistic mosquito breeding sites sometimes precede epidemics. The interaction between rainfall, evaporation, runoff, and temperature modulates the ambient air humidity which in turn affects the survival and activity of *Anopheles* mosquitoes. Mosquitoes can survive if relative humidity is at least 50 or 60 percent. Higher values lengthen the life-span of the mosquitoes and enable them to infect more people. As a proxy for humidity and rainfall, the vegetation index is shown to be a successful indicator (Thomson et al., 1997).

1.3 Malaria mortality, morbidity and immunity

The incubation period for *P. falciparum* malaria (the time between the inoculation of the parasite and the first medical symptoms) is around 8–15 days. The main symptoms in all malaria forms are (periodic) fever outbreaks. The most severe form of malaria morbidity is cerebral malaria, which is characterized by coma with detectable parasitemia, and it is accompanied by the obstruction of capillaries in the central nervous system. Cerebral malaria is a severe complication of clinical malaria in areas with a malaria transmission of 10–20 infectious bites per year. Other major complications are severe anaemia, acute renal insufficiency or failure, hepatic or pulmonary problems, jaundice and gastrointestinal symptoms such as abdominal pain, nausea, vomiting, diarrhea or constipation (Gilles and Warrell, 1993).

Acquired immunity is developed after repeated infections. Adults can tolerate parasites without developing symptoms. Infants are protected due to maternal antibodies in the first 3–6 months of life. Until they have built their own immunity, they are vulnerable to clinical malaria episodes. Infant mortality in high endemic malaria regions is high (Kalipeni, 1993; Smith et al., 2001). Pregnancy leads to suppression of immunity. High parasitemia is observed during the first pregnancy and is decreasing for further pregnancies (Brabin, 1983; McGregor, 1984; Steketee et al., 2001). The malaria infection of the mother is a major reason for abortion and stillbirth and reduces the survival chances of a newborn (McCormick, 1985; Bouvier et al., 1997).

1.4 Measures of malaria endemicity and transmission

Malaria prevalence is the most widely available measure of endemicity. Prevalence data are obtained by community surveys of individuals who are tested for the presence of parasites in their blood. The acquiring of partial immunity in older children and adults in endemic malaria areas leads to age-dependence of this measure. Prevalence is only an indirect measure of the amount of malaria transmission, because malaria infections may persist for varying length of time. A direct transmission measure is the incidence of the disease, that is the number of new cases of malaria diagnosed per unit time and person. Incidence data can be biased when collected in health centers, because it may reflect patients' access to these centers. They also depend on accurate estimates of the population at risk.

The most common entomological measure of malaria transmission is the entomological inoculation rate (EIR), which is defined as the number of sporozoite positive mosquito bites per person and time unit (typically year) and is the product of the anopheline density, the human biting rate and the sporozoite index (the number of infective mosquitoes) (Macdonald, 1957; Hay et al., 2000). The human biting rate can be measured by human bait catches or mosquito traps.

One of the best documented studies on malaria transmission was conducted in 1971–1973 in the Garki area of Northern Nigeria (Molineaux and Gramiccia, 1980). Using the Garki data, a mathematical model was formulated (Dietz et al., 1974) that makes

predictions of the age-specific prevalence of *P. falciparum* in humans as a function of the vectorial capacity. It can be used to link several measures of transmission (including the vectorial capacity and the entomological inoculation rate) and the malaria prevalence.

1.5 Measures of malaria mortality

There are basically four ways to measure mortality attributable to malaria: from clinical records, when the cause of death is identified; from observing the rise in mortality during malaria epidemics; from observing the fall in mortality when malaria is brought under control; or by calculating the mortality necessary to maintain the observed level of the sickling gene in a balanced polymorphism (Molineaux, 1985).

Clinical records in Africa hardly ever include post-mortem series and, more seriously, introduce bias because they are only derived from tertiary-care facilities and very rarely include young children and infants. The fact that most people die outside the hospital and the limitation of paediatric beds in Africa make clear that information on death certificates are a poor measure of malaria mortality (Snow and Marsh, 1998).

Interactions between malaria and other diseases in areas of high malaria endemicity make it difficult to quantify the mortality attributable to malaria. Malaria may be a relevant risk factor for many deaths even when it is not the immediate cause (Molineaux, 1985). Moreover, low birth weight is an important risk factor for infant mortality and it is known to arise because of both prematurity and intrauterine growth retardation resulting from malaria infection of the mother during pregnancy (Steketee et al., 2001). Molineaux (1985) emphasized that is as important to look at the relationship of malaria endemicity with all-cause mortality as it is to look at its relationship with malaria specific deaths.

1.6 Spatial epidemiology of malaria

Spatial epidemiology is the study of the spatial/geographical distribution of the incidence of disease and its relationship to potential risk factors. The origins of spatial epidemiology go back to 1855 with the seminal work of Snow on cholera transmission. He mapped the cholera cases together with the locations of water source in London, and showed that contaminated water was the major cause of the disease. Spatial analysis in the nineteenth and twentieth century was mostly employed by plotting the observed disease cases or rates (Howe, 1989). Recent methods make use of computer based cartographic methods, satellite derived data and modern statistical methods and allow an integrated approach to address both tasks; inference on the geographical distribution of a disease and its prediction at new locations.

Spatial epidemiological tools applied in malaria research can identify areas of high malaria transmission and assess potential environmental and other risk factors which can explain variation in space. Elucidating the relation between environment and malaria allows prediction of the impact environmental changes have on malaria risk, including

the effect of global warming and of man made interventions (dams, change in agriculture, urbanization, etc.). The understanding of environmental aspects of malaria is important for effective malaria interventions, which not only focus on the parasite directly, but also on the mosquito vector and its living conditions. Maps of malaria distribution provide estimates of the disease burden and assist in the evaluation of intervention programs.

1.6.1 GIS and remote sensing

Advances in computer cartography and the development of Geographic Information System (GIS) brought a new impetus to the field of spatial epidemiology. GIS is a computerized database management system for the capture, storage, retrieval, analysis and display of spatially referenced (geo-referenced) data. It classifies data coming from disparate sources into map layers, then linking these layers by spatially matching them, querying and analyzing them together to produce new information and hypotheses. In order to use survey information in GIS, the data must be geographically identified (geolocated). This is often accomplished by using the Global Positioning System (GPS) (August et al., 1994; Logsdon, 1992; Wells, 1988).

A general introduction to GIS and its use in tropical and malaria epidemiology is given by Robinson (2000). Fully descriptive malaria research using GIS software is done by Hightower et al. (1998). Omumbo et al. (1998) use GIS to quantify the relation between occurrence of anopheline and environmental variables and Carter (2000) use GIS to investigate in the geographical relation between malaria risk and its vector breeding sites. Schellenberg et al. (1998) spatially link malaria incidence to households using GIS, to investigate the relation between malaria related hospital admission rates and distance to hospital. A similar approach was chosen by van der Hoek et al. (2003) to investigate the malaria risk in relation to the distance between household locations and rivers. GIS has been used in combination with environmental data by Rogers et al. (2002) to predict entomological inoculation rates and the occurrence of different species of the *A. gambiae* complex in Africa. Suitability maps of malaria transmission in Africa based on climatic models using GIS have been produced by Snow et al. (1998) for Kenya and by Craig et al. (1999) for the whole of Africa.

The relation between malaria risk and environmental indices derived by remote sensing is described by Connor et al. (1998) and Thomson et al. (1996, 1997). These authors describe remote sensing databases which are publicly available and are proven to give useful contribution to malaria research. Hay et al. (2002) analyzed long-term meteorological data from four sites in high-altitude in East Africa and concluded that claimed associations between local malaria resurgence and regional changes in climate are overly simplistic.

1.6.2 Spatial statistical methods

Many analysis of remote sensed data in relation to malaria make little, no, or limited use of field data and few of them have allowed for the geographical structure of the data.

Geographical data are correlated in space. Data in close geographical proximity is more likely to be influenced by similar factors and thus affected in a similar way. In the case of malaria, spatial correlation is present at both, short and large scales, reflecting the transmission of malaria infection by the mosquitoes which fly over short distances and the effects of environmental factors which determine mosquito survival over large areas.

Standard statistical methods assume independence of observations. When using this methods to analyze spatially correlated data, the standard error of the covariate parameters is underestimated and thus the statistical significance is overestimated (Cressie, 1993, pp.20–21). This was demonstrated in a malaria application by Thomson et al. (1999).

Spatial statistical methods incorporate spatial correlation according to the way geographical proximity is defined. Proximity further depends on the geographical information, which can be available at areal level or at point-location level. Areal unit data are aggregated over contiguous units (countries, districts, census zones) which partition the whole study region. Proximity in space is defined by their neighboring structure. Point-referenced or geostatistical data are collected at fixed locations (households, villages) over a continuous study region. Proximity in geostatistical data is determined by the distance between sample locations.

Bayesian methods have been applied extensively in recent years for modelling both, areal unit and geostatistical data because they allow flexible modelling and inference and provide computational advantages via the implementation of Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990). The spatial structure is commonly introduced in a hierarchical fashion via the prior distribution of area of site-specific random effects, although spatial dependence can be built directly on Gaussian response data. The choice of prior distributions or spatial models depend on the type of spatial data.

In areal data, simultaneously autoregressive (SAR) models (Whittle, 1954), conditional autoregressive (CAR) models (Clayton and Kaldor, 1987) and modifications (Besag at al., 1991; Sun et al., 2000)) have been suggested as prior specifications in the Bayesian approach. In geographical mapping of disease and mortality rates spatially autoregressive models are employed assuming Poisson count data (Bernardinelli and Montomoli, 1992; Clayton et al., 1993; Waller et al., 1997). Smith et al. (1995) applied these models in malaria epidemiology to map the malaria vector density in a single village and Kleinschmidt et al. (2001b) have implemented CAR models for mapping malaria incidence rates data. Vounatsou et al. (2000) and Gelfand et al. (2003) extended CAR models for multinomial response data with application to geographical mapping of allele and haplotype frequencies.

Geostatistical models introduce spatial correlation in the correlation matrix of location-specific random effects which model a latent Gaussian spatial process (Cressie, 1993; Diggle et al., 1998). In case of isotropy, the covariance between any two sites depends only on the distance between them. Typical covariogram functions are the exponential, Gaussian, Cauchy, spherical and Bessel (Ecker and Gelfand, 1999). Under the assumption of stationarity, which postulates that the spatial correlation is a function of distance and independent of location, the covariance determines the well known variogram. Despite the usefulness of stationary spatial models, in many applications including those in malaria epidemiology, the spatial structure changes with the location especially over large geographical areas.

Modelling alternatives to take into account non-stationary spatial covariance include the spatial deformation approach (Sampson and Gottorp, 1992), kernel convolution approach (Higdon et al., 1998) and the spectral approaches (Nychka et al., 2002; Fuentes et al., 2002). An issue of practical concern is that the computation of the prior distribution of random effects requires the inversion of the covariance matrix of the spatial process. Moreover, implementation of the usual iterative model fit requires repeated inversions of this matrix which for large number of locations is not feasible within practical time constraints. Gelfand et al. (1999) suggested replacing matrix inversion with simulation using importance sampling. Christensen et al. (2002) suggest speeding MCMC implementation via Langevin-Hastings updates. Kim et al. (2002) use piecewise Gaussian processes to model non-stationary Gaussian permeability data. They overcome matrix inversion by partitioning the space in random tesselations and assuming separate spatial processes in the tiles and independence between the tiles of the tesselation.

In geostatistics, spatial prediction is referred to as kriging. Matheron (1963) coined this term in honor of the South African mining engineer D. G. Krige. Bayesian kriging (Diggle et al., 1998) allows estimation of the prediction error, a feature which is not possible in classical kriging estimators.

Geostatistical methods have occasionally been applied to disease mapping. Carrat and Valleron (1992) give an introduction to kriging for epidemiologists. A Bayesian spatial model using MCMC has been employed by Alexander et al. (2000) and applied to individual-level counts of the nematode *Wucheria bancrofti*, a parasite of humans which causes lymphatic filariasis. There is only little research done in using kriging in malaria mapping. Ribeiro et al. (1996) mapped the vector density in a single village, by fitting a standard regression model and applying classical kriging on the model residuals. A similar approach was revisited by Kleinschmidt for mapping malaria prevalence in Mali (Kleinschmidt et al., 2000) and for the whole of West Africa (Kleinschmidt et al., 2001a). The only approach so far for mapping malaria prevalence using Bayesian kriging has been presented by Diggle et al. (2002). These authors applied MCMC to map malaria in The Gambia but use only few surveys. The purpose of their analysis was thus rather the demonstration of the methodology. It needs to be further discussed how this approach can be extended to larger malariological dataset and such with non-stationary spatial structure.

1.7 Objectives of the thesis

The main objectives of this research were to a) develop Bayesian variogram models for the analysis of point-referenced prevalence and mortality data collected over a large number of locations and b) to validate and implement the developed models in the area of spatial malaria epidemiology in order to produce smooth maps of malaria transmission in Africa and assess relations between child mortality and malaria endemicity. The specific objectives in statistical methodology were

- assessment of existing geostatistical methods in modelling malaria data collected over a large number of locations. The methods were evaluated in terms of ease of

implementation, estimation accuracy and computational efficiency. This is addressed in chapter 2;

- development of geostatistical survival models for mapping mortality data. The analysis is reported in chapter 3;
- modelling geostatistical misaligned data for assessing the impact of site-specific malaria endemicity on child mortality collected at different set of locations. This is the topic of chapter 4;
- development of models for non-stationary, geostatistical malaria prevalence data. These models are describes in chapter 8;
- evaluating numerical algorithms to improve computation of geostatistical models using MCMC. This is addressed in chapter 7;
- development of models for mapping malaria transmission. The maps are presented in chapters 5 and 6.

The developed statistical methods were applied on data extracted from the MARA/ARMA and DHS databases in order to

- identify factors related with geographical differences in infant mortality risk in Mali and assess the effect of malaria endemicity on infant mortality;
- evaluate the impact of site-specific malaria endemicity on child mortality rate in Mali;
- produce smooth maps of malaria transmission and age-specific malaria risk in Mali allowing for the effect of environmental factors;
- map malaria transmission in West- and Central Africa adjusted for age, seasonality and environmental factors.

CHAPTER 2

Fitting generalized linear mixed models for point-referenced spatial data

Gemperli A. and Vounatsou P.
Swiss Tropical Institute, Basel, Switzerland

This paper has been published in *Journal of Modern Applied Statistical Methods* **2** 481–495, 2003.

Abstract

Non-Gaussian point-referenced spatial data are frequently modelled using generalized linear mixed models (GLMM) with location-specific random effects. Spatial dependence can be introduced in the covariance matrix of the random effects. Maximum likelihood-based or Bayesian estimation implemented via Markov chain Monte Carlo (MCMC) for such models is computationally demanding especially for large sample sizes because of the large number of random effects and the inversion of the covariance matrix involved in the likelihood. Sampling-Importance-Resampling (SIR) has been proposed to overcome matrix inversion. In this study, we review three fitting procedures, the Penalized Quasi Likelihood method, the MCMC and the SIR method. We assess these methods in terms of estimation accuracy, ease of implementation and computational efficiency using a spatially structured dataset on infant mortality from Mali. The objective of data analysis was to assess the effect of maternal and socio-economic parameters on infant mortality and produce a smooth map of mortality risk in Mali.

Keywords: geostatistics; infant mortality; kriging; Markov chain Monte Carlo; penalized quasi likelihood; risk mapping; sampling-importance-resampling.

2.1 Introduction

Point referenced spatial data arise from observations collected at geographical locations over a fixed continuous space. Proximity in space introduces correlations between the observations rendering the independence assumption of standard statistical methods invalid. Ignoring spatial correlation will result in underestimation of the standard error of the parameter estimates, and therefore liberal inference as the null hypothesis is rejected too often. A wide range of analytical tools within the field of geostatistics have been developed concerning with the description and estimation of spatial patterns, the modelling of data in the presence of spatial correlation and the kriging, that is the spatial prediction, at unobserved locations.

Statistical inference of point referenced data often assumes that the observations arise from a Gaussian spatial stochastic process and introduce covariate information and possibly trend surface specification on the mean structure while spatial correlation on the variance-covariance matrix, $\boldsymbol{\Sigma}$ of the process. Under second order stationarity, $\boldsymbol{\Sigma}$ determines the well-known variogram. When isotropy is also assumed, the elements of $\boldsymbol{\Sigma}$ are modelled by parametric functions of the separation between the corresponding locations. For non-Gaussian data, the spatial correlation is modelled on the covariance structure of location-specific random effects introduced into the model and assumed to arise from a Gaussian stationary spatial process.

For Gaussian data, the generalized least squares (GLS) approach can be used iteratively to obtain estimates $\hat{\boldsymbol{\beta}}$ of the regression coefficients conditional on the covariance parameters. The covariance parameters $\boldsymbol{\theta}$ can be estimated conditional on $\hat{\boldsymbol{\beta}}$ by fitting the semivariogram empirically or by maximum likelihood or restricted maximum likelihood

methods (Zimmerman and Zimmerman, 1991).

Statistical estimation for non-Gaussian data is based on the theory of generalized linear mixed models (GLMM). A common approach is to integrate out the random effects and proceed with maximum likelihood based approaches for estimating the covariate and covariogram parameters. This integration can be implemented numerically (Anderson and Hinde, 1988; Preisler, 1988; Lesaffre and Spiessens, 2001) when dimensionality is low or via approximations. Breslow and Clayton (1993) showed, that for known covariance parameters, the Laplace approximation leads to the same estimator for the fixed and random effects as the one arising by maximizing the penalized quasi-likelihood (PQL). Implementation of this approach requires iterating between iterated weighted least squares for estimating the fixed and random effects and maximizing the profile likelihood for estimating the covariance parameters. An extension of the PQL procedure is discussed by Wolfinger and O'Connell (1993). The PQL approach is implemented in some statistical packages due to its relative simplicity, however it provides biased estimates when the number of random effects increases (McCulloch, 1997; Booth and Hobert, 1999) or when the data are far from normal.

The generalized estimating equation methods developed by Liang and Zeger (1986) and Zeger and Liang (1986) estimate covariate effects under the assumption of independence, but correct their standard error to account for the spatial dependence. The method is unable to estimate the spatial random effects. The EM algorithm (Dempster et al., 1977) has been implemented in model fit by treating the spatial random effects as "missing" data. The intractable integration of the random effects which is required in the E-step is overcome by simulation, such as Metropolis-Hastings algorithm (McCulloch, 1997) or importance sampling/rejection sampling method (Booth and Hobert, 1999). For spatial settings, particular Pseudo-Likelihood approaches have been established which capture solely the site to site variation between pairs or groups of observations (Besag, 1974). For the special case of a binary outcome, Heagerty and Lele (1998) have proposed a thresholding model using a composite likelihood approach.

A drawback of the maximum likelihood-based methods employed in geostatistical modelling is the large sample asymptotic inference. For a spatial stochastic process $\{\boldsymbol{Y}(u); u \in D\}$, with $D \subset R^2$ the asymptotic concept can be applied either to the sample size within a fixed space D (infill asymptotics) or to the space D (increasing domain asymptotics). In the latter, observations are spaced far enough to be considered uncorrelated. The results can differ, depending on the type of asymptotics used (see Tubilla, 1975).

Bayesian hierarchical geostatistical models implemented via Monte Carlo methods avoid asymptotic inference as well as many computational problems in model fitting and prediction. Diggle et al. (1998) suggest inference on the posterior density via Markov chain Monte Carlo (MCMC). This iterative approach requires repeated inversions of the covariance matrix of the spatial process, which is involved in the likelihood. The size of this matrix increases with the number of locations. Inversions of large matrices can drastically slow down the running time of the algorithm and cause numerical instabilities affecting the accuracy of the estimates. To overcome this problem Gelfand et al. (1999) suggest non-iterative

simulation via the Sampling-Importance-Resampling (SIR) algorithm (Rubin, 1987). The quality of SIR hinge on the ability to formulate an easy-to-draw-from importance-density, which comes as close as possible to the true joint posterior distribution of the parameters.

In this article, we review three fitting procedures; the maximum likelihood-based PQL method, MCMC and the SIR. We assess these methods in terms of estimation accuracy, ease of implementation and computational efficiency using a spatially structured dataset on infant mortality from Mali collected over 181 locations. A description of the dataset and the applied questions which motivated this work are given in section 2.2. Section 2.3 describes the model as well as the three fitting approaches. Section 2.4 provides implementation details and presents the results. A discussion on the ease of implementation of each approach and a comparison of the inferences obtained is given in section 2.5.

2.2 Data

The data which motivated this work were collected within the Demographic and Health Surveys (DHS) program. The aim of the program is to collect and analyze reliable demographic and health data for regional and national family and health planning. Data are commonly collected in developing countries. DHS is funded by the U.S. Agency for International Development (USAID) and implemented by Macro International Inc. The standard DHS methodology involves collecting complete birth histories from women of childbearing age, from which a record of age and survival can be computed for each child. The data are available to researchers via the internet (www.measureDHS.com).

Birth histories corresponding to 35,906 children were extracted from the data of the DHS-III 1995/96 household survey carried out in Mali. Additional relevant covariates extracted were the year of birth, residence, mothers education, infant's sex, birth order, preceding birth interval and mothers age at birth. Using location information provided by Macro International, we were able to geo-locate 181 distinct sites by using digital maps and databases, such as the African data sampler (World Resources Institute, 1995) and the Geoname Gazetteer (GDE Systems Inc., 1995). The objective of data analysis was to assess the effect of birth and socio-economic parameters on infant mortality and produce smooth maps of mortality risk in Mali. These maps will help identifying areas of high mortality risk and assist child mortality intervention programs.

2.3 Generalized linear mixed model for point-referenced spatial data

Let Y_{ij} be a binary response corresponding to the mortality risk of child j at site $s_i, i = 1, \ldots, n$ taking value 1 if the child survived the first year of life and 0 otherwise, and let $\boldsymbol{X}_{ij}$ be the vector of associated covariates. Within the generalized linear model framework (GLM), we assume Y_{ij} are i.i.d. Bernoulli random variables with $\mathrm{E}(Y_{ij}) = \pi_{ij}$ and model predictors as $g(\pi_{ij}) = \boldsymbol{X}_{ij}^t \boldsymbol{\beta}$ where $g(\cdot)$ is a link function such as logit in our mortality risk

application. However the spatial structure of the data renders the independence assumption of Y_{ij} invalid, leading to narrower confidence intervals for $\boldsymbol{\beta}$ and thus to overestimation of the significance of the predictors.

One approach to take into account spatial dependence is via the generalized linear mixed model (GLMM) reviewed by Breslow and Clayton (1993). In particular, we introduce the unobserved spatial variation by a latent stationary, isotropic Gaussian process $\boldsymbol{U}$ over our study region, $\mathcal{D}$, such that $\boldsymbol{U} = (U_1, U_2, \ldots, U_n) \sim N(0, \boldsymbol{\Sigma})$, where Σ_{ij} is a parametric function of the distance d_{ij} between locations s_i and s_j. Conditional on the random term U_i, we assume that Y_{ij} are independent with $\mathrm{E}(Y_{ij} \mid U_i) = \pi_{ij}$. The U_i enters the model on the same scale as the predictors, that is

$$g(\pi_{ij}) = \boldsymbol{X}_{ij}^t \boldsymbol{\beta} + U_i \tag{2.1}$$

and captures unmeasured geographical heterogeneity (small scale variation).

A commonly used parameterization for the covariance $\boldsymbol{\Sigma}$ is $\Sigma_{ij} = \sigma^2 \rho(\phi; d_{ij})$ where σ^2 is the variance of the spatial process and $\rho(\phi; d_{ij})$ a valid correlation function with a scale parameter ϕ which controls the rate of correlation decay with increasing distance. In most applications a monotonic correlation function is chosen i.e. the exponential function which has the form $\rho(\phi; d_{ij}) = \exp(-\phi d_{ij})$. Ecker and Gelfand (1997) propose several other parametric correlation forms, such as the Gaussian, Cauchy, spherical and the Bessel.

A separate set of location-specific random effects, $\boldsymbol{W} = (W_1, \ldots, W_n)^t$ is often added in equation (2.1) to account for unexplained non-spatial variation (Diggle et al., 1998), where $W_i, i = 1, \ldots, n$ are considered to be independent, arising from a normal distribution, $W_i \sim N(0, \tau^2)$. The τ^2 is known in geostatistics as the nugget effect and introduces a discontinuity at the origin of the covariance function, $\Sigma_{ij} = \tau^2 \delta_{ij} + \sigma^2 \rho(\phi; d_{ij})$. δ_{ij} is the Kronecker delta and takes the value of one if $i = j$ and zero otherwise. A large number of repeated samples at the same location make the nugget identifiable, otherwise its use in the model is not justifiable since the extra-binomial variation is already accounted for by the spatial random effect.

2.3.1 Parameter estimation

The above GLMM is highly parameterized and maximum likelihood methods fail to estimate all parameters simultaneously. The estimation approach starts by integrating out the random effects and estimating the other parameters using the marginal likelihood $\int p(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\beta}, \sigma, \phi)\ p(\boldsymbol{U}|\sigma, \phi) d\boldsymbol{U}$. However, this integral has analytical solution only for Gaussian data. For non-Gaussian data the integrand can be approximated using a first-order Taylor series expansion around its maximizing value, after which the integration is feasible. This approach, known as the Laplace approximation, results in the penalized quasi-likelihood (PQL) estimator (Breslow and Clayton, 1993), which was shown in various simulation studies to produce biased results (Browne and Draper, 2000; Neuhaus and Segal, 1997). Breslow and Lin (1995) determined the asymptotic bias in variance component problems for first- and second-order approximations in comparison to McLaurin approximations.

Following the Bayesian modelling specification, we need to adopt prior distributions for all model parameters. We chose non-informative Uniform priors for the regression coefficients, i.e. $p(\boldsymbol{\beta}) \propto \mathbf{1}$, and vague inverse Gamma priors for the σ^2 and ϕ parameters: $p(\phi) \equiv \text{IG}(a_1, b_1)$ and $p(\sigma^2) \equiv \text{IG}(a_2, b_2)$. Bayesian inference is based on the joint posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi|\boldsymbol{Y}) \propto L(\boldsymbol{\beta}, \boldsymbol{U}; \boldsymbol{Y})p(\boldsymbol{\beta})p(\boldsymbol{U}|\sigma^2, \phi)p(\sigma^2)p(\phi)$, where $p(\boldsymbol{U}|\sigma^2, \phi)$ is the distribution of the spatial random effects, that is $p(\boldsymbol{U}|\sigma^2, \phi) \equiv N(0, \boldsymbol{\Sigma})$.

Markov chain Monte Carlo estimation

Diggle et al. (1998) suggest Markov chain Monte Carlo and in particular Gibbs sampling for fitting GLMM for point-referenced data. The standard implementation of the Gibbs algorithm requires sampling from the full conditional posterior distributions which in our application have the following forms:

$$p(\beta_k|\boldsymbol{\beta}_{-k}, \boldsymbol{U}, \boldsymbol{Y}) \propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{\exp(\boldsymbol{X}_{ijk}\beta_k \cdot Y_{ij})}{1 + \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta} + U_i)} \tag{2.2}$$

$$\begin{aligned} p(U_i|\boldsymbol{U}_{-i}, \sigma^2, \phi, \boldsymbol{Y}) \propto & \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{\exp(U_i Y_{ij})}{1 + \exp(\boldsymbol{X}_{ij}^t\boldsymbol{\beta} + U_i)} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \times \\ & \exp(-\frac{1}{2}(U_i - \boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{U}_{-i})^2(\boldsymbol{\Sigma}_i)^{-1}) \end{aligned} \tag{2.3}$$

$$p(\phi|\boldsymbol{U}, \sigma^2) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{U}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{U} + b_1/\phi)\right) \phi^{-(a_1+1)} \tag{2.4}$$

$$p(\sigma^2|\boldsymbol{U}, \phi) \sim \text{Inverse Gamma}(a_2 + \frac{n}{2}, b_2 + \frac{1}{2}\boldsymbol{U}^t\boldsymbol{R}\boldsymbol{U}) \tag{2.5}$$

where $\boldsymbol{\beta}_{-k} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_K)^t$, $\boldsymbol{U}_{-i} = (U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_n)^t$, $\boldsymbol{\Sigma}_{-i,i} = \boldsymbol{\Sigma}_{i,-i}^t = \text{Cov}(\boldsymbol{U}_{-i}, U_i)$, $\boldsymbol{\Sigma}_{-i} = \text{Cov}(\boldsymbol{U}_{-i}, \boldsymbol{U}_{-i})$, $R_{kl} = \rho(\phi; d_{kl})$ and $\boldsymbol{\Sigma}_i = \sigma^2 - \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{\Sigma}_{i,-i}$. Samples from $p(\sigma^2|\boldsymbol{U}, \phi)$ can be drawn easily as this is a known distribution. The conditionals of the other parameters do not have standard forms and a random walk Metropolis algorithm with a Gaussian proposal density having mean equal to the estimate from the previous iteration and variance derived from the inverse second derivative of the log-posterior could be employed for simulation.

The likelihood calculations in (2.3) and (2.4) require inversions of the $(n-1) \times (n-1)$ matrices, $\boldsymbol{\Sigma}_{-i}$, $i = 1, \ldots, n$ and the $n \times n$ matrix $\boldsymbol{\Sigma}$, respectively. Matrix inversion is an order 3 operation, which has to be repeated for evaluating the conditional distribution of all n random effects U_i and that of the ϕ parameter, within each Gibbs sampling iteration. This leads to an enormous demand of computing capacity and makes implementation of the algorithm extremely slow (or possibly infeasible), especially for large number of locations.

Sampling-Importance-Resampling

Gelfand et al. (1999) propose Bayesian inference for point-referenced data using non-iterative Sampling-Importance-Resampling (SIR) simulation. They replace matrix in-

version with simulation by introducing a suitable importance sampling density $g(\cdot)$ and re-write the joint posterior as

$$p^{\star}(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi|\boldsymbol{Y}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi|\boldsymbol{Y})}{g(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi; \boldsymbol{Y})} \cdot g(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi; \boldsymbol{Y}) \tag{2.6}$$

They construct the importance sampling density (ISD) by

$$g(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi; \boldsymbol{Y}) = g_s(\boldsymbol{\beta}|\boldsymbol{U}; \boldsymbol{Y}) g_s(\boldsymbol{U}|\sigma^2, \phi) g_s(\sigma^2, \phi) \tag{2.7}$$

which is easy to simulate from and then re-sample from $g(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi|\boldsymbol{y})$ according to the importance weights

$$w(\boldsymbol{\beta}, \sigma^2, \phi, \boldsymbol{U}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi|\boldsymbol{Y})}{g(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi; \boldsymbol{Y})}. \tag{2.8}$$

The density $g_s(\sigma^2, \phi)$ of the ISD could be taken as a product of independent inverse Gamma distributions $g_s(\sigma^2)g_s(\phi)$. It is however preferable to adopt a bivariate distribution which accounts for interrelations between the two parameters and thus it approximates closer the $p(\sigma^2, \phi \mid \boldsymbol{Y})$. We considered a bivariate t-distribution on $\log(\sigma^2)$ and $\log(\phi)$ with low degrees of freedom and mean around the maximum likelihood estimates of $\log(\sigma^2)$ and $\log(\phi)$. The spatial random effects can be simulated from a multivariate normal distribution, $g_s(\boldsymbol{U}|\sigma^2, \phi) \equiv N(\boldsymbol{0}, \sigma^2\rho(\phi, \cdot))$. This step requires matrix decomposition of $\sigma^2\rho(\phi, \cdot)$, repeatedly at every iteration. This is an operation of order 2 and the most expensive numerical part of the simulation from the ISD. The density $g_s(\boldsymbol{\beta}|\boldsymbol{U}; \boldsymbol{Y})$ can be a normal distribution, $g_s(\boldsymbol{\beta}|\boldsymbol{U}; \boldsymbol{Y}) \equiv N(\widehat{\boldsymbol{\beta}}_U, \widehat{\boldsymbol{\Sigma}}_\beta)$, with $\widehat{\boldsymbol{\beta}}_U$ equal to the regression coefficients estimated from an ordinary logistic regression with offset $\boldsymbol{U}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ equal to the covariance matrix of $\widehat{\boldsymbol{\beta}}_U$.

When the ISD approximates well the posterior distribution, one expects that the standardized importance weights are Uniformly distributed. When this not the case, the ISD would give rise to very few dominant weights leading to an inefficient and wrong sampler. A possible remedy would be to embed the Sampling-Importance-Resampling simulation in an iterative scheme which refines the initial guesses of the ISD and allows after few iterations more uniform weights.

Point estimates of the parameters should preferably be calculated from the importance weights using all sampled values, rather than from the re-sampled values, what leads to smaller bias. For example the mean and variance of β_i is estimated by $\overline{\beta_i} = \sum_k w_k \beta_i^{(k)} / \sum_k w_k$ and $\sum_k w_k (\beta_i^{(k)} - \overline{\beta_i})^2 / \sum_k w_k$ respectively, where $\beta_i^{(k)}$ is the kth sampled value of β_i from the ISD.

2.3.2 Spatial prediction

Modelling point-referenced data is not only useful for identifying significant covariates but for producing smooth maps of the outcome by predicting it at unsampled locations. Spatial prediction is usually refereed as kriging.

Let $\boldsymbol{Y}_0$ be a vector of the binary response at new, unobserved locations s_{0i}, $i = 1, \ldots, n_0$. Following the maximum likelihood approach, the distribution of $\boldsymbol{Y}_0$ is given by:

$$P(\boldsymbol{Y}_0 \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{U}}, \hat{\sigma}^2, \hat{\phi}) = \int P(\boldsymbol{Y}_0 \mid \hat{\boldsymbol{\beta}}, \boldsymbol{U}_0) P(\boldsymbol{U}_0 \mid \hat{\boldsymbol{U}}, \hat{\sigma}^2, \hat{\phi})\ d\boldsymbol{U}_0 \tag{2.9}$$

where $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ and $\hat{\phi}$ are the maximum likelihood estimates of the corresponding parameters. In PQL, $\hat{\boldsymbol{U}}$ is derived as part of the iterative estimation process (Breslow and Clayton, 1993). $P(\boldsymbol{Y}_0 \mid \hat{\boldsymbol{\beta}}, U_0)$ is the Bernoulli-likelihood at new locations and $P(\boldsymbol{U}_0 \mid \hat{\boldsymbol{U}}, \hat{\sigma}^2, \hat{\phi})$ is the distribution of the spatial random effects $\boldsymbol{U}_0$ at new sites, given $\hat{\boldsymbol{U}}$ at observed sites and is normal

$$P(\boldsymbol{U}_0 \mid \hat{\boldsymbol{U}}, \hat{\sigma}^2, \hat{\phi}) = \mathcal{N}(\boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\hat{\boldsymbol{U}}, \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10}) \tag{2.10}$$

with $\boldsymbol{\Sigma}_{11} = \mathrm{E}(\boldsymbol{U}\boldsymbol{U}^t)$, $\boldsymbol{\Sigma}_{00} = \mathrm{E}(\boldsymbol{U}_0\boldsymbol{U}_0^t)$ and $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^t = \mathrm{E}(\boldsymbol{U}_0\boldsymbol{U}^t)$. The mean of the Gaussian distribution in (2.10) is the classical kriging estimator (Matheron, 1963).

The Bayesian predictive distribution of $\boldsymbol{Y}_0$ is given by:

$$\begin{aligned} P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}) \;\; = \;\; & \int P(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{U}_0) P(\boldsymbol{U}_0 \mid \boldsymbol{U}, \sigma^2, \phi) \times \\ & P(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi \mid \boldsymbol{Y})\ d\boldsymbol{\beta}\ d\boldsymbol{U}_0\ d\boldsymbol{U} d\sigma^2 d\phi \end{aligned} \tag{2.11}$$

$P(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi \mid \boldsymbol{Y})$ is the posterior distribution of the parameters and obtained by the Gibbs sampler or the SIR approach. Simulation-based Bayesian spatial prediction is performed by consecutive drawing samples from the posterior distribution, the distribution of the spatial random effects at new locations and the Bernoulli-distributed predicted outcome. In SIR, drawing is performed from the set of all sampled parameters with weighting given in equation (2.8).

The maximum likelihood predictor (equation 2.9) can be interpreted as the Bayesian predictor (equation 2.11), with parameters fixed at their maximum-likelihood estimates. In contrast to Bayesian kriging, classical kriging does not account for uncertainty in estimation of $\boldsymbol{\beta}$ and the covariance parameters.

2.4 Results

A generalized linear mixed model was fitted to the infant mortality data in Mali using the three estimation approaches discussed in section 2.3, PQL, MCMC and SIR together with the ordinary logistic regression (GLM) which did not account for spatial dependence. The purpose of the analysis was to assess the effect of maternal and socio-economic factors on infant mortality, produce a smooth map of mortality risk in Mali and compare the results obtained from the above procedures. Univariate analysis based on the ordinary logistic regression revealed that the following variables should be included in the model: child's birthday, region type, mother's degree of education, sex, birth order, preceding birth interval and mother's age at birth.

Model	Initial sample size	Final sample size from posterior	No. of batches and size	Iterations to convergence	Thinning*	Time per 1,000 iterations
MCMC	50,000	1,720	-	7,000	25	7 hrs 14 min
SIR	400,000	1,600	800 batches with 500 values (2 draws per batch)	0	0	1 hr 23 min

* Minimum lag at which autocorrelation was not significant.

Table 2.1: Comparison of the computational costs for the Bayesian, simulation based approaches.

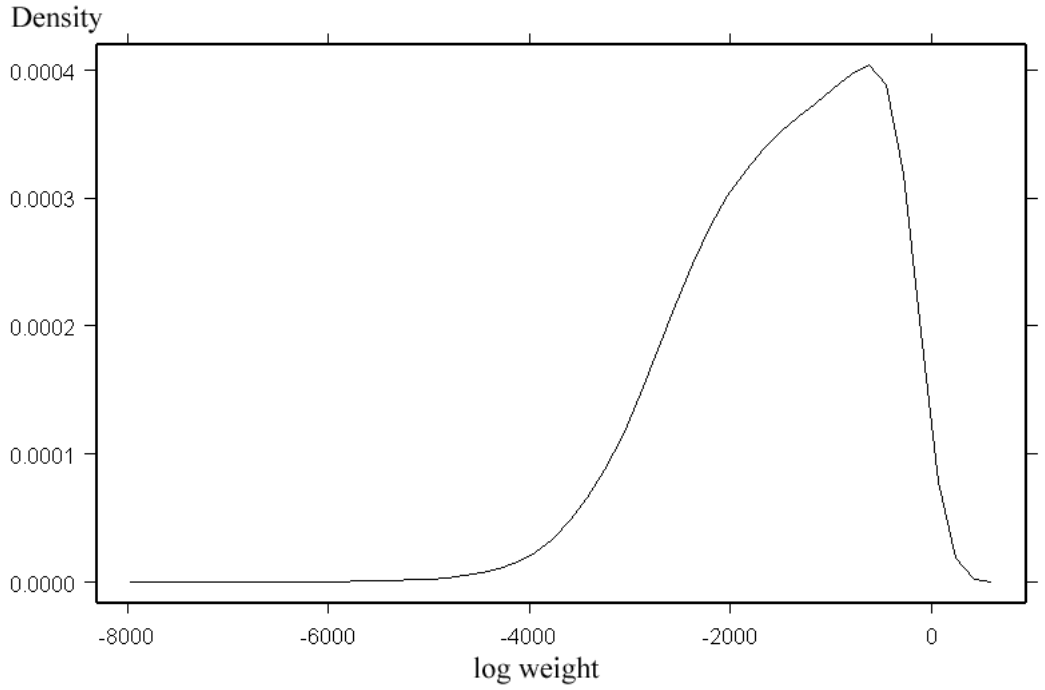


Figure 2.1: Distribution of the weights in the Sampling-Importance-Resampling (SIR) procedure.

We fitted the non-spatial logistic model (GLM) in SAS (SAS Institute Inc., Cary, NC, USA) using Proc Logistic. The spatial model with the PQL estimation method was also fitted in SAS using the `%GLIMMIX`-macro (see appendix). This macro is based on the approach of Wolfinger and O'Connell (1993) and does subsequent calls of Proc Mixed to iteratively estimate mixed models for non-normal data. It is supported by a collection of spatial correlation functions, such as the exponential, Gaussian, linear, power and spherical. In our application, we have chosen the exponential function. MCMC and SIR estimation were implemented in software written by the authors in Fortran 95 (Compaq Visual Fortran v6.6) and run on an Unix AlphaServer 8400. For small number of locations the freeware software WinBUGS (www.mrc-bsu.cam.ac.uk/bugs) can also be used to obtain MCMC based estimates. Proc Mixed for normal data supports Bayesian modelling by allowing spe-

					Birth year					Residency	Education	
Model	Estimate	σ^2	ϕ	Intercept	1966–71	1972–77	1978–83	1984–89	1990–96	Urban	No	Primary
GLM	MLE	-	-	1.81	-0.18	0.04	0.09	0.12	0.17	0.32	-0.56	-0.66
	95% CI	-	-	(1.43,2.11)	(-0.44,0.09)	(-0.22,0.29)	(-0.16,0.34)	(-0.13,0.37)	(-0.08,0.42)	(0.22,0.36)	(-0.75,-0.31)	(-0.83,-0.43)
PQL	MLE	1.05	2.07	2.59	-0.19	0.03	0.09	0.12	0.16	0.29	-0.54	-0.58
	95% CI	(0.72,1.81)	(0.54,4.63)	(1.43,3.74)	(-0.48,0.11)	(-0.26,0.31)	(-0.19,0.37)	(-0.17,0.40)	(-0.12,0.44)	(0.19,0.39)	(-0.75,-0.32)	(-0.78,-0.38)
MCMC	Mean	1.32	0.07	1.76	-0.20	0.01	0.07	0.10	0.15	0.30	-0.55	-0.60
	Median	0.91	0.04	1.75	-0.21	0.01	0.07	0.09	0.14	0.30	-0.54	-0.59
	95% CI	(0.22,3.89)	(0.008,0.24)	(1.47,2.09)	(-0.46,0.08)	(-0.25,0.27)	(-0.19,0.33)	(-0.16,0.36)	(-0.11,0.40)	(0.23,0.38)	(-0.74,-0.36)	(-0.78,-0.42)
SIR	Mean	0.91	0.005	1.77	-0.19	0.03	0.08	0.11	0.16	0.33	-0.5	-0.57
	Median	0.61	0.03	1.73	-0.18	0.03	0.08	0.11	0.16	0.34	-0.5	-0.57
	95% CI	(0.22,2.62)	(0.0004,0.015)	(0.34,3.25)	(-0.44,0.06)	(-0.21,0.27)	(-0.16,0.31)	(-0.13,0.34)	(-0.08,0.39)	(0.25,0.41)	(-0.68,-0.32)	(-0.75,-0.4)

		Sex	Birth order			Preceding birth interval in years		Mothers age at birth		
Model	Estimate	Male	2nd or 3rd	4th to 6th	7th or higher	2–4	> 4	20–29	30–39	40–49
GLM	MLE	-0.14	-1.90	-1.97	-2.10	2.34	2.71	0.24	0.31	0.19
	95% CI	(-0.16,-0.05)	(-2.40,-1.32)	(-2.48,-1.38)	(-2.62,-1.51)	(1.76,2.84)	(2.11,3.22)	(0.13,0.29)	(0.15,0.40)	(-0.02,0.42)
PQL	MLE	-0.14	-1.88	-1.95	-2.07	2.31	2.67	0.25	0.32	0.19
	95% CI	(-0.20,-0.09)	(-2.42,-1.34)	(-2.50,-1.40)	(-2.63,-1.52)	(1.77,2.85)	(2.12,3.22)	(0.17,0.33)	(0.19,0.44)	(-0.07,0.44)
MCMC	Mean	-0.14	-1.90	-2.00	-2.10	2.37	2.73	0.26	0.33	0.20
	Median	-0.14	-1.95	-2.02	-2.15	2.38	2.74	0.26	0.33	0.20
	95% CI	(-0.19,-0.1)	(-2.39,-1.44)	(-2.48,-1.51)	(-2.61,-1.63)	(1.87,2.82)	(2.2,3.22)	(0.19,0.32)	(0.23,0.43)	(-0.009,0.43)
SIR	Mean	-0.14	-1.88	-1.96	-2.09	2.31	2.65	0.25	0.32	0.21
	Median	-0.14	-1.88	-1.96	-2.09	2.30	2.65	0.25	0.32	0.21
	95% CI	(-0.19,-0.09)	(-2.34,-1.42)	(-2.43,-1.49)	(-2.56,-1.62)	(1.82,2.77)	(2.16,3.13)	(0.18,0.31)	(0.22,0.43)	(0.01,0.42)

Table 2.2: Comparison of parameter estimates from the binary spatial model using different estimation strategies. The binary outcome is the survival of the first year of life.

cification of prior distributions for the parameters and MCMC. However, this possibility is currently available only for variance component models and not for spatial covariances, which holds for the `%GLIMMIX` macro, too.

Convergence of the PQL approach to the global mode of the likelihood was highly dependent on the starting values. We suggest running the procedure with several starting values by using the `parms`-command. Computationally, the PQL is fast in comparison to the simulation-based procedures, MCMC and SIR, but it runs quickly out of workspace for a larger dataset. A comparison of the computational time required for the MCMC and SIR algorithms is given in table 2.1. MCMC estimation was applied using a single chain. Convergence was assessed using the Geweke (1992) criterion. The algorithm converged after 7,000 iterations. A final sample from the posterior distribution of size 1,720 was obtained by sampling every 25 iterations after convergence was reached. The SIR algorithm required extensive fine tuning in order to derive good estimates. We ran the sampler several times and adjusted the degrees of freedom and mean parameter in the bivariate t-distribution $g_s(\sigma^2, \phi)$, according to those values leading to large weights. Instead of resampling from the whole sequence of parameters according to their weights, we obtained better results by dividing the generated parameters into batches and drawing an equal number of samples with replacement from every batch. The implementation of the SIR algorithm was found to be difficult. Despite the effort applied to improve the SIR estimator, the derived weights show a highly skewed distribution, with a few dominating values (figure 2.1).

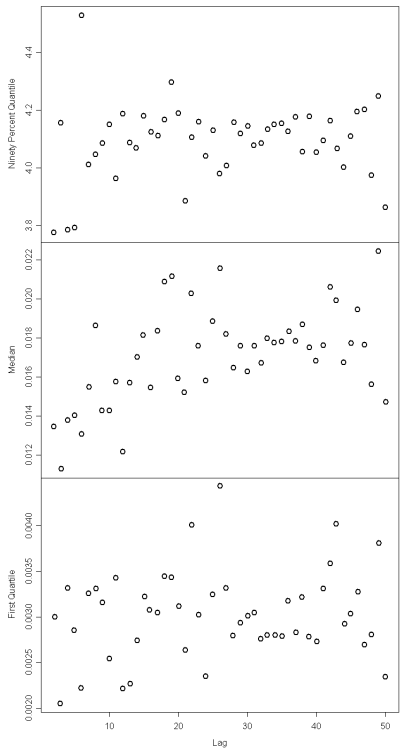


Figure 2.2: Variogram cloud of the residuals in a non-spatial model.

Table 2.2 gives the parameter estimates obtained by the four approaches. The fixed effect coefficients $\boldsymbol{\beta}$ show no fundamental difference in their point estimates between the competing models, with the exception of the intercept coefficient. The PQL estimate of the intercept is higher than that from the other estimators. The standard error of $\boldsymbol{\beta}$ estimated from GLM is narrower than in the spatial models, as we were expecting. Discrepancies between the fitting approaches are observed in the estimates of the covariance parameters σ^2 and ϕ. The posterior density of σ^2 obtained from MCMC was found to be highly skewed to the left. PQL overestimates ϕ suggesting a lower spatial variation than the Bayesian approaches. This confirms known results about bias in the PQL estimates especially for the covariance parameters σ^2 and ϕ due to the bad quality of the first-order approximation of the integrand. The SIR estimates are similar to those obtained from MCMC.

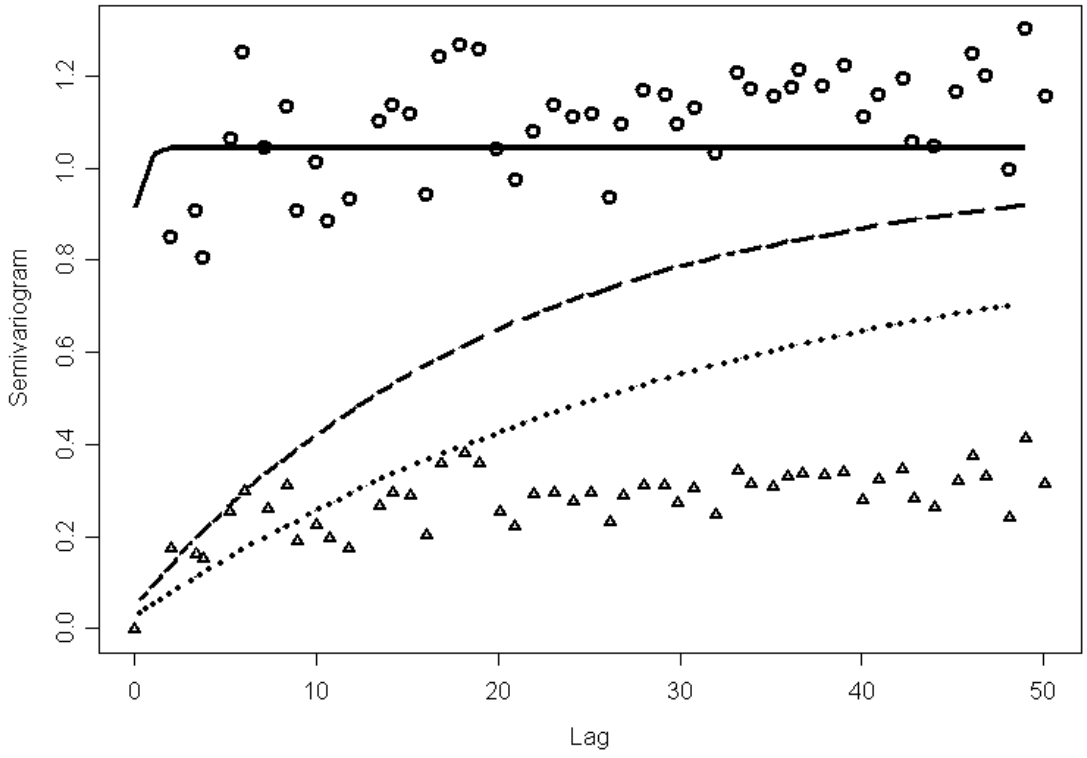


Figure 2.3: Semivariogram estimators: Classical semivariogram estimator by Matheron (circles), Robust version by Cressie and Hawkins (triangles), MCMC (long dashed line), SIR (short dashed line) and PQL (line) fit.

Figure 2.2 shows three plots of the semivariogram cloud based on the Anscombe residuals obtained after fitting the GLM model. The semivariogram cloud is a plot of half the squared difference of the residuals versus the distance between their sample locations.

The mean of the squared differences at each lag gives an estimator of the semivariogram. The three plots correspond to the 5 percent, 50 percent and 95 percent quartile of the squared difference of the residuals. The semivariogram cloud shows high variability and an increasing trend from the origin indicating lag-dependent variation. For a stationary spatial process, the semivariogram relates to the covariance of the random effects. Therefore we expect high variability in the covariance parameters.

Figure 2.3 depicts different semivariogram estimators. The classical estimator by Matheron (1963) was calculated by $\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|}\sum_{N(\boldsymbol{h})}(Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j))^2$, where $Z(\boldsymbol{s}_i)$ is the Anscombe residual at location $\boldsymbol{s}_i$, $N(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \| \boldsymbol{s}_i - \boldsymbol{s}_j \| = \boldsymbol{h} \pm \epsilon\}$ and $|N(\boldsymbol{h})|$ is its cardinality. This estimator is sensitive to outliers and a robust version was proposed by Cressie and Hawkins (1980), which is displayed in figure 2.3, too. The MCMC, SIR and PQL based estimators were calculated by replacing the estimates of σ^2 and ϕ obtained from the three approaches in $\gamma(\boldsymbol{h}) = \sigma^2(1 - \exp(-\phi\, \boldsymbol{h}))$. The MCMC and SIR estimators appear to be between the two other empirical semivariogram estimators. Since we have omitted the nugget term, they pass through the origin. Nevertheless, their values fit nicely into the graph. The PQL estimate does not capture the correlation present at large lags. It represents the classical semivariogram estimator well, but it is far off the robust version.

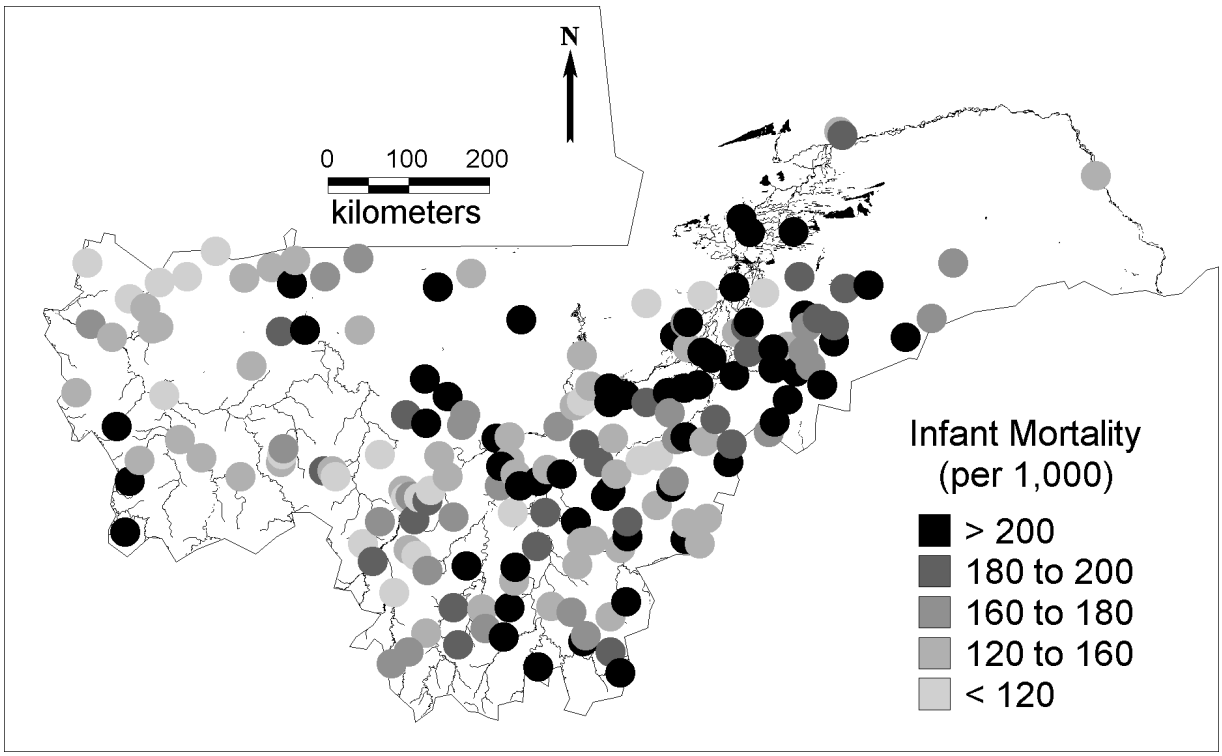


Figure 2.4: Observed mortality in 36,906 infants from the DHS surveys conducted in the years 1995 and 1996 at 181 distinct locations in Mali.

Regarding our application, figure 2.4 displays the locations of the DHS surveys and the observed infant mortality risk in Mali. The risk factors which were found to be statistically

significant related to infant mortality (table 2.2) confirm findings made by other authors. The negative association between maternal education and mortality has been described by Farah et al. (1982) and Cleland and Ginneken (1989). Higher education may result in higher health awareness, better utilization of health facilities (Jain, 1988), higher income and ability to purchase goods and services which improves infants health (Schultz, 1979).

The observed time trend, with higher infant survival for more recent years, was found not statistically significant. Longer birth intervals and low birth order reduce the risk of infant death. Mortality was related to the residency and sex of the infant with girls and urbanites being at lower risk of dying during the first year of life. The impact mothers age has on infant mortality shows the typical J-shape (Kalipeni, 1993) with lowest risk for age around thirty. The higher risk in young women may be explained by not fully developed maternal resources and that in older women by the effect of ageing. The MCMC-based estimate of the ϕ parameter revealed strong spatial correlation which reduces to less than 5 percent for distances longer than 75km.

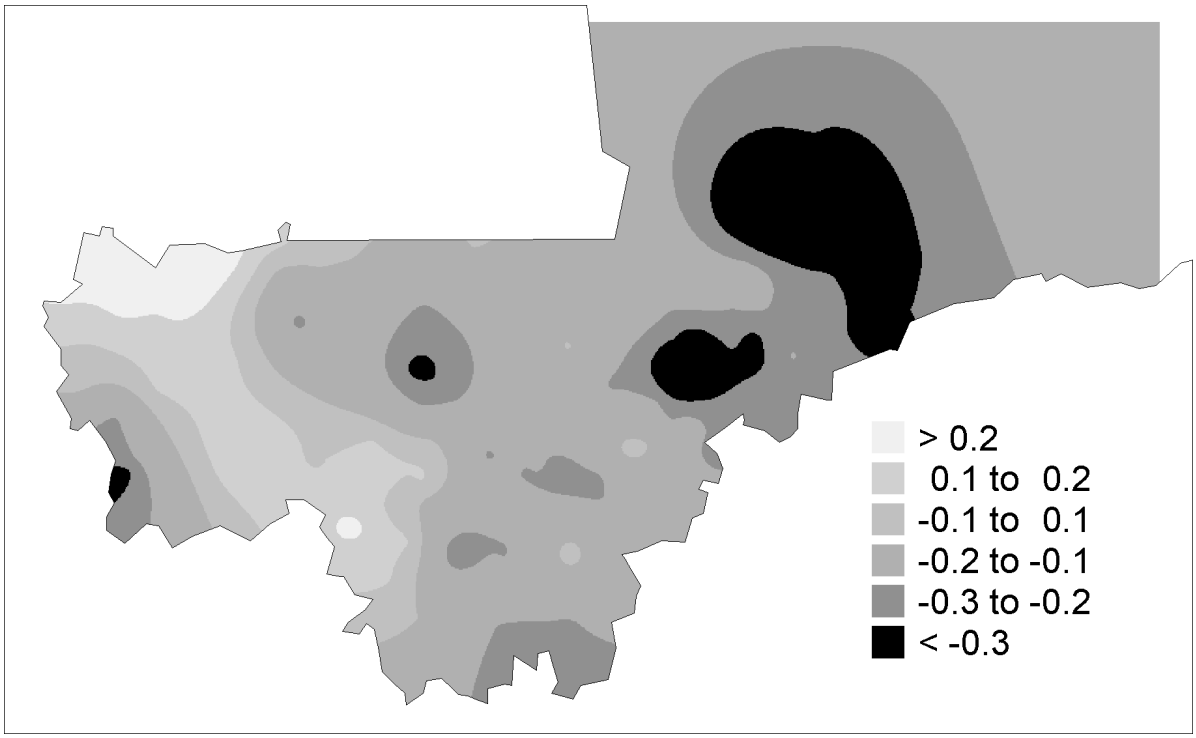


Figure 2.5: Predicted spatial random effects from the infant mortality model using MCMC. The darker the shading, the lower the survival.

Predictions of the child mortality risk using the MCMC approach were made at 600,000 new locations on a regular grid, covering the whole area of Mali south of 18 degrees latitude north. Since the covariates are infant-specific and can not be extrapolated for the new locations, we predict the random effects only. The map of predictions is displayed in figure 2.5. The map indicates a higher infant mortality risk mainly in the Northern part of the

Niger delta. This region has low population density and water availability is seasonal. The many lakes in this region are preferred breeding sites for the malaria mosquito. Low mortality is predicted in North-Western Mali at the border to Mauritania and Sénégal. In this region, the population is more active in migrating to other countries for business purposes, bringing money to the region. Health facility coverage is also reflected in the predictive map, where the coverage is low in the Northern Niger delta and high in the North-East.

2.5 Discussion

Generalized linear mixed models for large point-referenced spatial data are highly parameterized and their estimation is hampered by computational problems. Reliable estimation methods that can be applied in standard software or algorithms that can accurately estimate the model parameters within practical time constraints do not exist. In this paper we compared a few recent developments using a real dataset on infant mortality in Mali.

The advantage of the PQL method is that it can be applied in standard statistical software package. However estimates are biased especially those for the covariance parameters. The algorithm depends highly on the starting values and can easily converge to a local mode. For medium to large number of locations implementations of this algorithm is impeded by computer memory problems.

Bayesian methods can provide flexible ways of modelling point-referenced data, give unbiased estimates of the parameters and their standard error and have computational advantages for problems larger than the ones the maximum likelihood methods can handle. However, for very large number of locations, an implementation may be infeasible due to long computing time. The SIR runs considerably faster than MCMC, but it requires tedious tuning. Finding an ISD which approximates well the posterior distribution is difficult to develop and application-specific. Rigorous methods for evaluating the suitability of the ISD do not exist. This increases the possibility of drawing misleading inference.

MCMC is the most practical and, when it comes to prediction, accurate approach todate for fitting geostatistical problems. However, it is computationally intensive, especially for dataset with large number of locations. More research is required in ways of improving the convergence of the algorithm and the inversion of large matrices. Gilks and Roberts (1996), Mira and Sargent (2000) and Haran et al. (2003) have proposed general MCMC algorithms for improving convergence. Rue (2000) and Pace and Barry (1997) have applied innovative numerical methods using sparse matrix solvers for fitting areal data. In future, similar approaches need to be adapted and assessed for modelling point-referenced spatial data.

Acknowledgements

We are grateful to Macro International Inc. for providing the names of the villages of the DHS surveys. We would like to thank Tom Smith for discussing many aspects of the child mortality risk analysis. Many thanks also to Marcel Tanner for his encouragement and support. This work was supported by the Swiss National Science Foundation grant Nr. 3200–057165.99.

Appendix 2.A PQL estimation

The `%GLIMMIX` macro in SAS provides PQL estimates for non-normal geostatistical models. The macro fits a GLMM model of the form $g(\pi_i) = \boldsymbol{X}_i^t\boldsymbol{\beta} + Z_i\boldsymbol{U} + R_i$, where Z_i links the ith observation to its sample location. This is done by defining a variable (here called: `LOCATION`), that takes only distinct values for separate locations. There are two possibilities to include spatial correlation. Either $\boldsymbol{U}$ is set to zero and the spatial structure is incorporated in $\boldsymbol{R} = (R_1, \ldots, R_n)^t$ using the `repeated` statement. Alternatively, the vector $\boldsymbol{U}$ is taken to be multivariate normal with spatial covariance, which is introduced via the `random` statement. The R_i's are then considered to be independent and model the nugget effect. The two coding possibilities are shown below. The performance can substantially differ between the two approaches, depending on the size of the dataset and how the computing system allocates memory.

The listing below gives the `%GLIMMIX` code for our application. The SAS statements are set lowercase. The dataset used is called `InfantMortality` and includes the binary outcome `Y` and 18 predictors `X1` to `X18`. The coordinates are called `LAT` and `LONG` and `LOCATION` is the identifier of the distinct locations.

Code 1:

```
%glimmix(data=InfantMortality,
     procopt= ord covtest,
     stmts=%str(
     model Y = X1-X18 / cl solution notest /* outp=PREDICTED */;
     repeated / subject=intercept /* local */ type=sp(exp)(LAT LONG);
     /* parms (0 to 5 by 0.5) (1 to 50 by 5) (0.05 to 2.05 by 0.25) */;
               ),
     error=binomial,
     link=logit,
     options= pql);
```

Code 2:

```
%glimmix(data=InfantMortality,
     procopt= ord covtest,
     stmts=%str(
     model Y = X1-X18 / cl solution notest /* outp=PREDICTED */;
     random LOCATION / type=sp(exp)(LAT LONG);
     /* parms (0 to 5 by 0.5) (1 to 50 by 5) (0.05 to 2.05 by 0.25) */;
               ),
     error=binomial,
     link=logit,
     options= pql);
```

An explanation on the options can be found in the SAS Proc Mixed online help. It needs to note, that the range parameter ϕ estimated by `%GLIMMIX` corresponds to $1/\phi$ compared to our model. A nugget effect can be added in the repeated statement by including the `local` statement. If the dataset includes outcome with missing data, `%GLIMMIX` can predict its value. The predictions are written in a new dataset (here called `PREDICTED`), specified via the statement `outp=PREDICTED` in the model option section. It is advised to check convergence of the algorithm rigidly. We recommend to try several initial values for the PQL maximization. This is done via the `parms` statement, followed by a sequence of initial values.

CHAPTER 3

Spatial patterns of infant mortality in Mali; the effect of malaria endemicity

Gemperli A.[1], Vounatsou P.[1], Kleinschmidt I.[2],
Bagayoko M.[3], Lengeler C.[1] and Smith T.[1]

This paper has been published in *American Journal of Epidemiology* **159** 64–72, 2004.

[1] Swiss Tropical Institute, Basel
[2] Medical Research Council (South Africa), Durban
[3] Faculté de Médicine de Pharmacie et d'Otondo-Stomatologie, Université du Mali, Bamako

Abstract

A spatial analysis was carried out to identify factors related with geographical differences in infant mortality risk in Mali by linking data from two spatially structured databases; the Demographic and Health Survey (DHS) of 1995–96 and the Mapping Malaria Risk in Africa (MARA) databases in Mali. Socio-economic factors measured directly at individual level, and site-specific malaria prevalence predicted for the DHS locations by a spatial model fitted to the MARA database, were examined as possible risk factors. The analysis was carried out by fitting a Bayesian hierarchical geostatistical logistic model to infant mortality risk, by Markov chain Monte Carlo. It confirmed that mother's education, birth order and interval, sex of infant, residence and mother's age at birth had a strong impact on infant mortality risk in Mali. The residual spatial pattern of infant mortality showed a clear relationship with well known foci of malaria transmission, especially the inland delta of the Niger river. No effect of estimated parasite prevalence could be demonstrated. Possible explanations include confounding by unmeasured covariates, and sparsity of the source malaria data. Spatial statistical models of malaria prevalence are useful for indicating approximate levels of endemicity over wide areas and hence for guiding intervention strategies. However at points very remote from those sampled it is important to consider prediction error.

Keywords: bayesian hierarchical model; geostatistical data; infant mortality risk; kriging; malaria transmission.

3.1 Introduction

Malaria is an important cause of mortality in children in Africa, but the relationship between malaria transmission intensity and child mortality remains controversial (Bradley et al., 1991; Molineaux, 1985; Payne et al., 1976; Smith et al., 2001; Snow and Marsh, 1998). A review of published studies of malaria specific mortality show some evidence that the highest mortality may be at intermediate transmission intensities (Snow and Marsh, 1995). Rates of hospitalization with severe malaria in African children appeared to be highest at intermediate levels of transmission.

A difficulty with mortality studies is that the verbal autopsies used to assign a cause of death are not very reliable (Snow and Marsh, 1998). Many deaths in malaria endemic areas, assigned to other causes, are related to malaria infection (Molineaux, 1985). Moreover, low birth weight is an important risk factor for infant mortality and is known to arise because of both prematurity and intrauterine growth retardation resulting from malaria infection of the mother during pregnancy (Steketee et al., 2001). It follows that malaria may be a relevant risk factor for many deaths even when it is not the immediate cause. Hence it is as important to look at the relationship of malaria endemicity with all-cause mortality as it is to look at its relationship with malaria specific deaths.

Smith et al. (2001) linked published all-cause mortality rates and Entomological Inoculation Rates (EIR) across Africa and found an increase in infant mortality rate (IMR)

with EIR, but no clear trend with the child (12–59 month) mortality rate. Major shortcomings of the study were the small number of sites compared and the fact that they were a convenience sample. Geographical variation in factors independently affecting both malaria transmission and mortality (such as water availability) introduce ecological confounding. Analysis linking site specific mortality data with local malaria indices, that make appropriate adjustment for these ecological confounders, are clearly needed.

There are a few data sets which allow these type of analysis. The databases we have used for this study are the MARA (Mapping Malaria Risk in Africa) and the DHS (Demographic and Health Surveys) database. The MARA database consists of surveys recording malariological information, with over 10,000 collected data points all over Sub-Saharan Africa, to date. It is currently the most comprehensive database on malariological surveys in Africa (MARA/ARMA, 1998). The DHS (Demographic and Health Surveys) database, coordinated by Macro Systems Inc., provide nationally representative household surveys worldwide with large sample sizes of between 5,000 and 30,000 households, typically. It monitors indicators in the areas of demography, health, and nutrition.

We linked the two databases using their site-specific data and developed a geostatistical model which enabled us to investigate spatial patterns of infant mortality risk, assess its determinants and carry out an ecological analysis to examine the relationship between infant mortality risk and malaria. We demonstrated the methodology by applying the technique to data collected from Mali. The savanna and Sahel zones of the country represent an appropriate setting for implementing such a methodology, because of the tendency for living conditions to become generally more difficult in the more northerly, drier areas, while malaria transmission is generally expected to be more intense in the wetter, southern areas. Therefore we can not properly assess the effect of malaria transmission on infant mortality without adjustment for potential confounders. We believe our approach of linking the two databases to carry out an ecological analysis of infant mortality risk to be novel. Instead of a district specific approach (Rip et al., 1986; Kalipeni, 1993) we modelled the data at individual level, using a site dependent correlation structure, to estimate the various effects without relying on data that are aggregated by administrative boundaries, which would have lead to a loss of information. This approach provides estimates of the effects of various factors including malaria endemicity, taking into account confounding and spatial correlation. In addition, it allows us to produce maps of infant mortality risk adjusted for socio-economic factors.

3.2 Methods and materials

3.2.1 Data sources

MARA/ARMA is an international collaboration initiated to provide a database and an atlas of malaria in Africa, by collating both published and unpublished results of malariological surveys. Data on malaria endemicity were obtained from a model fitted to the MARA database by Kleinschmidt et al. (2000). This was a spatial logistic model of malaria

prevalence in children > 1 and ≤ 10 years of age using the results of prevalence surveys conducted between 1965 and 1998 at 101 different locations in Mali. This age group was chosen because it has the highest prevalence and shows the clearest distinction between regional malaria endemicity patterns. In addition, most of the surveys were conducted for this age range and thus allowed inclusion of most of the data. The model included temperature, rainfall (Hutchinson et al., 1996), normalized difference vegetation index (NDVI) obtained from satellite data collected by the NOAA/NASA (National Oceanic and Atmospheric Administration) Pathfinder AVHRR (Advance Very High Resolution Radiometer) Land Project and distance from the nearest water body. The MARA survey sites are shown in figure 3.1. We used predictions of this model to obtain estimates of malaria prevalence at the DHS sample sites (figure 3.2). To avoid the linearity assumption between malaria prevalence and infant mortality we converted these estimates to the following categories corresponding to different degrees of endemicity: 0–0.24, 0.25–0.49, 0.50–0.74, and 0.75–1.00.

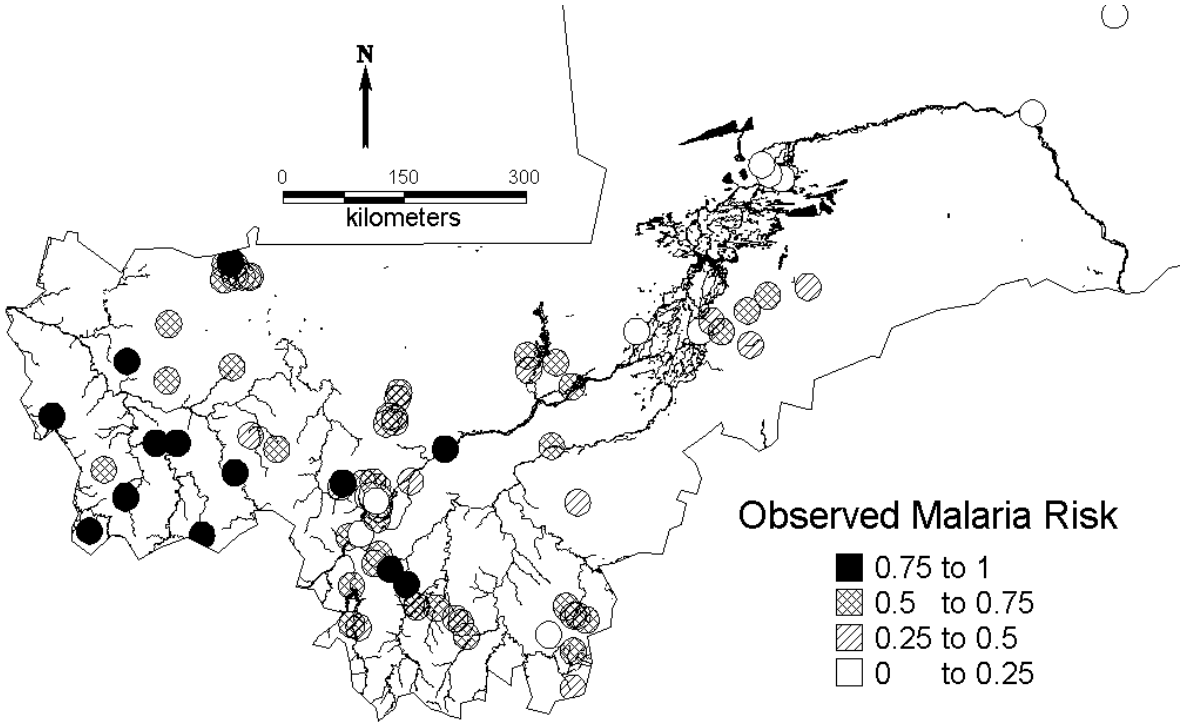


Figure 3.1: Observed malaria prevalence in 34,800 children 1 to 10 years old from the MARA surveys conducted in Mali between 1965 and 1998 at 101 locations. Rivers and lakes indicated in grey.

The DHS database is the most comprehensive database on child survival in Africa, compiled by a survey program in 23 countries funded by the United States Agency for International Development and coordinated by Macro International Inc. The aim of these surveys is to provide data for a range of monitoring and impact evaluation indicators within

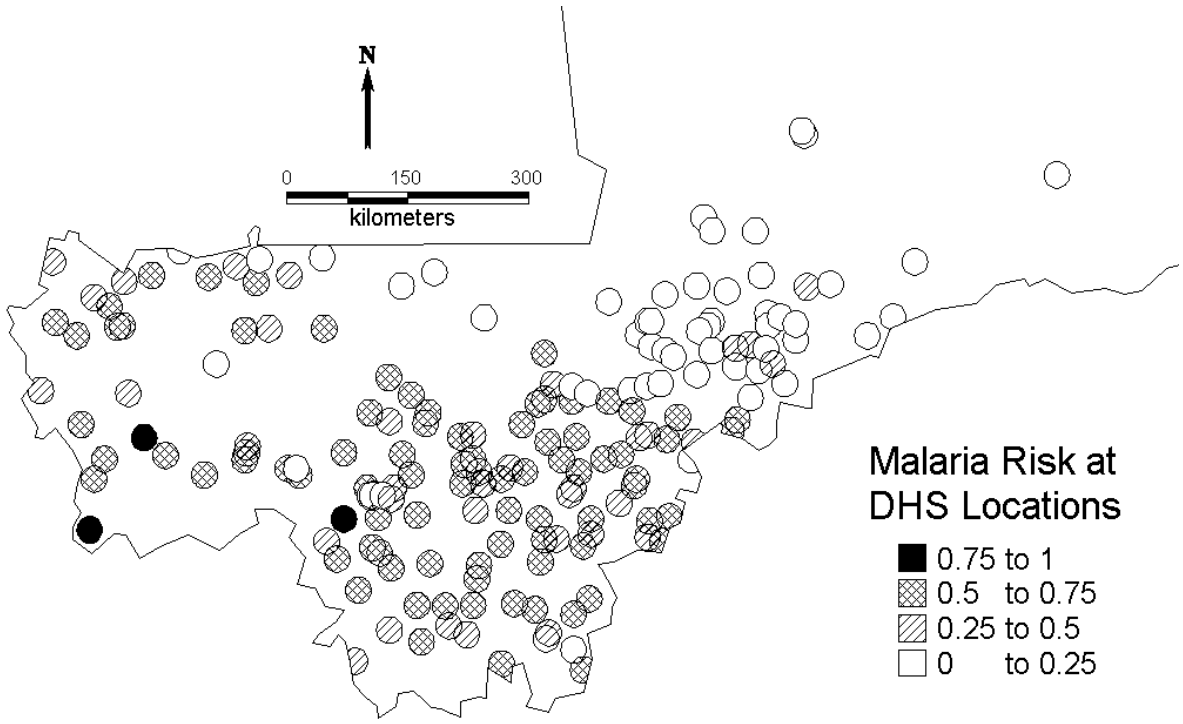


Figure 3.2: Estimated malaria prevalence at the 183 infant mortality sample locations in Mali obtained from the model of Kleinschmidt et al. (2000).

the population, health, and nutrition sector. Birth histories since 1960 corresponding to 35,906 infants were extracted from the DHS survey carried out from November 1995 to April 1996 in Mali. Additionally, year of birth, residence, mother's education, infant's sex, birth order, preceding birth interval and mother's age at birth were extracted. Using location information provided by Macro International we were able to geolocate the 183 distinct survey sites shown in figure 3.2 by using digital maps and databases, such as the African data sampler (World Resources Institute, 1995) and the Geoname Gazetteer (GDE Systems Inc., 1995).

3.2.2 Statistical analysis

Logistic regression models were fitted to infant mortality, using SAS v8.2 (SAS Institute Inc., Cary, NC, USA) to identify significant socio-economic, demographic and birth related covariates. The variables chosen were those analyzed by Coulibaly et al. (1996) in their report which summarizes the results of the 1995–1996 DHS survey in Mali in relation to mortality. Mother's age at birth and the date of birth of the child were included in all models. In addition, the following variables which showed a significant bivariate association with infant mortality were selected for the subsequent spatial multivariate analysis: region-type, mother's education, sex, birth order and preceding birth interval.

Bayesian hierarchical models were fitted to estimate the amount of spatial heterogeneity in infant mortality as well as associations between risk factors and infant mortality in the presence of spatial correlation. Ignoring this correlation would result in underestimation of the variance of the effects of risk factors (Cressie, 1993). We used logistic models with village-specific random effects to assess geographical heterogeneity and the effects of different covariates. Some covariates used were at individual level (socio-economic, demographic, birth-related covariates) and some were at village-level (malaria prevalence category). Malaria prevalence was not observed at the locations of the mortality data, but was estimated using the model by Kleinschmidt et al. (2000). These estimates were categorized because explanatory analysis revealed a non-linear relation between infant mortality and malaria prevalence. The cut-offs were chosen according to the frequency distribution of mortality data and on epidemiological considerations. Spatial correlation was modelled by assuming that the random effects are distributed according to a multivariate normal with a variance-covariance matrix related to the variogram of the spatial process (Diggle et al., 1998). We used Markov chain Monte Carlo (Gelfand and Smith, 1990; Smith and Roberts, 1993) to estimate the model parameters. Simulation-based Bayesian kriging was also applied to produce smoothed maps of mortality risk and of the variance of the map estimates (Gelfand et al., 1999). The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) was applied to assess the effect of malaria prevalence on infant mortality. The smaller the DIC values the better the fit of the model. Further details of this modelling approach are given in the appendix. The analysis was implemented using software written by the authors in Fortran 95 (Compaq Visual Fortran v6.1) using IMSL numerical libraries (Visual Numerics, Inc., Houston, Texas, USA).

3.3 Results

The median malaria prevalence estimated for the 183 locations was 49 percent ranging from 4 percent to 82 percent. 16.9 percent of the 35,906 infants sampled died before completing their first year of life. This is untypically high compared to the current census based estimates of 12.3 percent (World Health Organization, 2000). However in the DHS data, observations are retrospective over forty years, and there is an indication that infant mortality was higher in earlier years, although not statistically significant. Of those who died, 53.4 percent were male and 77.6 percent were living in rural parts of the country. Infant mortality was higher for mothers younger than 20 years (21.4 percent) and firstborn children (20.8 percent). Infant mortality was higher (17.5 percent) for mothers with no formal education than mothers with secondary or higher education (8.0 percent).

Three spatial Bayesian models were fitted. A baseline model (0) included no covariates but an overall constant and site-specific random effects. Model 1 was an extension of the baseline model by including year of birth, socio-economic and demographic variables as potential risk factors. Model 2 included the same parameters as model 1 but adjusted for levels of malaria endemicity. In addition, a Bayesian non-spatial analogue of model 2 was fitted for comparative purposes. Parameters estimates obtained from Models 1, 2 and the

	Crude data		Nonspatial		Model 1		Model 2	
	No of births	IMR per 1,000	OR	95% CI	OR	95% CI	OR	95% CI
Year of birth								
1960–1965	299	230.8	1.16	0.89, 1.56	1.16	0.88, 1.56	1.16	0.89, 1.56
1966–1971	1766	239.0	1.41	1.15, 1.79	1.42	1.09, 1.85	1.42	1.12, 1.79
1972–1977	4200	186.4	1.14	0.93, 1.41	1.15	0.90, 1.49	1.15	0.91, 1.42
1978–1983	7709	172.8	1.09	0.92, 1.35	1.08	0.84, 1.40	1.09	0.86, 1.34
1984–1989	11087	163.2	1.05	0.88, 1.30	1.05	0.81, 1.35	1.06	0.85, 1.30
1990–1996	10845	152.6	1		1		1	
Residence								
Rural	25615	183.9	1		1		1	
Urban	10291	132.1	0.68	0.64, 0.72	0.74	0.69, 0.80	0.73	0.68, 0.80
Mother's education								
Secondary or higher	1416	79.8	0.54	0.42, 0.73	0.55	0.43, 0.74	0.55	0.42, 0.73
Primary	3563	149.3	0.95	0.80, 1.11	0.95	0.79, 1.15	0.96	0.80, 1.13
None	30927	175.4	1		1		1	
Sex								
Female	17718	159.6	1		1		1	
Male	18188	178.3	1.15	1.10, 1.20	1.15	1.10, 1.21	1.16	1.10, 1.21
Birth order								
Firstborn	7680	208.3	1		1		1	
2nd or 3rd	11746	158.6	1.10	1.02, 1.17	1.10	1.01, 1.20	1.10	1.01, 1.21
4th to 6th	10851	154.3	1.17	1.06, 1.29	1.15	1.05, 1.31	1.17	1.06, 1.30
7th or higher	5629	165.7	1.34	1.20, 1.52	1.35	1.17, 1.52	1.34	1.17, 1.55
Preceding birth interval								
Below 2 Years	18149	213.8	1		1		1	
2-4 Years	15231	131.8	0.59	0.56, 0.62	0.58	0.53, 0.63	0.59	0.54, 0.63
Above 4 Years	2526	88.7	0.40	0.36, 0.44	0.40	0.35, 0.47	0.39	0.32, 0.46
Mother's age at birth								
Younger than 20 years	9070	213.9	1		1		1	
20–29 Years	19087	156.6	0.79	0.75, 0.83	0.78	0.72, 0.83	0.78	0.73, 0.83
30–39 Years	7163	146.2	0.73	0.67, 0.78	0.72	0.65, 0.80	0.72	0.65, 0.80
40–49 Years	586	160.4	0.82	0.69, 1.02	0.81	0.66, 1.01	0.83	0.66, 1.02
Malaria Endemicity								
0.0–0.15	9951	157.8	1				1	
0.16–0.35	5159	185.3	0.96	0.88, 1.04			0.97	0.87, 1.09
0.36–0.64	16824	170.9	0.88	0.83, 0.94			0.89	0.78, 1.10
0.65–1.0	3972	168.4	0.85	0.77, 0.92			0.92	0.80, 1.08
σ^2					0.88	0.21, 3.82	0.88	0.20, 3.56
ϕ					0.04	0.007, 0.24	0.05	0.008, 0.30
DIC			31794.21		31682.64		31755.73	

σ^2 is an estimate of the geographical variability and ϕ the smoothing parameter (see appendix). DIC is a measure of model fit for the comparison of models, with smaller values of DIC indicating superior fit.

Table 3.1: Infant mortality estimates in Mali for the DHS data 1995–96 in combination with malaria risk extracted from the MARA/ARMA database.

non-spatial model are shown in table 3.1. Estimates of the odds ratios indicate that in the non-spatial analysis infant mortality was related to estimated malaria prevalence after adjusting for the other risk factors. After taking into account the spatial correlation which was present in the infant mortality risk data, the effect of malaria transmission was no longer significant. In fact, the point estimates of the Odds Ratios change little, however the confidence intervals become wider, confirming the importance of taking into account spatial correlation when analyzing geographical data (Cressie, 1993). Model comparison revealed that the model with the smallest DIC value and therefore with the best fit to be the spatial model 1 which does not include malaria risk.

The fixed effects parameters of the best fit Model 1 showed well-known patterns and confirmed most of the results obtained from the crude data summaries (table 3.1). In particular, non-first born children are at higher risk than their first-born siblings (OR=1.10, 95 percent confidence interval: 1.01,1.20 for second or third born, OR=1.35, 95 percent confidence interval: 1.17,1.52 for seventh or higher in birth order). The discrepancies

between model-based estimates and observed frequencies in the estimates of the parameters for birth order and preceding birth interval can be explained by the high correlation of the two variables, which introduces confounding. The model allows adjustment for confounders and provides estimates of the effects of one factor in the presence of the other. Infants born to mothers with no education are at higher risk than those born to mothers with secondary education or higher (OR=0.55, 95 percent confidence interval: 0.43,0.74). Mortality was related to the sex of the infant with boys being at higher risk of dying during the first year of life than girls (OR=1.15, 95 percent confidence interval: 1.10,1.21). On the other hand, longer birth intervals reduce the risk of infant death (OR=0.40, 95 percent confidence interval: 0.35,0.47 above 4 years vs. less than two years). Infants born to older mothers and in urban areas have higher chances of surviving to their first birthday and there is an indication that infants born in recent years have better survival chances.

The parameters σ^2 and ϕ (table 3.1) measure the variance of the spatial process and the rate of correlation decay (smoothing parameter), respectively. Our dataset indicates a small value of ϕ with posterior median of 0.04 (95 percent credible interval: 0.007,0.24) suggesting a strong spatial correlation because this parameter measures the range of the geographical dependency, which is defined as the minimum distance at which spatial correlation between locations is below 5 percent. In our exponential setting it can be calculated as $3/\phi = 75$ degrees of longitude and latitude. This implies a non-vanishing correlation between all sampled points and results in very smooth maps for the predicted random effects.

Figure 3.1 displays the distribution of malaria surveys. The figure shows that most of the surveys were carried out in the south and south-west of the country. There are very few surveys in the center around the Niger river and no surveys in the north of Mali. Model predictions of malaria risk at the DHS locations (figure 3.2) indicate a low malaria risk zone north of the Niger which contradicts empirical evidence of high transmission, suggesting that the lack of data leads to imprecise estimates in this part of the country which may distort the relationship between malaria endemicity and infant mortality. The unadjusted map of infant mortality risk (figure 3.3), obtained by the predictions of the Bayesian model, reveals that the highest infant mortality rates are found in the central and central-east part of the country around the Niger river. Distinct foci of high mortality can be identified in the regions of Nara, Banamba, Dioila, Kadiolo, Kolondieba and Kenieba. Figure 3.4 represents the variation in infant mortality which is not explained by socio-economic differences (on a logit scale). This map is a measure of our estimate of the geographical variation of risk of infant mortality, independent of the particular socio-economic circumstances of the mother. It therefore reflects the marginal burden of infant mortality that is due to ecological factors such as malaria transmission intensity and other diseases at map locations. A component of this may be residual socio-economic factors that our co-variate data did not fully account for. In this map we can distinguish three zones of high risk; the one in the central and central-east border with Burkina Faso, a zone of South Mali which covers the regions from Nara and Diema to Kolondieba and Bougouni, and a zone in the south-west in the region of Kenieba. Estimates of the variance of the residual spatial variation (figure 3.5) show lower variance in estimates near locations with observed mortality.

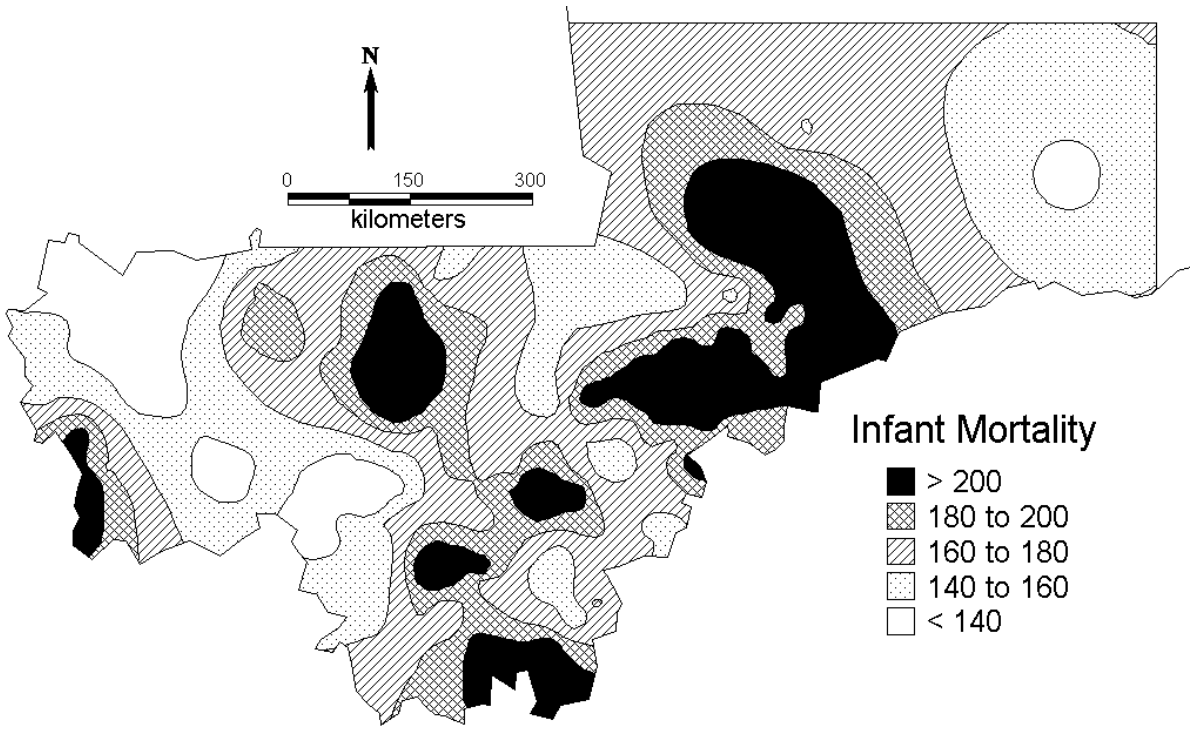


Figure 3.3: Smoothed map of the infant mortality rate (per 1,000) in Mali based on the baseline model without covariates.

3.4 Discussion

This analysis demonstrates the use of Bayesian geostatistical models in assessing risk factors and producing smooth maps of infant mortality risk from spatially correlated disease data on individuals, such as those available from DHS's. Results confirmed strong geographical differences in mortality risk and the importance of a number of risk factors such as maternal education and age, birth order and interval, sex and residence. Year of birth appeared not to be significantly associated with mortality during the first year of life, except of the period 1966–1971 in which a statistically significant increase in infant mortality was observed. There were no differences in infant mortality between the four categories of malaria endemicity defined using the model by Kleinschmidt et al. (2000), suggesting that the geographical distribution of malaria is not a major determinant of the pattern of infant mortality in Mali. This finding was not supported by the non-spatial analysis, since accounting for spatial correlation results in more precise estimates of the standard error and widens the confidence limits of the estimated odds ratios.

The risk factors which were found to be related to mortality are already well known. The negative association between maternal education and mortality has been previously described by Farah et al. (1982). Higher education may result in higher health awareness, better utilization of health facilities (Jain, 1988), higher income and ability to purchase

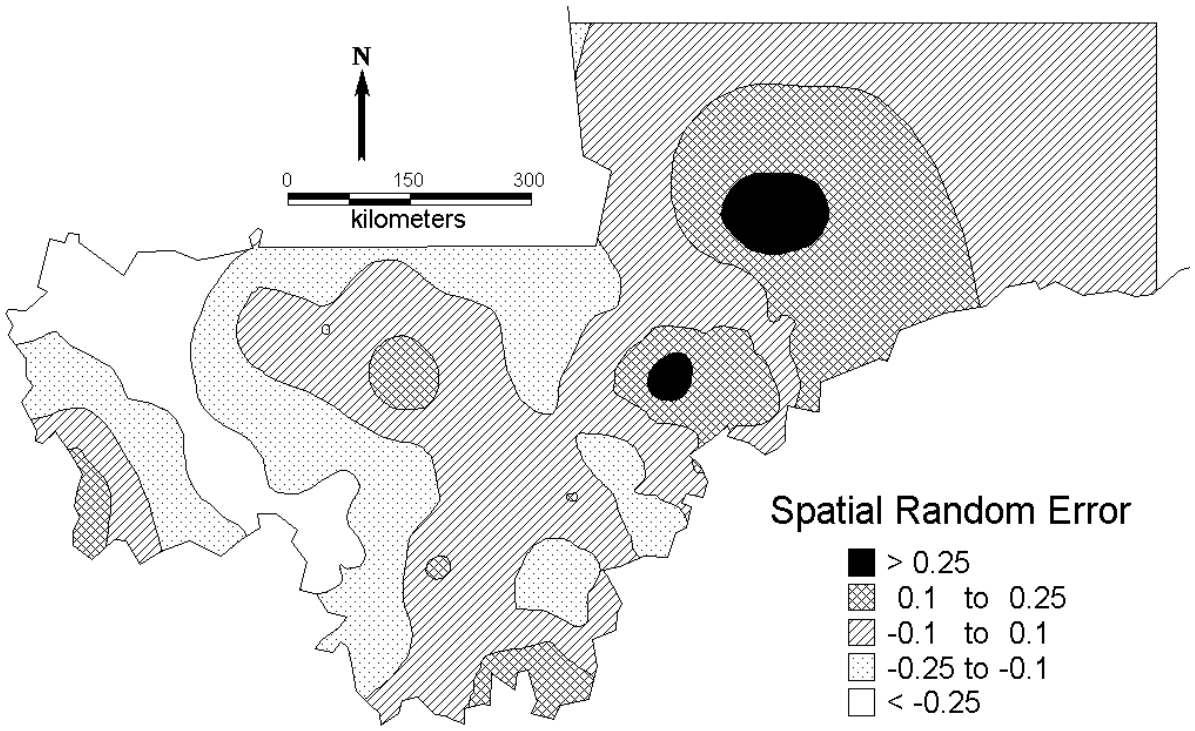


Figure 3.4: Map of Mali showing the spatial random effects at the log-odds scale for the socio-economic-adjusted model.

goods and services which improves infants health (Schultz, 1979), longer birth intervals and possibly higher maternal ages (Cleland and Ginneken, 1989).

While several studies have investigated the relationship between malaria transmission and child mortality, this is the first study to our knowledge which attempts to assess this relationship taking into account the geographical variation which is present for both parameters after adjusting for socio-economic confounders. In this analysis, we consider malaria prevalence as a measure of malaria transmission. There are alternative indicators of transmission intensity which have been used to study the effects of malaria on mortality, however the relationship between these indicators has not been fully investigated. The most usual measure is the entomological inoculation rate (EIR), which is the product of the vector biting rate times the proportion of mosquitoes infected with sporozoite-stage malaria parasites. Beier et al. (1999) reports that EIR is only weakly related to malaria prevalence. To our knowledge no studies have been carried out on the best measure of transmission to study mortality.

We have estimated for the first time the geographical distribution of the burden of infant mortality in Mali in addition to that which can be attributed to socio-economic differences. It is plausible that a large measure of this burden is due to the effect of malaria on infant mortality, even though we were not able to demonstrate this directly. The lack of a relationship between malaria risk and infant mortality could reflect unmeasured local

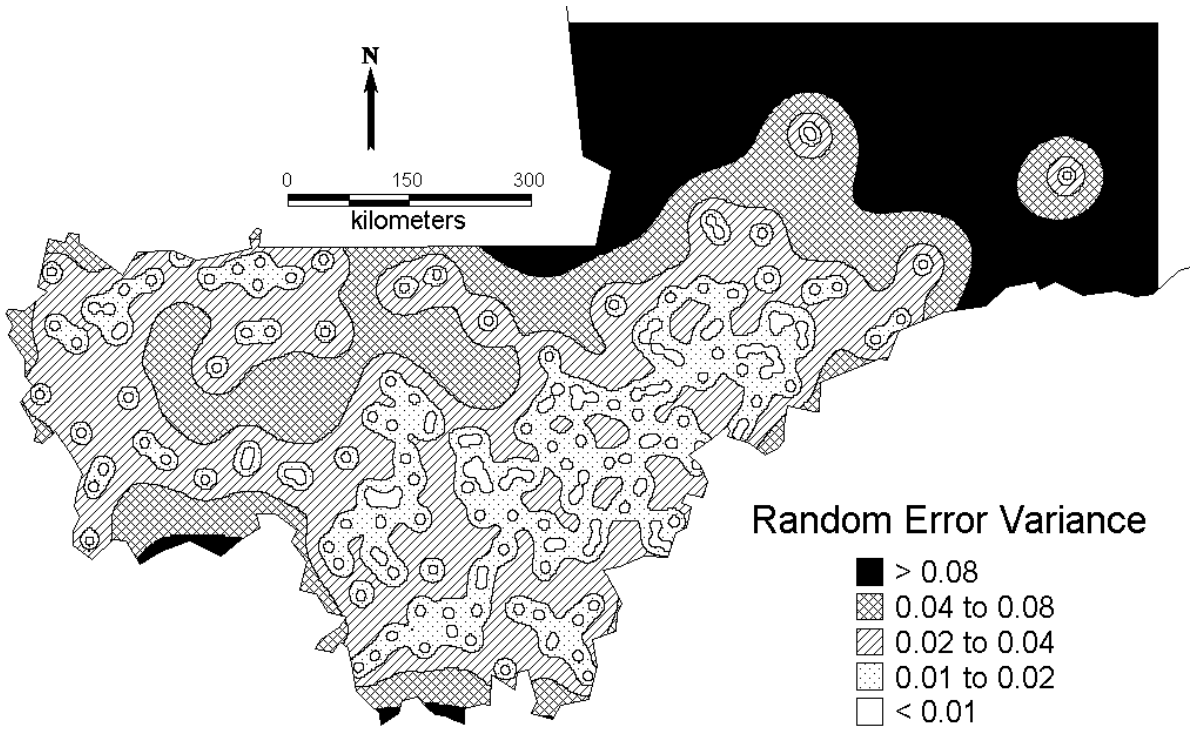


Figure 3.5: Map of Mali showing the variance of the residual spatial variation of the infant mortality risk (at logit scale) adjusted for socio-economic variables.

factors, for instance variations in health provisions or availability of water supply in the dry Sahel region, which could have a stronger influence than malaria risk on infant mortality. Such unmeasured covariates with spatial structure could also explain the residual spatial correlation in the data. Information missing from the database regarding malaria control measures taken at different locations, could also confound the analysis. Methodological problems related to the compilation of survey datasets, such as the MARA database constitute a further limitation to our analysis. The surveys, for example, were carried out at different seasons, and include different age groups at the various locations. It may well be that in areas with seasonal malaria the effect on mortality would be stronger than in areas with perennial malaria. The model for malaria prevalence of Kleinschmidt et al. (2000) did not take into account seasonal variation, although malaria transmission in Mali is known to be highly seasonal (Tanser et al., 2002).

An additional problem with the database is the sparsity of the surveys in the central-east part of the country. In this analysis we used a subset of the data in order to deal with the different age grouping of the surveys at the various locations. Currently we are working on alternative approaches which overcome the limitations of heterogenous age grouping, without omitting data.

An additional methodological problem is the misalignment of the DHS and MARA surveys in time and space. Our analysis is based on the assumption that spatial patterns

of infant mortality and malaria risk are relatively stable in time. Although this assumption can be questioned, statistical analysis of the temporal changes of malaria prevalence over the last forty years at country level (Snow et al., 1997), showed no significant patterns. In addition, our analysis of infant mortality rates indicates a statistically significant time trend only for the early years of 1966–1971, but it is quite stable for the last twenty years, when over 80 percent of all cases are recorded. To overcome geographical misalignment, we estimated malaria prevalence at the DHS locations using the spatial malaria model of Kleinschmidt et al. (2000). Although we believe that this modelling approach gives a good estimate of the general pattern of malaria prevalence in Mali and of overall populations at risk, we cannot be confident in local malaria predictions, especially in areas remote from sampled locations. In particular, the paucity of sampling points in areas of very high infant mortality, especially in the Niger delta, may have resulted in poor predictions in these areas. We propose to address this problem, by compiling the databases from a larger area of West Africa, and analyzing only data points where misalignment is minimal. Despite these limitations, our study has demonstrated considerable potential of spatial statistical methods for analyzing the DHS data. To our knowledge this is the first analysis of infant mortality employing geostatistical models. The methods presented are valuable both for producing smoothed (covariate adjusted) maps of mortality risk and assessing covariate effects. Such maps are particularly helpful to identify high mortality areas for most efficiently allocating limited resources in child survival programs.

Acknowledgements

The work of the first author was supported by Swiss National Foundation grant Nr. 3200–057165.99. We are grateful to the MARA collaboration for making the Mali malaria data available and to the Macro International Inc. for providing the names of the villages of the DHS surveys.

Appendix 3.A Statistical model

Let Y_{ij} be a binary response corresponding to the survival status of child i at site $s_j, j = 1, \ldots, n$ taking value 1 if the child is alive after the first year and 0 otherwise, and let $\boldsymbol{X}_{ij}$ be the vector of associated covariates. Following the modelling framework of Diggle et al. (1998), we introduce the unobserved spatial variation by assuming a latent stationary Gaussian process $U(\boldsymbol{s})$ over our study region, $\mathcal{D}$, such that $\boldsymbol{U} = (U(s_1), U(s_2), \ldots, U(s_n)) \sim N(0, \Sigma)$, where Σ_{ij} is a parametric function of the separation vector $s_i - s_j$. Conditional on $\boldsymbol{U}$ and the regression coefficients $\boldsymbol{\beta}$, the Y_{ij} are independent Bernoulli variates with survival probabilities p_{ij} given by $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + U(s_j)$. We assume an isotropic spatial process with $\Sigma_{ij} = \sigma^2\rho(s_i - s_j; \phi)$ and an exponential correlation function $\rho(s_i - s_j; \phi) = \exp(-\phi d_{ij})$ where d_{ij} measures the Euclidean distance between the sites s_i and s_j.

To complete the Bayesian model specification, we need to adopt prior distributions

for the model parameters. We chose noninformative Uniform priors for the regression coefficients, i.e. $\boldsymbol{\beta} \propto \mathbf{1}$, and the following vague priors for the σ^2 and ϕ parameters: $\sigma^2 \sim$ Inverse-Gamma(a_1, b_1) and $\phi \sim$ Gamma(a_2, b_2), with $a_1 = 0.01$, $b_1 = 0.01$, $a_2 = 2.01$, $b_2 = 1.01$. The model was fitted using Markov chain Monte Carlo and in particular Gibbs sampling (Gelfand and Smith, 1990). The posterior distribution of σ^2 then is conjugate Inverse-Gamma. From the non-standard one-dimensional conditional distributions of all components of $\boldsymbol{\beta}$, $\boldsymbol{U}$ such as ϕ, we sampled by employing a random walk Metropolis algorithm having a Gaussian proposal density with mean equal the estimate from the previous iteration and variance derived from the inverse second derivative of the log-posterior. We run a single chain sampler with a burn-in of 5,000 iterations with convergence assessed by inspection of ergodic averages of selected model parameters. The chain thereafter sampled every 60th iteration until a sample size of size 2,000 has been attained.

For model comparison we utilize the Deviance Information Criterion (DIC), as recently proposed by Spiegelhalter et al. (2002). For a vector of parameters $\boldsymbol{\theta}$, it is defined by DIC$= 2\overline{D} - D(\overline{\boldsymbol{\theta}})$, with $D(\cdot)$ being the deviance statistic $D(\boldsymbol{\theta}) = -2 \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + 2 \log f(\boldsymbol{y})$. $\overline{D}$ is the posterior expectation of D and $\overline{\boldsymbol{\theta}}$ the posterior expectation of $\boldsymbol{\theta}$, with both of them easily estimated from outputs of the MCMC sampler. Smaller values of the DIC indicate better fitting models.

To produce a smooth map of mortality risk we use Bayesian kriging (Cressie, 1993; Gelfand et al., 1999). In particular, we obtain estimates of the mortality risk, $\boldsymbol{Y}_0 = (Y(s_{01}), Y(s_{02}), \ldots, Y(s_{0l}))$ at any unsampled location $\boldsymbol{s}_0 = (s_{01}, s_{02}, \ldots, s_{0l})$ by the predictive distribution

$$\begin{aligned} P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}) \;\; = \;\; & \int P(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{U}_0) P(\boldsymbol{U}_0 \mid \boldsymbol{U}, \sigma^2, \phi) \times \\ & P(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi \mid \boldsymbol{Y}) \; d\boldsymbol{\beta} \; d\boldsymbol{U}_0 \; d\boldsymbol{U} d\sigma^2 d\phi \end{aligned} \tag{3.1}$$

where the distribution of $\boldsymbol{U}_0$ at new sites given $\boldsymbol{U}$ at observed sites is normal

$$P(\boldsymbol{U}_0 \mid \boldsymbol{U}, \sigma^2, \phi) = \mathcal{N}(\boldsymbol{\Sigma}_{01} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{U}, \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{10}) \tag{3.2}$$

with $\boldsymbol{\Sigma}_{11} = \mathrm{E}(\boldsymbol{U}\boldsymbol{U}^t)$, $\boldsymbol{\Sigma}_{00} = \mathrm{E}(\boldsymbol{U}_0 \boldsymbol{U}_0^t)$, $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^t = \mathrm{E}(\boldsymbol{U}_0 \boldsymbol{U}^t)$ and $p\left(Y(s_{0i}) \mid \boldsymbol{\beta}, U(s_{0i})\right) \sim Ber(p(s_{0i}))$, with $\mathrm{logit}(p(s_{0i})) = \boldsymbol{x}(s_{0i})\boldsymbol{\beta} + U(s_{0i})$. Equation (3.1) is the expectation $\mathrm{E}[P(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{U}_0) P(\boldsymbol{U}_0 \mid \boldsymbol{U}, \sigma^2, \phi)]$ over the posterior distribution $P(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi \mid \boldsymbol{Y})$, which is identified by the Gibbs sampler. Numerically this expectation is approximated by the average

$$\frac{1}{r} \sum_{k=1}^{r} P(\boldsymbol{Y}_0^{(k)} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{U}_0^{(k)}) P(\boldsymbol{U}_0^{(k)} \mid \boldsymbol{U}^{(k)}, \sigma^{2(k)}, \phi^{(k)}), \tag{3.3}$$

where $(\boldsymbol{\beta}^{(k)}, \boldsymbol{U}^{(k)}, \sigma^{2(k)}, \phi^{(k)})$ are samples drawn from the posterior $P(\boldsymbol{\beta}, \boldsymbol{U}, \sigma^2, \phi \mid \boldsymbol{Y})$. For mapping purposes, predictions were made for 600,000 pixels covering on a regular grid the whole area of Mali south of 18 degrees latitude north.

CHAPTER 4

Bayesian modelling of geostatistical survival data in relation to misaligned covariates: an application to mapping child survival in Mali

Gemperli A.[1], Vounatsou P.[1], Smith T.[1] and Gelfand A.E.[2]

This paper is being prepared for submission to *Statistics in Medicine*.

[1] Swiss Tropical Institute, Basel
[2] Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708–0251

Abstract

A Bayesian accelerated failure time model for spatially point referenced data was developed to assess the impact of site specific malaria endemicity on child mortality in Mali. The data on malaria endemicity were collected at sites different than the ones of the survival data. We solve this spatial misalignment problem by fitting a spatial logistic model to the malaria prevalence data and then predicting the malaria prevalence at the sites where the survival data were available, using Bayesian kriging. The estimation error of the malaria endemicity is taken into account in the survival model as a measurement error in the covariate. The fitting of the mortality data and the prediction of the malaria covariate were built within a single Bayesian hierarchical model. Markov chain Monte Carlo was used to estimate the model parameters. No significant effect of location specific malaria risk on child mortality was found.

Keywords: accelerated failure time model; bayesian hierarchical model; geostatistics; kriging; malaria; Markov chain Monte Carlo; misalignment.

4.1 Introduction

The effectiveness of malaria control in Africa in reducing child mortality depends not only on the extent to which malaria endemicity is reduced but also on the relationship between endemicity and mortality. A number of meta-analysis have recently reported on the relationship of malaria-specific mortality rates in children to the level of malaria exposure, however the results are not conclusive (Smith et al., 2001; Snow and Marsh, 2002). This is mainly due to methodological problems such as inconsistencies between studies in case definitions and in the systems for monitoring mortality, ecological bias, and small number of published studies on malaria specific mortality rates. One approach, which has not been fully explored, is to analyze site specific overall mortality data available from demographic and health surveys (DHS) across Africa with local malaria indices taking into account geographical variation in both malaria transmission and mortality.

In this work we developed a survival model for spatial point referenced data to assess the impact of site specific malaria endemicity on child survival in Mali. The data on malaria endemicity and infant survival were obtained from the Mapping Malaria Risk in Africa (MARA/ARMA, 1998) and the Demographic and Health Survey (DHS, Mali 1996) databases respectively. The two databases are misaligned, as the sites at which the malaria surveys were carried out do not match with the sites of the DHS survey and thus the malaria endemicity covariate is available at different locations than the survival outcome.

Gotway and Young (2002) provide a recent review of statistical approaches dealing with the spatial misalignment problem. Mugglin et al. (2000) developed Bayesian methodology for misaligned areal units data. The spatial dependence in their approach is modelled via Markov random field models. Gelfand et al. (2001) discuss spatial misalignment between point-referenced datasets within the context of Bayesian spatial prediction (Le and Zidek, 1992; Handcock and Stein, 1993; Gaudard et al., 1999), however they do not con-

cern with modelling misaligned covariates. Banerjee and Gelfand (2002) developed a fully Bayesian approach for modelling misaligned geostatistical data. Their work is based upon a multivariate stationary Gaussian process model, which takes into account the covariance structure between the outcome and the misaligned covariates and presumes a common degree of smoothness for both outcome and covariates. Furthermore, it does not allow for the covariate to depend on other predictors or inclusion into the model of aligned covariates. In the malaria-child mortality application, it is unreasonable to assume a common smoothness parameter for both the malaria risk and the infant mortality rates. In addition the malaria endemicity relates to environmental and climatic predictors.

In our approach, we predict the misaligned malaria risk covariate at the locations where the child survival outcome is observed and then fit the survival model, incorporating the prediction error as a measurement error in the covariate. The prediction of the malaria covariate and the fitting of the survival data are built in a single Bayesian hierarchical model. We employed variogram modelling via Bayesian inference (Diggle et al., 1998) due to the flexibility in the model fit via Markov chain Monte Carlo (MCMC). Furthermore MCMC inherently estimates the uncertainty in model parameters avoiding the asymptotic inference problem which lies in the distinction between increasing-domain versus infill asymptotics (Cressie, 1993).

Spatial lifetime models are considered by Crook et al. (2003), who analyzed areal survival data by restructuring them as binary longitudinal and fitting a binary regression with a probit link function. The authors adopted Markov random field priors for the spatially structured random effects. Banerjee at al. (2003) fitted the Cox proportional hazards model with a Weibull baseline hazard and assumed Gaussian random field priors for the spatially structured frailties. Our contribution is to extend this work in the case of the accelerated failure time model for point-referenced data in the presence of misaligned covariates.

The remainder of this article is structured as follows: In section 4.2 we present the child survival and the malaria prevalence which motivated this work. In section 4.3 we describe the accelerated failure time model for geostatistical data and extend the model in the presence of misaligned covariates. We discuss the results of our application in section 4.4. We conclude with final remarks in section 4.5.

4.2 Data

Data on child survival are available from the Demographic and Health Survey program (DHS, Mali 1996) conducted by Macro International Inc. (Coulibaly et al., 1996). The database includes nationally representative household surveys intended to provide data for a range of monitoring and impact evaluation indicators within the population, health, and nutrition sector. Birth histories corresponding to 35,906 children were extracted from the DHS database, collected during a survey carried out from November 1995 to April 1996 in Mali. The women's questionnaire contains information about their social (e.g. education, ethnicity) and economical (e.g. access to sanitation) being, among others, such as the information about their children's dates of birth and death. There is no information

available about the cause of the child's death. The sample sites were geo-located to add coordinates for every observation. The data we analyzed were collected at 181 distinct villages.

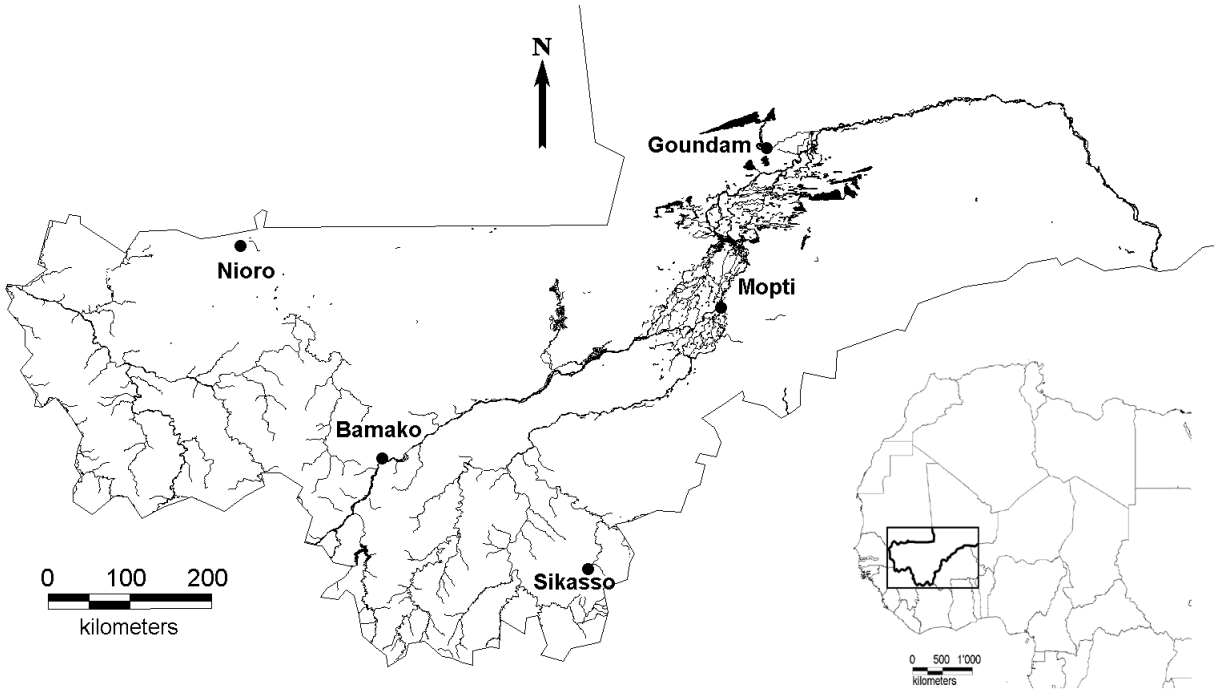


Figure 4.1: Southern, non-Sahara region of Mali with rivers and lakes plus larger towns, where most of the larger malaria surveys took place.

Data on malaria endemicity were obtained from the Mapping Malaria Risk in Africa (MARA) database (MARA/ARMA, 1998) which is the most comprehensive geo-referenced database of all available published and unpublished malariometric survey data in 44 African countries. These surveys record the presence of *Plasmodium falciparum* in blood smears. For this analysis, we extracted 153 survey prevalence data carried from 95 distinct villages in Mali, including children in the range of one to ten years old from 1965 up to 1999.

The environmental and climatic factors which we used to predict malaria endemicity were obtained from remote sensing. The normalized Vegetation Index (NDVI) was extracted from the NOAA/NASA Pathfinder AVHRR Land Project and considered as a measure of greenness. The temperature and rainfall data were obtained from the topographic and climate database for Africa (Hutchinson et al., 1996).

The MARA data for Mali and the topographic and climate database for Africa have been used by Kleinschmidt et al. (2000) to produce smooth maps of malaria risk. In a recent publication (Gemperli et al., 2004) we analyzed the relation between malaria risk and infant mortality in Mali using the same datasets. The two analyzed datasets were spatially misaligned, what was not taken into account.

4.3 Model specification

4.3.1 Spatial accelerated failure time model

Let $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)^t$, $s_i \in D \subset R^d$ be the set of locations with the observed mortality data, $T_j(s_i)$ the time to death for infant j at location s_i, $\boldsymbol{X}_j(s_i)$ a row-vector of associated covariates and $Z(s_i)$ the malaria prevalence covariate at the logit-scale which we consider for the time being, that it is observed at s_i. Assuming a Weibull distribution for $T_j(s_i)$ with shape and scale parameters, $\lambda_j(s_i)$ and γ, respectively, the survival function is given by $S(T_j(s_i)) = \exp(-\lambda_j(s_i)T_j(s_i)^\gamma)$. We introduce spatial correlation via location-specific, spatially structured frailties $\phi(s_i)$ and model covariates and random effect as a linear term into $\eta_j(s_i) = \boldsymbol{X}_j^t(s_i)\boldsymbol{\alpha}_1 + Z(s_i)\beta + \phi(s_i)$ where $\lambda_j(s_i) = \exp(-\gamma\eta_j(s_i))$. The hazard function of this model is given by $h(T_j(s_i) \mid \boldsymbol{X}_j(s_i), Z(s_i), \phi(s_i)) = \gamma T_j(s_i)^{\gamma-1}\lambda_j(s_i)$. A normal prior is adopted for the coefficients $\boldsymbol{\alpha}_1$ and β and an inverse gamma prior is selected for γ.

We incorporate the spatial structure on $\boldsymbol{\phi}$ by assuming a latent stationary Gaussian isotropic spatial process over D such that $[\boldsymbol{\phi} \mid \sigma_\phi^2, \delta_\phi] \sim N(0, \Sigma_\phi)$ where $(\Sigma_\phi)_{ij} = \sigma_\phi^2\rho(d(s_i, s_j); \delta_\phi)$ and $\rho(\cdot)$ is a valid (non-negative definite) correlation function in R^2. $d(s_i, s_j)$ is the Euclidean distance between s_i and s_j, δ_ϕ a correlation scale parameter capturing the scale of correlation decay with distance and σ_ϕ^2 corresponds to the sill of the spatial process. For the current example we chose the exponential form $\rho(d_{ij}; \delta_\phi) = \exp(-d(s_i, s_j)/\delta_\phi)$ as the correlation function. More general forms are discussed in Handcock and Wallis (1994) and Diggle et al. (1998). To complete Bayesian formulation of the model we specify an inverse gamma prior distributions for the variance parameters σ_ϕ^2 and a gamma distribution for the parameter δ_ϕ.

In our model specification, the spatial random effect alters the hazard multiplicatively by the factor $\exp(-\gamma\phi(s_i))$. Alternatively, Bolstad and Manda (2001) consider an exchangeable gamma frailty $\phi(s_i)$ with mean one, which acts multiplicatively on the hazard function, i.e. $h(T_j(s_i) \mid \boldsymbol{X}_j(s_i), Z(s_i), \phi(s_i)) = \phi(s_i)\gamma T_j(s_i)^{\gamma-1}\exp\left(-\gamma(\boldsymbol{X}_j^t(s_i)\boldsymbol{\alpha}_1 + Z(s_i)\beta)\right)$. Henderson et al. (2002) define gamma frailties and incorporate the spatial dependence by a multivariate Gaussian distribution on the mean parameter of the Gamma distribution. In this paper, the spatial correlation structure is included at a log-Gaussian scale in the hazard. The fixed and the random effects are then expressed at the same scale, rendering the analysis of interdependence more reliable. Assuming that the $T_j(s_i)$ are independent conditional on the covariates and spatial random effects, we have

$$[\boldsymbol{T}(\boldsymbol{s}) \mid \boldsymbol{X}(\boldsymbol{s}), Z(\boldsymbol{s}), \boldsymbol{\phi}(\boldsymbol{s}), \gamma] = \prod_{i,j}[T_j(s_i) \mid \boldsymbol{X}_j(s_i), Z(s_i), \phi(s_i), \gamma]$$

and the posterior distribution of the model will be

$$[\boldsymbol{\phi}(\boldsymbol{s}), \sigma_\phi^2, \delta_\phi, \boldsymbol{\alpha}_1, \beta, \gamma \mid \boldsymbol{T}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}), Z(\boldsymbol{s})] = [\boldsymbol{T}(\boldsymbol{s}) \mid \boldsymbol{X}(\boldsymbol{s}), Z(\boldsymbol{s}), \boldsymbol{\phi}(\boldsymbol{s}), \boldsymbol{\alpha}_1, \beta, \gamma][\boldsymbol{\phi} \mid \sigma_\phi^2, \delta_\phi][\boldsymbol{\alpha}_1, \beta, \gamma, \sigma_\phi^2, \delta_\phi]$$

The above model assumes that all covariates are observed, but in the mortality application, the malaria prevalence covariate is observed at a different set of locations. To align the locations of the malaria covariate with those of the mortality data we fit a spatial binomial model on the observed malaria prevalence considering environmental and ecological factors as predictors and then predict malaria at the mortality locations. The fitting of the mortality data and the prediction of the malaria covariate can be encompassed within a single Bayesian hierarchical model formulation.

4.3.2 Spatial accelerated failure time model with misaligned covariates

Let $\boldsymbol{s}^{'} = (s_1^{'}, s_2^{'}, \ldots, s_n^{'})^t$ be the set of locations with the observed malaria data which are different than the locations $\boldsymbol{s}$. Let also $N(s_i^{'})$ be the number of children screened during a particular survey at site $s_i^{'}$ and $Y(s_i^{'})$ be the number of those found positives to malaria parasites. We assume that $Y(s_i^{'})$ are conditionally independent given the $p(s_i^{'})$, that is $[Y(\boldsymbol{s}^{'}) \mid N(\boldsymbol{s}^{'}), p(\boldsymbol{s}^{'})] = \prod_i [Y(s_i^{'}) \mid N(s_i^{'}), p(s_i^{'})]$ where $[Y(s_i^{'}) \mid N(s_i^{'}), p(s_i^{'})] \sim Bn\left(N(s_i^{'}), p(s_i^{'})\right)$ and introduce the location-specific environmental covariates $\boldsymbol{\Psi}(s_i^{'})$ and spatial random effects $\omega(s_i^{'})$ on the logit scale via $\text{logit}(p(s_i^{'})) = \boldsymbol{\Psi}^t(s_i^{'})\boldsymbol{\alpha}_2 + w(s_i^{'})$, where $Z(s_i^{'}) = \text{logit}(p(s_i^{'}))$. Spatial correlation is captured in the $\boldsymbol{w}(\boldsymbol{s}^{'})$ by adopting a multivariate Gaussian process as described for the survival model, that is $[\boldsymbol{w}(\boldsymbol{s}^{'}) \mid \sigma_w^2, \delta_w] \sim N\left(0, \Sigma_{w^{'}}\right)$ where $(\Sigma_{w^{'}})_{i^{'}j^{'}} = \sigma_w^2 \exp(-d(s_i^{'}, s_j^{'})/\delta_w)$. Similarly with the prior specification of the mortality covariance parameters, we assume an inverse gamma prior distributions for σ_w^2, a gamma distribution for the parameter δ_w, and a normal prior for the coefficient parameter $\boldsymbol{\alpha}_2$.

Let $\boldsymbol{\Psi}(\boldsymbol{s}) = (\boldsymbol{\Psi}(s_1), \boldsymbol{\Psi}(s_2), \ldots, \boldsymbol{\Psi}(s_m))^t$ and $Z(\boldsymbol{s}) = (Z(s_1), Z(s_2), \ldots, Z(s_m))^t$ be the vectors of the malaria-related environmental covariates and predicted malaria covariate in the logit scale at the mortality locations $\boldsymbol{s}$, respectively. Then $Z(s_i) = \boldsymbol{\Psi}^t(s_i)\boldsymbol{\alpha}_2 + w(s_i)$ where the $\boldsymbol{w}(\boldsymbol{s}) = (w(s_1), w(s_2), \ldots, w(s_m))^t$ is the vector of malaria random effects predicted at the mortality locations $\boldsymbol{s}$. Conditional on $w(\boldsymbol{s}^{'}) = \left(w(s_1^{'}), w(s_2^{'}), \ldots, w(s_m^{'})\right)^t$ and the covariance parameters of the malaria spatial process, the $\boldsymbol{w}(\boldsymbol{s})$ have a Gaussian distribution and therefore the predicted malaria covariate $Z(\boldsymbol{s})$ at the mortality locations $\boldsymbol{s}$ will be also Gaussian

$$\left[Z(s) \mid \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma_w^2, \delta_w, \boldsymbol{\Psi}(\boldsymbol{s})\right] \sim N\left(\boldsymbol{\Psi}^t(s_i)\boldsymbol{\alpha}_2 + \Sigma_{w^{'}w}\Sigma_{w^{'}}^{-1}\boldsymbol{w}(\boldsymbol{s}^{'}), \Sigma_w - \Sigma_{w^{'}w}\Sigma_{w^{'}}^{-1}\Sigma_{ww^{'}}\right).$$

where $(\Sigma_{w^{'}w})_{ij} = (\Sigma_{ww^{'}})_{ji} = \sigma_w^2 \exp(-d(s_i^{'}, s_j)/\delta_w)$, $(\Sigma_w)_{ij} = \sigma_w^2 \exp(-d(s_i, s_j)/\delta_w)$, $(\Sigma_{w^{'}})_{ij} = \sigma_w^2 \exp(-d(s_i^{'}, s_j^{'})/\delta_w)$.

The posterior distribution $[\boldsymbol{\theta} \mid \boldsymbol{T}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}), Y(\boldsymbol{s}^{'}), N(\boldsymbol{s}^{'}), Z(\boldsymbol{s}^{'})]$ of the accelerated model taking into account the misaligned covariates can be factored as follows:

$$[\boldsymbol{T}(\boldsymbol{s}) \mid \boldsymbol{X}(\boldsymbol{s}), Z(\boldsymbol{s}), \boldsymbol{\phi}(\boldsymbol{s}), \boldsymbol{\alpha}_1, \beta, \gamma][Z(\boldsymbol{s}) \mid \boldsymbol{\alpha}_2, \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma_w^2, \delta_w, \boldsymbol{\Psi}(\boldsymbol{s})][\boldsymbol{w}(\boldsymbol{s}^{'}) \mid \sigma_w^2, \delta_w] \times$$

$$[Y(\boldsymbol{s}^{'}) \mid N(\boldsymbol{s}^{'}), Z(\boldsymbol{s}^{'})][\boldsymbol{\phi}(\boldsymbol{s}) \mid \sigma_\phi^2, \delta_\phi][\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta, \gamma, \sigma_\phi^2, \delta_\phi, \sigma_w^2, \delta_w]$$

where $\boldsymbol{\theta}$ denotes the parameter vector $\boldsymbol{\theta} = (Z(\boldsymbol{s}), \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta, \gamma, \boldsymbol{\phi}(\boldsymbol{s}), \sigma^2_\phi, \delta_\phi, \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma^2_w, \delta_w)^t$.

The above posterior distribution will from now on be denoted as the joint form. Its parameters are estimated via Gibbs sampling. Alternatively, we can write the posterior distribution as the product of three factors, which are the posterior of the survival-model parameters, the posterior of the malaria-model parameters and the predictive distribution, that links the two posteriors:

$$[\boldsymbol{\phi}(\boldsymbol{s}), \boldsymbol{\alpha}_1, \beta, \gamma \mid \boldsymbol{T}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}), Z(\boldsymbol{s})][Z(\boldsymbol{s}) \mid \boldsymbol{\alpha}_2, \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma^2_w, \delta_w, \boldsymbol{\Psi}(\boldsymbol{s})]\times$$
$$[\boldsymbol{\alpha}_2, \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma^2_w, \delta_w \mid Y(\boldsymbol{s}^{'}), N(\boldsymbol{s}^{'}), \Psi(\boldsymbol{s}^{'})]$$

As an addition to the joint form, parameter estimation can be accomplished by running a separate Gibbs sampling for all three factors in the above posterior distribution. In this approach $Z(\boldsymbol{s})$ is interpreted as the predicted malaria risk at the mortality locations $\boldsymbol{s}$, derived using environmental covariates. The parameter $\boldsymbol{\alpha}_2$ specifies the effect environmental factors have on malaria risk, $\boldsymbol{\alpha}_1$ gives information on the effect of individual covariates on mortality and $\boldsymbol{\beta}$ measures the impact malaria risk has on mortality. Those interpretations are lost if the parameters are estimated using the joint form. $\boldsymbol{\alpha}_1$ no longer only fits the environmental factors to the malaria risk, but is also the parameter optimizing the prediction of $Z(\boldsymbol{s})$. If $Z(\boldsymbol{s})$ is not kept fixed in the survival model, it forms a compromise between location specific frailty in the survival model and the malaria predictor from the malaria model. Its estimate, such as that of $\boldsymbol{\alpha}_1$, is affected by the difference in sample size between the malaria and the mortality data.

In the separate formulation, the malaria risk affects the mortality, but not vice-versa. In the joint model, however, the mortality does also affect the malaria risk. This can be realistic in various ways. A high mortality among health professionals and among people who take care of children can result in neglected protection and treatment.Moreover, if the mortality is very high, the disease can no longer spread out and its risk is diminishing. But this latter possibility of a situation where mortality affects the disease risk is certainly not the case for malaria, which has a very low case-fatality rate.

Predictions of the mortality outcome $\boldsymbol{T}(\boldsymbol{s}_0)$ at a new set of locations $\boldsymbol{s}_0$ can be made using the posterior predictive distribution which is given by

$$[\boldsymbol{T}(\boldsymbol{s}_0) \mid \boldsymbol{X}(\boldsymbol{s}_0), \boldsymbol{Z}(\boldsymbol{s}_0), \boldsymbol{\phi}(\boldsymbol{s}_0), \boldsymbol{\alpha}_1, \beta, \gamma][\boldsymbol{Z}(\boldsymbol{s}_0) \mid \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma_w, \delta_w, \boldsymbol{\Psi}(\boldsymbol{s}_0)]\times[\boldsymbol{\phi}(\boldsymbol{s}_0) \mid \boldsymbol{\phi}(\boldsymbol{s}), \sigma_\phi, \delta_\phi]\times$$
$$[\boldsymbol{\theta} \mid \boldsymbol{T}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}), Y(\boldsymbol{s}^{'}), N(\boldsymbol{s}^{'}), Z(\boldsymbol{s}^{'})].$$

$[\boldsymbol{Z}(\boldsymbol{s}_0) \mid \boldsymbol{w}(\boldsymbol{s}^{'}), \sigma_w, \delta_w, \boldsymbol{\Psi}(\boldsymbol{s}_0)]$ is the predictive distribution of malaria risk at the new locations and derived in the same way as for the mortality locations. $\boldsymbol{\phi}(\boldsymbol{s}_0)$ and $\boldsymbol{\phi}(\boldsymbol{s})$ have a joint distribution, which is Gaussian, and the conditional $[\boldsymbol{\phi}(\boldsymbol{s}_0) \mid \boldsymbol{\phi}(\boldsymbol{s}), \sigma_\phi, \delta_\phi]$ can be easily obtained.

4.4 Application

The environmental covariates, which enter the malaria model are those found to be statistically related at a significance level of 0.2 in an univariate non-spatial analysis. These

are the vegetation index (NDVI), the distance to the nearest water source, the average maximum temperature between March and May and the length of rainy season specified by number of month with more than 60mm rainfall.

Univariate non-spatial analysis has lead to the following significant variables to adjust for in the forthcoming multivariate survival setting: Region type, mothers education degree, sex, birth order, preceding birth interval and mothers age at birth. Child's birthday year appeared to be not significantly related to mortality, however was considered in the model to measure time-trends over decades.

The model parameters were estimated using Markov chain Monte Carlo and in particular Gibbs sampling. The conditional distribution of the parameters σ_ϕ^2 and σ_w^2 are conjugate inverse gamma from which we can easily draw samples. All the other parameters do not have a conditional distribution of standard form, and a Metropolis-Hastings step was performed for sampling. We adopt a Gaussian proposal for the covariate parameters and a gamma proposal distribution for the parameters δ_w and δ_ϕ. The proposal distributions were chosen with mean equal to the parameter estimate from the previous iteration. The variance of the proposal was adaptively adjusted during the convergence period to achieve an average acceptance rate in the Metropolis-Hastings sampler of around 0.4.

The Gibbs sampler was used to estimate the model parameters in both possible ways described in section 4.3.2. The model with malaria risk handled without measurement error, plus the joint Gibbs sampling approach are both presented to allow a complete discussion of the analysis. Since we believe, that they both do not lead to the model-interpretation we are interested in (see section 4.3.2), the estimated parameters are only discussed in terms of the separate model with measurement error.

The Gibbs sampling was implemented with a single chain. After convergence, which was assessed using the Raftery-Lewis criterion (Raftery and Lewis, 1992), we collected samples every 60'th iteration reducing the autocorrelation to almost zero. The total running time was 124 hours CPU time on an AlphaServer GS80 with 8 processors and code written in Fortran 95.

The parameter estimates of the spatial malaria model in table 4.1, indicate a significant effect of temperature, rainfall and vegetation to the malaria risk. Surprisingly, the distance to water covariate appears not to be related with the malaria prevalence. This could be due to limitations of the data, because the malaria surveys were carried out throughout the year during non-standardized and overlapping periods and therefore accounting for seasonality was not possible. On the other hand is the geographical appearance of water bodies known to be highly seasonal in Mali. The vegetation (NDVI) parameter estimates has the highest precision, reflecting the effort put into extracting this factor.

The 181 locations of the MARA surveys and the 95 locations of DHS surveys are displayed in figure 4.2. There was a relatively balanced spread of both surveys over the southern Savannah area of Mali, although more malaria surveys were carried out around high populated areas (Bamako, Mopti, Sikasso, see figure 4.1), than the DHS surveys. The figure shows also the estimates of malaria prevalence obtained from Bayesian kriging. The variability of prediction generally is low, even for remote locations (inter-quartile range of

Variable	Median	5% Quantile	95% Quantile
Intercept	-99.22	-100.24	-98.42
Temperature	16.22	15.95	16.39
Rainfall	2.25	1.68	3.09
Water	0.02	-0.09	0.15
Vegetation	0.57	0.51	0.63
σ_w^2	0.84	0.65	1.10
δ_w	0.033	0.015	0.059

Table 4.1: Posterior parameter estimates for $\boldsymbol{\beta}, \sigma_w$ and δ_w from the spatial logistic malaria model which was used to predict malaria risk at new locations. All environmental-variables are taken at the log-scale.

predicted prevalence is at maximum 0.3). This is explained by the low spatial correlation in the malaria prevalence data estimated by the correlation decay parameter δ_w and the high precision of the covariate coefficients. In fact δ_w was estimated to be 0.033. In our setting with the exponential spatial correlation function this translates to a minimum distance where correlation drops below 0.05, of around 10.8 kilometers (95 percent CI: 5.0–19.5 kilometers).

Parameter estimates of the survival model are given in table 4.2. The effect of malaria endemicity on child survival was found not statistically significantly related. The spatial correlation in the mortality data is higher than that in the malaria prevalence. In particular, $\delta = 0.52$ which implies that the minimum distance where correlation drops below 0.05, is 169.3 kilometers (95 percent CI: 95.9–377.3 kilometers). On the other hand, the estimate of the spatial variance parameter σ_ϕ in the survival model ($\sigma_\phi = 0.08$ with 90 percent CI: 0.06–0.12) is lower than in the logistic model ($\sigma_w = 0.84$ with 90 percent CI: 0.65–1.10).

Prediction of random effects in the survival data were carried out, using Bayesian kriging at resolution of 596 × 1,005 equally spaced locations. The individual-specific, socio-economic covariates at new locations $\boldsymbol{X}(\boldsymbol{s}_0)$ are not known in our study and prediction performed for the spatial random effects only. Estimates of those predictions together with their standard error are mapped in figure 4.3. Figure 4.3a shows high survival in Southern and Mid-Southwest Mali and low survival in the seasonally flooded area of the Niger delta in the north-east. The standard deviations of the predicted values (figure 4.3b) reflect the distribution of the survey locations.

4.5 Discussion

We have developed a Bayesian accelerated failure time model for geostatistical data to assess the impact of site-specific malaria endemicity on child survival in Mali. The data on malaria endemicity were collected at sites different from those of the survival data. We solve this spatial misalignment problem with a Bayesian hierarchical model, where a spatial random effect in the malaria prevalence and environmental covariates are introduced to

		Separate Model				Joint Model	
		No Measurement Error		Measurement Error		Measurement Error	
	Frequency	Median	95% CI	Median	95% CI	Median	95% CI
Year of birth							
1960–1965	264	1.16	(0.96,1.42)	1.15	(0.96,1.37)	1.07	(0.75,1.51)
1966–1971	1563	1.12	(1.03,1.22)	1.11	(1.03,1.20)	1.19	(0.97,1.44)
1972–1977	3809	0.92	(0.86,0.99)	0.91	(0.86,0.98)	1.02	(0.84,1.24)
1978–1983	7091	0.87	(0.83,0.94)	0.87	(0.82,0.92)	0.98	(0.83,1.17)
1984–1989	10214	0.90	(0.85,0.95)	0.90	(0.85,0.95)	0.90	(0.78,1.07)
1990–1996	12081	1		1		1	
Residence							
Rural	24702	1		1		1	
Urban	10320	0.70	(0.65,0.75)	0.79	(0.74,0.85)	0.78	(0.66,0.93)
Mothers education							
None	30014	1		1		1	
Primary	3544	0.81	(0.75,0.88)	0.89	(0.84,0.95)	0.79	(0.65,0.96)
Secondary or higher	1464	0.44	(0.38,0.52)	0.43	(0.37,0.50)	0.54	(0.34,0.81)
Sex							
Female	17522	1		1		1	
Male	17500	1.04	(1.01,1.08)	1.09	(1.06,1.13)	1.08	(0.97,1.19)
Birth order							
Firstborn	7088	1		1		1	
2nd or 3rd	11579	1.34	(1.25,1.43)	1.33	(1.26,1.41)	1.30	(1.09,1.52)
4th to 6th	10745	1.34	(1.25,1.44)	1.33	(1.25,1.42)	1.37	(1.11,1.66)
7th or higher	5610	1.26	(1.15,1.39)	1.25	(1.16,1.37)	1.31	(1.00,1.70)
Preceding birth interval							
Below 2 Years	16757	1		1		1	
2–4 Years	15493	0.74	(0.71,0.78)	0.74	(0.70,0.77)	0.79	(0.70,0.89)
Above 4 Years	2772	0.35	(0.37,0.43)	0.39	(0.36,0.44)	0.39	(0.28,0.53)
Mothers age at birth							
Younger than 20 years	8439	1		1		1	
20–29 Years	18703	0.83	(0.79,0.88)	0.84	(0.80,0.89)	0.76	(0.66,0.89)
30–39 Years	7257	0.86	(0.80,0.94)	0.86	(0.80,0.95)	0.85	(0.66,1.08)
40–49 Years	623	0.99	(0.84,1.18)	1.00	(0.83,1.20)	0.61	(0.33,1.12)
Malaria Endemicity							
0.0–0.30	9141	1		1		1	
0.31–0.45	9113	0.68	(0.58,0.80)	0.78	(0.59,1.06)	0.78	(0.51,1.11)
0.46–0.6	4160	0.61	(0.51,0.74)	0.89	(0.72,1.16)	0.83	(0.44,1.42)
0.61–1.0	12608	0.62	(0.53,0.72)	0.89	(0.79,1.03)	0.73	(0.56,1.11)
γ		0.544	(0.535,0.553)	0.544	(0.535,0.552)	0.560	(0.537,0.585)
σ^2_ϕ		0.07	(0.05,0.10)	0.08	(0.06,0.12)	0.06	(0.05,0.13)
δ_ϕ		0.57	(0.36,1.04)	0.52	(0.29,1.15)	0.55	(0.29,1.08)

Table 4.2: Posterior estimates of the hazard ratio for child survival in Mali. The frequencies for the malaria endemicity predictor are based on the median predicted values.

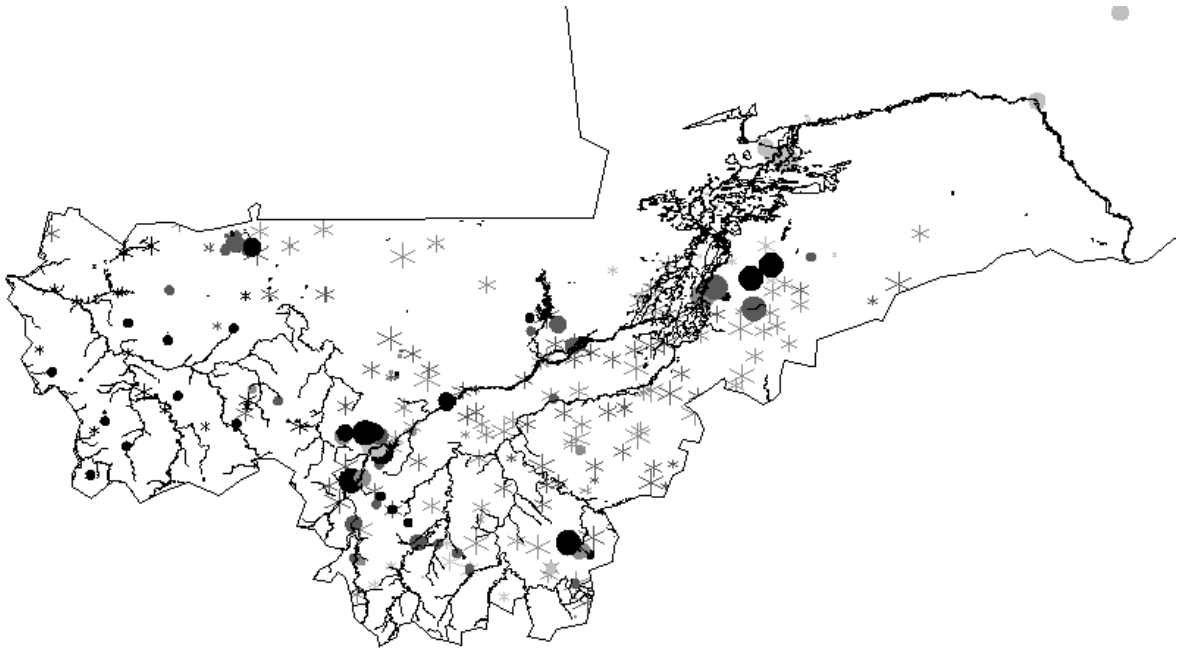
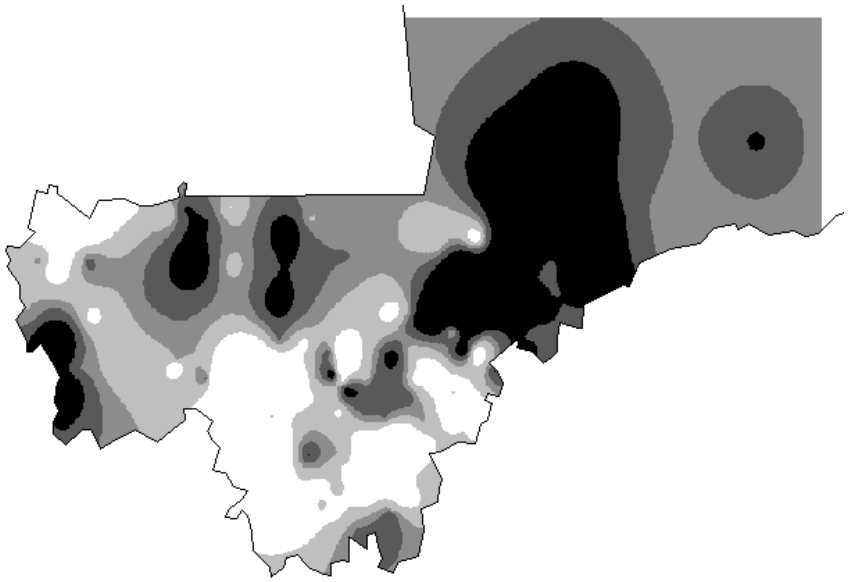


Figure 4.2: Malaria surveys were undertaken at the locations indicated by dots with dot-size proportional to the number of samples. The shading is drawn proportional to the observed prevalence. At the locations indicated by a star, the survival data are observed and the malaria risk is predicted via Bayesian kriging. The kriging variance is expressed by the stars size and its mean by the shading.

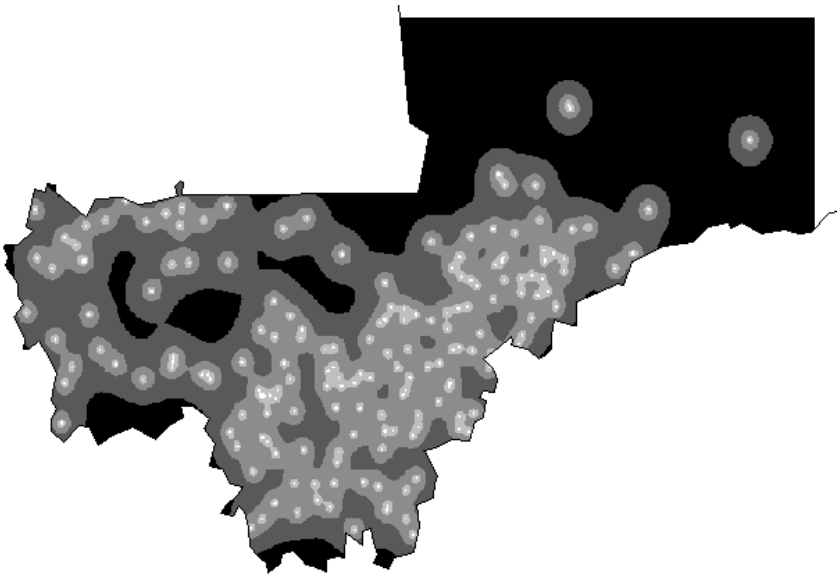
bridge the child mortality and the malaria risk data.

The malaria data has high spatial variation and low spatial interdependence, whereas the child survival data has low spatial variation and high spatial interdependence. The environmental predictors which are spatially related, were able to capture the spatial correlation present in the malaria risk data. The estimated hazard ratios for the fixed effects in the child survival model are close to those obtained from similar studies and reports (Ibrahim et al., 1996; Uchudi, 2001; Cleland and Ginneken, 1989). However the inclusion of spatial effects has reduced the statistical significance of the socio-economic predictors on child mortality. This is the first work which takes account of spatial correlation in analyzing these data and the standard error of estimates are assumed to be appropriately inflated by the introduction of spatial random effects.

The analysis appears to indicate that malaria endemicity is not related to child survival in Mali after adjusting for socio-economic, maternal factors and malaria exposure. The malaria risk appears to be higher in the more humid areas in the south which are relatively less poor and lower in the northern dry sub-saharan areas which are more affected by poverty due to the climatic conditions. Child survival also appears to be higher in Southern Mali. The wide low-survival region in the north-east is the flood plain of the river Niger, which is a preferred breeding sites of malaria mosquitoes due to many shallow temporal



a) Mean of prediction of spatial random effects from the survival model with categories - from white to black - of below -0.15, -0.15 to -0.05, -0.05 to 0, 0 to 0.05 and 0.05 to 0.15



b) Standard Deviance of prediction of spatial random effects from the survival model with categories - from white to black - of below 0.002, 0.002 to 0.01, 0.01 to 0.02, 0.02 to 0.04 and 0.04 to 0.06

Figure 4.3: Distribution of spatial random effects of the child survival model.

ponds in a hot area and therefore the river brings a high risk in a poor area.

The lack of relationship may also reflect unmeasured local factors, such as variations in health provisions or availability of water supply in the dry Sahel region, which could have a stronger influence than malaria risk on child mortality. Information missing from the database regarding malaria control measures taken at different locations, could also confound the analysis. An important limitation in the analysis of malaria prevalence data is that those data are obtained from surveys carried out at different locations with non-standardized and overlapping age-groups and seasons. We are currently working on using malaria transmission models to convert observed prevalence data to other transmission indicators adjusted for age and seasonality.

Acknowledgements

We are grateful to the Macro International Inc. for providing the names of the villages of the DHS surveys. The authors would also like to thank the NOAA/NASA Pathfinder AVHRR Land Project (University of Maryland) and the Distributed Active Archive Center (Code 902.2) at the Goddard Space Flight Center, Greenbelt, MD 20771 for the production and distribution of these data, respectively. The work of the first author was supported by Swiss National Foundation grant Nr. 3200–057165.99.

CHAPTER 5

Malaria mapping using transmission models: application to survey data from Mali

Gemperli A.[1], Vounatsou P.[1], Sogoba N.[2] and Smith T.[1]

This paper was submitted to *American Journal of Epidemiology*.

[1] Swiss Tropical Institute, Basel
[2] Malaria Research and Training Center, Faculté de Médicine de Pharmacie et d'Otondo-Stomatologie, Université du Mali, Bamako

Abstract

Geographical mapping of the distribution of malaria is complicated by the limitations of the available data. The most widely available data are from prevalence surveys, but these are generally carried out at arbitrary locations and include non-standardized and overlapping age groups. To achieve comparability between different surveys, we propose the use of transmission models, in particular that of the Garki model, to convert heterogeneous age prevalence data to a common scale of estimated entomological inoculation rates (EIR), vectorial capacity, or force of infection. We have applied this approach to the analysis of survey data from Mali extracted from the mapping malaria risk in Africa (MARA) database. We use Bayesian geostatistical models to produce smoothed maps of the EIR estimates obtained from the Garki model allowing for the effect of environmental covariates. Using again the Garki model, we converted kriged EIR values to age-specific malaria prevalence. The approach makes more efficient use of the available data than do previous malaria mapping methods, and produces highly plausible maps of malaria distribution.

Keywords: entomological inoculation rate; kriging; malaria; Markov chain Monte Carlo; parasite prevalence.

5.1 Introduction

Reliable maps of the prevalence or transmission intensity of malaria are urgently needed, especially in endemic areas of sub-Saharan Africa. Such maps are fundamental for estimating the scale of the problem, and hence of the resources needed to combat malaria. They provide benchmarks for assessing the progress of control and indicate which geographical areas should be prioritized.

Malariological measures that might be mapped include categories of endemicity (e.g. unstable, mesoendemic, holoendemic); vector-based measures (vector densities, vectorial capacity, entomological inoculation rate (EIR); incidence of disease; or the force of infection. However, although malaria endemicity can vary widely over only short distances, most of these measures have been studied only in a few widely separated localities, and in general the measurements available from distinct sites differ. The most widely available malariological measures are point prevalence data, assessed by microscopy. Estimates of malaria prevalence at unsampled locations can be made by incorporating information from environmental covariates (Hay et al., 2000). The precision of such estimates can be further improved by using spatial smoothing or geostatistical methods (Diggle et al., 2002; Kleinschmidt et al., 2000, 2001a,b).

Spatial statistical models have already made substantial contributions to the modelling of malaria risk (Diggle et al., 2002; Kleinschmidt et al., 2000, 2001a,b; Ribeiro et al., 1996; Thomas and Lindsay, 2000), and Bayesian geostatistical methods have demonstrated their value for this application in the work of Diggle et al. (2002) (see also Thomson et al., 1999) for mapping childhood malaria risk in the Gambia and in Gemperli et al. (2004) for relating infant mortality to malaria risk. Spatial statistical models have also been used to produce

malaria maps of the whole of West Africa (Kleinschmidt et al., 2001a) and specifically of Mali (Kleinschmidt et al., 2000). All these analysis modelled directly the prevalence data without taking into account age-dependence of the malaria risk.

Malaria prevalence data are usually reported by age group, but with different age-groupings used in different series of surveys. Direct mapping of age-prevalence data therefore involves choosing a target age-group (with some flexibility in the choice of age-category boundaries), and discarding data for other age-groups and for sites where data for the target age-group are not available.

We propose to replace this subjective and inefficient procedure by using a mathematical model to convert a set of heterogeneous malariological indices onto a common scale for mapping purposes. Mathematical models, such as that of the Garki project (Dietz et al., 1974), can be used to predict the relationships between different measures of malaria transmission and endemicity and the shape of the age-prevalence relationship. Statistical fitting of the Garki model can therefore be used to obtain estimates of any malariological parameter predicted by the model as a function of whatever community-based malariological data are available for a site. We recently used this approach to obtain an interval estimate of the EIR on the island of Príncipe from the age-prevalence curve for *Plasmodium falciparum* malaria (Hagmann et al., 2003).

We have now applied this approach to an assemblage of age-prevalence data from Mali. The data were extracted from the Mapping Malaria Risk in Africa (MARA/ARMA, 1998) database, the most comprehensive database on malaria in Africa, containing survey data since early sixties. Using the Garki model, we translated the raw prevalence data from each MARA survey into an (interval) estimate of the EIR. We then followed Bayesian geostatistical methods to generate smoothed maps of the EIR, allowing for the effects of environmental covariates.

We have used estimates from the fitted model to produce smooth EIR maps for Mali via Bayesian kriging. Using again the Garki model we have also converted the kriged EIR values to estimates of malaria prevalence in children under-5 years of age and in children 2–10 years of age and produced maps of these parameters.

5.2 Methods and materials

5.2.1 Data sources

Data on malaria prevalence were extracted from the MARA/ARMA database (MARA/ARMA, 1998). This is the most comprehensive database compiled by an international collaboration initiated to provide a database and an atlas of malaria in Africa by collating both published and unpublished results of malariological surveys since 1965. We selected data from 164 surveys carried out in 147 locations in Mali between 1965 and 1998 covering various ranges of age groups (table 5.1). Entomological inoculation rates (EIR) were estimated by fitting the Garki model on the malaria survey data. A geostatistical model was fitted to EIR estimates to produce smooth maps of malaria transmission.

The maps were adjusted for environmental and climatic covariates. In our analysis, we considered the same covariates used by Kleinschmidt et al. (2000) to produce smooth maps of malaria prevalence fitted directly to MARA prevalence data, that is the average maximum temperature from March to May, the length of rainy season defined as the number of month with more than 60mm rainfall, the distance from the nearest water source, and the normalized difference vegetation index (NDVI).

Age categories	Number of Surveys	Number of blood slides	Positives (Malaria Prevalence)
2 to 9	52	2787	1842 (66.1%)
5 to 9 & 10 to 14	15	3176	1786 (56.2%)
0 to 44	12	2488	230 (9.2%)
0 to 1 & 2 to 4 & 5 to 9	12	8842	4084 (46.2%)
0 to 1 & 2 to 4 & 5 to 9 & 10 to 15	11	2722	1528 (56.1%)
1 to 2 & 3 to 5 & 6 to 10	10	8284	3883 (46.9%)
0 to 1 & 2 to 4 & 5 to 9 & 10 to 14	9	3616	912 (25.2%)
5 to 9	9	1435	468 (32.6%)
0 to 12	6	360	160 (44.4%)
1 to 15	4	2715	736 (27.1%)
6 to 14	4	923	108 (11.7%)
0 to 15	3	800	481 (60.1%)
2 to 15 & 16 to 70	3	582	207 (35.6%)
0 to 1 & 2 to 9	2	129	66 (51.2%)
0 to 5 & 6 to 10	2	279	121 (43.4%)
1 to 9	2	712	93 (13.1%)
2 to 9 & 10 to 10	2	346	215 (62.1%)
8 to 14 & 15 to 19 & 20 to 29 & 30 to 39 & 40 to 49 & 50 to 59	2	2023	1063 (52.5%)
0 to 1 & 2 to 4 & 5 to 9 & 10 to 14 & 15 to 19	1	110	72 (65.5%)
1 to 4 & 5 to 9 & 10 to 14 & 15 to 24 & 25 to 34 & 35 to 44 & 45 to 54 & 55 to 64	1	476	308 (64.7%)
2 to 9 & 10 to 60	1	251	124 (49.4%)
6 to 9	1	300	77 (25.7%)

Table 5.1: Age range of the MARA surveys.

Data on temperature and length of rainy season were obtained from the "Topographic and Climate Data Base for Africa" Version 1.1 by Hutchinson et al. (1996). The database includes spatial estimates of monthly values averaged over years for the whole continent of Africa at a resolution of 0.05 degrees of longitude and latitude. The base data is collected from diverse research agencies and contains measurements between 1920 to 1980, averaged for at least five years. Daily maximum temperature is recorded at 1,499 stations and rainfall at 6,051 stations in Africa. The predictions are created using thin-plate splines (Hutchinson, 1991), where the standard errors are reported to lie below 0.5 degrees centigrade for the temperature and between 5 and 15 percent for the rainfall data.

The NDVI values were extracted from the NOAA/NASA Pathfinder AVHRR Land Project database (Agbu and James, 1994) which records daily observed emitted and reflected radiations in different channels of the electromagnetic spectrum, sent by a satellite

on a spatial resolution of 8 kilometers. To reduce distortion due to clouds and atmospheric contaminants, we used as a composite measure the maximum value over ten days since clouds reduce the reported NDVI value. The NDVI is derived from the reflectance rate of two (visual and near-infrared) channels. It is shown to be highly correlated with vegetation parameters (Justice et al., 1985) and used as a proxy for. Its values range from -1 to 1 with negative values standing for no-vegetation, which is not found in our study. In contrast to the other predictors used, the NDVI was able to express temporary variability.

5.2.2 Statistical analysis

Fitting the Garki model

We used the mathematical model of the Garki project to convert the observed prevalence at each location to estimated EIR values. The model comprises a set of linked difference equations describing transitions among seven categories of host distinguished by their infection and immunological status. This can be used to make predictions of the age-specific prevalence of *P. falciparum* in humans as a function of transmission measures, including the vectorial capacity, the entomological inoculation rate or the force of infection. To estimate the Y values of (EIR) from community parasitological survey data, the equilibrium age-prevalence curves for the Garki model were estimated for different values of the EIR, using a golden section search routine to locate the maximum likelihood estimate (Press et al., 1988). Asymptotic 95 percent confidence estimates were obtained by numerical estimation of the Fisher's information. Further details of the models used are reported in the appendix.

Spatial modelling of EIR

We assumed that the logarithmic transformed EIR estimates, Y_j, for location j are normally distributed, having a mean which is a function of the covariates $\boldsymbol{X}_j$. We model the spatial dependency by assuming that the covariance of the EIR values at two locations, say i and j decreases with their distance d_{ij}, that is $\Sigma_{ij} = \mathrm{Cov}(Y_i, Y_j) = \sigma^2 \exp(-d_{ij}/\rho)$ where σ^2 is the spatial variance and ρ is a parameter describing the degree of correlation decay. In addition, the variance of EIR at each location i is specified by $\Sigma_{ii} = \mathrm{Var}(Y_i) = \tau^2 + \sigma^2$ where τ^2 models the remaining non-spatial variation in EIR which is not explained by the covariates. Under the assumptions of second order stationarity the covariance matrix $\boldsymbol{\Sigma}$ determines the well known exploratory tool in geostatistics, the variogram. The τ^2 corresponds to the nugget parameter, the σ^2 estimates the partial sill and the ρ, is related to the range, that is the minimum distance that the spatial correlation is less than 5 percent which is 3ρ.

We choose Bayesian methods in model fit and prediction (kriging) because they allow estimation of the precision of model parameters and kriged EIR values without depending on asymptotic inference which can not be uniquely defined in the case of spatial data (Cressie, 1993, p. 350). We estimate the model parameters using Markov chain Monte

Variable	Median	5% Quantile	95% Quantile
Intercept	-13.46	-32.85	8.17
Distance to Water			
4 to 40km	-1.44	-2.18	-0.71
more than 40km	-1.07	-2.57	0.42
Duration of rainy season	0.31	-0.37	1.00
Temperature*	0.34	-0.20	0.82
NDVI†	9.73	-2.14	21.16
σ^2	3.341	2.733	4.153
$\rho^\ddagger$	9.795	2.170	28.163
τ^2	2.743	0.883	6.868

* Average maximum temperature March to May in degrees Celsius

† Normalized Difference Vegetation Index

‡ In distance units of latitude and longitude divided by 9

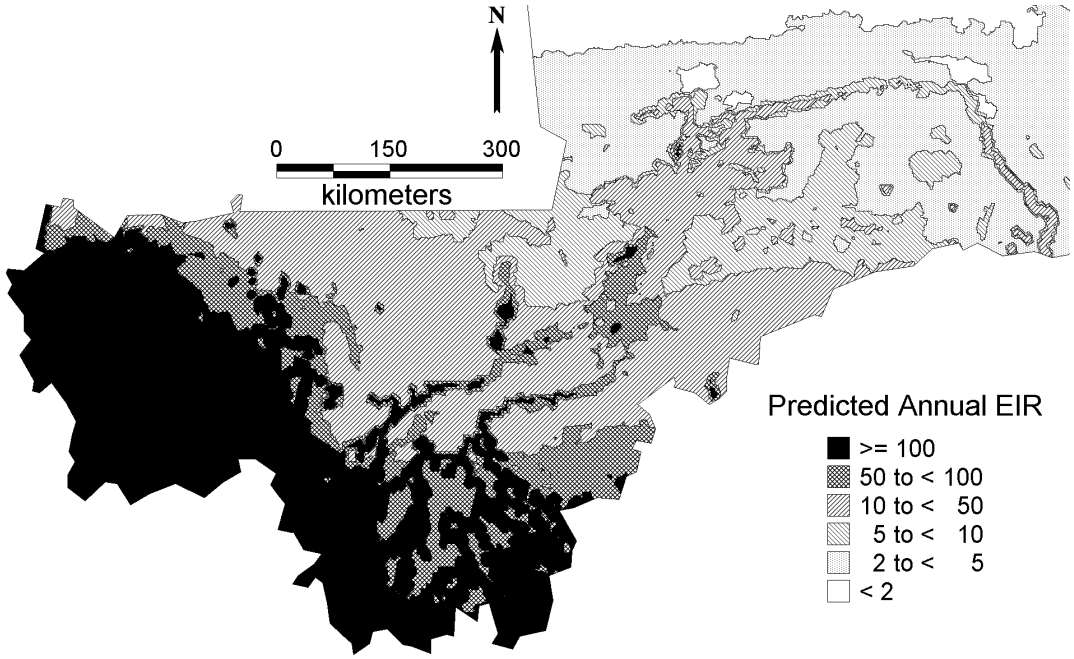
Table 5.2: Estimates of coefficients and covariance parameters in the regression model on the natural logarithm of the annual entomological inoculation rate EIR.

Carlo methods. Further details of this modelling approach are given in the appendix. The analysis was implemented using software written by the authors in Fortran 95 (Compaq Visual Fortran v6.6) using IMSL numerical libraries (Visual Numerics, Inc., Houston, Texas, USA).

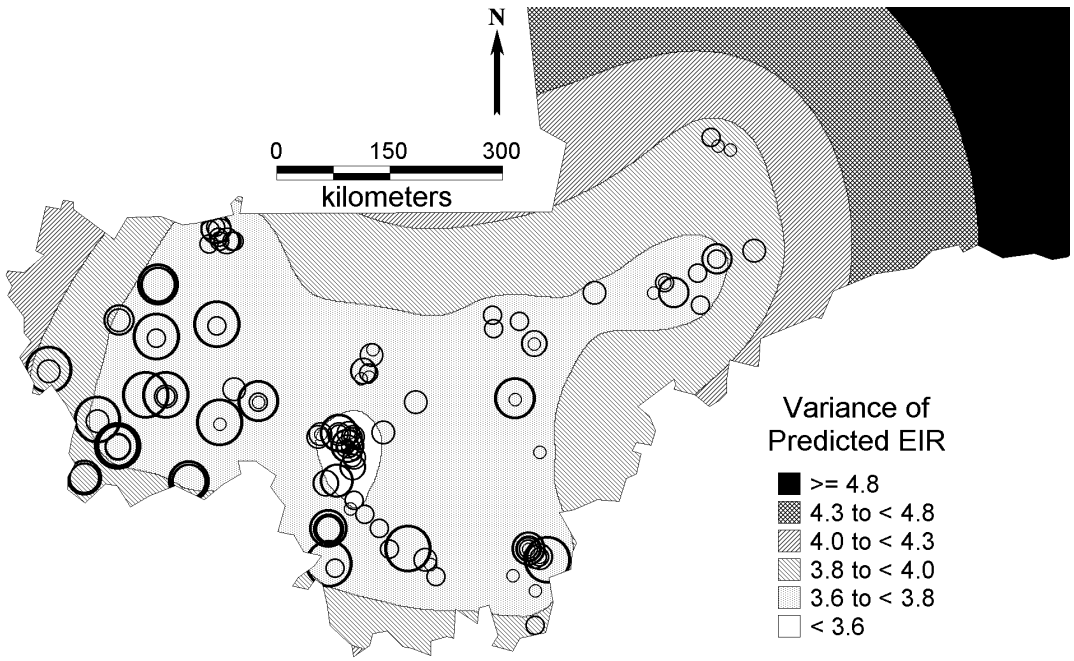
5.3 Results

Parameter estimates are summarized in table 5.2. The only environmental covariate significantly related to transmission intensity was the distance from the water, indicating high transmission in the areas within 4km away of the water source. The duration of rainy season, NDVI and temperature were not statistically significant related to the EIR. Estimates of the ρ suggest a strong spatial correlation reflected in the high median range distance, defined as the minimum distance between two points with correlation below 5 percent, of 356km (90 percent confidence interval: 78km, 1,023km). This corresponds to a median correlation of 65 percent for points 50 kilometers apart. In the survey data, 92.50 percent of all distances between pairs of locations are within the distance of the median range of 356km.

In addition the EIR show high variability estimated by τ^2. Maps of predicted EIR estimates are shown in figure 5.1. The map depicts a clear north-south and east-west pattern of transmission, ranging from disease free regions in the Saharan desert to high-prevalence areas in the south and west parts of Mali. The map is able to predict the high transmission areas along the Niger river, and in the Niger-delta which brings large



a) Mean Prediction



b) Prediction Error. Survey locations are indicated by circles with diameter proportional to the natural logarithm of the estimated annual EIR.

Figure 5.1: Spatial prediction of the entomological inoculation rate (EIR) in Mali.

water masses to otherwise low endemic areas. It can also identify distinct foci of high EIR around water sources. Estimates of the prediction error are shown in map 5.1b. The small prediction error in the regions around Bamako, Nioro and Mopti reflects the high density of surveys carried out in those regions. In contrast predictions in the north part of the country are not reliable because of the sparse malaria surveys in the Saharan desert.

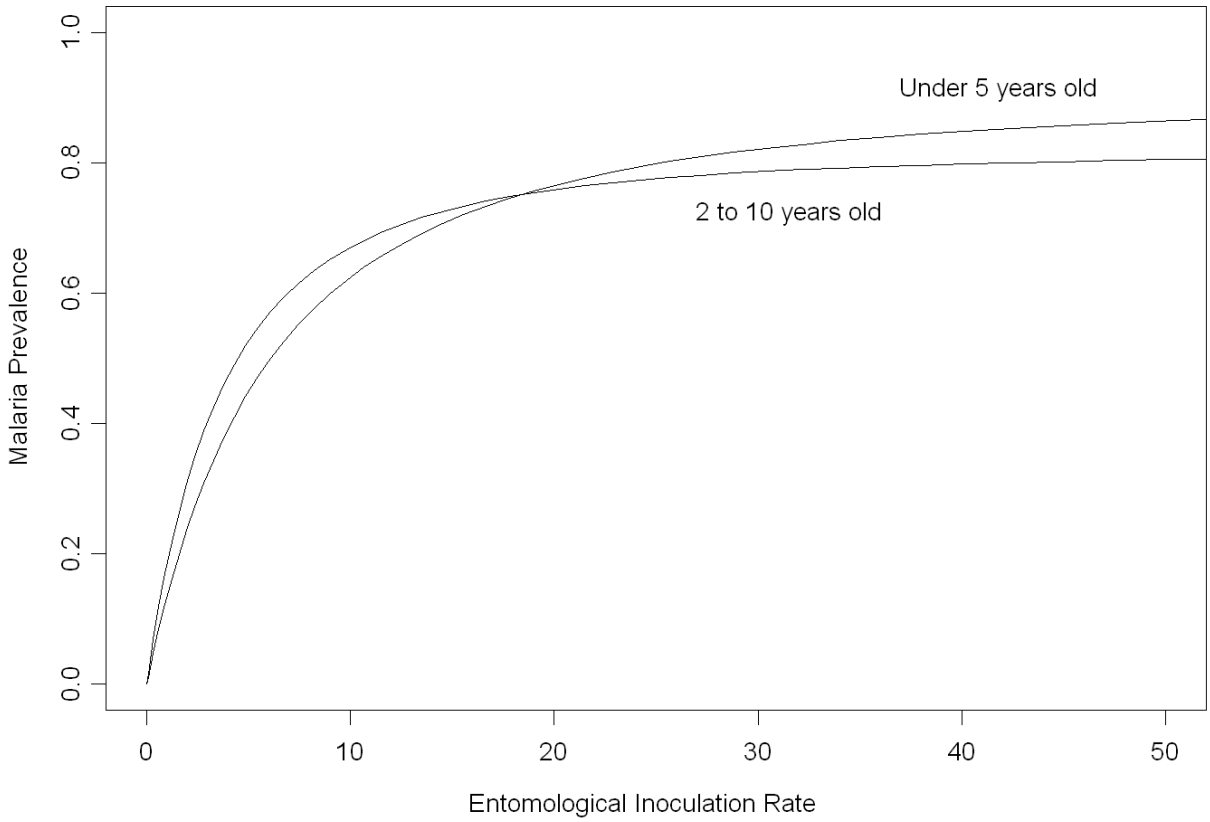
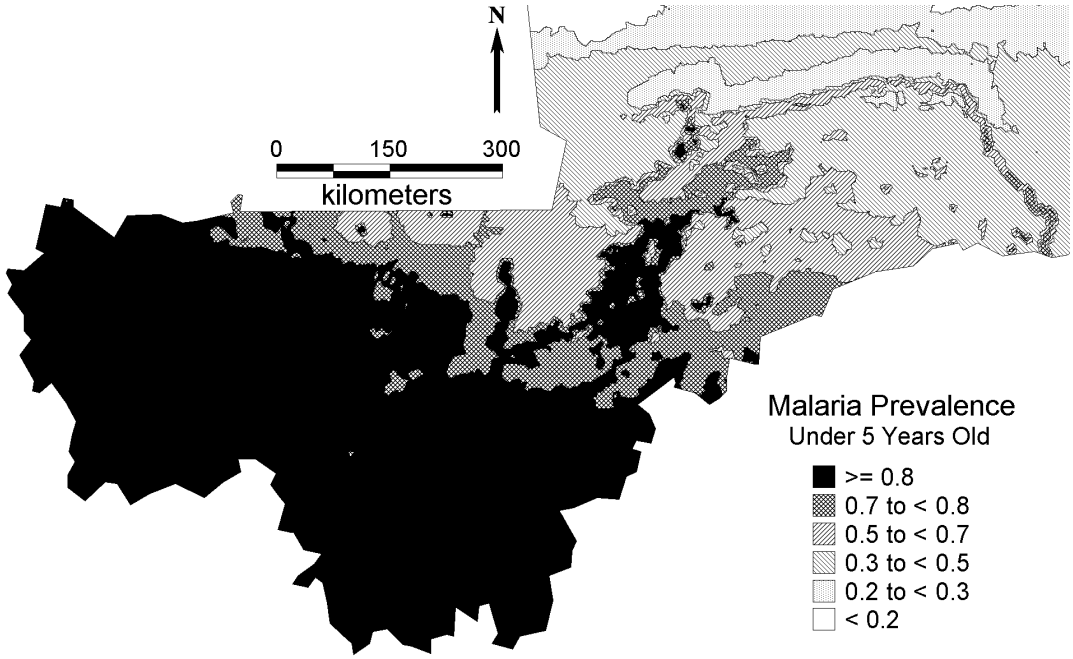
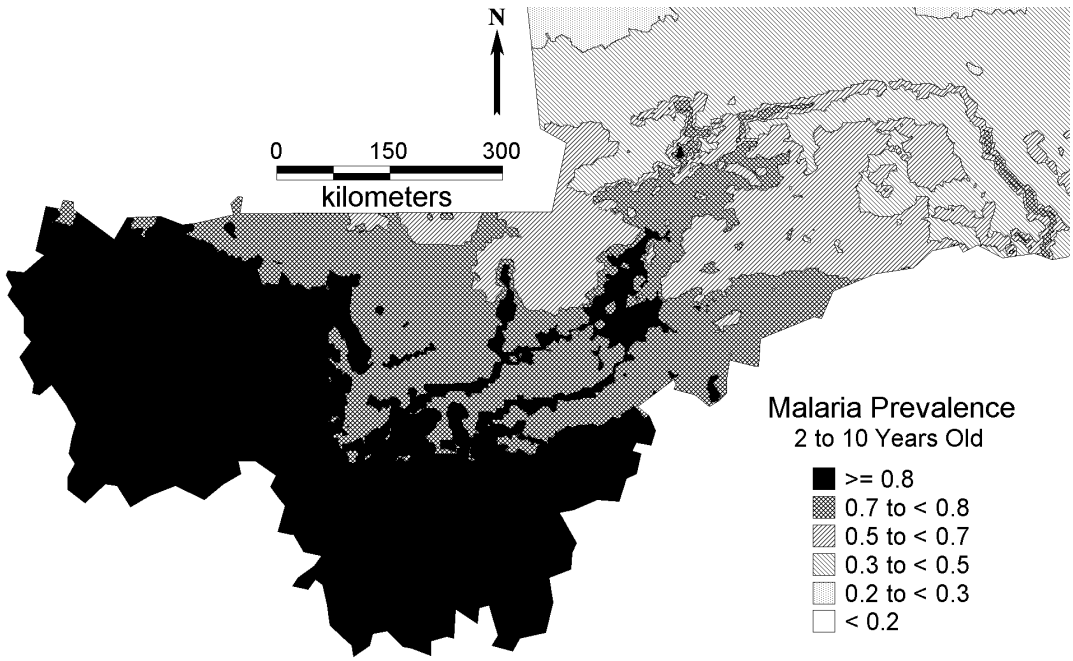


Figure 5.2: Relationship between malaria prevalence and entomological inoculation rate as estimated by the Garki model for two age categories (neglecting effects of seasonality).

Figure 5.2 displays the relation between the malaria prevalence and the transmission intensity which was estimated via fitting the Garki model to the malaria survey data. This model allows us to estimate this relation by age. We have chosen two age groups, those younger than 5 years old and those in the age range of 2 to 10 years old. The figure shows that at high levels of transmission, children younger than 5 years old tend to be at higher risk than older children. The opposite is observed at areas of low transmission. Maps of malaria prevalence for the two age groups of children are shown in figure 5.3. These maps were calculated by converting the EIR values to malaria prevalence using the



a) Under five years old



b) Two to ten years old

Figure 5.3: Spatial prediction of age specific malaria prevalence in Mali derived after transforming the predicted entomological inoculation rate.

EIR-prevalence relation of figure 5.2. The middle level of transmission in the districts of Koulikoro and Ségou in Central-West Mali contributes to higher malaria risk in the under 5 years old group than the 2 to 10 years old one. This corresponds to the change point in the prevalence-EIR relation for the chosen age categories at that level of transmission.

5.4 Discussion

The transmission model-based approach has the advantage, that it is age-adjusted and makes use of all the survey data available (as we did not have to discard any surveys because of inappropriate age groups). In principle this approach can exploit whatever population-based malariological data are available.

The transmission model-based approach also makes it possible to tailor the outputs of the mapping exercise to the specific needs of users, who may be interested in specific age-groups of hosts or predictions of malariological indices which are rarely measured in the field, as illustrated by our maps of malaria transmission intensity in Mali as well as of age-specific prevalence.

In contrast to earlier geostatistical malaria models, we fitted the Bayesian spatial model to the log transformed EIR values, a continuous outcome with an approximately normal distribution, estimated by fitting the Garki model to the available age-prevalence data. The Bayesian approach allows flexible model fitting and estimation and mapping of the prediction error. The method also allowed us to generate maps of the prediction error, demonstrating which geographical areas need further field investigation if the maps are to be uniformly reliable.

The maps we have produced broadly correspond to the known distribution of malaria in Mali and in particular indicate high transmission of malaria in the areas around the main rivers, the Niger and Sénégal. However, a specific difficulty

arises in modelling the relationship between malaria and distance to water bodies in West Africa. Undeniably the presence of water bodies and flooding in low-rainfall zones leads to malaria transmission in areas that would otherwise be malaria free and in general, mosquito numbers are highest near to water, especially areas prone to flooding. However recent studies in Niono in Mali (Dolo et al., 2000) have found that the highest malaria risk can be several kilometers away from the main Anopheline breeding sites. This may reflect a greater tendency of people exposed to very high mosquito densities to adopt protective measures, together with the lower average age of mosquitoes close to sites of emergence. In Mali, the center of the inland delta of the Niger is not considered to be zone of highest risk (Doumbo, 1992) and in the model of Kleinschmidt et al. (2000), the river system did not appear to strongly influence malaria distribution. Locations within 4km of the nearest water body were estimated to have lower risk than locations 4–40km from water. In the map of malaria across West Africa (Kleinschmidt et al., 2001a) broad zones of lower malaria risk close to rivers were estimated.

Most malaria surveys include people from areas of several square kilometers, so surveys close to water bodies may include some people from the riverbank and others from several

kilometers away. It is therefore not obvious what relationship with distance to water to expect. The exact relationship between proximity to rivers and malaria appears to be very sensitive to which datapoints are included and to the details of the model, especially when there are very few datapoints in the critical areas of the river flood plains. It may also be that the lack of adjustment for age in the earlier models biased some of the covariate effects. Because we were able to include data from 164 surveys rather than just the 101 analyzed by Kleinschmidt et al. (2000) we have some confidence that the present model provides an improved estimate of the broad geographical pattern of malaria, though possibly not of local variation within this.

We chose to use the Garki model to estimate EIR from age-prevalence curves because in its original development it was designed to accurately reproduce this relationship in field data from the savannah zone of Nigeria (Molineaux and Gramiccia, 1980). The Nigerian field site was in many ways similar to Southern Mali, and hence the model is likely to be most accurate for the range of conditions seen in our study. However the flexibility in the outputs is bought at the price of making many approximations. Like all mathematical models, the Garki model is a simplification. Some elements of it could probably be improved by using recent insights from molecular epidemiology studies and advances in statistical computation.

The main simplification inherent in our application to Malian data is that, following other exercises in empirical mapping of malaria (Kleinschmidt et al., 2000) we have ignored the seasonal patterns, although both the acquisition of the data, and the transmission of malaria itself were seasonal. Seasonality in transmission is an important consideration in the interpretation of the EIR map (figure 5.1a), because when many inoculations occur over a short period of time the proportion resulting in erythrocytic infections is reduced (Beier et al., 1994; Charlwood et al., 1998). Clustering of inoculations in the transmission season thus means that the average force of infection is lower than would result from the same number of entomological inoculations spread over the whole year. The Garki model does capture this phenomenon, but only if a seasonal input of vectorial capacity is assumed. Since, in the present analysis, we have assumed a constant vectorial capacity for each location the true EIR values in Mali must be higher than those we have estimated.

The prevalence maps (figure 5.3) are less affected by seasonality, because the transformation back to a scale of prevalence corrects the bias introduced by assuming uniform yearly transmission. Moreover, prevalence is known to show much less seasonality than does EIR (Molineaux and Gramiccia, 1980; Charlwood et al., 1998; Smith et al., 1993). For comparisons of malaria risk at regional level, and in zones where there is considerably variation in the degree of seasonality, it will, however, be essential to correctly allow for seasonality in the estimation of transmission parameters from age-prevalence data. In further developments of our model-based approach to malaria mapping we propose to use maps of seasonality in transmission as an input to the modelling procedure, in order to correct for the biases in EIR estimates.

Acknowledgements

The authors would like to thank the NOAA/NASA Pathfinder AVHRR Land Project (University of Maryland) and the Distributed Active Archive Center (Code 902.2) at the Goddard Space Flight Center, Greenbelt, MD 20771 for the production and distribution of these data, respectively. The work of the first author was supported by SNF grant Nr. 3200–057165.99. This work is a product of the MARA (Mapping Malaria Risk in Africa) collaboration, and the authors would also like to acknowledge the contributions of the many field, laboratory and office workers who carried out the surveys and compiled the malariological database.

Appendix 5.A The Garki model

The mathematical model of the Garki project makes predictions of the age-specific prevalence of *P. falciparum* in humans as a function of the vectorial capacity, C. It comprises a set of linked difference equations describing transitions among seven categories of host distinguished by their infection and immunological status. Two compartments, comprising proportions x_1 and x_3, of the population account for the uninfected individuals, two for those with prepatent infections (x_2 and x_4), and the remaining three, comprising proportions y_1, y_2, and y_3, represent those with blood-stage infections.

The model consists of an algorithm for predicting the proportion of human population at each age in each of these compartments and is defined by a set of difference equations (equations 1-7) that specify the change in each of these proportions from one time point to the next (i.e. $\Delta x_1 = x_1(t+1) - x_1(t)$).

$$\Delta x_1 = \delta + y_2 R_1(h) - (h+\delta)x_1 \tag{5.1}$$
$$\Delta x_2 = h x_1 - (1-\delta)^N + h(t-N)x_1(t-N) - \delta x_2 \tag{5.2}$$
$$\Delta x_3 = y_3 R_2(h) - (h+\delta)x_3 \tag{5.3}$$
$$\Delta x_4 = h x_3 - (1-\delta)^N + h(t-N)x_3(t-N) - \delta x_4 \tag{5.4}$$
$$\Delta y_1 = (1-\delta)^N + h(t-N)x_1(t-N) - (\alpha_1+\delta)y_1 \tag{5.5}$$
$$\Delta y_2 = \alpha_1 y_1 - (a_2 + R_1(h) + \delta)y_2 \tag{5.6}$$
$$\Delta y_3 = \alpha_2 y_2 - (1-\delta)^N + h(t-N)x_3(t-N) - (R_2(h)+\delta)y_3 \tag{5.7}$$

The meanings of the additional symbols are given in table 5.3. For simplicity the time points to which the proportions and the force of infection refer to are only indicated in the above equations when they differ from t.

To complete specification of the model, h, the force of infection, must be specified as a function of the vectorial capacity C. From the definition of C, it follows that each bite on an infective individual will result in C new inoculations, N days later (where N is the duration of sporogony). Since a proportion $y_1(t)$ of the population is infective; the entomological inoculation rate E is $E(t) = C(t-N)y_1(t-N)$.

To ensure that the model reproduces the observed saturation in the force of infection as E increases, $h(t)$ is assumed to be related to $E(t)$ via the equation $h(t) = g(1-\exp(-E(t)))$ where g then represents the upper limit of the force of infection, and is hence a parameter measuring host susceptibility. The problems of superinfection and acquired immunity are addressed by specifying the recovery rates, R_1, and R_2 as functions of h, using the relationship $R = h/\left(\exp\left(h/r\right) - 1\right)$ where r is the recovery rate for single clone infections. Non-immunes are assumed to recover at rate R_1, calculated from this equation by setting $r = r_1$. Immunes recover at rate R_2, calculated by setting $r = r_2$ where $r_2 > r_1$. In addition to this difference, the acquisition of immunity prevents transmission from the human host to the mosquito.

Symbol	Meaning	Default value
δ	Human birth and death rates	36.5 per 100-year
α_1	Rate at which non-immunes move into the non-infective category	0.002 per day
α_2	Rate at which non-immunes recovering from infection move into the immune category	0.00019 per day
h	Force of infection (Rate of infection of susceptibles)	to be estimated
N	Duration of pre-patent period	15 days
r_1	Recovery rate for individual clones (non-immune)	0.0023 per day
r_2	Recovery rate for individual clones (immune)	$10r_1$
$R_1(h)$	Recovery rate from infection in non-immunes y_2 (as a function of h)	to be estimated
$R_2(h)$	Recovery rate from infection in immunes y_3 (as a function of h)	to be estimated
g	Maximum value of force of infection	0.097 per 5 days
q_1	Detectability of parasites in infectives (y_1)	1
q_2	Detectability of parasites in non-immunes (y_2)	1
q_3	Detectability of parasites in immunes (y_3)	0.7

Table 5.3: Quantities appearing in the Garki model.

In order to examine the fit of the model to real parasite prevalence data, three detectability parameters, q_1, q_2, and q_3 are required to allow for imperfect detection of parasitaemia in each of the three infected classes (table 5.3). The overall observed prevalence is then $z(t) = q_1y_1(t) + q_2y_2(t) + q_3y_3(t)$. Predicting the age-specific parasite rate for any given vector $C(t)$ then involves a two-stage process. Initially the model (1)-(7) is simulated starting with arbitrary values of x_1 to x_4 and y_1 to y_3, until equilibrium is reached. The input, $C(t)$, may be constant, or may vary cyclically. Following the original implementation we use time intervals of five days. It is therefore natural to consider the input to follow a 365 day (73 time intervals) repeating cyclical pattern, and hence convergence is judged to have been achieved when each of x_1 to x_4 and y_1 to y_3 is equal to the value attained 73 time units previously. The life history of a cohort of individuals born into the non-immune susceptible category is then simulated by running the model with x_1 initialized to be 1, δ set to 0, and $C(t)$ set to the equilibrium values. To incorporate effects of the season of birth, a series of such cohorts are simulated with birthdates spread uniformly throughout the year.

Appendix 5.B The geostatistical model

Let Y_j be the logarithmic transformation of the EIR estimates, at location $j = 1, \ldots, n$. We assume that Y_j are normally distributed, having a mean which is a function of the covariates $\boldsymbol{X}_j$ at j. We model the spatial dependency by assuming that the covariance of the EIR values at two locations, say i and j, decreases with their distance d_{ij}, that is $\Sigma_{ij} = \text{Cov}(Y_i, Y_j) = \sigma^2 R_{ij}(\rho)$ with $R_{ij} = \exp(-d_{ij}/\rho)$ where σ^2 is the spatial variance and ρ is a parameter describing the degree of correlation decay. In addition, the variance of EIR at each location i is specified by $\Sigma_{ii} = \text{Var}(Y_i) = \tau^2 + \sigma^2$ where τ^2 models the remaining non-spatial variation in EIR which is not explained by the covariates. The likelihood function of the EIR is multivariate normal, that is $\boldsymbol{Y} \sim MVN(\boldsymbol{X}^t\boldsymbol{\beta}, \Sigma)$ where $\Sigma = \tau^2 I_n + \sigma^2 R(\rho)$ and I_n is the unity matrix of dimension n.

To complete Bayesian formulation of the model, we specify prior distributions for the model parameters, $\boldsymbol{\beta}, \rho, \sigma^2$ and τ^2. In particular, we adopt independent normal non-informative priors for the regression coefficients β_k, $\beta_k \sim \mathcal{N}(0, 10^6)$, inverse Gamma priors for the variance parameters, $\sigma^2 \sim IG(a_1, b_1)$ and $\tau^2 \sim IG(a_2, b_2)$ and a Gamma prior for the ρ parameter where $\rho \sim Ga(a_3, b_3)$ with hyper-priors $a_1 = a_2 = 2.01$, $b_1 = b_2 = 1.01$ and $a_3 = b_3 = 0.01$. Following the Bayesian paradigm, the full posterior distribution takes the form,

$$\left[\boldsymbol{\beta}, \rho, \sigma^2, \tau^2 \mid \boldsymbol{Y}\right] \propto det(\tau^2 I_n + \sigma^2 R(\rho))^{-1/2} \times$$

$$\exp\left(-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{\beta})^t(\tau^2 I_n + \sigma^2 R(\rho))^{-1}(\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{\beta})\right)\left[\boldsymbol{\beta}, \rho, \sigma^2, \tau^2\right].$$

We estimate the parameters of the model using Markov chain Monte Carlo (MCMC) and in particular Gibbs sampling (Gelfand and Smith, 1990). Implementation of the Gibbs sampler requires simulating from the conditional posterior distributions of all parameters.

The full conditional posterior distribution of $\boldsymbol{\beta}$ is a normal distribution and it is straightforward to simulate from. The conditional posterior distributions of σ^2, τ^2 and ρ have non-standard forms. We sampled from these distributions, by employing a random walk Metropolis algorithm having a Gaussian proposal density with mean equal the estimate from the previous iteration and variance derived from the inverse second derivative of the log-posterior.

To estimate the unobserved logarithm of EIR at a set of new locations $s_{01}, s_{02}, \ldots, s_{0l}$ we use Bayesian kriging. Let $\boldsymbol{Y}_0 = (Y(s_{01}), Y(s_{02}), \ldots, Y(s_{0l}))$ denote the values to predict. Then the predictive distribution

$$P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}) = \int P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}, \boldsymbol{\beta}, \sigma^2, \tau^2, \rho) P(\boldsymbol{\beta}, \sigma^2, \tau^2, \rho \mid \boldsymbol{Y})\, d\boldsymbol{\beta} d\tau^2 d\sigma^2 d\rho \tag{5.8}$$

is numerically approximated by the average $1/r \sum_{k=1}^{r} P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}, \tau^{2(k)}, \rho^{(k)})$. $\boldsymbol{\beta}^{(k)}$, $\sigma^{2(k)}$, $\tau^{2(k)}$ and $\rho^{(k)}$ are samples drawn from the posterior $P(\boldsymbol{\beta}, \sigma^2, \tau^2, \rho \mid \boldsymbol{Y})$ and $P(\boldsymbol{Y}_0 \mid \boldsymbol{Y}, \boldsymbol{\beta}, \sigma^2, \tau^2, \rho) = \mathcal{N}(\boldsymbol{X}_0^t\boldsymbol{\beta} + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{\beta}), \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10})$, when $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^t = \text{Cov}(\boldsymbol{Y}_0, \boldsymbol{Y})$, $\boldsymbol{\Sigma}_{11} = \text{Var}(\boldsymbol{Y})$ and $\boldsymbol{\Sigma}_{00} = \text{Var}(\boldsymbol{Y}_0)$.

CHAPTER 6

Mapping malaria transmission in West- and Central Africa

Interim Report on analyzes proposed at the ESHAW-2 malaria mapping workshop. Bamako, Mali, 2002.

Abstract

We have produced maps of malaria transmission in West and Central Africa from a database comprising all malaria prevalence data for these regions that we could geolocate (the Mapping Malaria Risk in Africa (MARA) Database). The malaria surveys were carried out at different seasons, and were reported using different age groupings of the human population. To allow for this we used the Garki malaria transmission model to convert the malaria prevalence data at each location to a single estimate of transmission intensity E, making use of a seasonality model based on Normalized Vegetation Index (NDVI), temperature and rainfall data. We fitted a Bayesian hierarchical variogram model to the E estimates, adjusting for environmental predictors extracted from remote sensing and applied Bayesian kriging to obtain smooth maps of malaria transmission intensity. The predicted E values were then for mapping purposes converted to age-specific estimates of malaria risk using again the Garki model. The resulting maps have been validated by expert opinion and confirmed known patterns of malaria transmission.

Keywords: entomological inoculation rate; kriging; malaria; Markov chain Monte Carlo; parasite prevalence; vectorial capacity.

6.1 Introduction

Plasmodium falciparum malaria is the most important parasitic disease of humans, with most of its burden of morbidity and mortality in Africa. A frequently quoted estimate of its impact is that there are around 1 million deaths and 220 million clinical episodes annually that are directly attributable to malaria in Sub-Saharan Africa (Snow et al., 1999). These figures are very uncertain however, since empirical maps of the distribution of malaria transmission and the numbers of affected individuals are not available for most of the African continent. Reliable maps of the geographical distribution of malaria are urgently needed for accurate estimation of disease burden, to identify geographical areas which should be prioritized in terms of resource allocations and for assessing the progress of intervention programs.

A number of malaria distribution maps are available for Africa based on climatic and other environmental predictors of malaria transmission (Craig et al., 1999; Snow et al., 1999; Rogers et al., 2002), however they make little or no use of the data of field surveys of malaria prevalence, which form much the largest body of relevant information. The Mapping Malaria Risk in Africa MARA/ARMA (1998) project was established in 1996 to provide estimates of the distribution of malaria in Africa. It is a collaborative network of key African scientists and institutions with the aim of providing an atlas of malaria for evidence-based and targeted malaria control in Africa. To date results of well over 10,000 malaria prevalence surveys have been collated from published and unpublished sources into a single, electronically accessible repository representing the most comprehensive database on malaria prevalence in Africa.

The MARA database has been used to produce malaria risk maps for Kenya (Snow et

al., 1998) and Mali (Kleinschmidt et al., 2000). A different data compilation has been used to construct an empirical malaria map for The Gambia (Diggle et al., 2002). Kleinschmidt et al. (2001a) used the MARA database to produce a regional map for the whole of sub-Saharan West Africa. These maps make use of both the prevalence data and relevant environmental data obtained from remote sensing and GIS databases. However there are a number of problems related to the limitations of using available data. In particular, compilations of prevalence data need to make use of data from surveys carried out at different seasons with non-standardized and overlapping age groups of the population. This constraint of the data makes it difficult to allow for seasonality and the age dependence of the malaria prevalence (Gemperli et al., 2003b). Most analysis of MARA data have chosen a target age-group and discarded data for other age-groups and for sites where data for the target age-group was not available. This usually results to waste of a large amount of data and thus estimates of malaria transmission for some geographical regions with sparse data are imprecise.

Mathematical models of malaria transmission provide an approach for converting a set of heterogeneous malariological indices onto a common scale for mapping purposes. The Garki model (Dietz et al., 1974) is a dynamic compartment model which considers basic characteristics of immunity to malaria and the dynamics of the interactions among humans, mosquitoes and malaria. Given entomological measures of transmission intensity as input, the model predicts age-specific prevalence. Conversely, it can be used to predict transmission from age-specific prevalence. Gemperli et al. (2003b) have used this model to convert the MARA prevalence data from Mali to a measure of entomological inoculation rates which in turn can be used for mapping purposes. However, that analysis treated malaria transmission as constant throughout the year; this leads to biases in the estimation of transmission rates as the length of transmission season varies between locations.

In this paper we produce age-specific maps of malaria risk for West and Central Africa, using an extension of the approach of Gemperli et al. (2003b) that allows for the seasonality in malaria transmission between locations. We base our estimates of seasonality on a seasonality map that makes use of temperature, rainfall and the Normalized Difference Vegetation Index (NDVI), based on an augmented version of the model of Tanser et al. (2000). Using both the seasonality map and the Garki model we estimate the transmission intensity (E) for each location from the age-specific malaria prevalence values. Then we fit a Bayesian geostatistical model on the E using as covariates a number of environmental and ecological variables obtained from Remote Sensing (RS) and Geographical Information Systems (GIS). We then produce smooth maps of E for the whole of West and Central Africa using Bayesian kriging. We back-transform this map to maps of age-specific malaria prevalence by re-applying the Garki model.

6.2 Methods and materials

6.2.1 Datasets

This analysis was based on datasets which were obtained from different sources and databases. Details on the data we used are given below.

Malaria data

The malaria prevalence data were extracted from the version of the MARA/ARMA (1998) database available in mid 2002. In addition, we included 2,760 datapoints which were extracted by literature research in MEDLINE. The augmented database contained 7,738 age-specific prevalence for West and Central Africa, collected during 2,371 surveys, carried out at 1,220 distinct locations. In this analysis, we only included surveys conducted in rural regions after the year 1950 and discarded data obtained from locations without transmission throughout the year. The final data set we analyzed was collected at 976 distinct locations over 1,846 surveys and comprised 294 different (overlapping) age categories (figure 6.1).

Climatic, environmental and population data

The temperature and rainfall data were obtained from the "Topographic and Climate Data Base for Africa" Version 1.1 by Hutchinson et al. (1996). The database reports predicted values based on thin-plate splines interpolation (Hutchinson, 1991) which was applied to data collected by various research agencies at 1,499 stations for temperature and at 6,051 stations for rainfall, between 1920 and 1980, averaged over at least five years. We calculated monthly estimates of temperature and rainfall by averaging over the years with available data.

The NDVI data were extracted from satellite information conducted by the NOAAA/NASA Pathfinder AVHRR Land Project (Agbu and James, 1994). This database records daily emitted and reflected radiations in five channels of different wavelengths of the electromagnetic spectrum at a spatial resolution of eight kilometers. The NDVI is calculated as the ratio of the contrast between the first two channels (0.58–0.68 and 0.73–1.10 micrometer wavelength). This ratio 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. In order to reduce distortion effects due to clouds and atmospheric contaminants, the maximum value for every month was considered. Monthly NDVI values for each location were derived by averaging the maximum monthly values for the eleven year period from 1985 to 1995.

Monthly estimates of soil water storage index (SWS) were obtained using the procedure given by Droogers et al. (2001). The SWS describes the amount of water that is stored in the soil within the plant's root zone. Data on population density was derived from the "African Population Database" (Deichman, 1996) and corresponds to the number of persons at a resolution of 3.7 by 4.8 square-kilometers.

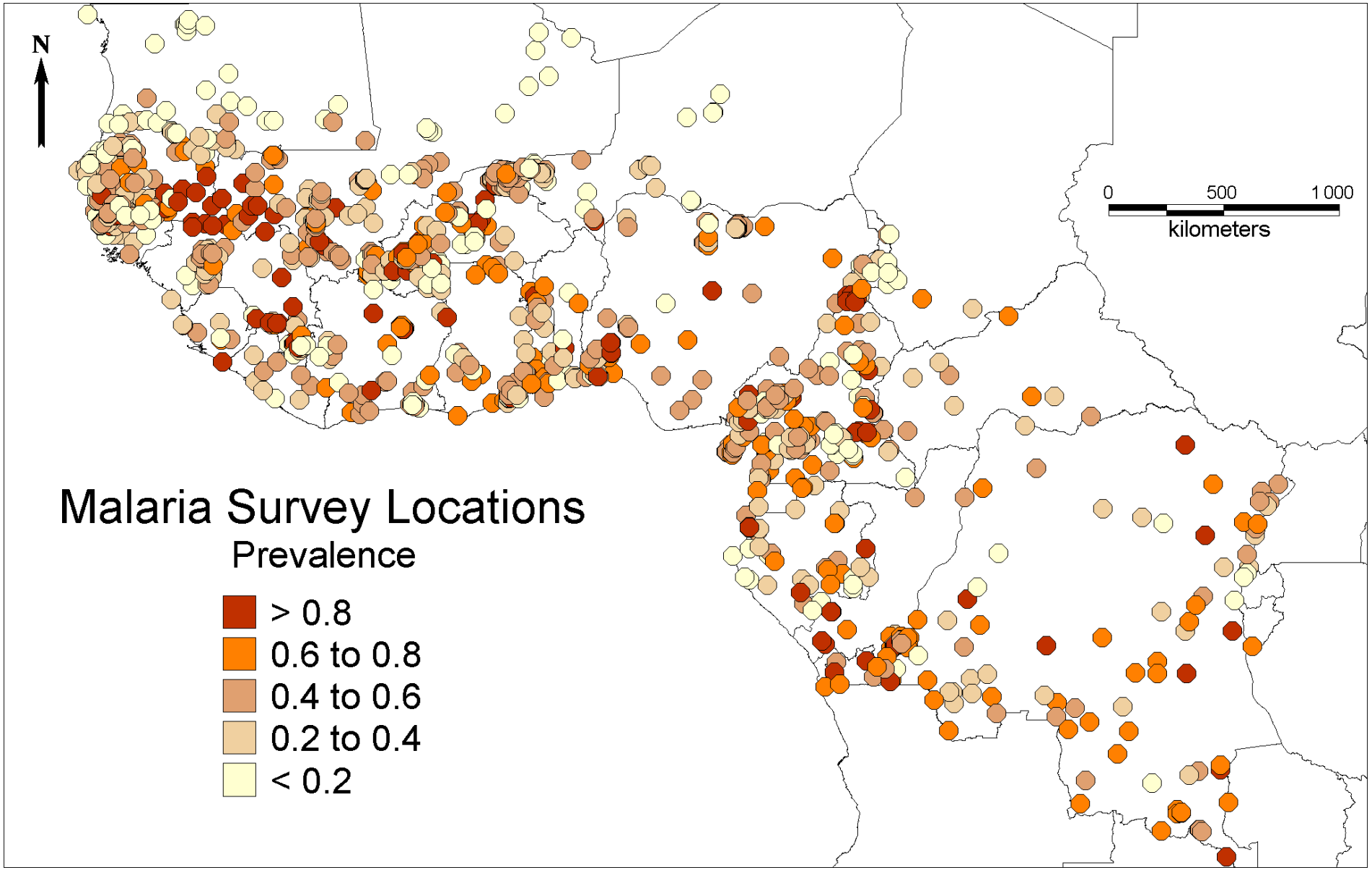


Figure 6.1: Sampling locations of the MARA surveys in West- and Central Africa.

The landuse classifier was extracted from the landuse/landcover database which is maintained by the United States Geological Survey and the NASA's distributed active archive center, that describes global land cover characteristics. We have chosen the 24 categories classification scheme which is described by Anderson et al. (1979) and regrouped it to 6 broad categories (water, very low transmission, under-average transmission, average transmission, higher-than average transmission and high transmission), using knowledge of the vector abundance in the different landuse types. Subsequently, the proportions of the six classes of landuse were calculated in a buffer area around the actual location. The size of the buffer was calculated by fitting models on the logarithmic transformed E values with various buffer size as predictors. It was found that the best fitted model arose for a buffer of 20×20 kilometers.

Factor	Resolution	Source
Temperature	$5km^2$	Hutchinson et al. (1996)
Rainfall	$5km^2$	Hutchinson et al. (1996)
NDVI	$8km^2$	NASA AVHRR Land data sets; (Agbu and James, 1994)
Landuse	$1km^2$	USGS-NASA;
Water bodies	$1km^2$	African Data Sampler; World Resources Institute (1995)
Soil Water Storage Index	$5km^2$	Droogers et al. (2001)
Agro-ecological Zone	Vector Coverage	FAO (1978)
Population Density	3.7×4.8 km^2	Deichman (1996)
Transmission Seasonality	$5km^2$	Calculated using criterions in table 6.2

Table 6.1: Spatial databases used in the analysis.

Permanent rivers and lakes were extracted from the "African Data Sampler" (World Resources Institute, 1995) while the nearest Euclidean distances of points on a grid of 1km resolution were calculated using the Idrisi software (Clark Labs, Clark University). Additionally to the "distance-to-water" variate, we estimated a "content-of-water" effect by calculating the proportion of water contained in a buffered area of 20×20 kilometers.

We divided the whole West and Central Africa in four agro-ecological zones (AEZ) which were determined as a function of precipitation, evaporation and availability of water stored in the ground, according to the procedure described in FAO (1978). A list of the environmental variables and the databases from which they were extracted is given in table 6.1.

For those environmental factors for which monthly values could be assigned (the minimum and maximum temperature, rainfall, the NDVI and the soil water storage index), summary statistics were calculated for each location for those months predicted by the seasonality map as having malaria transmission. The summary statistics computed were the total, the mean and the coefficient-of-variation.

Description	Climatic Effect	Rule
Frost	Minimum Annual Temperature	$> 5°C$
Vector Survival	Mean Monthly Temperature*	$> 19.5°C$ + annual standard deviation
Catalyst month	Annual Maximum Rainfall	> 80mm
Availability of breeding sites	NDVI‡ or Rainfall†	> 0.35 or > 60mm

∗: Average of minimum and maximum temperature. Moving average from two previous months and the current one.
†: Moving average from two previous months and the current one.
‡: NDVI value from preceding month.

Table 6.2: Criteria for suitability of stable *P. falciparum* malaria transmission. A month is suitable for transmission when all rules are fulfilled for the current month or for the immediate preceding and following months. The table extends the seasonality model by Tanser et al. (2000) by including the NDVI effect.

6.2.2 Seasonality model

The seasonality map of malaria transmission (figure 6.2) is an amended version of the map of Tanser et al. (2000). Tanser's original map makes use only of temperature and rainfall data to define suitability. In order to ensure that irrigated low rainfall areas were classified as suitable for transmission, we defined a region and month as suitable for stable malaria transmission when either it met the criteria set by Tanser et al. (2000) or when it met those criteria excluding the rainfall of 60mm but the NDVI values were higher than 0.35 (Hay et al., 1998). For each location and month we therefore calculated 1) the moving average over the current and the previous two months of the mean of minimum and maximum temperature 2) the moving average of the monthly temperatures of the current and previous two months and 3) the NDVI value of the previous month. In addition we calculated for each location the minimum and maximum annual temperatures. These criteria are presented in table 6.2.

6.2.3 Malaria transmission model

For each location, the raw MARA/ARMA prevalence data were converted to estimates of malaria transmission intensity (E) by fitting the Garki model (Dietz et al., 1974) using maximum likelihood (appendix 6.A). The Garki model is a dynamic compartmental model adjusted to field data from Northern Nigeria. It translates the age-dependence in the relation between malaria transmission and malaria prevalence into a set of curves. Each curve corresponds to a specific age and length of transmission season. Given entomological measures of transmission intensity, the model predicts age-specific prevalence. Conversely, it can be used to predict transmission from age-specific prevalence. E is a measure of entomological inoculation rate and it can be biased especially for large values of prevalence.

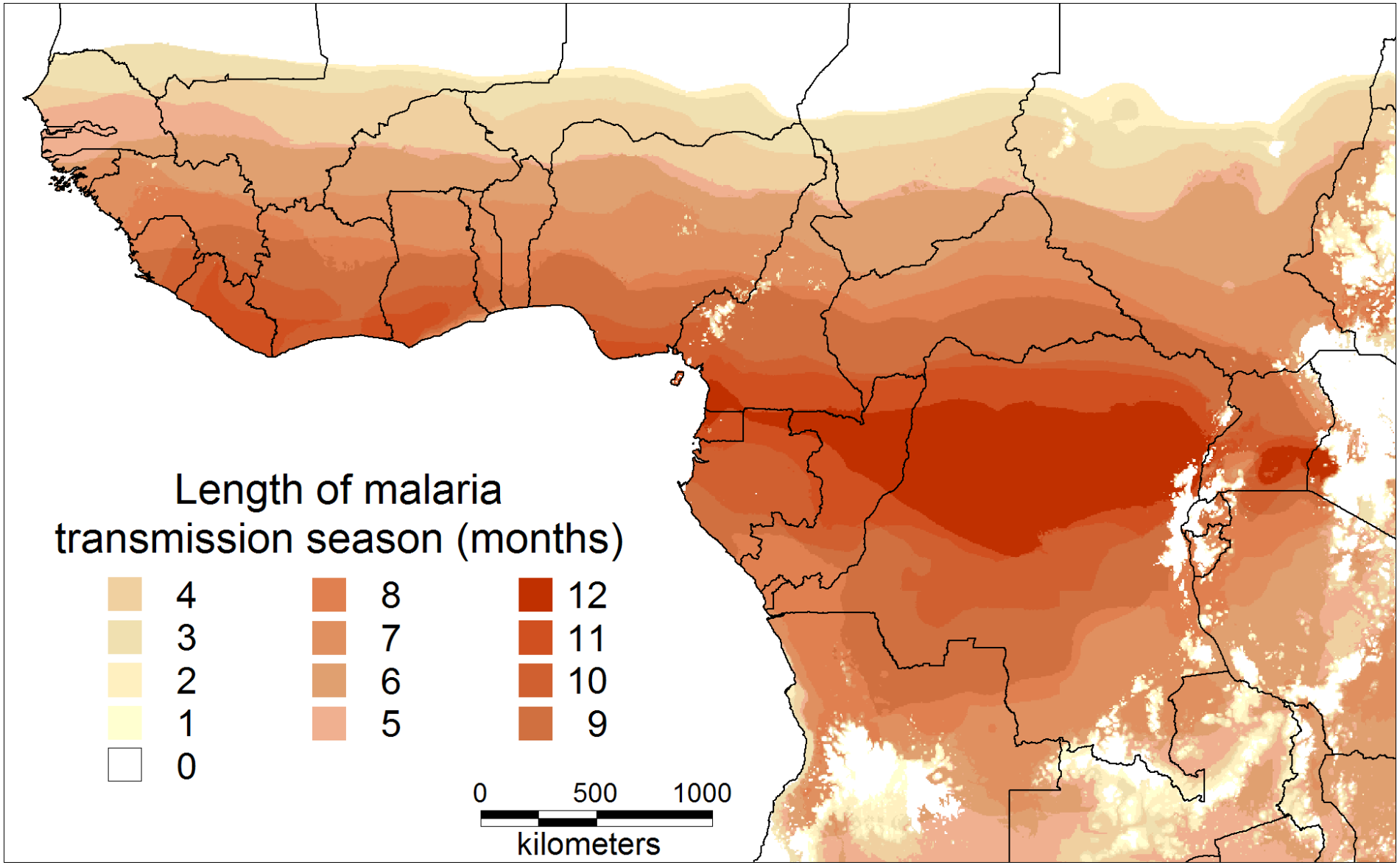


Figure 6.2: Map of the length of stable malaria transmission in West- and Central Africa.

This is because the prevalence curve has an upper bound and observed prevalence above this bound are subject to error when converted to entomological inoculation rate. The standard error of the E point estimates were obtained by calculating numerically, the Fisher's information.

6.2.4 Geostatistical model

A Bayesian linear geostatistical model was fitted on the E values taking into account a number of environmental predictors. In particular, the logarithm of the point estimate of E was assumed to be normally distributed, with mean being a non-linear function of the covariates. The spatial dependency among the log-E values Y_j for locations $j = 1, \ldots, m$ was modelled using the exponential correlation function $\mathrm{Cov}(Y_j, Y_k) = \sigma^2 \exp(-d_{jk}/\rho)$ for $j \neq k$ and $\mathrm{Var}(Y_j) = \sigma^2 + \tau^2\omega_j$, where d_{jk} is the Euclidean distance between the locations of observation Y_j and Y_k. ω_j is a weight introduced to account for uncertainty in estimates derived from the Garki model and equal to the reciprocal of the variance of the estimated log-E. The parameter σ^2 captures the variation attributable to spatial dependency and τ^2 the remaining variation. The decay of spatial variation as a function of the distance between sample points is expressed by the parameter ρ. Markov chain Monte Carlo was applied for model fitting. Bayesian kriging was employed to produce a smooth map of the E in West and Central Africa. The smoothed E map was back-transformed to age-specific maps of malaria risk in children using the Garki model. Details on the spatial Bayesian model and kriging are given in appendix 6.B.

Before fitting the spatial model, a number of possible predictors of E such as NDVI, rainfall, minimum/maximum temperature, soil water storage index, distance from nearest water source, population density, proportion of water, agro-ecological zone, year of survey and length of transmission season were screened univariately to select those which were statistically significantly related to E. Some of these covariates were used in earlier spatial malaria risk models by Kleinschmidt et al. (2000, 2001a); Thomson et al. (1999) and Diggle et al. (2002), however the proportion of water, land-use classifier, soil water storage index and the climatic suitability indicator were not considered in previous models.

We fitted various multiple, non-spatial models to identify the best subset of predictors and their best (possibly non-linear) functional form based on the bias-corrected Akaike's information criterion AICC (Hurvich and Tsai, 1989) which was used to assess model fit. The functional forms of predictors which we screened include polynomials up to second order, interaction terms of first order, logarithmic, inverse and exponential forms with different parameterizations and combinations of those. Only one parameter was found to enter the best model non-linearly. For ease of application, this parameter was fixed at its optimal estimate to end up with a purely linear model.

The non-spatial analysis of the E was carried out in The SAS System (SAS Institute, Cary, NC). The software used for fitting the Bayesian model was written by the authors in Fortran 95 (Compaq Visual Fortran v6.6) using using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.).

6.3 Results

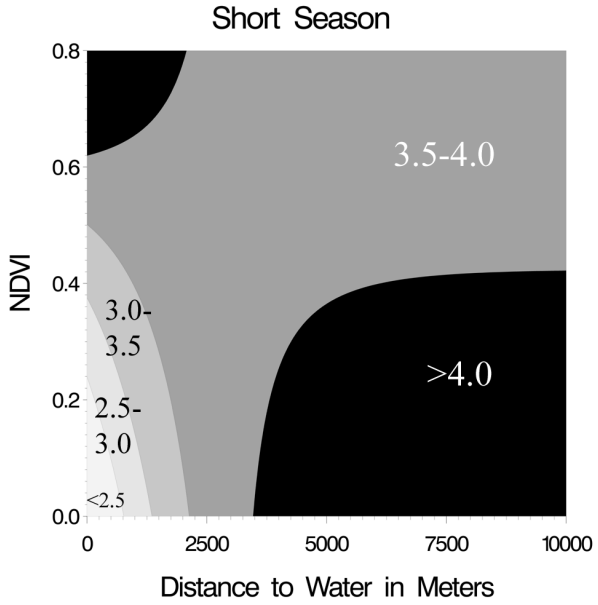
The univariate non-spatial analysis indicated that the following environmental factors were related to the E: year of survey, NDVI, distance from water, length of season, rainfall, soil water storage index, agro-ecological zone, minimum and maximum temperature. In section 6.2.1, we mentioned that temporal variables such as NDVI, rainfall and temperature whose values change from month to month were summarized for each location by total, mean, and coefficient of variation (CV) over the months with stable transmission during the year. Univariate analysis revealed that the mean leads to a better model fit than the total and the CV. No statistically significant univariate relation was found between the logarithm of E and either the land-use or population density.

Variable	Median	95% Confidence Interval
Intercept	1.296	(0.970, 1.637)
Year*	0.048	(0.0056, 0.0908)
Log(NDVI)†×Water proximity	2.107	(0.406, 3.821)
Water proximity	-0.875	(-1.477, -0.278)
Log(Length of Season)× Log(NDVI)†	-0.381	(-0.753, -0.0155)
1/Rainfall*	0.0538	(0.0051,0.1031)
Maximum Temperature*	0.072	(-0.0866, 0.236)
Maximum Temperature*× Log(Length of Season))	-0.048	(-0.150, 0.047)
Maximum Temperature^{2*}	-0.093	(-0.152, -0.040)
τ^2	41.98	(38.475, 47.804)
σ^2	0.398	(0.310, 0.495)
ρ	0.294	(0.156, 0.472)

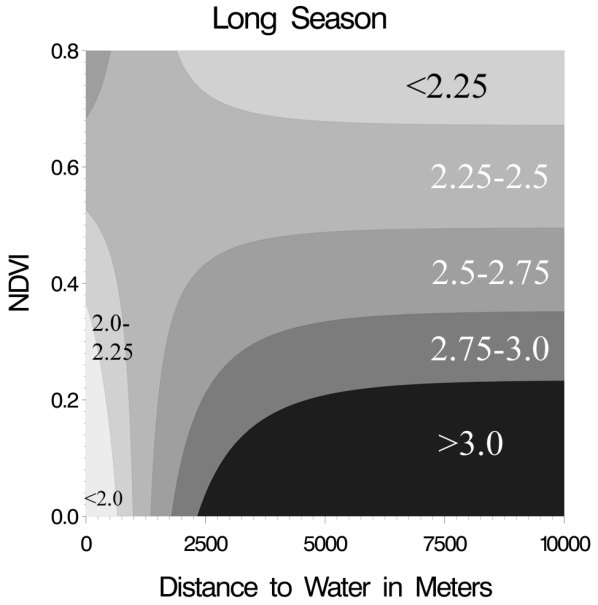
$*$: Standardized variables.
† : Shifted to be strictly positive.
Water proximity = exp(−Distance to the closest waterbody in meters/1500).
The predictors NDVI, rainfall and maximum temperature are the annual mean-values over those months estimated to be suitable for stable malaria transmission. The NDVI is increased by one, prior to taking the logarithm.

Table 6.3: Parameter estimates for the environmental covariates.

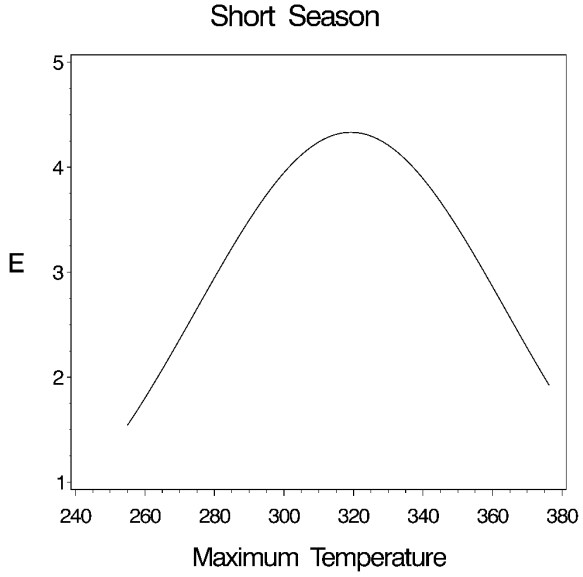
The best fitted model included NDVI and length of season on a logarithmic scale. The distance to water entered the model scaled as an exponential function. The scaling factor was chosen to optimize model fit. The relation with rainfall was best described by a reciprocal transformation. The parameter estimates obtained after fitting the spatial Bayesian model are presented in table 6.3. The results indicate an increase of 0.049 in the log-E every 14 years. Rainfall was also associated with transmission. Particularly in that higher amounts of rain lead to higher transmission intensities. The minimum temperature,



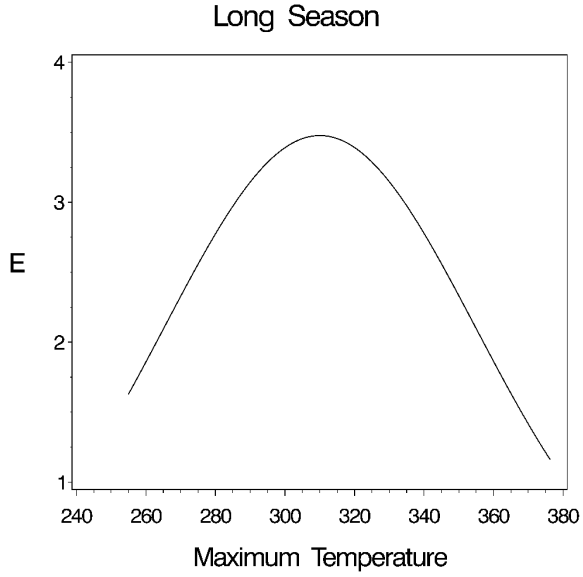
a) NDVI and Distance to Water



b) NDVI and Distance to Water



c) Maximum monthly temperature (in degrees decigrades)



d) Maximum monthly temperature (in degrees decigrades)

Figure 6.3: Contour plots indicating the effect of environmental factors on E estimated from the spatial Bayesian model. Short season corresponds to two months malaria transmission per year. Long season indicates perennial transmission.

agro-ecological zone and the soil water storage index were not retained in the multivariate model.

The interactions in the model capture the differences in the effects of environmental factors on E in the climatic zones. Some of these interactions which were estimated by the model are graphically depicted in figure 6.3. The higher the NDVI values, the higher the transmission except at locations far away from water and with perennial (long) malaria transmission (figure 6.3b). The distance to water is negatively associated with transmission for regions with high NDVI above 0.6 (figures 6.3a–6.3b). Malaria transmission increases as the maximum monthly temperature increases. It reaches a peak at around 32 degrees Celsius and then it reduces with higher temperatures (figures 6.3c and 6.3d). The above relation is not statistically significantly associated with the length of transmission season (table 6.3), in a spatially adjusted model, but the length of transmission season is significantly negatively associated to log-E in interaction with the NDVI.

The spatial correlation present in the data is measured by the parameter ρ which corresponds to the minimum distance between locations with correlation below 5 percent. This distance is estimated to be 87 kilometers (95 percent confidence interval: 45km, 141km). This large value probably arises because of large scale spatial effects due to unobserved ecological factors. The spatial correlation for locations 3 kilometers apart (mosquito flight range) is 90 percent and decreases to 81 percent for locations 6 kilometers apart. The spatial variation is very small ($\sigma^2 = 0.388$) compared to the residual non-spatial variation ($\tau^2 = 42.02$).

We were not able to fit the Garki model to data for locations where there is no single month of stable malaria transmission. In our data, we had 42 such locations in southern-Sahara regions, mainly in Mauritania (figure 6.1). The 69 surveys carried out at these locations were omitted from the analysis, which implicitly assumes that malaria is epidemic at these locations. The raw prevalence at 6 of these locations was zero and at 28 of these it was low (below 0.1). The recorded prevalence at 12 of these locations was between 0.1 and 0.25, and two sites in southern Mauritania close to the river Sénégal had prevalence values of 0.39 and 0.58.

The map with spatially predicted E values was converted to age-specific prevalence maps using the relationships assumed in the Garki model. The relation between the malaria prevalence and the transmission intensity for different lengths of transmission season and for two age groups (younger than five years old and 1 to 10 years old) is shown in figure 6.6. At high levels of transmission, children younger than five years old tend to be at higher risk than children 1 to 10 years old. The opposite is observed at areas of low transmission. In addition, as the length of transmission season increases the prevalence increases for areas with the same estimate of E.

The map of log-E for West- and Central Africa (figure 6.4) shows high transmission for most of sub-Saharan West Africa. The lowest transmission in that part of the continent was observed in the north-west of the Ivory Coast, the province of Sissili and most of the east part of the Poni province in Burkina Faso, the south-east region of Borgon in Benin, the south and central-east of Cameroon and the north of the Plateau region of Nigeria. Additionally there are large areas along the Atlantic Ocean estimated to have relatively low

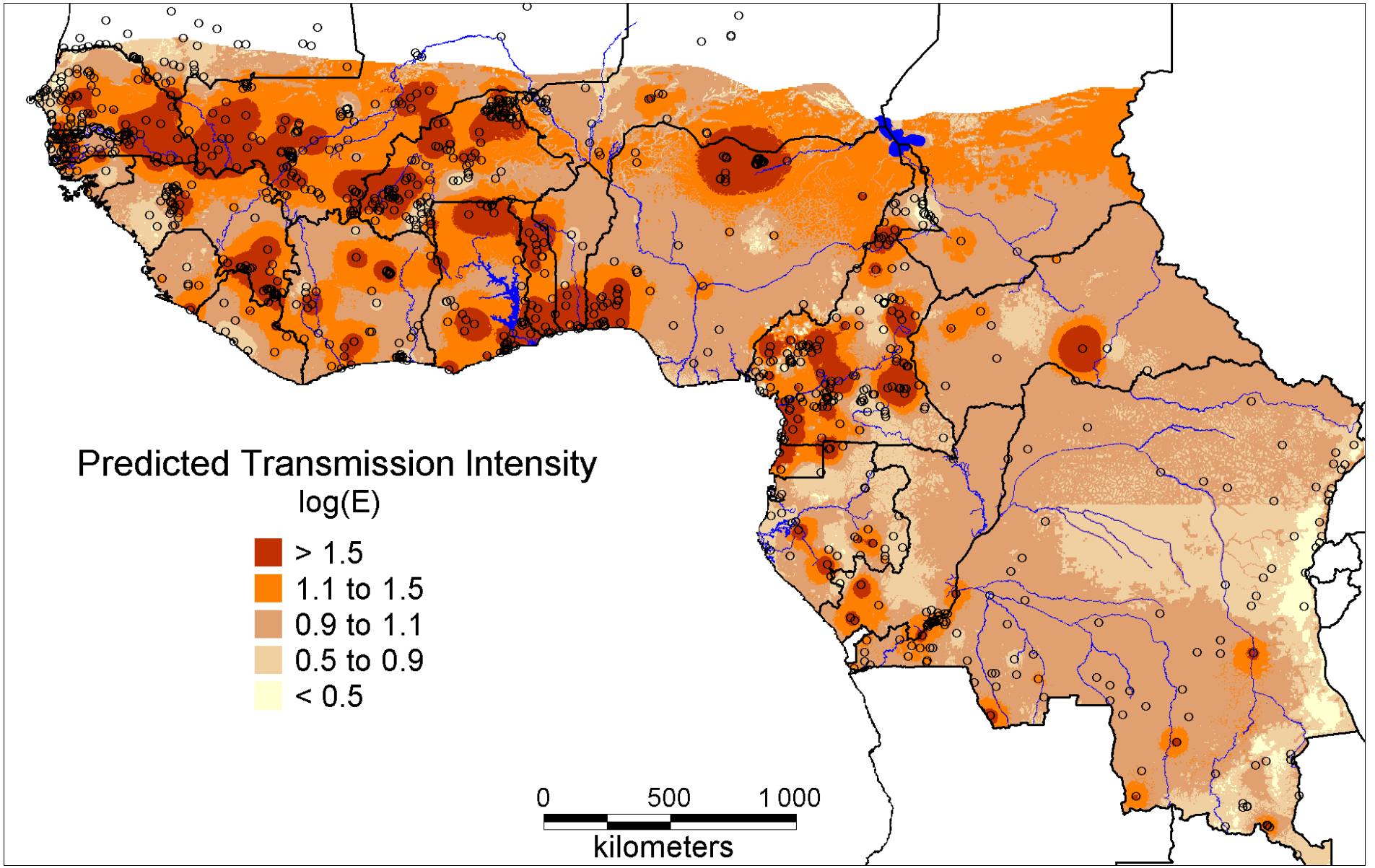


Figure 6.4: Predicted log(E) (Median) for West- and Central Africa. Sampling locations are shown by circles.

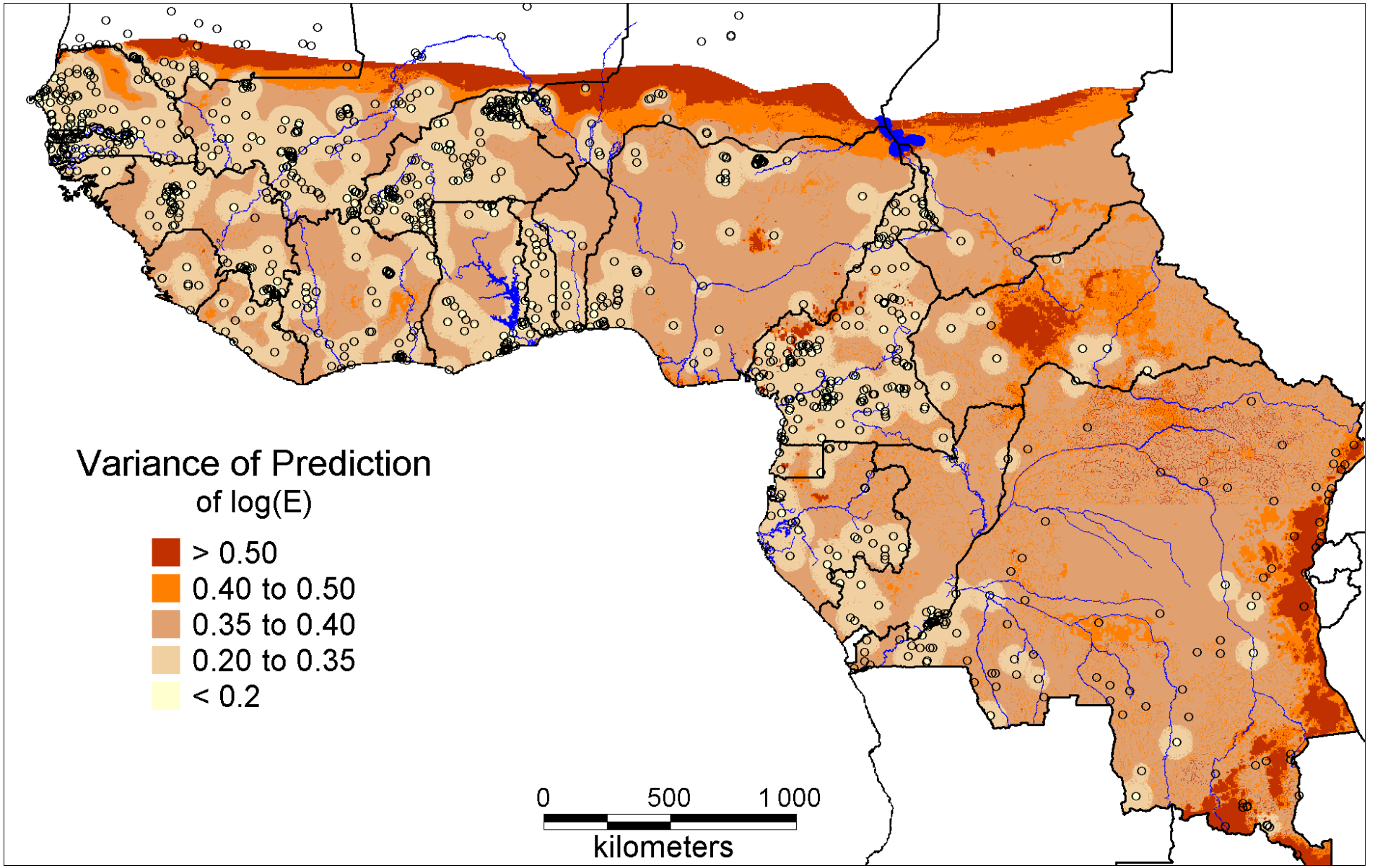


Figure 6.5: Variance of predicted log(E) for West- and Central Africa.

malaria transmission, such as the northern part of Sénégal, Guinea, Liberia and the region around Abidjan in Ivory Coast. Central Africa is estimated to have a low level of malaria transmission with few focal regions of high transmission around Bambari and Bossangoa (Central African Republic), South Gabon, South Republic of Congo and a few nodes in the Democratic Republic of Congo (Yamfu-Nunga, Dilolo, Kamina, Lubumbashi, Kabalo). Prediction in the northeastern part of the map is considered not reliable because of the sparse malaria surveys conducted in this region (figure 6.1).

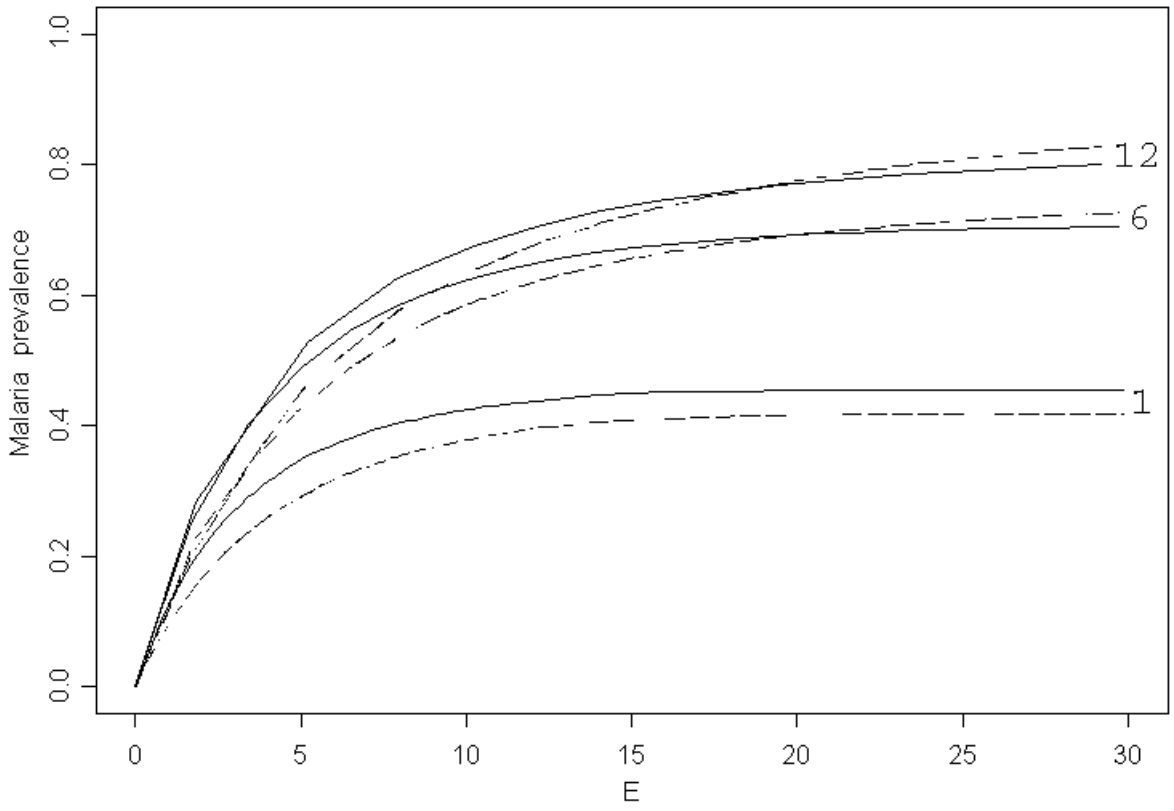


Figure 6.6: Estimated prevalence-E relationship for different length of malaria season and two age-groups, 1 to 10 years old (solid line) and less than five years old (dashed line). The length of season in units of months is attached to every curve.

The maps of malaria prevalence for the two age groups are shown in figures 6.7 and 6.8. A band of relatively high malaria prevalence was predicted for the West African Sudan-Savanna zone, including the Northern Guinea Savanna (see Kleinschmidt et al., 2001a, for the definition of the zonation). Some larger regions of relatively low levels of prevalence were estimated in the forest zone and in most of Central Africa. Exceptions of high prevalence in the forest zone are found in the south of Ghana, Togo, Benin and Nigeria on a coastal strip between Accra and Lagos, and for South Guinea at the border to Liberia.

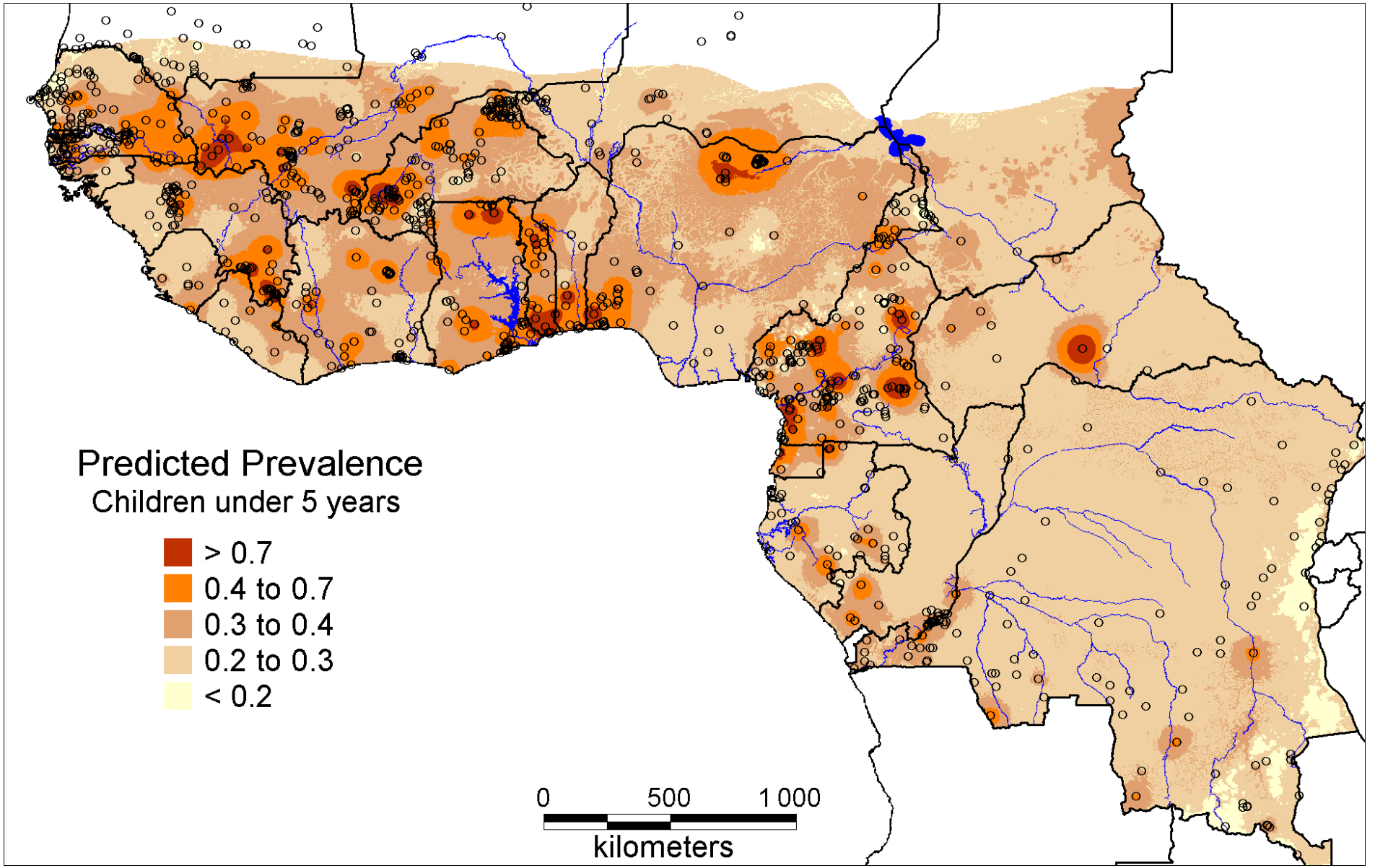


Figure 6.7: Predicted prevalence in children under five years for West- and Central Africa.

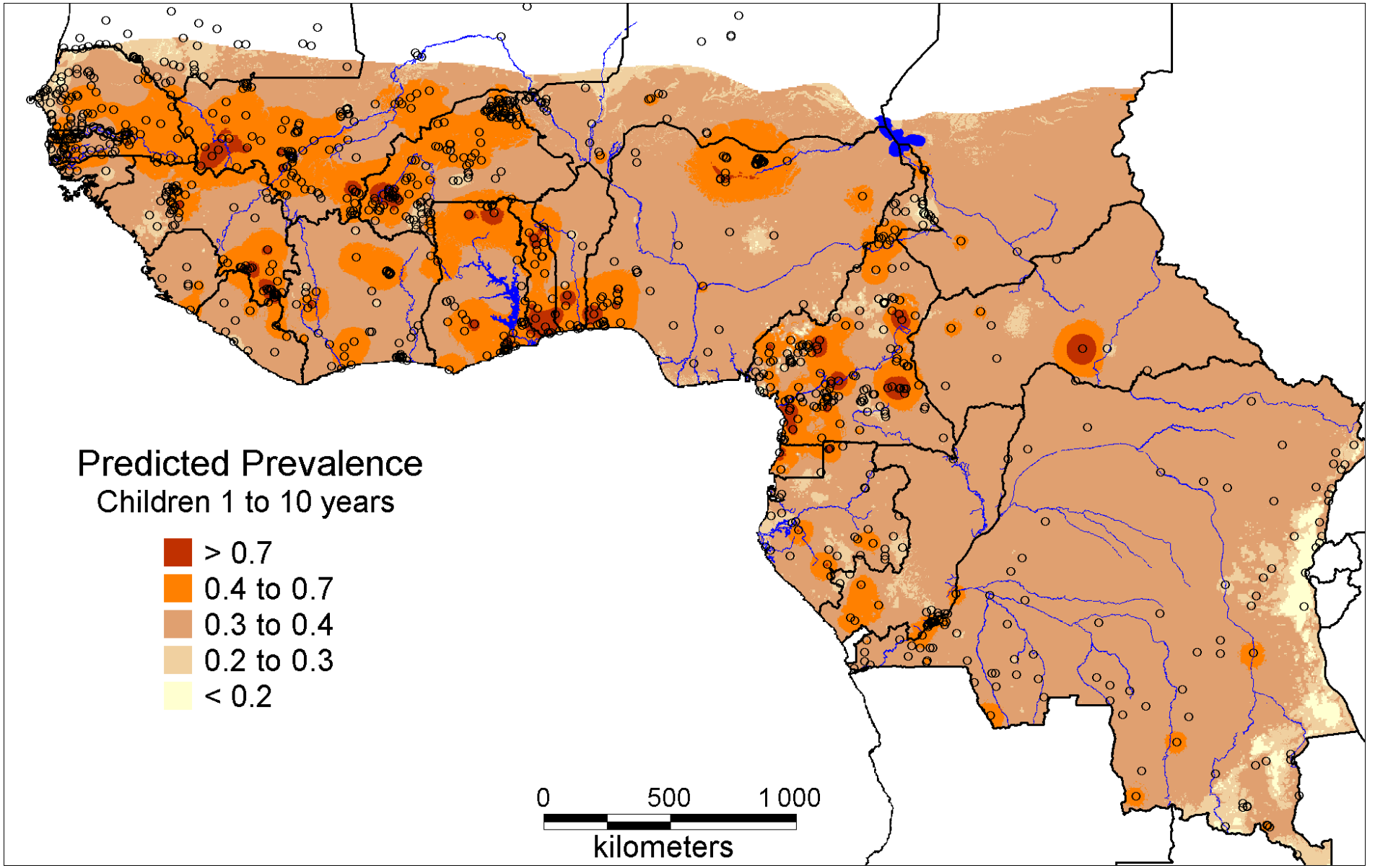


Figure 6.8: Predicted prevalence in children one to ten years for West- and Central Africa.

The two prevalence maps for the different age groups show only slight differences in the spatial distribution of prevalence, but for children one to ten the prevalence is estimated uniformly remarkably higher than for children under five years old.

6.4 Discussion

In this study, the Garki model was employed in a novel way to convert malaria prevalence data extracted from the MARA database to malaria transmission intensity for each survey location. In our recent work (Gemperli et al., 2003b), we used the Garki model to draw maps of malaria transmission and prevalence for Mali. However, we did not consider the seasonality in the malaria transmission and assumed that transmission season was the same at all locations. In this analysis, we have employed a modified approach which takes into account the length of transmission season at each location and thus the seasonality in the relation between transmission intensity and age-prevalence curves. Our model requires as inputs the length of transmission season for each location which was calculated by a modified version of the seasonality map of Tanser et al. (2000). A Bayesian variogram model was applied on the malaria estimates to obtain smooth maps of malaria transmission intensity for West and Central Africa adjusted for environmental covariates which were obtained from remote sensing.

Seasonality in transmission is an important, but neglected, consideration in malaria mapping, both because the season at which the data were collected may be important, and because the malaria maps themselves may be season specific. At very high transmission levels, malaria prevalence is generally not very seasonal (Smith et al., 1993), but at low transmission levels, surveys carried out in the dry season generally have much lower prevalence than wet season surveys. Many surveys are deliberately carried out during the peak transmission season, and this introduces a bias in the maps unless it is allowed for. Seasonality also affects the relationship between prevalence and inoculation rates, since when many inoculations occur over a short period of time the proportion resulting in erythrocytic infections is reduced (Beier et al., 1994; Charlwood et al., 1998). The Garki model adjusts automatically for this effect when a seasonal input of vectorial capacity is assumed. However it would have been preferable to use a seasonality model that predicted quantitative variation in transmission between months, rather than simply classifying them into months of transmission/no transmission. Moreover, there is a clear need for empirical maps of seasonality based on fitting models to local data on seasonality of either entomological or clinical indices. Despite our attempt to augment the seasonality map using NDVI data, it has clearly failed to correctly assign areas of endemic transmission in Southern Mauritania, and probably also in other areas where rivers flow north into dry zones.

The Garki model enabled us to convert malaria prevalence data collected from surveys from non-standardized age groups of the population to an age-independent transmission measure. Previous mapping efforts attempted to overcome the problem of age-adjustment by discarding inappropriate age groups. This resulted in a vast waste of available malaria data. The model can then be further applied to obtain age-specific prevalence. The

mapping of outputs of malaria transmission models provides a general framework to derive malaria prevalence estimates for any desired age group. It can be also used to derive other measures of transmission, different from the E, which are not measured in the field. However, the Garki model was developed on field data from the savannah zone of Nigeria (Molineaux and Gramiccia, 1980). It needs to be verified how accurately it can be adapted for other regions in West- and Central Africa, with different environmental conditions and malaria endemicity.

The Bayesian variogram modelling approach takes into account the spatial dependence present in the data in a flexible way. The method calculates inherently the standard error of the parameter estimates as well as the prediction error without relying on approximations or asymptotic results. Maps of the prediction error indicate the confidence we can have on the model predictions for the study area.

In a previous study to map malaria in West Africa, Kleinschmidt et al. (2001a) modelled interactions between the environmental predictors and agro-ecological zones by a separate analysis for each ecological zone. The resulting map showed discontinuities around the borders of the zones, which were further smoothed. This additional step applied after kriging made inference on the prediction error unfeasible. To avoid the separation into geographical zones, we considered interaction amongst the environmental predictors which capture space-varying functional relationships between the predictors and malaria transmission. This approach produces no discontinuities and avoids arbitrary geographical partitioning. Our modelling approach goes further beyond that of Kleinschmidt et al. (2001a), because, we could include all survey information, irrespective of their age group, and the Bayesian model applied allowed correct adjustment for estimation uncertainty and prediction error.

A comparison of our estimated malaria prevalence maps with those produced by Kleinschmidt et al. (2001a) for West-Africa reveals similar patterns, but the predicted prevalence in our map shows fewer regions with prevalence above 70 percent or below 30 percent. Both maps identify the same areas with high malaria prevalence (border of Sénégal-Mali-Guinea, North Ivory-Coast, Togo, North Nigeria, West Cameroon) and with low malaria prevalence (Guinea-Bissau, South-East Burkina Faso, Central Nigeria, and Central-North and North Cameroon). There are discrepancies between the two maps in the region of Central Nigeria which the map of Kleinschmidt et al. (2001a) shows to be a high risk area and in the border region between Burkina Faso and Mali and in South Guinea which was found to be a low risk area by Kleinschmidt et al. (2001a). Our map estimates much lower malaria prevalence for the whole country of Ghana (with the exception of the coastal strip). The two areas, Central Ghana and Central Nigeria, where the two maps depict their largest differences are also regions where the sampling density is relatively low (see figure 6.1). More surveys in this two regions are needed to assess the quality of the maps and help to improve them.

Surveys conducted in urban areas were omitted in our analysis. Thus, the produced maps may depict too high a malaria estimate for large urban areas (especially in Nigeria). In order to estimate the population at risk, based on our malaria risk map, a separate prevalence estimate for urban areas is required.

The MARA data includes surveys conducted as early as the 1950's. In our analysis,

we adjusted for temporal changes in E by a linear trend term for the year of the survey. This implies that we allow a linear increase or decrease of malaria prevalence with time in the whole map irrespective of location. Changes in malaria endemicity levels, however, may occur due to drug resistance of parasites, eradication programs (residual spraying with insecticides) or urbanization and may be location-dependent and not linear in time. The assumptions of a linear temporal evolution and a constant geographical structure of malaria transmission needs to be verified by fitting spatio-temporal models to account for non-linear trends in malaria risk and space-time interactions.

6.5 Acknowledgements

The authors would like to thank the NOAA/NASA Pathfinder AVHRR Land Project (University of Maryland) and the Distributed Active Archive Center (Code 902.2) at the Goddard Space Flight Center, Greenbelt, MD 20771 for the production and distribution of these data, respectively. The work of the first author was supported by Swiss National Science Foundation grant Nr. 3200–057165.99. This work is a product of the MARA/MARA (Mapping Malaria Risk in Africa) collaboration, and the authors would also like to acknowledge the contributions of the many field, laboratory and office workers who carried out the surveys and compiled the malariological database.

Appendix 6.A Garki model

The Garki model (Dietz et al., 1974) is a mathematical model of malaria transmission which can be used to predict age-specific malaria prevalence as a function of the vectorial capacity C. C is defined to be the number of potentially infective contacts induced by the mosquito population per infectious person per day. The Garki model describes transitions among seven categories of hosts distinguished by their infection and immunological status, (figure 6.9). The proportions x_1 and x_3, account for uninfected individuals, and x_2 and x_4 are compartments with prepatent infections. y_1, y_2, and y_3, represent proportions of humans with blood-stage infections. The model predicts the proportion of human population at each age in each of the compartments. It is defined by a set of linked difference equations that specify the change in each of these proportions from one time point to the next. Let Δ be the change in proportion from one time point to the next one i.e. $\Delta x_1 = x_1(t+1) - x_1(t)$,

then the equations are defined as below:

$$
\begin{aligned}
\Delta x_1 &= \delta + y_2 R_1(h) - (h+\delta)x_1 \\
\Delta x_2 &= h x_1 - (1-\delta)^N + h(t-N)x_1(t-N) - \delta x_2 \\
\Delta x_3 &= y_3 R_2(h) - (h+\delta)x_3 \\
\Delta x_4 &= h x_3 - (1-\delta)^N + h(t-N)x_3(t-N) - \delta x_4 \\
\Delta y_1 &= (1-\delta)^N + h(t-N)x_1(t-N) - (\alpha_1+\delta)y_1 \\
\Delta y_2 &= \alpha_1 y_1 - (a_2 + R_1(h) + \delta)y_2 \\
\Delta y_3 &= \alpha_2 y_2 - (1-\delta)^N + h(t-N)x_3(t-N) - (R_2(h)+\delta)y_3
\end{aligned}
$$

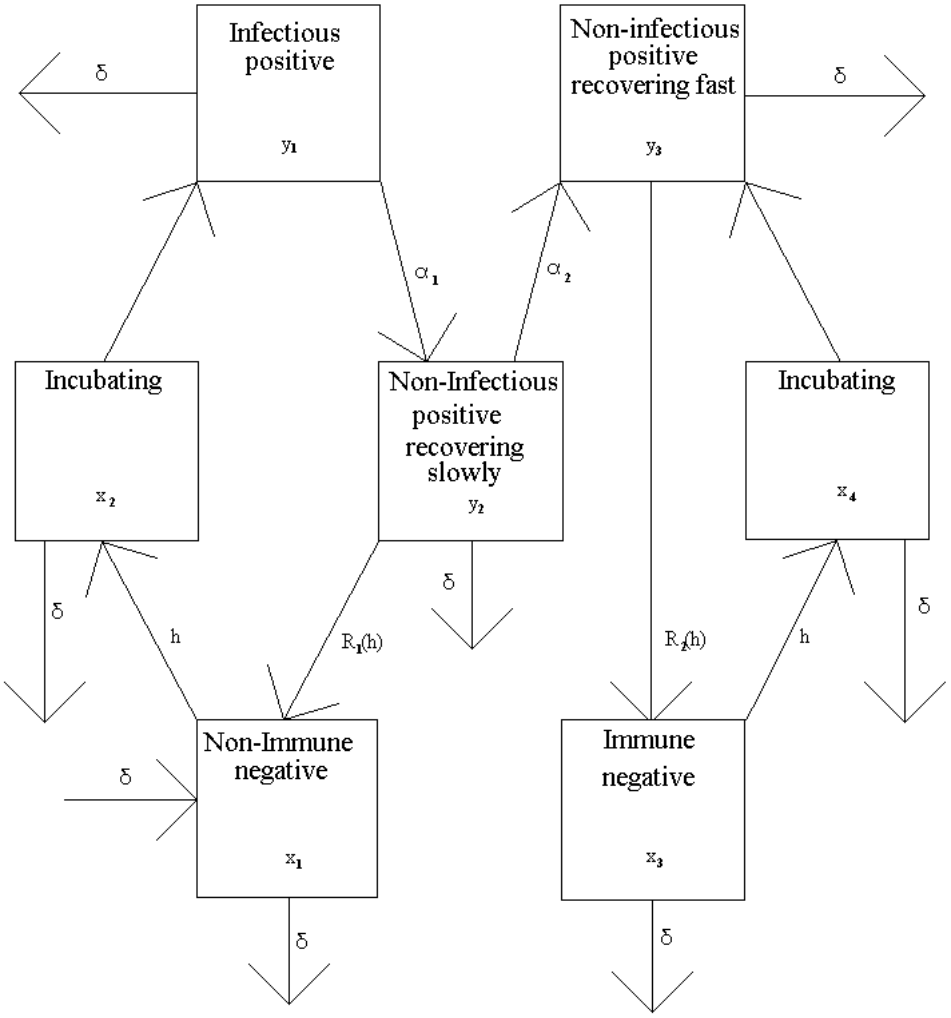


Figure 6.9: States and transitions in the Garki model.

The meanings of the additional symbols are given in table 6.4. The time points to which the proportions and the force of infection (h) refer to, are only indicated in the

above equations when they differ from t. h is the probability per unit time, that a given susceptible individual becomes infected. Here it is defined as a function of C.

Symbol	Meaning	Default value
δ	Human birth and death rates	36.5 per 100-year
α_1	Rate at which non-immunes move into the non-infective category	0.002 per day
α_2	Rate at which non-immunes recovering from infection move into the immune category	0.00019 per day
h	Force of infection (Rate of infection of susceptibles)	to be estimated
N	Duration of pre-patent period	15 days
r_1	Recovery rate for individual clones (non-immune)	0.0023 per day
r_2	Recovery rate for individual clones (immune)	$10r_1$
$R_1(h)$	Recovery rate from infection in non-immunes y_2 (as a function of h)	to be estimated
$R_2(h)$	Recovery rate from infection in immunes y_3 (as a function of h)	to be estimated
g	Maximum value of force of infection	0.097 per 5 days
q_1	Detectability of parasites in infectives (y_1)	1
q_2	Detectability of parasites in non-immunes (y_2)	1
q_3	Detectability of parasites in immunes (y_3)	0.7

Table 6.4: Quantities appearing in the Garki model.

In order to account for seasonal variation, C is considered to depend on the month and its suitability for malaria transmission, as estimated in table 6.2. Each bite on an infective individual will result in C new inoculations after N days, where N is the duration of sporogony. Dependent on the proportion of the population being infective, the E is defined as $E(t) = C(t-N)y_1(t-N)$.

$h(t)$ is assumed to be related to $E(t)$ via $h(t) = g(1 - \exp(-E(t)))$, which introduces an upper limit in the force of infection, when E increases. g specifies this upper limit and is interpreted as a parameter measuring host susceptibility. The recovery rates R_1 and R_2 are defined as $R = h/\left(\exp\left(h/r\right) - 1\right)$, where r is the recovery rate for a single clone infection. Non-immunes are assumed to recover at rate R_1, calculated from this equation by setting $r = r_1$. Immunes recover at rate R_2, calculated by setting $r = r_2$ where $r_2 > r_1$. q_1, q_2, and q_3 are introduced to allow for imperfect detection of parasitaemia in each of the three infected classes y_1, y_2 and y_3. Hence, the prevalence is estimated by $z(t) = q_1y_1(t) + q_2y_2(t) + q_3y_3(t)$.

The Garki model was developed to make predictions of the age-specific prevalence in humans as a function of C. We reversed the calculations and estimated E from the observed prevalence data, by using the golden section search routine (Press et al., 1988) to identify the E which fits better to the observed prevalence data. In particular starting with an arbitrary value of E (and for the given $C(\cdot)$ at the survey location) we estimated the age-dependent prevalence curve $z(\cdot)$ via simulating the model with arbitrary starting values of x_1 to x_4 and y_1 to y_3, until equilibrium was reached. The golden search routine searches for values of E which minimize the deviance goodness of fit (of the binomial likelihood)

between the observed prevalence and the estimated $z(\cdot)$ by the Garki model. We run the simulation with a time interval of 5 days. $C(\cdot)$ varies seasonal, depending on the estimated seasonality map (table 6.2). The effect of season of birth was accounted for by assuming uniformly random birthdates throughout the year. The starting values we have chosen are shown in table 6.4.

Appendix 6.B Spatial statistical model

Let Y_j denote the logarithm of E at location s_j, $j = 1, \ldots, m$. We assumed that Y_j is normally distributed and introduce spatial dependency between two measures Y_j and Y_k by defining a spatial exponential covariance $\mathrm{Cov}(Y_j, Y_k) = \sigma^2 \exp(-d_{jk}/\rho)$ for $j \neq k$. d_{jk} is the Euclidean distance that separates Y_j and Y_k, σ^2 quantifies the amount of spatially structured variation and ρ the spatial dependency. A parameter τ^2 is introduced to measure non-spatial variation at the origin and to add extra variability to those values with imprecise estimates from the Garki model. The variance in Y_j is then given by $\mathrm{Var}(Y_j) = \sigma^2 + \tau^2/\omega$, where w_j is a weight, formed by the reciprocal of the variance of the log-E estimate at location s_j from the Garki model. The mean of Y_j is modelled via a parametric function $\mu(\boldsymbol{x}_j, \boldsymbol{\beta})$ of the covariates $\boldsymbol{x}_j$ and a parameter vector $\boldsymbol{\beta}$.

The model for $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^t$ is written in matrix notation as $\boldsymbol{Y} \sim \mathcal{N}(\mu(\boldsymbol{X}, \boldsymbol{\beta}), \sigma^2 R(\rho) + \tau^2 \boldsymbol{W})$. $(R)_{jk} = \exp(-d_{jk}/\rho)$ and $\boldsymbol{W}$ is the weight matrix with elements $W_{jj} = 1/w_j$ and $W_{jk} = 0$ for $j \neq k$. The specification above holds if all m locations are distinct. In case of $n > m$ observations $Y_1, \ldots, Y_n$ at m distinct locations, an $m \times n$ incidence matrix Z is formed with $Z_{ji} = 1$ if observation i is observed at location j and $Z_{ji} = 0$ otherwise. Then $\boldsymbol{Y} \sim \mathcal{N}(\mu(\boldsymbol{X}, \boldsymbol{\beta}), \sigma^2 \boldsymbol{Z}^t R(\rho) \boldsymbol{Z} + \tau^2 \boldsymbol{Z}^t \boldsymbol{W} \boldsymbol{Z})$.

The following prior distributions are adopted for the parameters involved in the model: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, b_\beta \boldsymbol{I})$, $\sigma^2 \sim IG(a_{\sigma^2}, b_{\sigma^2})$, $\tau^2 \sim IG(a_{\tau^2}, b_{\tau^2})$ and $\rho \sim G(a_\rho, b_\rho)$. $G(\cdot)$ indicates the Gamma and $IG(\cdot)$ the Inverse-Gamma distribution. The hyperpriors are fixed to $b_\beta = 100$, $a_{\sigma^2} = a_{\tau^2} = 2.01$, $b_{\sigma^2} = b_{\tau^2} = 1.01$ and $a_\rho = b_\rho = 0.01$. This leads to a prior mean of one for all the covariance parameters and a large variance of 100.

Parameter are estimated using Markov chain Monte Carlo (MCMC) (Gelfand and Smith, 1990). The joint posterior distribution of the parameters is simulated using Gibbs sampling, what requires to generate random numbers from the conditional distribution of the parameters individually. For $\mu(\boldsymbol{X}, \boldsymbol{\beta})$ linear, the conditional distribution of $\boldsymbol{\beta}$ is normal and easy to sample from. The conditional distribution of the covariance parameters σ^2, τ^2 and ρ, are identified to have no standard forms and are sampled using a random walk Metropolis-Hastings algorithm having a log-Gaussian proposal density with mean equals the estimate from the previous iteration and variance iteratively altered to reach an acceptance rate of 0.4.

The log-E can be predicted at new locations $s_{01}, \ldots, s_{0l}$, once the spatial correlation between locations is estimated and the environmental covariates $\boldsymbol{X}_{\mathrm{new}}$ at the new locations are known. The algorithm for Bayesian kriging iteratively draws independent values from the predictive distribution. At iteration r, the algorithm starts by drawing values from

the joint posterior distribution of τ^2, σ^2 and ρ, which is given empirically as the output of the Gibbs sampler described above. The sampled values are used to form the covariance matrix $\boldsymbol{\Sigma}^{(r)} = \sigma^{2(r)} R(\rho^{(r)}) + \tau^{2(r)} \boldsymbol{W}$. $R(\rho^{(r)})_{jk} = \exp(-d_{jk}/\rho^{(r)})$, with d_{jk} the Euclidean distance between location s_j and location s_k. There are three matrices formed this way. $\boldsymbol{\Sigma}^{(r)}_{\text{old}}$ is build by including only the old locations $s_1, \ldots, s_m$, $\boldsymbol{\Sigma}^{(r)}_{\text{new}}$ takes only new locations $s_{01}, \ldots, s_{0l}$ and $\boldsymbol{\Sigma}^{(r)}_{\text{old-new}}$ describes covariances between old and new locations. That is, the $m \times l$ matrix $(\Sigma^{(r)}_{\text{old-new}})_{jk}$ includes locations $s_1, \ldots, s_m$ for j and locations $s_{01}, \ldots, s_{0l}$ for k. For new locations the weights in the diagonal of $\boldsymbol{W}$ are set to one.

Subsequently the parameter $\boldsymbol{\beta}^{(r)}$ is drawn form its posterior distribution to form the vector $\mu(\boldsymbol{X}_{\text{new}}, \boldsymbol{\beta}^{(r)})$. Finally, a single vector from the predictive distribution of $\boldsymbol{Y}_0$ is drawn from a multivariate normal with mean $\mu(\boldsymbol{X}_{\text{new}}, \boldsymbol{\beta}^{(r)}) + \boldsymbol{\Sigma}^{(r)t}_{\text{old-new}} \boldsymbol{\Sigma}^{(r)-1}_{\text{old}} (\boldsymbol{Y} - \mu(\boldsymbol{X}_{\text{old}}, \boldsymbol{\beta}^{(r)}))$ and variance $\boldsymbol{\Sigma}^{(r)}_{\text{new}} - \boldsymbol{\Sigma}^{(r)t}_{\text{old-new}} \boldsymbol{\Sigma}^{(r)-1}_{\text{old}} \boldsymbol{\Sigma}^{(r)}_{\text{old-new}}$.

The map with predicted log-E is back-transformed to age-related prevalence by applying the relations estimated by the Garki model. The back-transformation considers the location specific season-length.

CHAPTER 7

Strategies for fitting large, geostatistical data in MCMC simulation

Gemperli A. and Vounatsou P.
Swiss Tropical Institute, Basel, Switzerland

This paper was submitted to *Communications in Statistics - Simulation and Computation.*

Abstract

Models for geostatistical data introduce spatial dependence in the covariance matrix of location-specific random effects. This is usually defined to be a parametric function of the distances between locations. Bayesian formulations of such models overcome asymptotic inference and estimation problems involved in maximum likelihood-based approaches and can be fitted using Markov chain Monte Carlo (MCMC). The MCMC implementation however requires repeated inversions of the covariance matrix which makes the problem computationally intensive, especially for large number of locations. In the present work, we propose to convert the spatial covariance matrix to a sparse matrix and compare a number of numerical algorithms especially suited within the MCMC framework in order to accelerate large matrix inversion. The algorithms are assessed empirically on simulated datasets of different size and sparsity. We conclude that the band solver applied after ordering the distance matrix reduces the computational time in inverting covariance matrices substantially.

Keywords: band system solver; Bayesian model; generalized linear mixed model; geostatistics; Gibbs-Poole-Stockmeyer algorithm; incomplete factorization; Markov chain Monte Carlo; quotient-minimum-degree algorithm; sweep operator.

7.1 Introduction

Models for geostatistical data are embedded within the framework of generalized linear mixed models (GLMM). Geographical dependence is introduced via the covariance structure of location-specific random effects which is specified to be a parametric function of the distances between locations. Under the assumption of stationarity, the covariance matrix determines the variogram and the GLMM is also known as variogram model. Maximum likelihood-based estimation has major shortcomings. Asymptotic inference is not uniquely defined (Cressie, 1993) and the competing approaches may lead to different results (Tubilla, 1975; Stein, 1999). When the aim of modelling is kriging, that is prediction at un-sampled locations, the parameter uncertainty in maximum likelihood estimation is not fully accounted for and the standard error of predicted values underestimates the true variability (Prasad and Rao, 1990; Zimmerman and Cressie, 1992; Booth and Hobert, 1998).

Diggle et al. (1998) formulated the variogram model as a Bayesian hierarchical model and provided full Bayesian inference using Markov chain Monte Carlo (MCMC) estimation. However MCMC estimation is hampered by the repeated inversions of the covariance matrix of the random effects which for large number of locations can be infeasible within practical time constraints. To overcome the computational problems that arise from the inversion of large covariance matrices, Gelfand et al. (1999) suggested a non-iterative estimation procedure implemented via Sampling-Importance-Resampling (SIR) (Rubin, 1987). In contrast to MCMC-based sampling, SIR is not generally applicable and needs to be tailored to every spatial model and dataset (Gemperli and Vounatsou, 2003).

Christensen et al. (2000) adopted the modelling approach of Diggle et al. (1998) and proposed speeding up the computation of MCMC by jointly updating the whole vector of random effects using the Metropolis-Langevin algorithm instead of updating each random effect separately within an MCMC iteration. There are also a number of approaches aiming to improve MCMC in general and not specifically for fitting variogram models. There are methods which accelerate the convergence time (Liu, 2003) or reduce correlations between the parameters and autocorrelation in the samples drawn from the Markov chain by over-relaxation or centering (Gelfand et al., 1996). Other sampling schemes reduce the number of rejections made by the Metropolis-Hastings algorithm within Gibbs sampling (Green and Mira, 2002; Liu et al., 2000). None of these techniques overcome or improve the process of matrix inversion, but they make it possible to sample significantly shorter chains to reach the same accuracy.

In a series of papers Kelley and Barry (1997a,b) and Rue (2000) suggested methods to accelerate conditional autoregressive spatial models for areal data using the sparseness of the proximity matrix. In areal data settings, a proximity matrix with zero values for all areas which do not share a common border can be easily defined. This can be also adapted for variogram modelling by choosing appropriate covariance matrices. There are valid spatial covariance definitions with zero value for all covariances defined at lag distances higher than a specific parameter, called the range (Cressie, 1993). The sparseness is not a priori defined for these models but it depends on the range parameter. Barry and Pace (1997) demonstrated the computational advantages in solving the kriging equations using sparse spatial covariance matrices.

A sparse covariance matrix does not necessarily facilitate matrix inversion, since the factor of a sparse matrix may no longer be sparse. In most cases, incomplete factorization (Markowitz, 1957) can ensure also sparsity of the matrix factor. Incomplete factorization is primarily intended for use as a preconditioner in an iterative approach and it is well suited in the Gibbs sampling framework. This is because iterative methods can be computationally more efficient than non-iterative ones when starting values are good. Within the Gibbs sampling framework good starting values for the incomplete factorization can be obtained from the previous Gibbs sampling iteration.

In this work, we assess various numerical algorithms for fast matrix inversion within Markov chain Monte Carlo (MCMC) estimation of variogram models. In section 7.2.1 we present the Bayesian formulation of the model and in 7.2.2 we discuss the computational steps within MCMC which can delay model fit. In section 7.3 we review algorithms for fast inversion of large matrices to speed up the MCMC computations. These algorithms include methods for sparse matrices, iterative solvers and other more specialized methods. The computational speed of the suggested algorithms is empirically assessed on simulated datasets of different size and sparsity in section 7.4. Simulations were run for MCMC schemes updating either each location-specific random effect separately or the whole vector as a block. We provide final concluding remarks in section 7.5.

7.2 Variogram model

7.2.1 Bayesian formulation

To describe the basic model, let $\boldsymbol{Y}(\boldsymbol{s}) = (Y(s_1), Y(s_2), \ldots, Y(s_m))^t$ be the response data observed at the set of locations $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)^t$ where $s_i \in \mathcal{D} \subset \mathcal{R}^2$, $i = 1, \ldots, m$ and let $\boldsymbol{X}(s_i)$ be the vector of covariates associated with location s_i for $j = 1, \ldots, n_i$. Following the generalized linear mixed models framework of Diggle et al. (1998), we introduce location-specific random effects $\boldsymbol{\phi}(\boldsymbol{s}) = (\phi(s_1), \ldots, \phi(s_m))^t$ and assume that conditional on $\phi(s_i)$, the $Y(s_i)$ are independent with $\mathrm{E}(Y(s_i) \mid \phi(s_i)) = \mu(s_i)$. Covariates and spatial random effects are modelled on $g(\mu(s_i))$ where $g(\cdot)$ is the link-function and $g(\mu(s_i)) = \boldsymbol{X}^t(s_i)\boldsymbol{\beta} + \phi_i$. For simplicity, we drop s in the subscripts and write $\boldsymbol{Y}$, Y_i, $\boldsymbol{X}_i$, $\boldsymbol{\phi}$ and ϕ_i instead of $\boldsymbol{Y}(\boldsymbol{s})$, $Y(s_i)$, $\boldsymbol{X}(s_i)$, $\boldsymbol{\phi}(\boldsymbol{s})$ and $\phi(s_i)$, respectively and introduce $\boldsymbol{X} = (\boldsymbol{X}_1^t, \ldots, \boldsymbol{X}_m^t)^t$.

Spatial dependence is captured by the location-specific random effects. It is assumed that they model a latent stationary and isotropic Gaussian spatial process over the study region, $\mathcal{D}$, such that $\pi(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2) \equiv N(0, \boldsymbol{\Sigma})$, where Σ_{ij} is a parametric function of the distance between the corresponding locations s_i and s_j, that is $\Sigma_{ij} = \sigma^2 \varrho(\parallel s_i - s_j \parallel_2 ; \delta) + \tau^2 1_{\{i=j\}}$ where $\parallel \cdot \parallel_2$ denotes the Euclidean distance. $\varrho(\cdot)$ corresponds to a valid correlation function, σ^2 measures the spatial variance, δ quantifies the rate of correlation decay and τ^2 captures non-spatial variation. The exponential form $\varrho(\parallel s_i - s_j \parallel_2; \delta) = \exp(- \parallel s_i - s_j \parallel_2 / \delta)$ is the most commonly used correlation function. Alternative forms are proposed by Cressie (1993). Under the assumption of stationarity, $\boldsymbol{\Sigma}$ determines the variogram and the parameters σ^2, δ and τ^2 are directly related to the sill, range and nugget parameters of the variogram (Ecker and Gelfand, 1997).

A Bayesian formulation of the variogram model requires specification of prior distributions for the parameters, $\boldsymbol{\beta}$, σ^2, δ and τ^2. Typically we assume a non-informative distribution, such as a uniform or a normal with large variances, for the covariate coefficients $\boldsymbol{\beta}$ and inverse Gamma distributions for σ^2, δ and τ^2. Inference is based on the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \delta, \tau^2 \mid \boldsymbol{Y}) \propto L(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2) \pi(\boldsymbol{\beta}) \pi(\sigma^2) \pi(\delta) \pi(\tau^2)$ where $L(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\phi})$ is the likelihood and $\pi(\cdot)$ denote the prior distributions of the parameters. This model is highly parameterized especially for large number of locations and model fit is only feasible via Markov chain Monte Carlo.

7.2.2 Markov chain Monte Carlo computations

The Gibbs sampling algorithm is the most common Markov chain-based simulation algorithm (Gelfand and Smith, 1990). Its standard implementation requires sampling from the one-dimensional conditional distributions of all parameters, iteratively until convergence. Variations of the algorithm include sampling from multivariate conditional distributions of block of parameters, however these distributions have rarely known forms and sampling is not straightforward.

In the variogram model specified above, the one-dimensional conditional distributions of the ϕ_i, $i = 1, \ldots, m$, parameters have the following form:

$$p(\phi_i|\boldsymbol{\phi}_{\text{-}i}, \boldsymbol{\beta}, \sigma^2, \delta, \tau^2, \boldsymbol{Y}) \propto p(\phi_i|\boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2) \cdot L(Y_i \mid \phi_i, \boldsymbol{\beta}).$$

The first term on the right hand side is a univariate normal having mean $\mathrm{E}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2) = \boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}^{-1}_{\text{-}i,\text{-}i}\boldsymbol{\phi}_{\text{-}i}$ and variance $\mathrm{Var}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2) = \sigma^2 + \tau^2 - \boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}^{-1}_{\text{-}i,\text{-}i}\boldsymbol{\Sigma}_{\text{-}i,i}$, with $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{\phi}\boldsymbol{\phi}^t)$, $\boldsymbol{\phi}_{\text{-}i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_m)^t$, $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} = \mathrm{E}(\boldsymbol{\phi}_{\text{-}i}\boldsymbol{\phi}^t_{\text{-}i})$ and $\boldsymbol{\Sigma}^t_{\text{-}i,i} = \boldsymbol{\Sigma}^t_{i,\text{-}i} = \mathrm{E}(\phi_i\boldsymbol{\phi}^t_{\text{-}i})$. The Metropolis-Hastings algorithm can be used to simulate from the above conditional distribution. This requires computer intensive matrix inversions to calculate the term $\boldsymbol{c}_i^t = \boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}^{-1}_{\text{-}i,\text{-}i}$. The $(m-1) \times (m-1)$ matrix $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ is inverted m times during one iteration of the Gibbs Sampler. This step is time consuming, since inversion is an operation of order 3, and it can substantially suffer from numerical errors. The inversion is calculated via the solution of the linear system $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{\Sigma}_{\text{-}i,i}$. It is computed commonly by first decomposing $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ into its Cholesky factors $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} = \boldsymbol{Q}_i\boldsymbol{Q}_i^t$ and then using forward and backward substitution in $\boldsymbol{Q}_i\boldsymbol{x}_i = \boldsymbol{\Sigma}_{\text{-}i,i}$ and $\boldsymbol{Q}_i^t\boldsymbol{c}_i = \boldsymbol{x}_i$.

As an alternative to updating $\boldsymbol{\phi}$ componentwise, we could update the whole vector as a block using Metropolis-Hastings. Liu et al. (1994) discuss the advantages of block updating in the parameters, however it is difficult to find a good multivariate proposal distribution which is easy to simulate from and resembles the conditional distribution of $\boldsymbol{\phi}$. Neal (1996) proposes the Hamiltonian dynamics and provides a general method for constructing multivariate proposal distributions. A special case of the Hamilton method is the Langevin diffusion Metropolis-Hastings algorithm which has been applied by Christensen et al. (2000) for updating the spatial random effects of variogram models. The Langevin algorithm constructs a multivariate Gaussian proposal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{\phi}^{(t)}}$ and variance $a^2 \cdot \boldsymbol{I}_m$ and simulates $\boldsymbol{\phi}^{(t)\star} \sim N(\boldsymbol{\mu}_{\boldsymbol{\phi}^{(t)}}, a^2 \cdot \boldsymbol{I}_m)$ at a Gibbs iteration t where,

$$\boldsymbol{\mu}_{\boldsymbol{\phi}^{(t)}} = \boldsymbol{\phi}^{(t-1)} + \frac{a}{2}\nabla \log p(\boldsymbol{\phi}^{(t-1)}; \boldsymbol{Y})$$

and a is a constant which should be adjusted according to the acceptance rate. $p(\boldsymbol{\phi}; \boldsymbol{Y})$ is the full conditional distribution of $\boldsymbol{\phi}$ which has the form

$$p(\boldsymbol{\phi}; \boldsymbol{Y}) = p(\boldsymbol{\phi} \mid \boldsymbol{\beta}, \sigma^2, \delta, \tau^2, \boldsymbol{Y}) \propto \pi(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2) \cdot L(\boldsymbol{Y} \mid \phi, \boldsymbol{\beta}).$$

At iteration t the $\boldsymbol{\phi}^{(t-1)}$ is updated by $\boldsymbol{\phi}^{(t)*}$ with probability

$$\min\left\{\frac{p(\boldsymbol{\phi}^{(t)\star}; \boldsymbol{Y})}{p(\boldsymbol{\phi}^{(t-1)}; \boldsymbol{Y})}\frac{\exp\left(-\frac{1}{2a^2}\sum_{i=1}^m(\phi_i^{(t-1)} - \mu_{\boldsymbol{\phi}^{(t)\star}})^2\right)}{\exp\left(-\frac{1}{2a^2}\sum_{i=1}^m(\phi_i^{(t)\star} - \mu_{\boldsymbol{\phi}^{(t-1)}})^2\right)}, 1\right\}$$

Computation of $p(\boldsymbol{\phi}; \boldsymbol{Y})$ and in particular of $\pi(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2)$ requires inversion of the $m \times m$ matrix $\boldsymbol{\Sigma}$.

The conditional distribution of σ^2, δ and τ^2 includes also terms with $\boldsymbol{\Sigma}^{-1}$. It is proportional to

$$det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})\pi(\sigma^2)\pi(\delta)\pi(\tau^2)$$

where $\boldsymbol{\Sigma}$ is a function of σ^2, δ and τ^2. The Cholesky decomposition of $\boldsymbol{\Sigma}$ is required to solve the bilinear form $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$, and to calculate the determinant of $\boldsymbol{\Sigma}$. When $\boldsymbol{Q}$ is the Cholesky factor of $\boldsymbol{\Sigma}$, then $det(\boldsymbol{\Sigma}) = \prod_{i=1}^{m} \boldsymbol{Q}_{ii}^2$ and $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi} = \boldsymbol{x}^t\boldsymbol{\phi}$ where $\boldsymbol{x}$ is computed by forward and backward substitution of the systems $\boldsymbol{Q}\boldsymbol{z} = \boldsymbol{\phi}$ and $\boldsymbol{Q}^t\boldsymbol{x} = \boldsymbol{z}$.

7.3 Algorithms for fast matrix inversions

The most computationally intensive part in the implementation of MCMC is the calculation of the quantities $\boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}^{-1}, i = 1, \ldots, m$ and $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$ which involve the inversion of the $(m-1)\times(m-1)$ and $m \times m$ matrices $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ and $\boldsymbol{\Sigma}$ respectively. Next we present numerical algorithms which have been used for speeding the inversion of large matrices.

7.3.1 Sweeping

The $\mathrm{E}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2)$ and $\mathrm{Var}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2)$ involved in the implementation of the Metropolis-Hastings algorithm, can be computed by applying the sweep operator (Goodnight, 1979). For each parameter ϕ_i, $i = 1, \ldots, m$ the matrix $A_i^{(t)} = (a_{i,jk}^{(t)})$ is built, where

$$\boldsymbol{A}_i^{(0)} = \begin{pmatrix} \boldsymbol{\Sigma}_{\text{-}i,\text{-}i} & \boldsymbol{\Sigma}_{\text{-}i,i} & -\boldsymbol{\phi}_{\text{-}i} \\ \boldsymbol{\Sigma}_{i,\text{-}i} & \sigma^2+\tau^2 & 0 \\ -\boldsymbol{\phi}_{\text{-}i}^t & 0 & 0 \end{pmatrix}$$

and $a_{i,ss}^{(t)} = \frac{1}{a_{i,ss}^{(t-1)}}$, $a_{i,js}^{(t)} = -\frac{a_{i,js}^{(t-1)}}{a_{i,ss}^{(t-1)}}$, $a_{i,sk}^{(t)} = \frac{a_{i,sk}^{(t-1)}}{a_{i,ss}^{(t-1)}}$ and $a_{i,jk}^{(t)} = a_{i,jk}^{(t-1)} - \frac{a_{i,js}^{(t-1)} a_{i,sk}^{(t-1)}}{a_{i,ss}^{(t-1)}}$, $t, s = 1, \ldots, (m-1)$, $t = s$ and $j, k = 1, \ldots, (m+1)$, $j, k \neq s$. The conditional mean $\mathrm{E}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2)$ and variance $\mathrm{Var}(\phi_i \mid \boldsymbol{\phi}_{\text{-}i}, \sigma^2, \delta, \tau^2)$ are immediately available by the $a_{i,m(m+1)}^{(m-1)}$ and $a_{i,mm}^{(m-1)}$ respectively.

Sweeping can also be used to calculate $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$. In this case

$$\boldsymbol{A}^{(0)} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\phi} \\ \boldsymbol{\phi}^t & 0 \end{pmatrix}$$

and $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi} = \boldsymbol{\phi}^t(a_{1(m+1)}^{(m)}, \ldots, a_{m(m+1)}^{(m)})^t$.

7.3.2 Sequential decomposition

We propose to compute the matrix $\boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}^{-1}$ required for the update of ϕ_i by re-using quantities calculated during the update of ϕ_{i-1} for $i = 2, \ldots, m$. Let

$$\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} = \begin{pmatrix} \boldsymbol{\Sigma}_{1,\ldots,i-1;1,\ldots,i-1} & \boldsymbol{\Sigma}_{1,\ldots,i-1;i+1,\ldots,m} \\ \boldsymbol{\Sigma}_{1,\ldots,i-1;i+1,\ldots,m}^t & \boldsymbol{\Sigma}_{i+1,\ldots,m;i+1,\ldots,m} \end{pmatrix}$$

be the matrix $\boldsymbol{\Sigma}$ without the ith row and column and $\boldsymbol{Q}_{1,\ldots,i-1;1,\ldots,i-1}$ be the Cholesky factor of $\boldsymbol{\Sigma}_{1,\ldots,i-1;1,\ldots,i-1}$, a matrix with the first $i-1$ rows and columns of $\boldsymbol{\Sigma}$ with $\boldsymbol{Q}_{1,1} =$

$\sqrt{(\boldsymbol{\Sigma}_{1,1})} = \sqrt{\sigma^2 + \tau^2}$. During the update of ϕ_i we augment $\boldsymbol{\Sigma}_{1,\ldots,i-1;1,\ldots,i-1}$ to $\boldsymbol{\Sigma}_{1,\ldots,i;1,\ldots,i}$ such as

$$\boldsymbol{\Sigma}_{1,\ldots,i;1,\ldots,i} = \begin{pmatrix} \boldsymbol{\Sigma}_{1,\ldots,i-1;1,\ldots,i-1} & \boldsymbol{b}_{i,1} \\ \boldsymbol{b}_{i,1}^t & \sigma^2 + \tau^2 \end{pmatrix}$$

and $\boldsymbol{b}_{i,1} = \boldsymbol{\Sigma}_{1,\ldots,i-1;i}$. The Cholesky factor $\boldsymbol{Q}_{1,\ldots,i;1,\ldots,i}$ of $\boldsymbol{\Sigma}_{1,\ldots,i;1,\ldots,i}$ is computed by updating the $\boldsymbol{Q}_{1,\ldots,i-1;1,\ldots,i-1}$ by

$$\boldsymbol{Q}_{1,\ldots,i;1,\ldots,i} = \begin{pmatrix} \boldsymbol{Q}_{1,\ldots,i-1;1,\ldots,i-1} & \boldsymbol{0} \\ \boldsymbol{\omega}_i & \sqrt{\sigma^2 + \tau^2 - \boldsymbol{\omega}_i^t \boldsymbol{\omega}_i} \end{pmatrix}$$

where $\boldsymbol{w}_i$ is the solution of $\boldsymbol{Q}^{(i-1)}\boldsymbol{w}_i = \boldsymbol{b}_{i,1}$. Then the Cholesky factor of $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ will be

$$\boldsymbol{Q}_{\text{-}i,\text{-}i} = \begin{pmatrix} \boldsymbol{Q}_{1,\ldots,i;1,\ldots,i} & \boldsymbol{0} \\ \boldsymbol{W}^t & \boldsymbol{Q}_{i+1}^W \end{pmatrix}$$

where $\boldsymbol{Q}_{i+1}^W$ is the Cholesky factor of $\boldsymbol{\Sigma}_{i+1,\ldots,m;i+1,\ldots,m} - \boldsymbol{W}^t\boldsymbol{W}$ and $\boldsymbol{W}$ is the solution of $\boldsymbol{Q}_{1,\ldots,i;1,\ldots,i}\boldsymbol{W} = \boldsymbol{\Sigma}_{1,\ldots,i-1;i+1,\ldots,m}$. To calculate the $\boldsymbol{\Sigma}_{i,\text{-}i}\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}^{-1}$, we solve the linear system $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{\Sigma}_{\text{-}i,i}$, using forward and backward substitution in the equations $\boldsymbol{Q}_{1,\ldots,i-1;1,\ldots,i-1}\boldsymbol{x}_1 = \boldsymbol{b}_{i,1}$, $\boldsymbol{Q}_{i+1}^W\boldsymbol{x}_2 = \boldsymbol{b}_{i,2} - \boldsymbol{W}^t\boldsymbol{x}_1$ and $(\boldsymbol{Q}_{i+1}^W)^t\boldsymbol{c}_{i,2} = \boldsymbol{x}_2$, $\boldsymbol{Q}_{1,\ldots,i-1;1,\ldots,i-1}^t\boldsymbol{c}_{i,1} = \boldsymbol{x}_1 - \boldsymbol{W}\boldsymbol{x}_2$ where $\boldsymbol{b}_{i,2} = \boldsymbol{\Sigma}_{i+1,\ldots,m;i}$ and $\boldsymbol{c}_i = (\boldsymbol{c}_{i,1}^t, \boldsymbol{c}_{i,2}^t)^t$.

Sequential decomposition has the advantage that the size of the matrix $\boldsymbol{\Sigma}_{i+1,\ldots,m;i+1,\ldots,m} - \boldsymbol{W}^t\boldsymbol{W}$ decreases for larger i and its Cholesky factor $\boldsymbol{Q}_{i+1}^W$ is derived faster. The method cannot be used for calculating $\boldsymbol{\phi}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$.

7.3.3 Sparse solvers

Data observed at locations which are very far apart will have negligible spatial dependence. Fixing the spatial correlation to zero for distances beyond the range of the spatial process will result in sparse symmetric matrices which can be inverted faster using sparse matrix methods. One example of such a valid correlation structure is the spherical form:

$$\varrho(d_{ij};\delta) = \begin{cases} 1 - \frac{3d_{ij}}{2\delta} + \frac{1}{2}(\frac{d_{ij}}{\delta})^3 & d_{ij} \le \delta \\ 0 & d_{ij} > \delta \end{cases}$$

or the less frequently used cubic type:

$$\varrho(d_{ij};\delta) = \begin{cases} 1 - (\frac{d_{ij}}{\delta})^2 \left[7 - \frac{d_{ij}}{\delta}\left[\frac{35}{4} - \left(\frac{d_{ij}}{\delta}\right)^2\left(\frac{7}{2} - \frac{3}{4}\left(\frac{d_{ij}}{\delta}\right)^2\right)\right]\right] & d_{ij} \le \delta \\ 0 & d_{ij} > \delta \end{cases}$$

where $d_{ij} = \| s_i - s_j \|_2$. An alternative nonparametric method which allows setting small covariances to zero was presented by Hall et al. (1994) and applied in variogram modelling by Bjørnstad and Falck (2001).

The factorization of sparse matrices however tends to introduce additional non-zero entries, and these so-called fill-ins result in partially lost sparsity. Fill-ins increase the storage space,increase the number of operations to perform and contribute to error propagation. A solution to reduce fill-ins is to reorder the matrix $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ (or $\boldsymbol{\Sigma}$) by multiplying it with a permutation matrix $\boldsymbol{P}_i$ (or $\boldsymbol{P}$) which is chosen to reduce the number of fill-ins. A general reordering algorithm does not exist, but heuristic solutions (quotient minimum degree QMD algorithm, multiple minimum degree algorithm) efficiently reduce the number of fill-ins (George and Liu, 1981). An alternative approach to minimize fill-ins is to find a permutation matrix which transforms the covariance matrix to a band diagonal form. The advantage of the band diagonal matrix over the general reordering is that the Cholesky factor is guaranteed to have the same bandwidth as the original matrix and thus the same sparsity outside the diagonal band. There are various methods which can be used for bandwidth minimization such as the (reverse) Cuthill-McKee algorithm (Cuthill and McKee, 1969) or its modification known as the Gibbs-Poole-Stockmeyer (GPS) algorithm (Lewis, 1982; Gibbs et al., 1976).

Using sparse matrix solvers or band-solvers, we can invert say $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ by solving $\boldsymbol{\Sigma}^{\text{Perm}}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{P}_i\boldsymbol{\Sigma}_{\text{-}i,i}$ to obtain $\boldsymbol{c}_i$ where $\boldsymbol{\Sigma}^{\text{Perm}}_{\text{-}i,\text{-}i} = \boldsymbol{P}_i\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}\boldsymbol{P}_i^t$. The band Cholesky factorization takes for $m \gg p$ around $m(p^2 + 3p)$ flops (floating point operations) and m square roots operations in comparison to the non-banded version which requires $m^3/3$ flops (Golub and Van Loan, 1996). The solution of the linear system $\boldsymbol{\Sigma}^{\text{Perm}}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{P}_i\boldsymbol{\Sigma}_{\text{-}i,i}$ takes $m(p^2+7p+2)$ flops and m square roots operations, instead of $2m^2$ for the dense linear system solver. The square root computation can be avoided using the $\boldsymbol{LDL}^t$ technique where $\boldsymbol{\Sigma}^{\text{Perm}}_{\text{-}i,\text{-}i} = \boldsymbol{LDL}^t$, $\boldsymbol{Ly} = \boldsymbol{P}_i\boldsymbol{\Sigma}_{\text{-}i,i}$, $\boldsymbol{Dz} = \boldsymbol{y}$, and $\boldsymbol{L}^t\boldsymbol{c}_i = \boldsymbol{z}$, which takes $m(p^2 + 8p + 1)$ flops and no square root calculations. In case of a small bandwidth p this can lead to a reduced computational effort.

Many variations of the sparse or band matrix solvers can be employed for inverting $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$. For example, the permutation can be applied directly on $\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ and then we can extract the $\boldsymbol{\Sigma}^{\text{Perm}}_{\text{-}i,\text{-}i}$ from $\boldsymbol{\Sigma}^{\text{Perm}}$ where $\boldsymbol{\Sigma}^{\text{Perm}} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}^t$. This reduces the computational time required for calculating $\boldsymbol{P}_i$ separately for each $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$. Another simplification would be to permute the distance matrix $\mathcal{D} = \{d_{ij}\}$ once instead of the $\boldsymbol{\Sigma}$ matrices at every MCMC iteration. This must be done in a way, that produces off-diagonal corner elements of $\mathcal{D}$ as large as possible. The resulting covariance may not have minimal bandwidth, but band solvers can still be used without the need for numerically expensive bandwidth minimizer search routines. The data (locations) could be ordered according to the first principal component of latitude and longitude.

7.3.4 Iterative solvers

The inversion of the covariance matrices which are involved in the implementation of Gibbs sampling can be obtained by iterative solvers. The advantage of this approach is that iterative methods require less memory and arithmetic operations than direct methods especially when starting values are good. The Gibbs sampling framework is suitable to the iterative solver approach as the initial values for the solver can be chosen to be the

estimates of the previous Gibbs iteration.

One such algorithm solves the linear system $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{\Sigma}_{\text{-}i,i}$, by iteratively minimizing the function $\phi(\boldsymbol{c}_i) = \frac{1}{2}\boldsymbol{c}_i^t \boldsymbol{\Sigma}_{\text{-}i,\text{-}i}\boldsymbol{c}_i - \boldsymbol{c}_i^t \boldsymbol{\Sigma}_{\text{-}i,i}$, using the conjugate gradient method (CGM) (Hestenes and Stiefel, 1952) or a Lanczos type method (Paige and Saunders, 1975). For positive-definite matrices the CGM is considered more efficient than the Lanczos type method. Faster convergence can usually be achieved by using a preconditioner, which transforms the linear system in order to obtain a well conditioned matrix to invert. Preconditioning is typically based on incomplete factorizations (Barrett et al., 1994), which discards fill-in elements arising in the Cholesky factors. The accuracy of the factorization is controlled by the amount of fill-ins to be discarded. The algorithm can also be applied to invert $\boldsymbol{\Sigma}$.

An alternative iterative linear solver approach was presented by Harville (1999) who inverts large matrices using Gibbs sampling. Let $\boldsymbol{\Gamma}$ be the inverse of $\boldsymbol{\Sigma}$, $\boldsymbol{x} = (x_1, \ldots, x_{m-1})^t$ and $\boldsymbol{x}_{\text{-}k} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{m-1})^t$, such that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$ where $\Gamma_{k,j} = \mathrm{E}(x_k x_j)$, $\boldsymbol{\Gamma}_{k,\text{-}k} = \mathrm{E}(x_k \boldsymbol{x}_{\text{-}k})$ and $\boldsymbol{\Gamma}_{\text{-}k,\text{-}k} = \mathrm{E}(\boldsymbol{x}_{\text{-}k}\boldsymbol{x}_{\text{-}k}^t)$. It follows that the conditional distributions $[x_k|\boldsymbol{x}_{\text{-}k}]$ are normal distributions such that $[x_k|\boldsymbol{x}_{\text{-}k}] \sim \mathcal{N}(\boldsymbol{\Gamma}_{k,\text{-}k}\boldsymbol{\Gamma}_{\text{-}k,\text{-}k}^{-1}\boldsymbol{x}_{\text{-}k}, \Gamma_{k,k} - \boldsymbol{\Gamma}_{k,\text{-}k}\boldsymbol{\Gamma}_{\text{-}k,\text{-}k}^{-1}\boldsymbol{\Gamma}_{\text{-}k,k})$. Taking into account the relations $\Gamma_{k,k} - \boldsymbol{\Gamma}_{k,\text{-}k}\boldsymbol{\Gamma}_{\text{-}k,\text{-}k}\boldsymbol{\Gamma}_{\text{-}k,k} = \Sigma_{k,k}^{-1}$ and $\boldsymbol{\Gamma}_{k,\text{-}k}\boldsymbol{\Gamma}_{\text{-}k,\text{-}k} = -\Sigma_{k,k}^{-1}\boldsymbol{\Sigma}_{\text{-}k,k}$ (see for example Harville, 1997, p. 99), $[x_k|\boldsymbol{x}_{\text{-}k}]$ can be written as $[x_k|\boldsymbol{x}_{\text{-}k}] \sim \mathcal{N}(-\Sigma_{k,k}^{-1}\boldsymbol{\Sigma}_{k,\text{-}k}\boldsymbol{x}_{\text{-}k}, \Sigma_{k,k}^{-1})$ or $[x_k|\boldsymbol{x}_{\text{-}k}] \sim \mathcal{N}(-\Sigma_{k,k}^{-1}\sum_{j\neq k}\Sigma_{k,j}x_j, \Sigma_{k,k}^{-1})$, respectively. Gibbs sampling can be applied to simulate $\boldsymbol{\Gamma}$ where $\Gamma_{k,j}$ is estimated by $s^{-1}\sum_{t=1}^{s} x_k^{(t)} x_j^{(t)}$ and $\boldsymbol{x}^{(t)}$ are samples obtained after convergence of the algorithm and s is the size of the sample drawn from the posterior. Using Gibbs sampling outputs, we can solve the linear system $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i} \cdot \boldsymbol{c}_i = \boldsymbol{\Sigma}_{\text{-}i,i}$ by taking

$$\widehat{c}_{i,k} = s^{-1}\sum_{t=1}^{s} x_k^{(t)} \sum_{j=1}^{m-1} x_j^{(t)} u_{i,j}.$$

where $\widehat{c}_{i,k}$ is an estimate of the k-th element of $\boldsymbol{c}_i$ and $u_{i,j}$ is the j-th element of $\boldsymbol{\Sigma}_{\text{-}i,i}$. Harville (1999) proposes ways of reducing the variance of the estimator of $\Gamma_{k,j}$, however it becomes computationally more expensive to obtain $\boldsymbol{\Gamma}$.

7.4 Simulation results

We assess the performance of the above methods on simulated geostatistical data. We have chosen three levels of sparsity of the spatial covariance matrix (10 percent of zero entries, 40 percent and 80 percent) and three different numbers of locations (300, 600 and 1,000). For each sparsity level and number of locations we simulated two datasets with normal (study I) and Poisson (study II) response variables respectively as well as four covariates. Details on the simulated data are given in the appendix.

7.4.1 Study I

In this study the response data was simulated from a normal distribution, that is $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \tau^2 \boldsymbol{I}_m)$. We follow closely the model specification of section 7.2.1, and take $\boldsymbol{\mu} = \boldsymbol{X}^t\boldsymbol{\beta} + \boldsymbol{\phi}$ and $\boldsymbol{\phi} \sim N(0, \boldsymbol{\Sigma})$. In our first stage normal model, the τ^2 parameter in the variance of $\boldsymbol{Y}$ corresponds to the nugget effect and thus $\boldsymbol{\Sigma}$ models only the spatial variance, that is $\Sigma_{ij} = \sigma^2 \varrho(\parallel s_i - s_j \parallel_2; \delta)$. We adopt the spherical correlation function for $\varrho(\parallel s_i - s_j \parallel_2; \delta)$ discussed in section 7.3.3 and assume a vague normal prior distribution for $\boldsymbol{\beta}$. The full conditional posterior distribution of $\boldsymbol{\phi}$ is multivariate normal, therefore sampling the whole vector $\boldsymbol{\phi}$ is straightforward. In this study, we choose however to update the components $\phi_i, i = 1, \ldots, m$ separately and sample them from the corresponding conditional distributions $p(\phi_i | \boldsymbol{\phi}_{\text{-}i}, \boldsymbol{\beta}, \sigma^2, \delta, \tau^2, \boldsymbol{Y})$ which are also normal with mean $(\psi_i(y_i - \boldsymbol{X}_i\boldsymbol{\beta}) + \tau^2 \boldsymbol{\Sigma}_{i,\text{-}i} \boldsymbol{\Sigma}^{-1}_{\text{-}i,\text{-}i} \boldsymbol{\phi}_{\text{-}i})/(\psi_i + \tau^2)$ and variance $\psi_i \tau^2/(\psi_i + \tau^2)$ where $\psi_i = \sigma^2 - \boldsymbol{\Sigma}_{i,\text{-}i} \boldsymbol{\Sigma}^{-1}_{\text{-}i,\text{-}i} \boldsymbol{\Sigma}_{\text{-}i,i}$.

Our objective was to compare the algorithms discussed earlier for inverting the matrices $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$. In particular, we compared the computational efficiency of the following algorithms: Inversion without acceleration; SWEEP operator (section 7.3.1); Sequential factorization (section 7.3.2); Band solvers on pre-ordered data (section 7.3.3); Band solver using the GPS-algorithm at every Gibbs iteration (section 7.3.3); Band solver using the GPS-algorithm once per Gibbs iteration (section 7.3.3); QMD-algorithm to reduce fill-ins (section 7.3.3); Gibbs-sampling inversion (section 7.3.4) and iterative solver using incomplete Cholesky factorization (section 7.3.4).

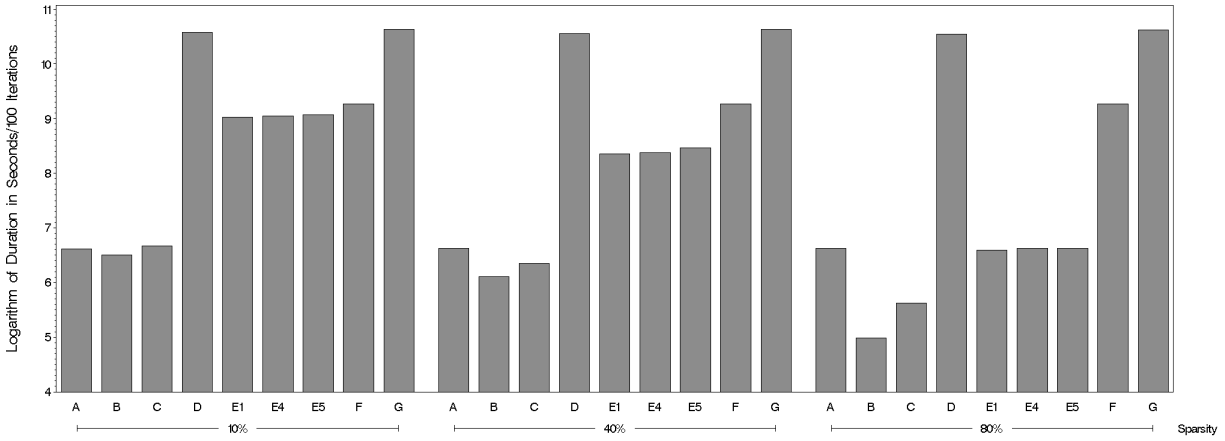


Figure 7.1: Processing time for the algorithms: Ordinary MCMC (A); Band-solver with ordered distance matrix (B); Band solver with GPS once per Gibbs iteration (C); Band solver with GPS for every random effect (D); Incomplete factorization (E1, E4, E5) with tol=0.1, 0.04 and 0.005 respectively, Sweeping (F) and QMD-algorithm (G). Simulated response data from normal distributions with sparsity 10 percent, 40 percent and 80 percent over 300 locations.

The code to perform the Gibbs sampling simulations was written in Fortran 95 and made use of the numerical libraries of the Numerical Algorithms Group (NAG).

It was run on a AlphaServer 8400 with eight processors and three gigabytes of memory. Iterative linear system solvers using the sparseness and incomplete Cholesky factorization are supported by the NAG routines F011. Permutations for fill-in or bandwidth reduction were implemented in SPARSPAK available under www.psc.edu/∼burkardt/src/sparspak/sparspak.html which uses TOMS libraries 508 and 509 from ACM Collected Algorithms (CALGO). More modern fill-in minimizers are implemented in the METIS (www-users.cs.umn.edu/∼karypis/metis/) routine METIS_NodeND. Band solvers are included in the NAG routines group F07.

The CPU-time for running all the different algorithms, for a sample size of 300 locations at various levels of sparsity is summarized in figure 7.1. The algorithm which applies a band-solver to the spatially-ordered dataset is seen to be the fastest solution (Sampler B). The re-ordering of $\mathbf{\Sigma}$ at every Gibbs iteration (Sampler C) results in a computationally less efficient algorithm in comparison to ordering the distance matrix once before applying MCMC. However, both band-solvers (B and C) are faster than ordinary MCMC (Sampler A) especially for high level of sparsity. The re-ordering of the covariance matrix for every single component ϕ_i $i = 1, \ldots, m$, is clearly very slow (Sampler D). Similarly, sweeping (Sampler F) and the QMD algorithm (Sampler G) do not substantially profit from sparsity and perform badly.

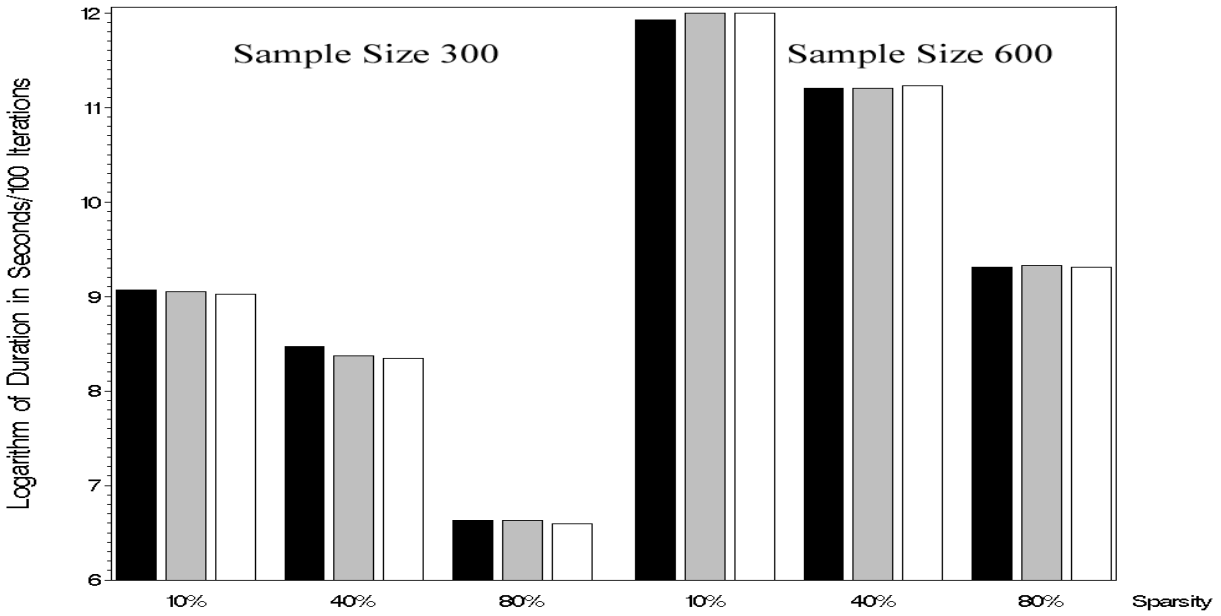


Figure 7.2: Iterative solver using incomplete Cholesky factorization for three different fill-in tolerance levels, 0.1 (black), 0.04 (gray) and 0.005 (empty). Simulated response data from normal distributions over 300 and 600 locations.

We have applied the iterative solver via incomplete Cholesky factorization algorithm using the NAG routine F11JAF. The routine reduces fills-in by replacing the elements a_{ij} of the factor matrix to zero if $\mid a_{ij} \mid < \texttt{tol}\sqrt{\mid a_{ii}a_{jj} \mid}$, where `tol` is a pre-specified tolerance parameter. We have chosen three tolerance levels of 0.005 (Sampler E5), 0.04 (Sampler E4)

and 0.1 (Sampler E1), respectively. The algorithms show good performance, if applied on a highly sparse covariance matrix. Although the incomplete-factorization sampler performs only slightly better than a crude, non-improved algorithm (Sampler A) in this example, the performance is clearly superior for larger sample sizes and high degree of sparsity as shown in figure 7.2. It also appears that the fill-in tolerance parameter `tol` does not affect much the computational time.

The iterative, incomplete-factorization sampler is the only non-exact inversion algorithm assessed in this study. A comparison of the estimates of the parameters of the geostatistical model fitted using the various solvers with the true values which generated the data, revealed no discrepancies. In addition the time-to-convergence which was assessed using the Raftery-Lewis convergence criterion (Raftery and Lewis, 1992) was similar in all compared algorithms.

The sequential decomposition algorithm of section 7.3.2 was seen to be not competitive in comparison to other improved algorithms and therefore no longer mentioned in this assessment. Similar holds for the Gibbs-sampling inversion algorithm of section 7.3.4, which is proposed for inverting very large covariance matrices. This algorithm requires extensive tuning of the number of iterations, the burn-in time and the thinning. The choice of this tuning-parameters influences the computing speed, but does not reduce it to a competing level for matrices with size up to 1,000 we considered in this study.

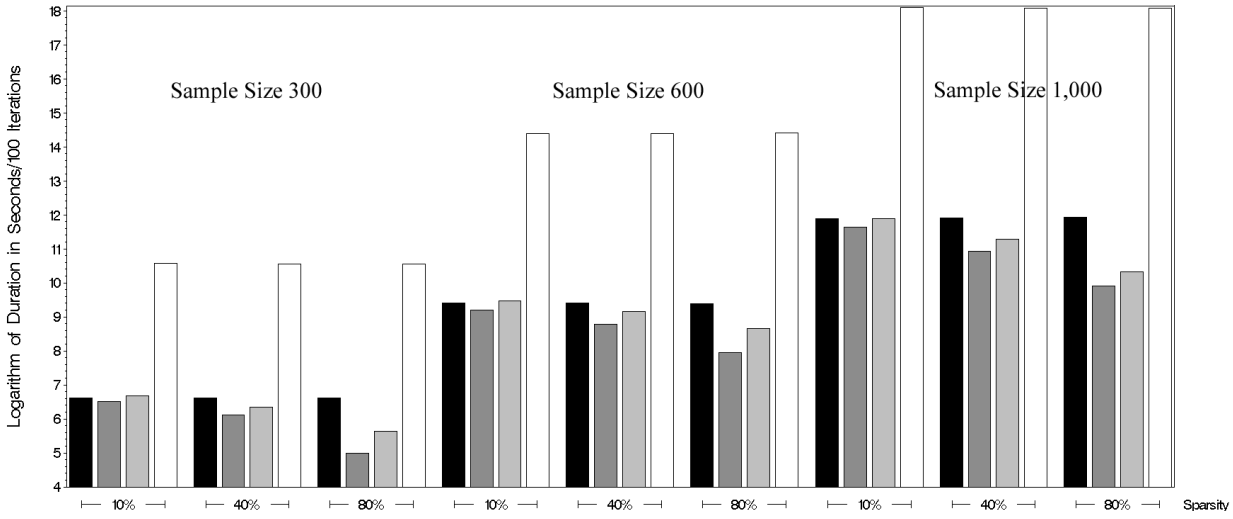


Figure 7.3: Comparison of ordinary MCMC algorithm (black) with those enhanced by three different band-solvers with ordered distance matrix (dark gray), with the GPS algorithm once per Gibbs iteration (light gray) and with the GPS algorithm in every random effect update (empty bar). Simulated response data from a normal distribution over 300, 600 and 1,000 locations.

The band-solvers which rely on sparse matrix techniques are compared in figure 7.3. A re-ordering of the covariance matrix for every ϕ_i $i = 1, \ldots, m$, is seen to be very slow (empty bar). Ordering the distance matrix before implementing MCMC leads to faster

computations than ordering the covariance matrix in each MCMC iteration. The gain in speed increases with the sparsity level and it is independent of the sample size.

7.4.2 Study II

In the previous simulation study we compared empirically the performance of the various algorithms for inverting $\boldsymbol{\Sigma}_{\text{-}i,\text{-}i}$ when the response data are normally distributed. The normal distribution of the spatial random effects is conjugate and it leads to conditional posterior distributions of known forms only in the case of the normal first stage model. In this study, we simulated data $\boldsymbol{Y}(\boldsymbol{s})$ from a Poisson distribution and assessed the benefit of implementing matrix inversion with the band-solvers when sampling of the whole vector of $\boldsymbol{\phi}$ at once was performed. We followed again the same model specification described in section 7.4.1, where $\boldsymbol{Y}(\boldsymbol{s}) \sim Po(\exp(\boldsymbol{\mu}))$ and $\boldsymbol{\mu} = \boldsymbol{X}^t\boldsymbol{\beta} + \boldsymbol{\phi}$. For the non-normal data we have in this model, the nugget effect τ^2 can be only specified in the covariance matrix of $\boldsymbol{\phi}$ as described in7.2. Instead of updating the parameter $\boldsymbol{\phi}$ component-wise, we update the whole vector as block. The conditional distribution of $\boldsymbol{\phi}$ is given by $p(\boldsymbol{\phi}; \boldsymbol{Y}) = L(\boldsymbol{Y}; \boldsymbol{\phi}, \boldsymbol{\beta}) \times p(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2)$ where $p(\boldsymbol{\phi} \mid \sigma^2, \delta, \tau^2) \equiv N(\mathbf{0}, \boldsymbol{\Sigma})$ and $L(\boldsymbol{Y}; \boldsymbol{\phi}, \boldsymbol{\beta}) \propto \prod_{i=1}^{m} (\boldsymbol{X}^t\boldsymbol{\beta} + \phi_i)^{Y_i} \exp(-(\boldsymbol{X}^t\boldsymbol{\beta} + \phi_i))$, which has not standard form. We sampled from this distribution using the Langevin-Hastings algorithm, discussed in section 7.2.2 and we compared the speed of updating $\boldsymbol{\phi}$ using the ordinal Langevin-Hastings without acceleration in the inversion of $\boldsymbol{\Sigma}$ and that of Langevin-Hastings with the band solver applied on the ordered distance matrix which was shown to be the fastest algorithm in the evaluation of study I (section 7.4.1). The score vector required for computing the mean of the multivariate normal distribution according to the Langevin-Hastings algorithm is: $\nabla \log p(\boldsymbol{\phi}; \boldsymbol{Y}) = \sum_i Y_i - \exp(\boldsymbol{X}_i^t\boldsymbol{\beta} + \phi_i) - \boldsymbol{\Sigma}^{-1}\phi_i$.

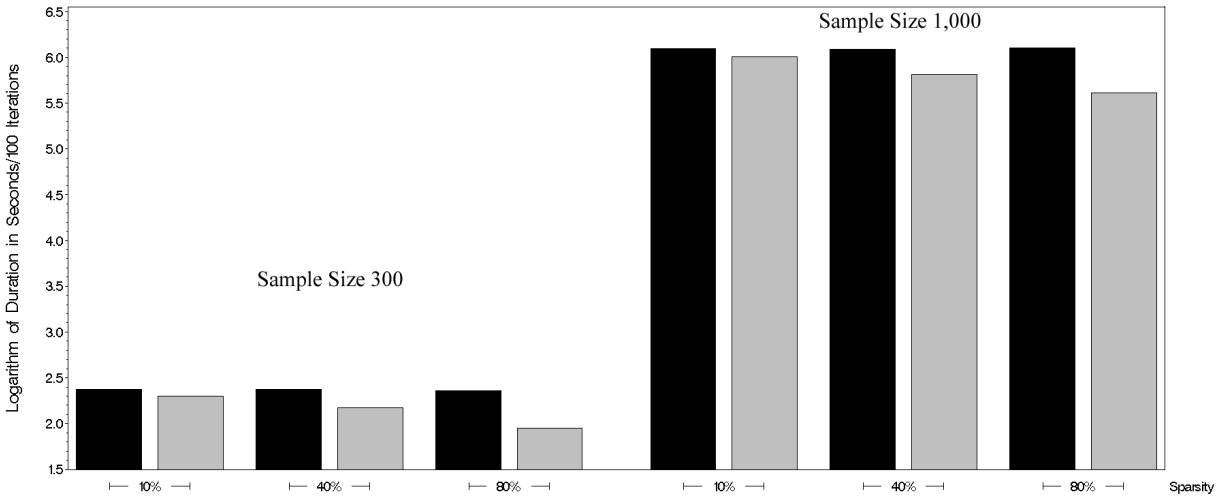


Figure 7.4: Comparison of the ordinary Langevin-Hastings algorithm (black) and the Langevin-Hastings with band solver applied on the ordered distance matrix (gray). Simulated data from a Poisson distribution over 300 and 1,000 locations.

Figure 7.4 displays results of the comparison for simulated datasets over 300 and 1,000

locations and with three levels of sparsity (10 percent, 40 percent and 80 percent, respectively). The results show that the sparse version saves 33.6 percent processing time for 300 locations and 38.4 percent for 1,000 locations in comparison to the ordinary the Langevin-Hastings sampler when the distance matrix has 80 percent sparsity. For very low sparsity of 10 percent the reduction in processing time is 7.1 percent and 7.9 percent for 300 and 1,000 locations, respectively. Therefore there is a computational advantage in implementing the algorithm together with the band solver especially for large datasets. Langevin-Hastings however requires extensive tuning by the user to achieve a reasonable Metropolis-Hastings acceptance ratio. This is controlled by the parameter a (see section 7.2.2) and it was found that small changes in a greatly affect the acceptance rate.

7.5 Discussion

In this article, we describe several algorithms to invert the large covariance matrices involved in fitting Bayesian geostatistical models. We relied on ideas from several authors (Barry and Pace, 1997; Rue, 2000) using sparse matrix techniques for areal data and extended those to models for geostatistical data. Using simulation studies, we tried to find out what can be gained by an improved procedure in terms of computer-processing speed and how much time will be spent by the researcher to implement an improved algorithm. In terms of computing speed, the results are extensively discussed. The algorithms are comparable in terms of time and effort to implement, however it was more difficult to implement those algorithms with the slower computational speed.

The sparse structure of geostatistical covariance matrices makes the computation via sparse linear system solvers efficient in MCMC estimation. A simple reordering of the dataset by latitude and longitude and the choice of a specific correlation function with zero correlation beyond the range already saves a substantial amount of CPU-time. Single random effect updating is seen to be not competitive with the block update of random effects, as demonstrated by the Langevin-Hastings sampler. Although the block updating strategy makes the largest contribution to computational speed, the sparsity of the dataset should further be considered. A sparse solver can be embedded in the block sampler without much effort and reduces the computational time substantially.

With the exception of the iterative solver of section 7.3.4, all other algorithms are exact, therefore neither the accuracy of the results nor the time to the MCMC convergence are affected. In fact we controlled the length of burn-in time, autocorrelation and time to convergence. In all examples, mixing was good, resulting in a negligible burn-in period and vanishing autocorrelation, even at small lag. The parameter were always estimated correctly and there were no differences in the estimates between the various samplers.

In applications to real geostatistical data, the aggregation of near locations needs to be considered. This reduces the computational burden, proportionally to the number of spatial random effects. It can further prevent instabilities in the computation, because a correlation matrix with some very small distances can be substantially ill-conditioned. Only if all distances are at a similar scale, the range parameter δ, can be accurately estimated.

Acknowledgement

This work was supported by the Swiss National Science Foundation grant Nr. 3200–057165.99.

Appendix 7.A Details on simulated datasets

The locations of the points in the simulated datasets were chosen to form an elliptic shape, with higher density of points in its center. For m locations, this was achieved by specifying $n >> m$ points on a spiral with latitude$(i) = \gamma i \cos(i)$ and longitude$(i) = \gamma\alpha i \sin(i)$, $i = 1, \ldots, n$. Then a random subset of m points was chosen to form the final locations. We believe this design comes close to spatial sampling structures seen in field studies.

The distance matrix and spherical correlation function matrix for all locations was calculated based on a fixed value of the range parameter δ. The sparsity is evaluated by counting the number of zero values in the correlation function. To achieve the desired sparsity of 10, 40 and 80 percent, the parameter γ was adjusted iteratively to find the appropriate spatial design with the exactly desired sparsity-percentage.

For every location, we simulated an intercept and four covariates. They were all drawn from a normal distribution with specified mean and variance. We simulated σ^2 and τ^2 from a Gamma distribution and multiplied the correlation function by σ^2 and added τ^2 to compute the covariance. The spatial process was simulated from a multivariate normal with mean zero and spatial covariance matrix, which was computed by multiplying a vector of independent random normal variates by the Cholesky factor of the covariance matrix. For the study in section 7.4.2, the response variable was simulated from a Poisson distribution with mean parameter (in the log scale) set to be equal to the regression equation defined by the covariates and the spatial process.

CHAPTER 8

Modelling non-stationary geostatistical data using random tesselations; an application in mapping malaria risk

Gemperli A.[1], Vounatsou P.[1] and Gelfand A.E.[2]

This paper is being prepared for submission to *Journal of the Royal Statistical Society, Series C: Applied Statistics*.

[1] Swiss Tropical Institute, Basel

[2] Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708–0251

Abstract

Variogram models are based on stationary spatial processes. The stationarity assumption may not be justifiable when data are collected over wide areas because local characteristics could introduce a location-dependence in spatial correlation. Moreover fitting variogram models for non-Gaussian data involves repeated inversions of the spatial covariance matrix. For large number of locations, inversion may be slow and model fit infeasible within time constraints. To avoid the assumption of stationarity and facilitate the inversion of the covariance matrix, we developed a variogram model which is based on random Voronoi tesselations. In particular, we partition the area in subregions (tiles) and assume a separate stationary spatial process in each tile and independence between tiles. We follow a hierarchical modelling specification and obtain full Bayesian inference using reversible jump Markov chain Monte Carlo computation. The methodology is applied to describe variation in malaria endemicity in Mali, using data from the Mapping Malaria Risk in Africa (MARA) database.

Keywords: bayesian inference; geostatistics; malaria risk mapping; non-stationarity; reversible jump Markov chain Monte Carlo; variogram model; Voronoi tesselation.

8.1 Introduction

Plasmodium falciparum malaria remains the most important parasitic disease of humans. Its transmission is influenced by interactions between the parasite, the mosquito vector, the human host and the environment. Sub-Saharan Africa carries most of the burden of malaria disease with nearly one million deaths and 300–500 million clinical cases every year. Mapping the endemicity in different areas is essential for accurate estimation of disease burden, for purposes of resource allocation and for assessing intervention programs.

A widely used measure of malaria endemicity is the parasite prevalence estimated from human populations by surveys carried out at various locations. The most comprehensive database on malaria transmission is the Mapping Malaria Risk in Africa (MARA) database which includes malaria prevalence data since 1950 at over 10,000 locations in Sub-Saharan Africa, extracted from articles published in scientific journals, ministry reports and from unpublished work done by research institutions. Malaria prevalence data collected at survey locations are typically binomial geostatistical data. The geographical proximity introduces correlation between the observations which violates the independence assumption of standard statistical methods. Spatial dependence is present at short scales due to the transmission of malaria infection by the mosquitoes which fly over short distances as well as at large scales due to the effects of environmental factors which influence mosquito survival and thus malaria transmission.

Geostatistical models can introduce spatial correlation in the variance-covariance matrix $\mathbf{\Sigma}$ of location-specific random effects via a latent Gaussian spatial process. Under second order stationarity, $\mathbf{\Sigma}$ determines the variogram. For isotropic processes, the elements of $\mathbf{\Sigma}$ are specified by parametric functions of the distance between the corresponding

locations. Maximum likelihood-based estimation has major shortcomings. In particular, the asymptotic inference is not uniquely defined (Cressie, 1993; Stein, 1999), prediction at un-sampled locations (kriging) does not fully account for the parameter uncertainty, and the standard error of predicted values underestimates the true variability (Prasad and Rao, 1990; Zimmerman and Cressie, 1992; Booth and Hobert, 1998). Diggle et al. (1998) formulated the variogram model as a Bayesian hierarchical model and provided full Bayesian inference using Markov chain Monte Carlo (MCMC) computation. However MCMC implementation requires repeated inversions of the covariance matrix of the spatial process which for large number of locations can be infeasible within practical time constraints.

Few maps of malaria risks have been produced based on field prevalence data, partly because of lack of readily available statistical software to fit non-Gaussian, geostatistical data collected over large number of locations. Thomson et al. (1999) model malaria prevalence amongst children in the Gambia. They estimate spatial variation from the residual of a regression model to obtain asymptotically valid inference about marginal regression parameters. Spatial variation is only handled as nuisance parameter and it does not allow smooth spatial interpolation. Kleinschmidt et al. (2000) use kriging on the residuals of a logistic regression model to get a smooth malaria prevalence map in West Africa, but they do not account for estimation uncertainty in the regression parameters and spatial random effects. Fully Bayesian inference have been adopted by Diggle et al. (2002) who predict childhood malaria risk in the Gambia and by Gemperli et al. (2003b) who use malaria transmission models to map age-specific malaria risk and other transmission measures in Mali.

All geostatistical modelling of malaria so far has been based on the assumption of a stationary spatial process implying that the spatial correlation is a function of the distance and independent of location. This assumption cannot be justified when mapping malaria risk over wide areas since local characteristics related to human activities, landuse, environment and vector ecology influence spatial correlation differently at the different locations. Moreover, the degree of precision in which malaria surveys can be "geolocated" is likely to vary over the map. Geostatistical methods for modelling non-stationary data have received little attention. Sampson and Gottorp (1992) and Damian et al. (2001) relax stationarity by using thin-plate splines for space transformation to reach stationarity on the deformed plane. Recent approaches use a Gaussian white noise kernel convolution to build the non-stationary process where either a kernel (Higdon et al., 1998) or a Gaussian process (Fuentes et al., 2002) was formed as a smooth function of the locations. Another approach is to partition the space into random tiles and estimate for every tile either a constant, as done by Ferreira et al. (2002), or an independent Gaussian stationary process, as demonstrated by Kim et al. (2002). Kim et al. (2002) use Gaussian data with conjugate priors and show an application on permeability data.

In this paper, we extend the work of Kim et al. (2002) and model non-Gaussian malaria prevalence data using random tesselations. The number and locations of the tiles are unknown and we assume an independent spatial process in each tile. This approach effectively addresses non-stationarity as well as computational problems in inverting large covariance matrices since the covariance matrix is reduced to a block-diagonal form. This model has

a variable number of parameters and estimation can be handled via a reversible jump Markov chain Monte Carlo sampler (RJMCMC) (Green, 1995). In section 8.2 we describe the malaria data and the environmental predictors derived from remote sensing or local stations. The Bayesian model formulation together with implementation details is given in section 8.3. The results of the application on mapping malaria risk in Mali are presented in section 8.4. The computational efficiency of the proposed approach is assessed on simulated data in section 8.5. A discussion with final remarks and suggestions for future work is given in section 8.6.

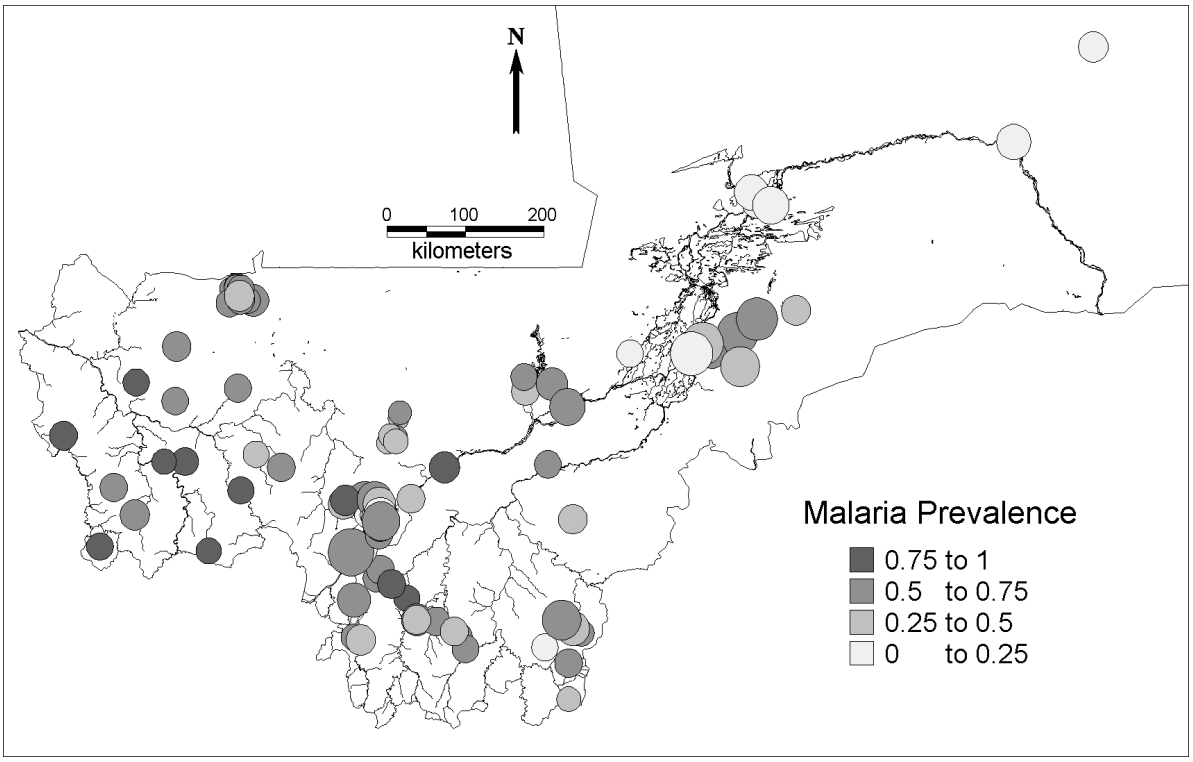


Figure 8.1: Sampling locations in Non-Saharan Mali with dot size related to the number of tested persons and its shading indicating the observed parasite ratio.

8.2 Data

The data which motivated this work were extracted from the "Mapping Malaria Risk in Africa" (MARA/ARMA, 1998) database. This is a unique database on malariological data in Africa collated from published and unpublished surveys carried out during the last 50 years in 44 countries in Africa. Todate, it contains malaria prevalence data over 10,000 locations. In this work, we analyzed prevalence data from malaria surveys carried out at 89 sites in Mali on children between 1 to 10 years old. The size of the surveys varied from 43 to 3,774 children. Children with presence of *Plasmodium falciparum* in blood smears

were considered as malaria positive. The distribution of the sampling locations in Mali is shown in figure 8.1.

Additional data on environmental predictors were collated from different sources. We used the same predictors as Kleinschmidt et al. (2000) who analyzed malaria endemicity in the same region. These include the distance to the nearest water source, the average maximum temperature between March and May, the length of rainy season specified by number of month with more than 60mm rainfall and the Normalized Difference Vegetation Index (NDVI). Data on rainfall and temperature were obtained at 5km resolution from the "Topographic and Climate Data Base for Africa" maintained by Hutchinson et al. (1996). NDVI data were obtained from the NOAA/NASA Pathfinder AVHRR Land Project (Agbu and James, 1994) at 8km resolution. We used the ten days composite NDVI values to avoid cloud-distortion and for each location we calculated an average NDVI over the eleven years period 1985–1995.

8.3 Model Specification

8.3.1 Stationary spatial process

Let N_i be the number of children screened during a particular survey at site $s_i \in D \subset R^2, i = 1, \ldots, m$, Y_i be the number of those found positive to *P. falciparum* parasitaemia and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^t$ be a vector of associated environmental covariates observed at s_i. Following the modelling framework of Diggle et al. (1998), we introduce unobserved spatial variation by assuming a latent spatial process $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_m)^t$ over the study region, and assume that conditional on ϕ_i the Y_i are independent random variables from a binomial distribution $Y_i \sim Bn(N_i, p_i)$ with parameter p_i measuring the malaria prevalence at s_i such as $\text{logit}(p_i) = \boldsymbol{X}_i^t \boldsymbol{\beta}^t + \phi_i$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$.

Furthermore we assume that $\boldsymbol{\phi}$ comes from an isotropic stationary Gaussian process over $\mathcal{D}$ with $\text{E}(\phi_i) = 0$ and $\text{Cov}(\phi_i, \phi_j) = \Sigma_{ij} = \sigma^2 \eta(\|s_i - s_j\|; \rho) + \delta_{ij}\tau^2$, where $\eta(\cdot)$ being a valid (non-negative definite) correlation function in R^2 and $\|s_i - s_j\|$ is the Euclidean distance between s_i and s_j and δ_{ij} the Kronecker delta. The parameter ρ measures the rate of correlation decay and it is known as the range parameter of the spatial process. We chose an exponential correlation function $\eta(\|\cdot\|; \rho) = \exp(-\|\cdot\| / \rho)$ because it is simple and leads to a nice epidemiological interpretation for ρ being one third of the minimum distance where the spatial correlation between locations falls below 0.05.

To complete the Bayesian model formulation, we adopt a vague normal prior distribution $\pi(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ with a pre-defined, large variance and inverse gamma priors for the parameters of the covariance matrix σ^2, ρ and τ^2. Bayesian inference is based on the joint posterior distribution of all parameters given by:

$$p(\boldsymbol{\phi}, \boldsymbol{\beta}, \sigma^2, \rho, \tau^2; \boldsymbol{Y}, \boldsymbol{N}) \propto L(\boldsymbol{Y}, \boldsymbol{N}; \boldsymbol{\beta}, \boldsymbol{\phi}) \det(\boldsymbol{\Sigma})^{-1} \exp(-\frac{1}{2}\boldsymbol{\phi}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}) \pi(\boldsymbol{\beta}) \pi(\sigma^2) \pi(\rho) \pi(\tau^2)$$

where $L(\boldsymbol{Y}, \boldsymbol{N}; \boldsymbol{\beta}, \boldsymbol{\phi})$ corresponds to the binomial likelihood, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)^t$ and $\boldsymbol{N} = (N_1, N_2, \ldots, N_m)^t$. We estimate parameters using Gibbs sampling, where the full

conditional distributions do not have standard forms. Calculation of the conditional distribution of $\boldsymbol{\phi}$ as well as that of ρ, σ^2 and τ^2 involve inversion of the $m \times m$ covariance matrix $\boldsymbol{\Sigma}$ which can slow down estimation, especially for large a number of locations m.

8.3.2 Non-stationary spatial process

The assumption of stationarity made in section 8.3.1 implies that spatial correlation between locations of the same distance remains the same throughout the region. This may not be true when we study large areas because local characteristics can influence spatial correlation differently in different parts of the map. In this work, we relax the assumption of stationarity, by partitioning the whole region in subregions or tiles and assume a separate stationary spatial process for each tile and independence between regions. Partitioning is based on Voronoi tesselation and it is random, allowing thus the data to choose the number and locations of the tiles.

Let $T_k, k = 1, \ldots, K$ be a subregion after partitioning the whole space into K regions and let ξ_k be the centroid of the Voronoi tile T_k. We define the tile T_k conditional on the size K of the partition as the set of all locations $s_i \in T_k$, such that $d(s_i, \xi_k) < d(s_i, \xi_l)$ $\forall l \neq k$, where $d(\cdot)$ is a distance measure (i.e. Euclidean distance). We re-arrange the vector of location-specific random effects $\boldsymbol{\phi}$, such that $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^t, \ldots, \boldsymbol{\phi}_K^t)^t$ where $\boldsymbol{\phi}_k$ is the sub-vector corresponding to the locations within T_k. In each tile T_k, we assume a Gaussian stationary process with covariance matrix $\mathrm{Var}(\boldsymbol{\phi}_k) = \Sigma_k$, that is $\boldsymbol{\phi}_k \sim N(\mathbf{0}, \Sigma_k)$ and $(\boldsymbol{\Sigma}_k)_{ij} = \mathrm{Cov}(\phi_{k,i}, \phi_{k,j}) = \sigma_k^2 \eta(\|s_i - s_j\|; \rho_k) + \delta_{ij}\tau_k^2$ for $s_i, s_j \in T_k$. We further assume independence between tiles, that is $\mathrm{Cov}(\boldsymbol{\phi}_k, \boldsymbol{\phi}_l) = \mathbf{0}$ for $k \neq l$. Thus $\mathrm{Var}(\boldsymbol{\phi}) = \boldsymbol{\Sigma}$ becomes a block diagonal matrix $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$.

The number and centroids of the partition are random and thus K and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^t$ are unknown parameters of the geostatistical model. We adopt a prior distribution for K which has the form, $\pi(K) \propto \alpha^{K-1}, \alpha \in (0, 1]$ and the hyperparameter α is pre-defined. This corresponds to a geometric distribution and is chosen in order to penalize large values of K (Cappé, 2002). For the parameter $\boldsymbol{\xi}$ we choose a uniform prior over the area of interest $\mathcal{A}$ which is $\pi(\boldsymbol{\xi}) \propto 1\{\boldsymbol{\xi} \in \mathcal{A}^K\}$. Small tiles with few sampling locations will lead to imprecise estimates of the corresponding random effects and spatial process parameters. In this case, estimation is driven by the priors rather than the data.

By combining the likelihood and prior distribution, the posterior will be

$$p(\boldsymbol{\phi}, \boldsymbol{\beta}, K, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2; \boldsymbol{Y}, \boldsymbol{N}) \propto L(\boldsymbol{Y}, \boldsymbol{N}; \boldsymbol{\beta}, \boldsymbol{\phi}) \prod_{k=1}^{K} \left[\det(\boldsymbol{\Sigma}_k)^{-1}\right] \exp(-\frac{1}{2} \sum_{k=1}^{K} \boldsymbol{\phi}_k^t \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\phi}_k)$$
$$\pi(\boldsymbol{\beta}) \prod_{k=1}^{K} \pi(\sigma_k^2) \prod_{k=1}^{K} \pi(\rho_k) \prod_{k=1}^{K} \pi(\tau_k^2).$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)^t$, $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_K^2)^t$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^t$. Parameter estimation is employed using Markov chain Monte Carlo (MCMC) conditional on the tesselation. Birth- and death-steps are added to allow moves between models of varying number of

tiles and they are estimated by reversible jump MCMC (Green, 1995) in order to guarantee reversability of the moves. Additionally we consider a step which updates the tesselation form conditional on the number of tiles. The sampling strategy is illustrated in the next section and a summary is given in the appendix.

Reversible Jump MCMC Computations

For a given tesselation with K tiles and the centroids $\boldsymbol{\xi} \mid K$, we proceed by suggesting one of four moves: 1) Keep the number of tiles K and their centers $\boldsymbol{\xi} \mid K$ at the current values and use Gibbs-Sampling to simulate from the conditionals of the rest of the parameters (Stay); 2) Keep K but choose one element ξ_k from $\boldsymbol{\xi}$ to move to a new location (Shift); 3) Add a new centroid ξ_{K+1} and increase K by one (Birth); 4) Reduce K by one by deleting a centroid ξ_k (Death). Stay, shift, birth and death steps are chosen with pre-specified probabilities Q_S, Q_H, Q_B and Q_D respectively.

The shift, birth and death steps alter the tesselation and change the dimension of the parameter space. The RJMCMC facilitates dimension-changing transitions by including a dimension-matching parameter and a function which deterministically relates the parameters between spaces of different dimensions in successive MCMC iterations. Thus, in the birth step at a given iteration t, we propose the new parameters $\boldsymbol{\theta}_{K+1}^{(t)} = (\sigma_{k+1}^{2(t)}, \tau_{k+1}^{2(t)}, \rho_{k+1}^{(t)}, \xi_{k+1}^{(t)})^t$ for the new tile $(K+1)$ as a weighted average of the parameters from the old tesselation, that is:

$$\sigma_{K+1}^{2(t)} = \frac{u_{\sigma^2}^{(t-1)}}{m^{(t)}} \sum_{k=1}^{K} w_k^{(t)} \sigma_k^{2(t-1)}, \tau_{K+1}^{2(t)} = \frac{u_{\tau^2}^{(t-1)}}{m^{(t)}} \sum_{k=1}^{K} w_k^{(t)} \tau_k^{2(t-1)}, \rho_{K+1}^{(t)} = \frac{u_{\rho}^{(t-1)}}{m^{(t)}} \sum_{k=1}^{K} w_k^{(t)} \rho_k^{(t-1)}, \xi_{K+1}^{(t)} = u_{\xi}$$

The $w_k^{(t)}$ are weights defined as the number of locations which are in tile T_k in iteration $t-1$ and fall into T_{K+1} in iteration t and $m^{(t)} = \sum_{k=1}^{K} w_k^{(t)}$. The weights $w^{(t)}$ and sum of weights $m^{(t)}$ are dependent on the current tesselation structure and therefore have to be calculated at every birth- or death-step. $\boldsymbol{u}^{(t-1)} = (u_{\sigma^2}^{(t-1)}, u_{\tau^2}^{(t-1)}, u_{\rho}^{(t-1)}, u_{\xi}^{(t-1)})^t$ is the dimension matching parameter which links the two parameter spaces via the function

$$g_{t-1,t}(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\theta}_{1\ldots k}^{(t-1)}, \boldsymbol{u}^{(t-1)}) = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\theta}_{1\ldots k}^{(t)}, \boldsymbol{\theta}_{k+1}^{(t)})$$

where $\boldsymbol{\theta}_{1\ldots k}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)})$. The parameters $u_{\sigma^2}^{(t-1)}, u_{\tau^2}^{(t-1)}, u_{\rho}^{(t-1)}$ are simulated from a log-normal distribution $b(\cdot)$ with mean equal to one and large variance. A new centroid $\xi_{K+1}^{(t)}$ is drawn with probability $1/m \sum_{i=1}^{m} h(\xi_{K+1}^{(t)} - s_i)$ and constructed by putting normal kernels $h(\cdot)$ around the sample points $s_i : i = 1, \ldots m$. This prevents the sampling of tile-centroids too far from the data locations and it increases the performance of the sampler in relation to a uniform proposal over the area $\mathcal{A}$.

A birth step, once proposed, is accepted with probability $\alpha_{\text{birth}} = \min(1, R_{\text{birth}})$ where

$$R_{\text{birth}} = \frac{p(\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\theta}_{1\ldots k}^{(t)}, \boldsymbol{\theta}_{k+1}^{(t)}; \boldsymbol{Y}, \boldsymbol{N}) Q_D d(\boldsymbol{\theta}_{k+1}^{(t)})}{p(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\theta}_{1\ldots k}^{(t-1)}; \boldsymbol{Y}, \boldsymbol{N}) Q_B b(\boldsymbol{u}^{(t-1)})} \left| \frac{\partial g_{t-1,t}(\boldsymbol{\theta}_{1\ldots k}^{(t)}, \boldsymbol{\theta}_{k+1}^{(t)})}{\partial(\boldsymbol{\theta}_{1\ldots k}^{(t-1)}, \boldsymbol{u}^{(t-1)})} \right|$$

$b(\boldsymbol{u}^{(t-1)})$ is the probability to sample the new vector $\boldsymbol{u}^{(t-1)}$ and $d(\boldsymbol{\theta}_{k+1}^{(t)})$ the probability that $\boldsymbol{\theta}_{k+1}^{(t)}$ is removed in a death step. The jacobian $\left|\frac{\partial g_{t-1,t}(\boldsymbol{\theta}_{1\ldots k}^{(t)}, \boldsymbol{\theta}_{k+1}^{(t)})}{\partial(\boldsymbol{\theta}_{1\ldots k}^{(t-1)}, \boldsymbol{u}^{(t-1)})}\right|$ used in the reversible jump Metropolis-Hastings ratio takes the simple form $\frac{1}{m^{(t)3}}(\sum_{k=1}^{K} w_k^{(t)} \sigma_k^{2(t-1)})(\sum_{k=1}^{K} w_k^{(t)} \tau_k^{2(t-1)})(\sum_{k=1}^{K} w_k^{(t)} \rho_k^{(t-1)})$. This design has the advantage that the parameters of a new tile are drawn using information from the previous tesselation instead of being drawn from an arbitrary distribution.

A death step is accepted with probability $\alpha_{\text{death}} = \min(1, R_{\text{death}})$ where R_{death} is the reciprocal of R_{birth} except that the indices t and $t-1$ switch places.

A proposal $\xi_k^{(t)}$ in the shift move $\xi_k^{(t-1)} \rightarrow \xi_k^{(t)}$ is drawn from a bivariate normal, centered at $\xi_k^{(t-1)}$, where k is selected uniformly from $\{1, \ldots, K\}$. Although a shift move may lead to a considerable change in the tesselation structure, the parameters $\sigma_k^{2(t-1)}$, $\tau_k^{2(t-1)}$ and $\rho_k^{(t-1)}$, $k = 1, \ldots, K$ are not altered in iteration k. The new tile and all its neighboring tiles, inherit all the information from its parent tile. This approach is therefore more likely to accept shifts to new tile centroids not too far from the old location. The transition probability for a shift move is $\alpha_{\text{shift}} = \min(1, R_{\text{shift}})$ where

$$R_{\text{shift}} = \frac{p(\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\theta}_{1\ldots k}^{(t)}; \boldsymbol{Y}, \boldsymbol{N})}{p(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\theta}_{1\ldots k}^{(t-1)}; \boldsymbol{Y}, \boldsymbol{N})}$$

8.3.3 Prediction

Prediction of malaria risk at unsampled locations can be obtained by Bayesian kriging. In particular, estimates of malaria prevalence $\boldsymbol{Y}_0 = (Y_{01}, Y_{02}, \ldots, Y_{0l})^t$ are obtained at a new set of locations $\boldsymbol{s}_0 = (s_{01}, s_{02}, \ldots, s_{0l})^t$ by the predictive distribution

$$\begin{aligned} p(\boldsymbol{Y}_0 \mid \boldsymbol{Y}, \boldsymbol{N}) \;=\; & \int p(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{\phi}_0) p(\boldsymbol{\phi}_0 \mid \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K) \times \\ & p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K \mid \boldsymbol{Y}, \boldsymbol{N}) d\boldsymbol{\beta} d\boldsymbol{\phi}_0 d\boldsymbol{\phi} d\boldsymbol{\sigma}^2 d\boldsymbol{\rho} d\boldsymbol{\tau}^2 d\boldsymbol{\xi} dK \end{aligned} \tag{8.1}$$

where $p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K \mid \boldsymbol{Y}, \boldsymbol{N})$ is the posterior distribution and $\boldsymbol{\phi}_0$ is the vector of random effects at $\boldsymbol{s}_0$. The distribution of the random effects $\boldsymbol{\phi}_0$ given $\boldsymbol{\phi}$ is normal

$$p(\boldsymbol{\phi}_0 \mid \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K) \equiv \mathcal{N}(\boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\phi}, \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10})$$

with $\boldsymbol{\Sigma}_{11} = \mathrm{E}(\boldsymbol{\phi}\boldsymbol{\phi}^t)$, $\boldsymbol{\Sigma}_{00} = \mathrm{E}(\boldsymbol{\phi}_0\boldsymbol{\phi}_0^t)$, $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^t = \mathrm{E}(\boldsymbol{\phi}_0\boldsymbol{\phi}^t)$ and $p(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{\phi}_0) = \prod_{i=1}^{l} p(Y_{0i} \mid \boldsymbol{\beta}, \phi_{0i})$ where $p(Y_{0i} \mid \boldsymbol{\beta}, \phi_{0i}) \sim Be(\pi_{0i})$, with $\mathrm{logit}(\pi_{0i}) = \boldsymbol{X}_{0i}^t\boldsymbol{\beta} + \phi_{0i}$ and $\boldsymbol{X}_{0i}$ is the vector of environmental covariates at location s_{0i}. Equation (8.1) is the expectation $\mathrm{E}[p(\boldsymbol{Y}_0 \mid \boldsymbol{\beta}, \boldsymbol{\phi}_0) p(\boldsymbol{\phi}_0 \mid \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K)]$ over the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K \mid \boldsymbol{Y}, \boldsymbol{N})$, which is estimated by the Gibbs sampler. Numerically this expectation is approximated by the average

$$\frac{1}{r}\sum_{q=1}^{r}\left[\prod_{i=1}^{l} p(Y_{0i}^{(q)} \mid \boldsymbol{\beta}^{(q)}, \phi_{0i}^{(q)})\right] p(\boldsymbol{\phi}_0^{(q)} \mid \boldsymbol{\phi}^{(q)}, \boldsymbol{\sigma}^{2(q)}, \boldsymbol{\rho}^{(q)}, \boldsymbol{\tau}^{2(q)}, \boldsymbol{\xi}^{(q)}, K^{(q)}),$$

where $(\boldsymbol{\beta}^{(q)}, \boldsymbol{\phi}^{(q)}, \boldsymbol{\sigma}^{2(q)}, \boldsymbol{\rho}^{(q)}, \boldsymbol{\tau}^{2(q)}, \boldsymbol{\xi}^{(q)}, K^{(q)})$ are samples drawn from the posterior $p(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\xi}, K \mid \boldsymbol{Y}, \boldsymbol{N})$. In the case of a stationary model, K is equal to 1 and the parameter $\boldsymbol{\xi}$ is not required.

8.4 Application

The spatial logistic model described in section 8.3 was applied to the malaria prevalence data from Mali. We use the Metropolis-Hastings algorithm to sample from the conditional distributions for all model parameters. Normal proposal distributions were used to simulate from the one-dimensional conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ parameters. Gamma proposal distributions were chosen for the elements of the parameter vectors $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}^2$ and $\boldsymbol{\rho}$. The mean of the proposal distributions were taken to be the values of the corresponding parameters drawn in the previous Gibbs iteration and the variance was iteratively adapted during the Gibbs implementation to optimize the Metropolis-Hastings acceptance probabilities.

We ran a single chain for 100,000 iterations. The starting values for the fixed effect parameters $\boldsymbol{\beta}$ were set equal to the estimates from an ordinary logistic model. The $\phi_i, i = 1, \ldots, m$ were initiated with zero. For the tile specific parameters $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}^2$ and $\boldsymbol{\rho}$ random numbers between 0.1 and 1 were given. We started the Gibbs sampler with 3 tiles and with randomly chosen centroids. The move probabilities Q_S, Q_B, Q_D, Q_H of the RJMCMC were chosen equal to 0.4, 0.15, 0.15 and 0.3, respectively. The parameter α in the prior distribution of K was set to one in order not to penalize a large value of K. All Metropolis-Hastings acceptance probabilities were between 0.28 and 0.72 (birth: 0.28; death: 0.35). Convergence in the log-likelihood was assessed by the Raftery-Lewis diagnostic (Raftery and Lewis, 1992) and was reached after 5,000 iterations. After convergence, samples from the posterior distribution were extracted during the stay move. To avoid autocorrelation in the samples, we have considered only every 19th of those values, and so obtained a final sample of size 2,000 from the joint posterior distribution.

Variable	Median	5% Quantile	95% Quantile
Intercept	-64.3326900	-64.8159800	-63.7351100
Maximum temperature[1]	10.3490400	10.2341300	10.4330600
Rainfall[1]	2.0309500	1.8081000	2.1641700
Water	0.0875700	0.0094600	0.2012900
Vegetation[1]	0.4446000	0.2830400	0.6029000

[1] Average of monthly values during malaria transmission season.

Table 8.1: Posterior estimates of the fixed effect parameters $\boldsymbol{\beta}$, with corresponding covariate at the log-scale.

Table 8.1 presents posterior estimates of the environmental covariates. As anticipated, the higher the average of maximum monthly temperature during the transmission season,

the higher the malaria risk. Similarly a positive association was found between malaria risk and the average amount of rainfall and vegetation during the malaria transmission period. The distance from the nearest permanent water source is also related to malaria transmission. The risk of malaria appears to be less in areas closer to water. The same result was found by Kleinschmidt et al. (2000) who analyzed the MARA data from Mali.

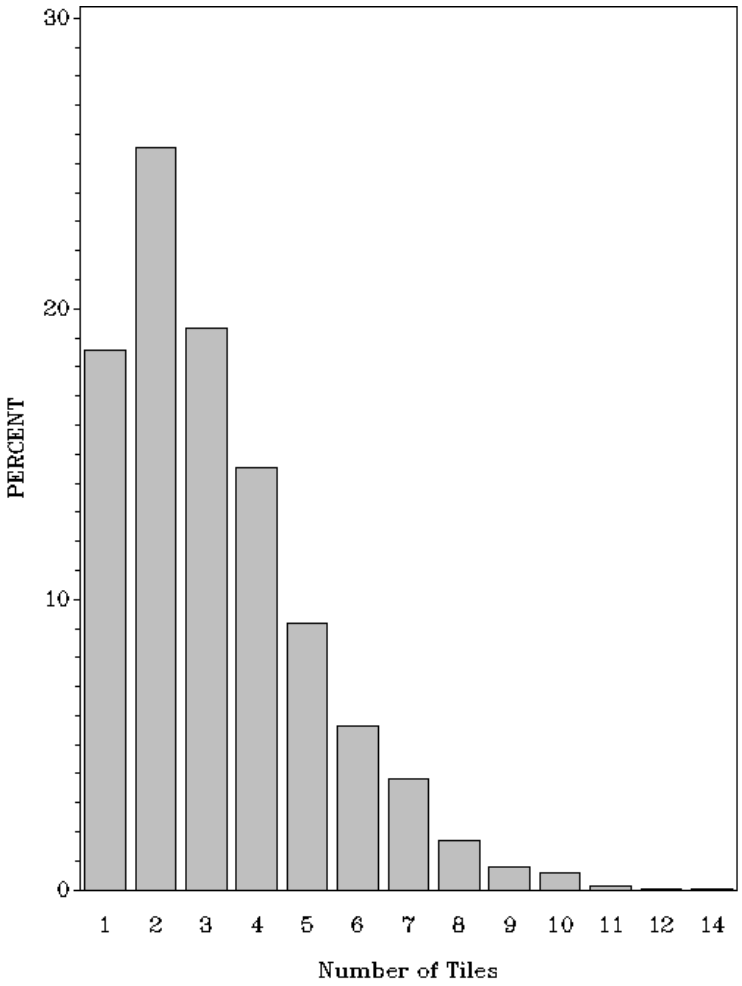


Figure 8.2: Frequency of the number of tiles.

Figure 8.2 depicts the posterior frequency distribution of the number of tiles parameter K. The most frequent tesselation favors two tiles, suggesting two separate spatial processes. Figure 8.3 displays summaries of the posterior distribution of the spatial covariance parameters $\boldsymbol{\sigma}^2, \boldsymbol{\tau}$ and $\boldsymbol{\rho}$. For each location on the map, the posterior distribution of the covariance parameters was obtained by the average of the tile-specific posterior distributions of the corresponding parameters. The range of $\boldsymbol{\rho}$ varies over the map from 0.038 to 2.74, indicating that the spatial correlation reduces to less than 5 percent at distances which vary from 1.14 to 82.2 kilometers. The spatial correlation is high in the densely populated area in Central-South Mali around Bamako (figure 8.3) with a smaller peak of

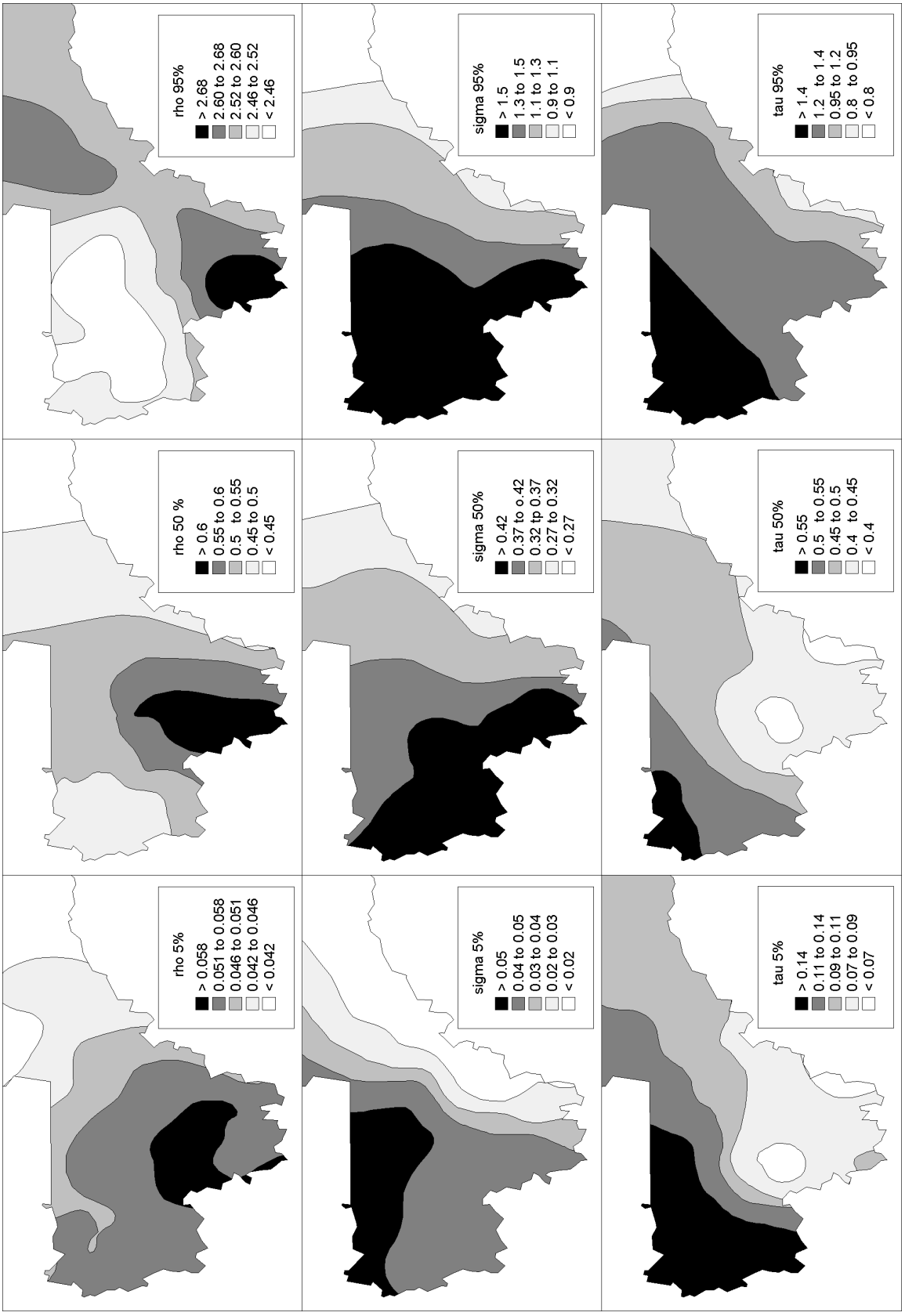


Figure 8.3: Distribution of covariance parameters.

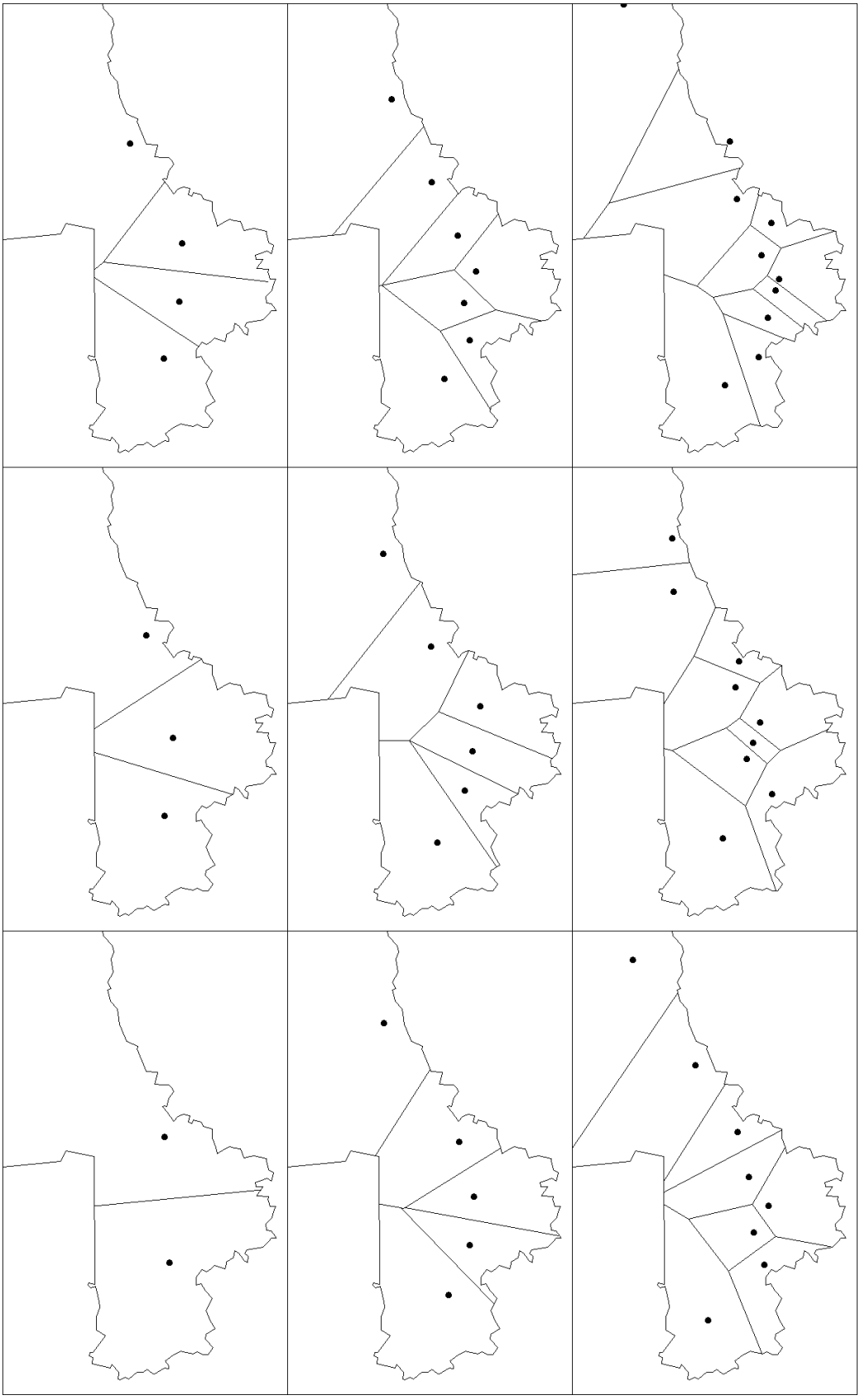


Figure 8.4: Average tesselation structure for 2 to 10 number of tiles.

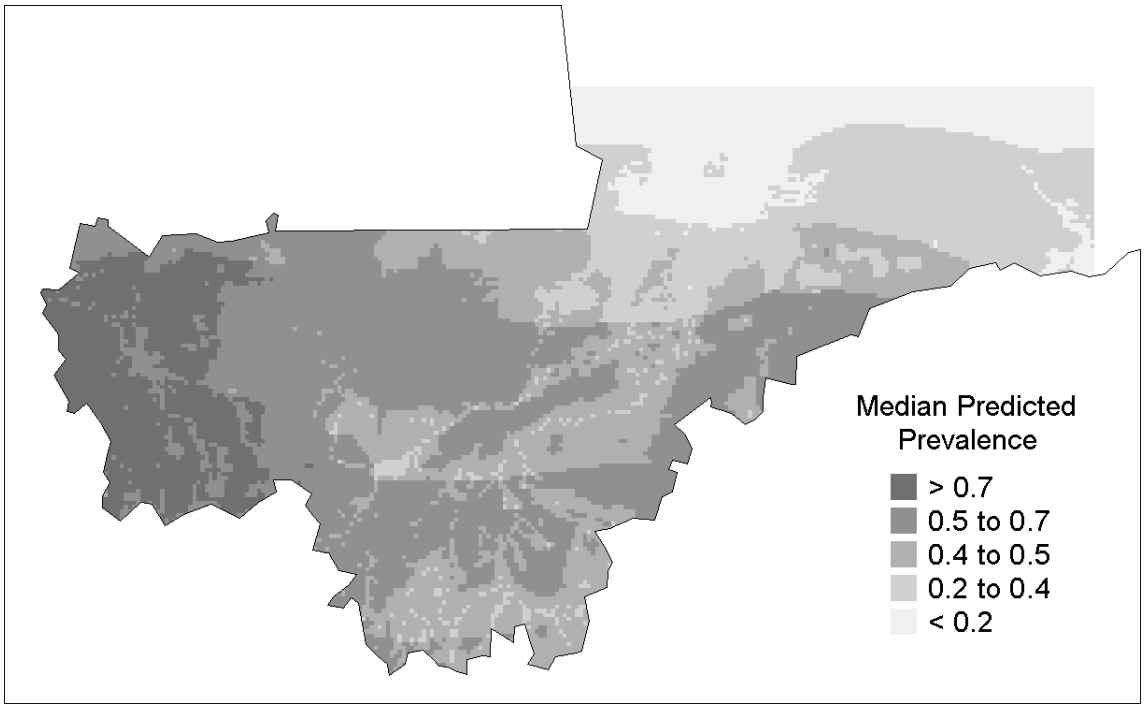
high correlation in the Niger delta region in the north (rho 95 percent) where conditions are known to be suitable for mosquito breeding sites. The spatial variance σ^2 ranges over the map from 0.002 to 2.4. It is high in the western part (sigma 95 percent) and low in the east part (sigma 5 percent). The non-spatial variance $\boldsymbol{\tau}^2$ is large in the central-west region (tau 95 percent) reflecting partly sparse and small surveys (figure 8.1). The areas with lowest non-spatial variation are those around Bamako where many and large surveys took place.

A smooth map of malaria risk in sub-saharan Mali is shown in figure 8.5a. The map is based on predictions over a regular grid of 40,000 locations. The malaria risk appears to become lower as we move from west towards east and from north towards south. The discontinuities in the malaria risk along the North-South directions parallels those of the length of the rainy season. The prediction error of the malaria prevalence in the logit scale is depicted in figure 8.5b. The error is larger at locations remote from the sampling locations in the Sahara-desert, in the North of the study area. Predictions have lower variance at locations near the sampling locations, such as around Nioro (North-West), Bamako (Central-South) and those close to the river Niger (Ségou, Mopti).

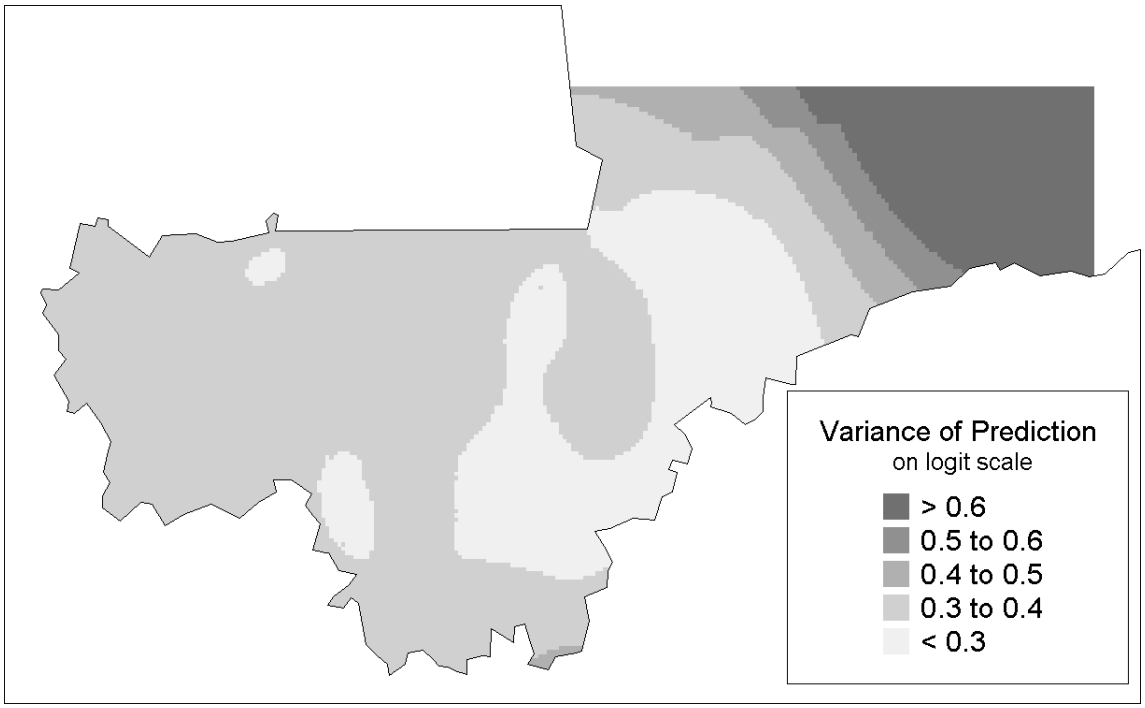
8.5 Assessing the computing performance on simulated data

We assessed the computing performance of the tesselation model on simulated geostatistical binomial data with 50, 100 and 200 locations, randomly chosen over a grid. We partitioned the grid in tiles with approximately equal number of locations and generated location-specific random effects from Gaussian distributions with spatial covariance matrices having different parameters between tiles. For each size of the dataset, we have chosen 10 tesselations with 1 to 10 tiles, respectively. In total we simulated 30 datasets, each included two covariates. For every location, we generated the survey size N from a Uniform distribution $U(10, 600)$. The binomial proportion was obtained on the logit scale as the sum of the covariates- and random-effects.

Table (8.2) shows the processing time for 1,000 iterations, for each dataset. The duration is reported relative to the baseline, which is the processing time required for the data set with 50 locations and one tile. The baseline time was 23.27 CPU-seconds for 1,000 iterations computed on a Pentium 4 PC with a 1.4 GHz processor and 386 MB RAM. The results demonstrate that the larger the number of tiles the smaller the computing time. In particular, a tesselation with 2 tiles reduces the computing time by 75 percent or 82 percent for data over 50 and 200 locations, respectively. This is because the numerical inversion of the covariance matrix of the spatial process for a dataset with m locations requires $\frac{1}{3}m^3 + 2m^2$ flops (Golub and Van Loan, 1996). A partition to K tiles converts the matrix to a block diagonal from, with blocks of size m_j by $m_j, j = 1, \ldots, K$ where m_j is the number of locations in the jth tile T_j. Thus the inversion of the block-diagonal covariance matrix requires $\frac{1}{3}\sum_{j=1}^{K} m_j^3 + 2\sum_{j=1}^{K} m_j^2$ flops.



a) Median



b) Variance

Figure 8.5: Predicted malaria prevalence.

Relative CPU-time in %			
Tiles	50 locations	100 locations	200 locations
1	100.00	850.25	12711.41
2	36.70	222.54	2293.01
3	22.31	116.69	1056.85
4	19.45	74.89	621.86
5	15.46	61.41	317.89
6	11.78	46.99	253.18
7	11.66	41.02	214.78
8	10.08	33.44	177.53
9	9.06	30.71	155.18
10	9.19	30.42	139.01

Table 8.2: Comparison of the computing performance for simulated dataset of different sizes. The baseline time is 23.27 CPU seconds for 1,000 iterations for a dataset with 50 locations and no partition (one tile) computed on a PC Pentium 4 with 1.4 Ghz.

8.6 Discussion

We have developed a geostatistical model for non-Gaussian response data which takes into account non-stationarity and facilitates model fit implemented via MCMC. The model divides the area in tiles and assumes a separate stationary spatial process in each subregion and independence between tiles. The assumption of independence converts the spatial covariance matrix to block diagonal form facilitating matrix inversion as the blocks have small size. The number and configuration of the tiles is random. The parameters of the model are estimated via RJMCMC. Maps of the distribution of the spatial covariance parameters can be produced by averaging the covariance parameters over all partitions. Model prediction can be obtained in a similar way. Averaging over the partitions prevents from estimating discontinuities in the predicted map.

A difficulty in the implementation of RJMCMC is the specification of proposal values for covariance parameters of new tiles introduced in birth moves. This is because there is no information about those parameters from previous MCMC iterations as these parameters did not exist. In our application the posterior distribution of all involved covariance parameters was rather widespread and we ended up with good Metropolis-Hastings acceptance rates. Finding good proposal distributions for new parameters derived in birth moves is a topic of current research in RJMCMC computation (Green and Mira, 2002; Rotondi et al., 2002; Brooks et al., 2003). However, more work is required to adapt general strategies to tesselation-based variogram models.

Further research is also needed in order to fully understand the behavior of the model. Applying it on simulated data derived from known tesselation designs would reveal whether

the model is able to capture correctly the different spatial processes. Model comparison between a stationary variogram model and a model which is based on random tesselations will show if parsimony is preferred over complexity.

The tesselation-based variogram model was applied in mapping malaria prevalence data in Mali. The non-stationary feature present in malaria data observed over large areas was never addressed previously. Local characteristics such as human activities, land use or malaria interventions can alter spatial correlation in different parts of the region. It is more likely to think of a mixture of spatial processes affecting large areas rather than a single process. Ignoring non-stationarity may partly explain differences between the various malaria maps produced so far. Proper modelling of the spatial process will lead to more accurate parameter estimation and prediction.

The random tesselation approach allows the data to decide on the number of tiles and thus the spatial processes. As long as we have large amount of data, inference are driven by the data rather than the prior specification. Maps of the spatial covariance parameters can be useful for control interventions. Regions where the spatial correlation reduces rapidly over short distances may indicate local unmeasured factors which influence malaria risk. Without additional information the model will not be able to find causal explanations. It will only identify the areas which have geographical dependencies over larger or shorter distances.

Acknowledgements

The authors would like to thank the MARA/ARMA collaboration for making the malaria prevalence data available. Spatial thanks to Nafomon Sogoba, MARA co-ordinator for the West Africa region. We are also thankful to NOAA/NASA Pathfinder AVHRR Land Project (University of Maryland) and the Distributed Active Archive Center (Code 902.2) at the Goddard Space Flight Center, Greenbelt, MD 20771 for the production and distribution of these data, respectively. This work was supported by the Swiss National Science Foundation grant Nr. 3200–057135.99.

Appendix 8.A RJMCMC sampler specification

Let K be the number of tiles at iteration $t-1$. Start iteration t by choosing one of the four moves, stay (S), birth (B), death (D) or shift (H) with probabilities Q_S, Q_B, Q_D, Q_H, respectively.

If S

- Employ a Gibbs sampling iteration, conditional on the tesselation ($K^{(t-1)}$ and $\boldsymbol{\xi}^{(t-1)}$), to update the parameters $\boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\phi}^{(t-1)}$ and $\boldsymbol{\theta}_{1\ldots K}^{(t-1)}$.
- Set $K^{(t)} = K^{(t-1)}$ and $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(t-1)}$.

If B

- Sample a new centroid $\xi_{K+1}^{(t)}$ and new parameters $\boldsymbol{\theta}_{K+1}^{(t)}$.
- Accept $\boldsymbol{\theta}_{K+1}^{(t)}$ and $\xi_{K+1}^{(t)}$ with probability α_{birth} and set $K^{(t)} = K^{(t-1)} + 1$, if rejected set $K^{(t)} = K^{(t-1)}$.
- Set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$ and $\boldsymbol{\phi}^{(t)} = \boldsymbol{\phi}^{(t-1)}$.

If D

- Remove a tile T_k by choosing k from a Uniform distribution over the set $\{1, \ldots, K\}$ and delete the corresponding centroid $\xi_k^{(t-1)}$ from $\boldsymbol{\xi}^{(t-1)}$.
- Accept the death step with probability α_{death} and set
  $K^{(t)} = K^{(t-1)} - 1$ and
  $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}_{-k}^{(t-1)}$, where $\boldsymbol{\xi}_{-k}^{(t-1)} = (\xi_1^{(t-1)}, \ldots, \xi_{k-1}^{(t-1)}, \xi_{k+1}^{(t-1)}, \ldots, \xi_K^{(t-1)})^t$.
  if rejected set $K^{(t)} = K^{(t-1)}$ and $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(t-1)}$.
- Set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$ and $\boldsymbol{\phi}^{(t)} = \boldsymbol{\phi}^{(t-1)}$.

If H

- Choose randomly a centroid $\xi_k^{(t-1)}$ to move to a new location $\xi_k^{'(t)}$.
- Accept the move with probability α_{shift} and set
  $\xi_j^{(t)} = \xi_j^{(t-1)}$, $\forall j \neq k$ and $\xi_k^{(t)} = \xi_k^{'(t)}$.
  if rejected set $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(t-1)}$.
- Set $K^{(t)} = K^{(t-1)}$, $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\phi}^{(t)} = \boldsymbol{\phi}^{(t-1)}$ and $\boldsymbol{\theta}_{1\ldots K}^{(t)} = \boldsymbol{\theta}_{1\ldots K}^{(t-1)}$.

CHAPTER 9

Conclusions

Motivated by epidemiological questions, novel statistical methods have been developed in this thesis for: 1) modelling non-Gaussian, large, geostatistical data; 2) modelling spatial survival data with misaligned covariates; 3) modelling non-stationary spatial data; 4) improving computational efficiency in fitting large geostatistical data. These methods have been applied to address important epidemiological questions, such as 1) mapping malaria prevalence data collected over a large number of locations and during different surveys; 2) assessing relationships between environmental factors and malaria endemicity; 3) assessing relations between malaria endemicity and child or infant mortality. The data for these analysis were extracted from established databases, i.e. the Mapping Malaria Risk in Africa (MARA/ARMA, 1998) database with survey information on malaria prevalence and the Demographic and Health Survey (DHS) database with mortality and socio-economic factors.

A detailed discussion on the findings was given in each chapter previously. Here we provide a summary of the main contributions and an outlook and recommendations for future research.

The statistical methods for geostatistical data developed in this thesis have contributed to: a) facilitating estimation of large, non-Gaussian geostatistical data; b) combining spatial information collected at non-matching locations; c) estimating geostatistical survival models; d) estimating geostatistical models with error-in-covariates; e) estimating and predicting non-stationary, non-normal spatial data.

In chapter 2, Bayesian methods for estimating non-normal geostatistical models are compared with maximum likelihood based techniques. This assessment confirmed the advantages of the Bayesian approach implemented via MCMC estimation. This approach was followed in the remainder of the thesis.

The main limitation in using MCMC estimation for non-normal, geostatistical data is the processing time required for inverting spatial covariance matrices for large number of

locations. To overcome this drawback, a number of numerical techniques appropriate for MCMC were investigated (chapter 7). This problem was further addressed via a tesselation-based geostatistical model (chapter 8), which partitions the region in random tiles and assumes a separate spatial process in each subregion and independence between tiles. This converts the spatial covariance matrix to block diagonal and reduces the size of the matrices to be inverted. The model allows in addition modelling of non-stationarity a feature which is likely to be present in mapping malaria data. There are other modelling approaches to account for non-stationarity (Fuentes et al., 2002; Higdon et al., 1998) but they do not facilitate the computation of the MCMC sampler.

Further research is needed in ways of summarizing the outputs of the tesselation-based variogram model. A visual presentation of the shape, size and configuration of the tiles would add information about the areas of the map with similar spatial dependencies. However the tesselations are random and summaries from the posterior distribution of the tesselation structure are not straightforward to obtain. Extensions of the model could allow for spatial interaction between neighboring tiles. This could be done by introducing tile-specific random effects modelled by conditional autoregressive models.

Country specific malaria maps have been obtained by Ribeiro et al. (1996), Kleinschmidt et al. (2000, 2001b) and Diggle et al. (2002). A map of malaria prevalence for West Africa has been produced by Kleinschmidt et al. (2001a). Continent wide maps of malaria transmission based on climate data only are available by the work of Craig et al. (1999) and Rogers et al. (2002). Although these mapping efforts have described malaria endemicity patterns, the derived maps have some drawbacks. They are either based on few sample points and cover a small area only, or they make use of a crude statistical techniques, without proper adjustment for prediction error.

The most complete work so far, from a methodological point of view, is comprised in the map of Kleinschmidt et al. (2001a). These authors combine climate information with local malariological measurements to produce smooth maps of malaria risk for West Africa. The map was divided into segregated zones and a separate model was fitted for every zone. This procedure better respects local characteristics, than fitting a single model for the whole area. But the separation of West Africa into four ecological zones seems arbitrary and non-smooth. Additionally, the post-processed smoothing along the edges of this zones, as suggested by Kleinschmidt et al. (2001a), hampers the calculation of prediction error.

The number of available malariological and environmental information for Africa has increased a lot during recent years and repeated updates of the maps are desired. The age dependency of malaria prevalence lead many researchers following the approach of leaving out surveys which do not report prevalence for a specific age-range (often 2 two 10 years old). The various documented surveys are not standardized by age of the study participants, and the procedure of omitting whole surveys with ineligible age categories, can results in a low sample size for the analysis. It should be the aim of a study to include as much as possible of the available data, by combining the information among distinct age categories.

The thesis makes an number of novel contributions to malaria mapping. Smooth maps of malaria transmission and prevalence have been produced for Mali (chapter 5) and for

West- and Central Africa (chapter 6). The maps were adjusted for environmental factors derived from remote sensing. An innovation in our work was the spatial modelling of parameters of malaria transmission models. This enabled us to overcome problems related to the data themselves. The surveys although report age-specific prevalence data, the age categories varied between locations making age-adjustment problematic. Previous attempts to map these data excluded all surveys with overlapping age groups. Using the Garki transmission model (Molineaux and Gramiccia, 1980), the prevalence data were converted to an estimate of entomological inoculation rate E. Smooth maps of E were converted to age-specific malaria risk maps using the age-dependent relation between prevalence and the E measure, described in the Garki model.

The Mali map assumed a constant transmission season for the whole country. For the West and Central Africa map a climatic suitability model was developed which allowed to estimate the length of transmission at each location. Seasonality was taken into account when the prevalence data converted to the E measure.

In this thesis, several maps of Mali are produced showing similar measures of malaria risk, based on different malariological and environmental datasets. Moreover, they are produced using distinct statistical techniques. Disparities are likely to by present because of the varying handling of the age groups or due to edge effects, when comparing local maps to maps produced on a larger scale. To assess the various proposed statistical methods, they need to be applied to the same dataset. Then a crossvalidation will reveal the best strategy when it comes to prediction. It further needs to be assessed, why different models found different relation between the distance to water bodies and malaria risk.

The produced maps in this thesis are found to be highly plausible when discussing them with local experts. Comparison with existing maps showed many areas with similar malaria patterns as well as a few zones with important disagreements. The disagreements could partly be explained by the different methodologies and different data used by the researchers. Additional field surveys will be required in areas with sparse data and large disagreement between our map and that of Kleinschmidt et al. (2001a). Large undersampled zones are found in the Democratic Republic of Congo, in the Central Africa Republic and in Nigeria, where especially the central part of Nigeria is an area of big differences between the two produced maps. A Nigerian MARA coordinator is currently examine data from the grey literature to help complete this part of the database.

entering data on forms, at the moment.

In this thesis malaria maps are drawn on a high resolution with environmental covariates on a resolution between one and eight kilometers and spatial smoothing on a resolution of one kilometer (chapters 5, 6 & 8). This resolution allows to pinpoint small regions to identify its predicted malaria risk. The methods quantify the risk even at remote locations, where no data are available or surveys are difficult to conduct. The maps help in the evaluation of health service provision, allow identification of appropriate malaria control tools and enable rational budgeting and timing of malaria control measures.

The established relations between environmental factors and malaria risk could predict changes in malaria risk in the presence of climatic changes of ecological transformations (i.e. building of dams, change in landuse). On the other hand the maps allow identification

of high risk areas, where man made interventions (i.e. drainage) could effectively reduce malaria transmission.

Spatial prediction can be further improved by the use of newly available remote sensing data. The European Space Administration (ESA) launched the Envisat satellite in March 2002. Envisat is the most powerful earth observation satellite and has begun making the most complete set of observations of our planet so far. In May 2002 NASA launched the Aqua satellite, which collects information about earth's water cycles, including water vapor in the atmosphere, clouds and precipitation. Once this data are available to researchers they may improve the predictive ability of the spatial malaria models.

The rainfall and temperature data used in this thesis were estimates from local stations and do not possess yearly differences. In order to derive temporal weather data, several researchers have suggested to estimate temperature and rainfall from satellite information (Thomson et al., 1996; Connor et al., 1998). It is possible to derive estimates of rainfall by measuring the cloud-top temperatures, which is measured by satellites. At a certain threshold temperature, the clouds will precipitate out into rainfall. By measuring the length of time a cloud is at this critical threshold temperature, known as the cold-cloud duration (CCD), it is possible to estimate the actual amount of rainfall (Milford et al., 1996). A complication is that this threshold temperature is depending on the location and a specific rainfall model only locally applicable. The land surface temperature (LST) can be estimated from the thermal channels measured by satellites (Franca and Cracknell, 1994; Prata et al., 1995; Coll and Caselles, 1997). This estimate may substantially differ from ambient (air) temperature and the relationship between these two indices is not straightforward. Both indices, the cold cloud duration and the land surface temperature are supposed to be highly correlated to the NDVI (Davenport and Nicholson, 1993; Thomson et al., 1996). Hay et al. (1998) uses these two indices together with the NDVI to model malaria season in Kenya and found only the NDVI to be an important predictor.

For information on water bodies, we used coordinates of rivers and lakes indicated as perennial in the African data sampler (World Resources Institute, 1995). Ideally that information should be available on a temporal base, too, what can be accomplished with a model which brings together water runoff, evaporation and precipitation such as landscape features (Patz et al., 1998). But the development of soil-wetness models is complex. The same holds true for population density estimates. In principle malaria risk should be related to population density, but the database we have used to model this (Deichman, 1996) could not confirm this assumption. We anticipate that a number of improved estimates will be available in the near future.

The maps for West- and Central Africa presented in chapter 6 are based on a seasonality model (Tanser et al., 2002; Hay et al., 1998) which defines the months suitable for stable malaria transmission for every location. The seasonality criterion based on the NDVI was developed based on data for Kenya only, and its generalization for West Africa is in question. Furthermore, make current malaria seasonality models only use of climate data and do not take into account clinical malaria data. There is an increasing amount of malaria incidence data available in Africa. Currently the MARA database contains malaria incidence mostly for Southern Africa. In areas where these data are available, they could

be used to improve the seasonality map.

The Garki model (Dietz et al., 1974) was used in chapters 5 and 6 to convert malaria prevalence data at each location to an estimate of transmission intensity. The Garki model was developed on field data from the savannah zone of Nigeria (Molineaux and Gramiccia, 1980). It is not clear how accurate are the predictions of this model in other regions in West- and Central Africa which have different environmental conditions and different levels of malaria endemicity. There is ongoing research in developing improved malaria transmission models. Parameters of those models could be used to improve the existing malaria maps. As an alternative, direct modelling of age-specific malaria risk could be accomplished. Such age-period-cohort models are complicated by the fact, that not only smoothing between locations, but also between age-categories need to be modelled (Lagazio et al., 2003).

The produced maps in this thesis were based on the assumption of no temporal changes in the spatial patterns of malaria. This assumption, although widely assumed in malaria mapping (Snow et al., 1997), is unlikely to hold true and needs justification. A second underlying assumption in the maps presented in chapters 5 and 6 is that of spatial stationarity. In stationary spatial models the spatial correlation in malaria risk is thought to be dependent on the distance between survey locations, but not on location. In chapter 8 a novel statistical approach for modelling non-stationary, geostatistical data was introduced. The model was applied to map malaria prevalence in Mali. Kleinschmidt et al. (2001a) separated West Africa into four ecological zones, following the directives of FAO (1978). These authors fitted for every zone a different model with distinct environmental factors, what results in a partly non-stationary model. An application of completely non-stationary spatial models for continent wide data, without reliance on an arbitrary, non-smooth space separation has not yet been done but is likely to improve the existing maps. The model presented in chapter 8 allows for non-stationarity in the spatial process only and does not handle varying influence of the environmental factors over space. In chapter 6, space-varying coefficients are modelled via interaction effects. It still needs to be assessed, if a model with non-stationary spatial variation and space-varying covariates can be combined, using the tesselation approach of chapter 8, to produce a smooth continent wide map.

The effectiveness of malaria control in Africa, in reducing child and infant mortality depends not only on the extent to which malaria endemicity is reduced but also on the relationship between endemicity and mortality. The relationship of malaria-specific mortality rates in infants and children has been compared to the level of malaria exposure by Smith et al. (2001) and Snow and Marsh (2002). These studies analyzed only a small number of published estimates of mortality, at a few specific locations and were not properly adjusted for ecological confounding.

A problem in relating malaria risk to death is that malaria may be a relevant risk factor for many deaths even when it is not the immediate cause (Molineaux, 1985). To account for this and the fact that verbal autopsies used to assign a cause of death are not very reliable (Snow and Marsh, 1998), we looked at the relationship of malaria endemicity with all-cause mortality in Mali. Local malaria indices were compared with overall infant mortality (chapter 3) and with overall child mortality (chapter 4). We analyzed mortality

data available from the demographic and health surveys (DHS) of 1995/96 and linked them to malaria prevalence from the MARA (MARA/ARMA, 1998) database. The estimates were adjusted for environmental factors in malaria risk, socio- and maternal factors in mortality as well as for geographical variation. The results did not clarify the relationship between mortality and malaria risk in Mali. No statistically significant relation was found and possible explanations were provided in chapters 3 and 4.

A main complication with the data was that the malaria prevalence and the mortality risk were not measured at the same locations. To overcome this spatial misalignment problem, the malaria risk was predicted at the locations with observed mortality data. This approach introduced additional uncertainty into the estimates due to the sparsity of the malaria data. The question regarding the relation between malaria and mortality may not be answered adequately without reducing this uncertainty.

In the analysis of child and infant mortality, a subset only of the existing malaria data was used in order to overcome the problem of age-heterogenous reported prevalence at the different locations. Future analyzes could utilize all available malaria data by using malaria transmission models to convert the prevalence data to a common age category as demonstrated in the chapters 5 and 6.

The spatial misalignment between locations with malaria prevalence and those with mortality data introduced extra variability in our model. To overcome this limitation, it is suggested to compile databases from a larger area of West Africa and analyze only data points were misalignment is minimal.

APPENDIX A

Databases used in the present work

A.1 The Mapping Malaria Risk in Africa database

The MARA/ARMA collaboration was initiated to provide an Atlas of malaria for Africa, containing relevant information for rational and targeted implementation of malaria control and was launched with its first workshop in 1996. Its main product is the MARA database with, to date, over 10,000 collected data points all over Sub-Saharan Africa. Malariological information is collected from published and unpublished sources, through literature searches and country visits. This data is entered into the MARA/ARMA database and checked via a double-entry validation system. Five regional centers, at existing institutions, are responsible for gathering malaria data in their region.

The MARA/ARMA initiative is non-institutional and runs in the spirit of an open collaboration. A group of dedicated African scientists, based at institutions across the continent, work co-operatively towards achieving the overall objectives. The Swiss Tropical Institute regularly makes contributions to the running data collection process by reporting data found in literature search or by presenting own survey results.

Detailed mapping of malaria risk and endemicity has never been done in Africa. Accurate estimates of the burden of malaria at regional or district level remain largely unknown. In the absence of such data it is impossible to rationalize allocation of limited resources for malaria control. The MARA/ARMA initiative intends not only to collect the data used for malaria mapping, but to foster scientific discussion on the topic, coordinate eduction and develop methodological work. Spatial statistical methods for use in the MARA/ARMA collaboration are to a great extent either developed or supervised by the Swiss Tropical Institute. Additionally, the Swiss Tropical Institute is an influential contributor of education in spatial statistical methods, with clear focus in malaria mapping for Africa.

As a result MARA/ARMA has provided the first continental maps of malaria distribution and the first evidence-base burden of disease estimates. There is currently hardly any major document on malaria in Africa that does not make use of MARA maps and the BOD figures produced by MARA/ARMA are now universally used.

A.2 The Demographic and Health Survey database

The U.S. Agency for International Development's (USAID) Bureau for Global Programs, Field Support and Research, Center for Population, Health and Nutrition (PHN Center) supports a 10-year results package entitled Monitoring and Evaluation to Assess and use Results (MEASURE). The strategic objective of MEASURE is to improve and institutionalize the collection and utilization of data by host countries for program monitoring and evaluation of and for policy development decisions. MEASURE activities support family planning, reproductive health, maternal health, child survival, and HIV/AIDS/STI control/prevention through data collection, analysis, and evaluation designed to improve program performance and to better understand program impact in these areas. As a key participant in this new MEASURE program, Demographic and Health Survey (DHS+) is specifically charged with the task of collecting and analyzing reliable demographic and health data for regional and national family planning and health programs. The DHS+ approach to data collection emphasizes integration, coordination, and cost-effectiveness. The Demographic and Health Surveys program is funded by USAID and implemented by Macro International Inc.

Historically the Demographic and Health Surveys (DHS) program is established 1984 at the Institute for Resource Development, Inc. (IRD), a subsidiary of the Westinghouse Electric Company. The DHS combines the qualities of the WFS and the CPS and adds important questions on maternal and child health and nutrition. The program can be subdivided into the phases DHS I (1984–1989), DHS II (1988–1993) and DHS III (1992–1999). In 1989 the Institute for Resource Development, Inc. was acquired by Macro International Inc. The name DHS has been changed to DHS+ in 1997 to reflect a new mandate under the MEASURE program. MEASURE DHS+ incorporates traditional DHS features with expanded content on maternal and child health.

To date, the DHS+ program has provided technical assistance for more than 100 surveys in Africa, Asia, the Near East, Latin America, and the Caribbean.

MEASURE DHS+ seeks to increase the utilization of population, health, and nutrition data for the monitoring and evaluation of programs. This is achieved through building of data collection systems in developing countries through formal and on-the-job training in research design and implementation, sampling, data processing, analysis, and dissemination.

Demographic and Health Surveys (DHS) are nationally representative household surveys with large sample sizes of between 5,000 and 30,000 households, typically. DHS surveys provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition. The core questionnaire for MEASURE DHS+

emphasizes basic indicators and flexibility. It allows for the addition of special modules so that questionnaires can be tailored to meet host-country and USAID data needs. The standard DHS survey consists of a household questionnaire and a women's questionnaire. A nationally representative sample of women ages 15–49 are interviewed. The household questionnaire contains information on the following topics:

- Household listing: For every usual member of the household and visitor, information is collected about age, sex, relationship to the head of the household, education, and parental survivorship and residence.

- Household characteristics: Questions ask about the source of drinking water, toilet facilities, cooking fuel, and assets of the household. There are additional questions about the use of bednets in the household.

- Nutritional status and anemia: The height and weight of women age 15–49 and young children are measured to assess nutritional status. For the same individuals, the level of hemoglobin in the blood is measured to assess the level of anemia.

It further comprises a large women's questionnaire with information on many topics, such as: Reproductive behavior and intentions; Contraception; Antenatal, delivery, and postpartum care; Breastfeeding and nutrition; Children's health; Status of women; AIDS and other sexually transmitted infections; and Husband's background.

The various types of survey conducted by Macro Inc. can be broadly subdivided into:

Demographic and Health Surveys (DHS)	Nationally representative large surveys as described above.
Interim Surveys	Shorter questionnaires than DHS with focus on key performance monitoring parameters
Baseline/Follow-up Surveys	Subnational surveys designed to cover a limited number of indicators for specific projects; less standardized.
Service Provision Assessments (SPA)	Surveys conducted in health facilities and communities to obtain information about the health and family planning services available in a country.
Geographic Data Collection	Collection of geographic locations (Latitude/Longitude), started in 1996, for the communities of the DHS.
Additional Surveys	Country specific surveys designed to obtain specialized information from a population subgroup. May be required to link surveys.

A.3 The NOAA/NASA pathfinder AVHRR land data sets

The NOAA/NASA Pathfinder AVHRR Land data sets contain global, land surface parameters derived from the Advanced Very High Resolution Radiometers (AVHRR) on the "afternoon" NOAA operational meteorological satellites (NOAA-7, -9, -11). The Pathfinder Program, initiated by NOAA and NASA, produces long-term data sets processed in a consistent manner for global change research. The data cover the period from July 1981 through the present.

There are three types of data provided by the Pathfinder AVHRR Land production system. These are the Daily, 10-Day Composite, and Climate data sets.

The Daily Data Set contains global, 8km data mapped to an equal area projection. Geophysical parameters contained in the data set include: Normalized Difference Vegetation Index (NDVI), cloud and quality-control flags, solar and scan geometry, reflectances derived from the AVHRR channels 1 and 2, brightness temperatures derived from the AVHRR channels 3, 4, and 5, and date and hour of observation. NDVI is a ratio of the contrast between the responses of the two reflective channels. Data over oceans, large inland water bodies, and in areas of twilight are derived directly from the AVHRR level 1B orbital data. There is one file per day for the entire Pathfinder processing period (June 25, 1981, to present). The Daily Data Set is useful for studies of many terrestrial variables (e.g., vegetation, temperature, snow cover) as well as for producing a variety of composite data sets, but each day a significant portion of the Earth's surface is covered by clouds.

The Composite Data Sets are 10-day composites of the same geophysical parameters as the Daily Data, and these data are mapped to the same global, 8km, equal area projection as the Daily Data. To minimize the effects of clouds and atmospheric contaminants, the composite selects the observation for each 8km x 8km bin within a 10 day period that has the fewest clouds, as identified by the highest NDVI value. Only data within 42 degrees of nadir are used in the composite to minimize spatial distortion and bidirectional effect biases at the edge of a scan. There are three composites per month. The first composite of each month is for days 1 through 10, the second is for days 11 through 20 and the third is for the remaining days.

A Composite Data Set is useful for studies of temporal and interannual behavior of surface vegetation and for developing surface background characteristics for use in climate modelling. The NDVI of the Composite Data Set was primarily used in the current work. Sometimes it was further processed or aggregated.

The Climate Data Set contains global NDVI data derived from mean Channel 1 and 2 reflectances. They are equal angle data at 1 degree latitude by 1 degree longitude resolution for each 8- to 11-day composite period. These data are derived from the Composite Data, and there are 36 climate data files for each year. This data set is intended primarily for use in Global Climate Models (GCM), Simple Biosphere Models, and other global time series studies.

The nominal orbit parameters for the NOAA-series satellites (NOAA-7, -9, -11) are:

Launch Date	6/23/81 (NOAA-7), 12/12/84 (NOAA-9), 9/24/88 (NOAA-11)
Orbit	Sun-synchronous, near-polar
Nominal Altitude	833 kilometers
Orbit Inclination	98.8 degrees
Orbital Period	102 minutes
Equator Crossing Time	14.30 (NOAA-7), 14.20 (NOAA-9), 13.40 (NOAA-9) LST
Nodal Increment	25.3 degrees

The NOAA-series satellites carry the AVHRR instruments. The orbital period of about 102 minutes produces 14.1 orbits per day. Because the daily number of orbits is not an integer, the suborbital tracks do not repeat daily, although the local solar time of the satellite's passage is essentially unchanged for any latitude. The 110.8 degrees cross-track scan equates to a swath of about 2700 km. This swath width is greater than the 25.3 degrees separation between successive orbital tracks and provides overlapping coverage (side-lap).

The spectral band widths and Instantaneous Field of View (IFOV) of the AVHRR instrument are given in the following table:

Channel	Wavelength (micrometer)	IFOV (milliradian)
1	0.58–0.68	1.39
2	0.73–1.10	1.41
3	3.55–3.93	1.51
4	10.3–11.3	1.41
5	11.5–12.5	1.30

A more detailed, comprehensive description of the NOAA series satellites, the AVHRR instrument, and the AVHRR GAC 1B data can be found in the "NOAA Polar Orbiter Data User's Guide" (Kidwell, 1991), which can be obtained from NOAA's National Environmental Satellite Data and Information Service (NESDIS).

All Pathfinder AVHRR Land data are stored in the Hierarchical Data Format (HDF) (Brown et al., 1993). HDF allows data (scientific data and metadata) to be implemented in several ways including Scientific Data Sets (SDS) and 8-bit Raster Image (RIS8) data sets. The SDS implementation has more flexibility in including metadata and allows data of a variety of word sizes (8- to 64-bit data).

Detailed information on data organization and interpretation is available in the "NOAA/NASA Pathfinder AVHRR Land Data Sets User's Manual" (Agbu and James, 1994).

The Daily Data Set contains data mapped to the Goode Interrupted Homolosine equal area projection (Steinwand et al., 1992). The table below shows the Daily Data Set structure and lists the parameter names, units, and field widths in bits. To obtain geophysical values from the data, take the value in the data, subtract the offset, and multiply by the gain.

Parameter	Unit	Field Width (bits)	Offset	Gain
NDVI	-	8	128	.008
CLAVR Flag*	-	8	1	1
Quality Control Flag$^{\sharp}$	-	8	1	1
Scan Angle	Radians	16	10481.98	.0001
Solar Zenith Angle	Radians	16	10	.0001
Relative Azimuth Angle	Radians	16	10	.0001
Ch1 Reflectance	%	16	10	.002
Ch2 Reflectance	%	16	10	.002
Ch3 Brightness Temp.	Kelvin	16	-31990	.005
Ch4 Brightness Temp.	Kelvin	16	-31990	.005
Ch5 Brightness Temp.	Kelvin	16	-31990	.005
Day of Year	DDD.HH	16	10	.01

* Note: The table below shows the CLAVR flag values and the conditions that relate to them.

Value	Condition
0	No decision
1–11	Cloudy
12–21	Mixed
22–30	Clear

$^{\sharp}$ The table below shows the Quality Control flag values and the conditions that relate to them.

Value	Condition	Condition
0	Normal	
1	Channel 1, 2 processing nonstandard	Ozone values unavailable, so climatology was used
2	Channel 3, 4, 5 processing nonstandard	Calibration coefficients unavailable
4	Filled data gap	A data gap resulting from forward transform used in binning has been filled with adjacent pixel
8	Range Check failure	Calculated values were outside the range of values
16	NOAA QC flag set	See "NOAA Polar Orbiter Data User's Guide" (Kidwell, 1991)

The Composite Data are implemented as HDF SDS with the same 8km x 8km dimensions as the Daily Data. But the Climate Data, also implemented as HDF SDS, has dimensions of 1 degree by 1 degree. In the Daily, Composite, Climate, and Ancillary Data

a binary flag value of 1 indicates ocean data, a value of 2 indicates the interrupted space in the equal area projection, and 0 indicates land or missing data over land.

Additionally there is the Pathfinder Ancillary File which contains land/sea flags, elevation (meter), and bin center latitude (degree) and longitude (degree), for global 8km bins that have all been coregistered to the same 8km equal area projection as the daily and composite data.

The production and distribution of this data set are being funded by NASA's Mission To Planet Earth Program. The data are not copyrighted, however it is obligatory to acknowledge the source for publications.

A.4 A topographic and climate data base for Africa

The "Topographic and Climate Data Base for Africa" Version 1.1 (Hutchinson et al., 1996) contains gridded values of elevation (DEM) and monthly mean climate for the African continent at a spatial resolution of 0.05 degrees of longitude and latitude. The climate consists of monthly mean and annual mean values of rainfall, daily minimum temperature and daily maximum temperature.

The DEM and the climate grid files were created using spatial analysis and interpolation techniques developed by the Center for Resource and Environmental Studies (CRES) at the Australian National University.

The climate grids were obtained by first fitting topographically dependent climate surfaces to point climate data using procedures in the ANUSPLIN package (Hutchinson, 1991; Hutchinson and Gessler, 1994). The surfaces were then interrogated using elevations from the DEM using the ANUCLIM package (McMahon et al., 1995).

Both elevation and climate data were subjected to comprehensive error detection and correction procedures based on ANUDEM and ANUSPLIN. Accurate geocoding (longitude, longitude and elevation) of climate station data was completed by CRES for many stations.

Monthly mean values of rainfall, daily minimum temperature and daily maximum temperature at a sufficient spatial density to support reliable spatial interpolation were compiled. In addition to data already obtained by CRES from miscellaneous sources, monthly climate data were acquired from research agencies including CIMMYT, FAO, East Anglia Climate Research Unit, CSIRO Division of Forestry, Texas A&M University and from the national meteorological services of Djibouti, Gambia, Ghana, Kenya, Malawi, Morocco, Nambia, Rwanda, Seychelles, Sudan, Tanzania, Uganda and Zaire.

Data were collected over all available years of record to maximize spatial coverage, subject to the condition that rainfall averages were for at least five years or record. Most data were collected between about 1920 and 1980 for both temperature and rainfall, so the fitted climates grids can be interpreted as estimates of standard means for the period 1920 to 1980.

The number of accurately geocoded stations for which monthly mean climate data were obtained were as follows:

Climate Variable	Number of Stations
Daily minimum temperature	1,504
Daily maximum temperature	1,499
Rainfall	6,051

The error of the climate grids depends mainly on the accuracy of the underlying climate surfaces. In using the DEM to calculate the climate grids, the stated errors in the DEM of up to a few hundred meters make only a minor additional contribution to errors in the climate grids. The standard errors of the temperature are about 0.5 degrees centigrade. The standard errors of the rainfall grids range between about 5 and 15 per cent, depending on data density and the spatial variability of the actual monthly mean rainfall.

Bibliography

Agbu P.A.,James M.E. (1994) *NOAA/NASA Pathfinder AVHRR Land Data Set User's Manual*, Goddard Distributed Active Archive Center, NASA Goddard Space Flight Center, Greenbelt

Alexander N.,Moyeed R.,Stander J. (2000) Spatial modelling of individual-level parasite counts using the negative binomial distribution, *Biostatistics* **1** 453–463

Anderson J.R.,Hardy E.E.,Roach J.T.,Witmer R.E. (1979) A land use and land cover classification system for use with remote sensor data, *US Geological Survey Professional Paper* **964**

Anderson D.A.,Hinde J.P. (1988) Random effects in generalized linear models and the EM algorithm, *Communication in Statistics: Theory and Methods* **17** 3847–3856

August P.J.,Labash M.C.,Smith C. (1994) GPS for environmental applications: accuracy and precision of location data, *Photogrammatic Engineering and Remote Sensing* **60** 41–45

Banerjee S.,Gelfand A.E. (2002) Prediction, interpolation and regression for spatially misaligned data, *Sankhya* **64** 227–245

Banerjee S.,Wall M.M.,Carlin B.P. (2003) Frailty modelling for spatially correlated survival data, with application to infant mortality in Minnesota, *Biostatistics* **4** 123–142

Barrett R.,Berry M.,Chan T.F.,Demmel J.,Donato J.,Dongarra J.,Eijkhout V.,Pozo R.,Romine C.,van der Vorst H. (1994) *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, Philadelphia: SIAM

Barry R.,Pace K.R. (1997) Kriging with large data sets using sparse matrix techniques, *Communications in Statistics: Simulation and Computation* **26** 619–629

Beier J.C.,Killeen G.F.,Githure J.I. (1999) Short report: entomologic inoculation rates and *Plasmodium falciparum* malaria prevalence in Africa, *American Journal of Tropical Medicine and Hygiene* **61** 109–113

Beier J.C.,Oster C.N.,Onyango F.K.,Bales J.D.,Sherwood J.A.,Perkins P.V.,Chumo D.K., Koech D.V.,Whitmire R.E.,Roberts C.R.,Diggs C.L.,Hoffman S.L. (1994) *Plasmodium falciparum* incidence relative to entomologic inoculation rates at a site proposed for testing malaria vaccines in Western Kenya, *American Journal of Tropical Medicine and Hygiene* **50** 529–536

Bernardinelli L.,Montomoli C. (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, *Statistics in Medicine* **11** 983–1007

Besag J. (1974) Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, Series B* **36** 192–236

Besag J.,York J.,Mollie A. (1991) Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43** 1–59

Bjørnstad O.N.,Falck W. (2001) Nonparametric spatial covariance functions: estimation and testing, *Environmental and Ecological Statistics* **8** 53–70

Bolstad W.M.,Manda S.O. (2001) Investigating child mortality in Malawi using family and community random effects: A Bayesian analysis, *Journal of the American Statistical Association* **96** 12–19

Booth J.G.,Hobert J.P. (1998) Standard errors of prediction in generalized linear mixed models, *Journal of the American Statistical Association* **93** 262–272

Booth J.G.,Hobert J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B* **61** 265–285

Bouvier P.,Breslow N.,Doumbo O.,Robert C.F.,Picquet M.,Mauris A.,Dolo A.,Dembele H.K.,Delley V.,Rougemont A. (1997) Seasonality, malaria, and impact of prophylaxis in a West African village. II Effect on birthweight, *American Journal of Tropical Medicine and Hygiene* **56** 384–389

Brabin B.J. (1983) An analysis of malaria in pregnancy in Africa, *Bulletin of the World Health Organization* **61** 1005–1016

Bradley D.J. (1991) Malaria, In: *Disease and Mortality in Subsaharan Africa*, Feachem R.G.,Jamison D.T., eds. New York: Oxford University Press 190–202

Breslow N.E.,Clayton D.G. (1993) Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88** 9–25

Breslow N.E.,Lin X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82** 81–91

Brown S.A.,Folk M.,Goucher G.,Rew R. (1993) Software for portable scientific data management, *Computers in Physics* **7** 304–308

Browne W.J.,Draper D. (2000) A comparison of Bayesian and likelihood methods for fitting multilevel models, Submitted

Brooks S.P.,Giudici P.,Roberts G.O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions, *Journal of the Royal Statistical Society, Series B* **65** 1–37

Cappé O. (2002) A Bayesian approach for simultaneous segmentation and classification of count data, *IEEE Transaction on Signal Processong* **50** 400–410

Carrat F.,Valleron A.-J. (1992) Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France, *American Journal of Epidemiology* **135** 1293–1300

Carter R.,Mendis K.N.,Roberts D. (2000) Spatial targeting of interventions against malaria, *Bulletin of the World Health Organization* **78** 1401–1411

Charlwood J.D.,Smith T.,Lyimo E., Kitua A.Y.,Masanja H.,Booth M.,Alonso P.L.,Tanner M. (1998) Incidence of *Plasmodium falciparum* infection in infants in relation to exposure to sporozoite-infected anophelines, *American Journal of Tropical Medicine and Hygiene* **59** 243–251

Christensen O.F.,Møller J.,Waagepetersen R. (2000) Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo, *Technical Report*, Departement of Mathematical Sciences, Aalborg University

Christensen O.F.,Waagepetersen R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models, *Biometrics* **58** 280–286

Clayton D.G.,Kaldor J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43** 671–681

Clayton D.G.,Bernardinelli L.Montomoli C. (1993) Spatial correlation in ecological analysis, *International Journal of Epidemiology* **43** 1193–1202

Cleland J.,van Ginneken J.K. (1989) Maternal education and child survival in developing countries: the search for pathways of influence, *Social Science and Medicine* **27** 1357–1368

Coll C.,Caselles V. (1997) A split-window algorithm for land surfaces temperature from advanced very high-resolution radiometer data: validation and algorithm comparison, *Journal of Geophysical Research* **102** 16697–16713

Connor S.J.,Thomson M.C.,Flasse S.P.,Perryman A.H. (1998) Environmental information systems in malaria risk mapping and epidemic forecasting, *Disasters* **22** 39–56

Coulibaly S.,Dicko F.,Traoré S.M.,Sidibe O.,Seroussi M.,Barrere B. (1996) *Enquête Démographique et de Santé Mali 1995–1996*, Calverton, Maryland, USA: Celule de Planification et de Statistique du Ministère de la Santé, Direction Nationale de la Statistique et de l'Informatique et Macro International Inc.

Craig M.H.,Snow R.W.,le Sueur D. (1999) A climate-based distribution model of malaria transmission in sub- Saharan Africa, *Parasitology Today*, **15** 105–111

Cressie N.,Hawkins D.M. (1980) Robust estimation of the variogram, *Mathematical Geology* **12** 115–125

Cressie N.A.C. (1993) *Statistics for Spatial Data*, New York: Wiley

Crook A.,Knorr-Held L.,Hemingway H. (2003) Measuring spatial effects in time to event data: a case study using months from angiography to coronary artery bypass graft, *Statistics in Medicine* **22** 2943–2961

Cuthill E.,McKee J. (1969) Reducing the bandwidth of sparse symmetric matrices, *Proceedings of the 24th National Conference of the ACM* 157–172

Damian D.,Sampson P.D.,Guttorp P. (2001) Bayesian estimation of semi-parametric non-stationary spatial covariance structures, *Environmetrics* **12** 161–178

Davenport M.,Nicholson S.E. (1993) On the relation between rainfall and normalized difference vegetation index for diverse vegetation types in East Africa, *International Journal of Remote Sensing* **12** 2369–2389

Deichman U. (1996) *Africa Population Database*, National Center for Geographic Information, United Nations Environmental Program, World Resources Institute

Dempster A.P.,Laird N.M,Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39** 1–38

Denison D.G.T.,Holmes C.C.,Mallick B.K.,Smith A.F.M. (2002) *Bayesian methods for non-linear classification and regression* Wiley Series in Probability and Statistics, Chichester: Wiley

Dietz K.,Molineaux L.,Thomas A. (1974) A malaria model tested in the African savannah, *Bulletin of the World Health Organization* **50** 347–357

Diggle P.J.,Tawn J.A.,Moyeed R.A. (1998) Model-based geostatistics, *Applied Statistics* **47** 299–350

Diggle P.J.,Moyeed R.A.,Rowlinson B.,Thomson M. (2002) Childhood malaria in the Gambia: A case-study in model-based geostatistics, *Applied Statistics* **51** 493–506

Dolo G.,Briët O.J.T.,Dao A.,Traoré S.F.,Bouaré M.,Sogoba N.,Niaré O.,Bagayoko M.,Sangaré D.,Doumbo O.K.,Touré Y.T. (2000) Rice cultivation and malaria transmission in the irrigated Sahel of Mali, West Africa, *Cahiers d'études et de recherches francophones Agricultures* (Cahiers Agricultures) **9** 425

Doumbo O. (1992) Epidemiologie du paludisme au Mali, étude de la chloroquinoresistance, essai de strategie de controle basée sur l'utilisation de rideaux impregnés de permethrine associée au traitement systematique des accès febriles (french), *PhD Thesis*, University of Montpellier

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

Ecker M.,Gelfand A.E. (1997) Bayesian variogram modelling for an isotropic spatial process, *Journal of Agricultural, Biological and Environmental Statistics* **2** 347–369

Ecker M.,Gelfand A.E. (1999) Bayesian modelling and inference for geometrically anisotropic spatial data, *Matematical Geology* **31** 67–83

FAO (1978) Report on the agro-ecological zones project, Vol. 1: *Methodology and Results for Africa; World Soil Resources Report* **48** 32–41

Farah A.A.,Preston S.H. (1982) Child mortality differentials in Sudan, *Population and Development Review* **8** 365–383

Ferreira J.T.A.S.,Denison D.G.T.,Holmes C.C. (2002) Partition modelling, In: *Spatial Cluster Modelling*, Lawson A.B.,Denison D.G.T., eds. London: Chapman & Hall

Franca G.B.,Cracknell A.P. (1994) Retrieval of land and sea surface temperature using NOAA-11 AVHRR data in North-Eastern Brazil, *International Journal of Remote Sensing* **15** 1695–1712

Fuentes M.,Smith R.L. (2002) A new class of nonstationary spatial models, *Technical Report*, Statistics Department, North Carolina State University

Gaudard M.,Karson M.,Linder E.,Sinha D. (1999) Bayes spatial prediction, *Environmental and Ecological Statistics* **6** 147–172

GDE Systems Inc. (1995) *Geoname Digital Gazetteer*, (CD-ROM) Version I

Gelfand A.E.,Smith A.F.M. (1990) Sampling-based approach to calculating marginal densities, *Journal of the American Statistical Association* **85** 398–409

Gelfand A.E.,Sahu S.K.Carlin B.P. (1996) Efficient parametrizations for generalized linear mixed models, *Bayesian Statistics* **5** 165–180

Gelfand A.E.,Ravishanker N.,Ecker M. (1999) Modeling and inference for point-referenced binary spatial data, In: *Generalized Linear Models: A Bayesian Perspektive*, Dey D.,Ghosh S.,Mallick B., eds. Marcel Dekker Inc. 373–386

Gelfand A.E.,Zhu L.,Carlin B.P. (2001) On the change of support problem for spatio-temporal data, *Biostatistics* **2** 31–45

Gelfand A.E.,Vounatsou P.,Smith T. (2003) Spatial modelling of gene frequencies in the presence of undetectable alleles, *Journal of Applied Statistics* **30** 49–62

Gemperli A.,Vounatsou P.,Sogoba N.,Smith T. (2003) Malaria mapping using transmission models: application to survey data from Mali, submitted to *The American Journal of Epidemiology*

Gemperli A.,Vounatsou P. (2003) Fitting generalized linear mixed models for point-referenced data, *Journal of Modern Applied Statistical Methods* **2** 481–495

Gemperli A.,Vounatsou P.,Kleinschmidt I.,Bagayoko M.,Lengeler C.,Smith T. (2004) Spatial patterns of infant mortality in Mali; the effect of malaria endemicity, *American Journal of Epidemiology* **159** 64–72

George A.,Liu J.W.H. (1981) *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Series in Computational Mathematics

Geweke J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, In: *Bayesian Statistics* **4**, Bernardo J.M.,Berger J.O.,Dawid A.P.,Smith A.F.M., eds. Oxford University Press 169–193

Gibbs N.E.,Poole W.G, Stockmeyer P.K. (1976) An algorithm for reducing the bandwidth and profile of a sparse matrix, *SIAM Journal on Numerical Analysis* **13** 236–250

Gilks W.R.,Roberts G.O. (1996) Strategies for improving MCMC, In: *Markov chain Monte Carlo in Practice*, Gilks W.R.,Richardson S.,Spiegelhalter D.J., eds. Chapman and Hall 89–114

Gilles H.M.,Warrell D.A. (1993) *Bruce-Chwatt's Essential Malariology*, London: Edward Arnold

Golub G.H.,Van Loan C.F. (1996) *Matrix Computations*, 3rd ed. Johns Hopkins University Press

Goodnight J.H. (1979) A tutorial on the SWEEP operator, *The American Statistician* **33** 149–158

Gotway C.A.,Young L.J. (2002) Combining incompatible spatial data, *Journal of the American Statistical Association* **97** 632–648

Green P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **81** 711–732

Green P.J.,Mira A. (2002) Delayed rejection in reversible jump Metropolis-Hastings, *Biometrika* **88** 1035–1053

Hagmann R.,Charlwood J.D.,Gil V.,Do Rosario V.,Smith T. (2003) Malaria and its possible control on the island of Príncipe, *Malaria Journal* **2** 15

Hall P.,Fisher N.I.,Hoffmann B. (1994) On the nonparametric estimation of the covariance function, *The Annals of Statistics* **22** 2115–2134

Handcock M.S.,Stein M.L. (1993) A Bayesian analysis of kriging, *Technometrics* **35** 403–410

Handcock M.S.,Wallis J.R. (1994) An approach to statistical spatio-temporal modelling of meteorological fields, *Journal of the American Statistical Association* **98** 368–390

Haran M.,Hodges J.S.,Carlin B.P. (2003) Accelerating computation in Markov random field models for spatial data via structured MCMC, *Journal of Computational & Graphical Statistics* **12** 249–264

Harville D.A. (1997) *Matrix Algebra from a Statistician's Perspective*, New York: Springer

Harville D.A. (1999) Use of the Gibbs sampler to invert large, possibly sparse, positive definite matrices, *Linear Algebra and its Application* **289** 203–224

Hay S.I.,Snow R.W.,Rogers D.J. (1998) Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92** 12–20

Hay S.I.,Rogers D.J.,Toomer J.F.,Snow R.W. (2000) Annual *Plasmodium falciparum* entomological inoculation rates (EIR) across Africa: literature survey, internet access and review, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94** 113–127

Hay S.I.,Cox J.,Rogers D.J.,Randolph S.E.,Stern D.I.,Shanks G.D.,Myers M.F.,Snow R.W. (2002) Climate change and the resurgence of malaria in the East African higlands, *Nature* **415** 905–909

Heagerty P.J.,Lele S.R. (1998) A composite likelihood approach to binary spatial data, *Journal of the American Statistical Association* **93** 1099–1111

Henderson R.,Shimakura S.,Gorst D. (2002) Modelling spatial variation in leukemia survival data, *Journal of the American Statistical Association* **97** 965–972

Hestenes M.,Stiefel E. (1952) Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards* **49** 409–436

Higdon D.,Swall J.,Kern J. (1998) Non-stationary spatial modelling, In: *Bayesian Statistics* **6**, Bernardo J. M. et al. edts., Proceedings of the 6th Valencia international meeting, Alcoceber near Valencia, Spain, June 6–10, 1998, Oxford: Clarendon Press 761–768

Hightower A.W.,Ombok M.,Otieno R.,Odhiambo R.,Oloo A.J.,Lal A.A.,Nahlen B.L.,Hawley W.A. (1998) A geographic information system applied to a malaria field study in Western Kenya, *American Journal of Tropical Medicine and Hygiene* **58** 266–272

Howe G.M. (1989) Historical evolution of disease mapping in general and specifically of cancer mapping, *Recent Results in Cancer Research* **114** 1–21

Hurvich C.M.,Tsai C.L. (1989) Regression and time series model selection in small samples, *Biometrika* **77** 709–719

Hutchinson M.F. (1991) The application of thin plate splines to continent-wide data assimilation, In: *Data Assimilation Systems*, Jasper J.D., ed. BMRC Res. Rep. **27** Bureau of Meteorology, Melbourne 104–113

Hutchinson M.F.,Gessler P.E. (1994) Splines - more than just a smooth interpolator, *Geoderma* **62** 45–67

Hutchinson M.F.,Nix H.A.,McMahon J.P.,Ord K.D. (1996) *Africa - A Topographic and Climate Database* (CD-ROM), The Australian National University, Canberra, ACT 0200, Australia

Ibrahim M.M.,Omar H.M.,Persson L.A.,Wall S. (1996) Child mortality in a collapsing African society, *Bulletin of the World Health Organization* **74** 547–52

Jain A. (1988) Determinants of regional variation in infant mortality in rural India. In: *Infant Mortality in India: Differentials and Determinants*, Jain A.,Visaria L., eds. Sage Publications 127–167

Justice C.O.,Townshend J.R.G.,Holben B.N,Tucker C.J. (1985) Analysis of the phenology of global vegetation using meteorological satellite data, *International Journal of Remote Sensing* **6** 1271–1318

Kalipeni E. (1993) Determinants of infant mortality in Malawi: a spatial perspective, *Socical Science and Medicine* **37** 183–198

Kelley P.R.,Barry R. (1997) Fast CARs, *Journal of Statistical Computation and Simulation* **59** 123–147

Kelley P.R.,Barry R. (1997) Quick computation of spatial autoregressive estimators, *Geographical Analysis* **29** 232–247

Kidwell K. (1991) *NOAA Polar Orbiter Data User's Guide*, NCDC/SDSD, National Climatic Data Center, Washington, DC

Kim H.-M.,Mallik B.K.,Holmes C.C. (2002) Analyzing non-stationary spatial data using piecewise Gaussian processes, *Technical Report*, Texas A&M University, Corpus Christi, TX 78412

Kleinschmidt I.,Bagayoko M.,Clarke G.P.Y.,Craig M.,LeSueur D.A. (2000) A spatial statistical approach to malaria mapping, *International Journal of Epidemiology* **29** 355–361

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

Kleinschmidt I.,Sharp B.L.,Clarke G.P.Y.,Curtis B.,Fraser C. (2001) Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa, *American Journal of Epidemiology* **153** 1213–1221

Lagazio C.,Biggeri A.,Dreassi E. (2003) Age-period-cohort models and disease mapping, *Environmetrics* **14** 475–490

Le N.D.,Zidek J.V. (1992) Interpolation with uncertain spatial covariance: A Bayesian alternative to kriging, *Journal of Multivariate Analysis* **43** 351–374

Lesaffre E.,Spiessens B. (2001) On the effect of the number of quadrature points in a logistic random-effects model: an example, *Applied Statistics* **50** 325–335

Lewis J.G. (1982) Implementing of the Gibbs-Poole-Stockmeyer and Gibbs-King algorithms, *ACM Transactions on Mathematical Software* **8** 180–189

Liang K.Y.,Zeger S.L. (1986) Longitudinal data analysis using generalized linear models, *Biometrika* **73** 13–22

Liu J.S.,Wong W.H.,Kong A. (1994) Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes, *Biometrika* **81** 27–40

Liu J.S.,Liang F.,Wong W.H. (2000) The multiple-try method and local optimization in Metropolis sampling, *Journal of the American Statistical Association* **95** 121–134

Liu C. (2003) Alternating subspace-spanning resampling to accelerate Markov chain Monte Carlo simulation, *Journal of the American Statistical Association* **98** 110–117

Logsdon T. (1992) *The Navstar Global Positioning System*, New York: Van Nostrand Reinhold

Macdonald G. (1957) *The Epidemiology and Control of Malaria*, London: Oxford University Press

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

Markowitz H.M. (1957) The elimination form of the inverse and its application to linear programming, *Management Science* **3** 255–269

Matheron G. (1963) Principles of geostatistics, *Economic Geology* **58** 1246–1266

McCormick M.C. (1985) The contribution of low birth weight to infant mortality and childhood mortality, *New England Journal of Medicine* **312** 82–90

McCulloch C.E. (1997) Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92** 162–170

McGregor I.A. (1984) Epidemiology, malaria and pregnancy, *American Journal of Tropical Medicine and Hygiene* **33** 517–525

McMahon J.P.,Hutchinson M.F.,Nix H.A.,Ord K.D. (1995) *ANUCLIM User's Guide*, Draft Report, Centre for Resource and Environmental Studies, Australian National University, Canberra

Milford J.R,Dugdale G.,McDougall V.D. (1996) Rainfall estimation from cold cloud duration: experience of the TAMSAT group in West Africa. In: *Validation Problems of Rainfall Estimation by Satellite in Intertropical Africa*, Guillot B., edt., Proc. Niamey workshop 1–3 Dec., 1994, Paris: ORSTOM

Mira A.,Sargent D.J. (2000) Strategies for speeding Markov chain Monte Carlo algorithms, *Technical Report*, University of Insubria, Varese

Molineaux L.,Gramiccia G. (1980) *The Garki Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa*, Geneva: World Health Organization

Molineaux L. (1985) La lutte contre les maladies parasitaires: le problème du paludisme, notamment en Afrique, In: *La Lutte Contre la Mort*, Vallin J.,Lopez A., eds. Travaux et Documents No. 108 Paris: Presses Universitaires de France 111–140

Mugglin A.S.,Carlin B.P.,Gelfand A.E. (2000) Fully model-based approaches for spatially misaligned data, *Journal of the American Statistical Association* **95** 877–887

Neal R.M. (1996) *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics **118**, New York: Springer

Neuhaus J.N.,Segal M.R. (1997) An assessment of approximate maximum likelihood estimators in generalized linear mixed models, In: *Modelling Longitudinal and Spatially Correlated Data*, Gregoire T.G.,Brillinger D.R.,Diggle P.J.,Russek-Cohen E.,Warren W.G.,Wolfinger R.D., edts. Lecture Notes in Statistics **122** 11–22

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

Omumbo J.A.,Ouma J.,Rapuoda B.,Craig M.H.,Le Sueur D.,Snow R.W. (1998) Mapping malaria tranmission intensity using geographical information systems (GIS); an example from Kenya, *Annals of Tropical Medicine and Parasitology* **92** 7–21

Pace K.R.,Barry R. (1997) Quick computation of the regressions with spatially autoregressive dependent variable, *Geographical Analysis* **29** 232–247

Paige C.C.,Saunders M.A. (1975) Solution of sparse indefinite systems of linear equations, *SIAM Journal on Numerical Analysis* **12** 617–629

Patz J.A.,Strzepek K.,Lele S.,Hedden M.,Greene S.,Noden B.,Hay S.,Kalkstein L.,Beier J.C. (1998) Predicting key malaria transmission factors, biting and entomological inoculation rates, using modelled soil moisture in Kenya, *Tropical Medicine and International Health* **3** 818–827

Payne D.,Grab B.,Fontaine R.E.,Hempler J.H.G. (1976) Impact of control measures on malaria transmission and general mortality, *Bulletin of the World Health Organization* **54** 369–377

Prasad N.G.N.,Rao J.N.K. (1990) The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association* **85** 163–171

Prata A.J.,Caselles V.,Coll C.,Sobrino J.A.,Ottle C. (1995) Thermal remote sensing of land surface temperature from satellites: current status and future prospects, *Remote Sensing Review* **12** 175–224

Preisler H.K. (1988) Maximum likelihood estimates for binary data with random effects, *Biometrical Journal* **3** 339–350

Press W.H.,Flannery B.P.,Teukolsky S.A.,Vetterling W.T. (1988) *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge: Cambridge University Press

Raftery A.E.,Lewis S. (1992) How many iterations in the Gibbs sampler? In: *Bayesian Statistics* **4**, Bernardo J.M.,Berger J.O.,Dawid A.P.,Smith A.F.M. edts. Oxford University Press 763–773

Ribeiro J.M.C.,Seulu F.,Abose T.,Kidane G.,Teklehaimanot A. (1996) Temporal and spatial distribution of anopheline mosquitos in an Ethiopian village: Implications for malaria control strategies, *Bulletin of the World Health Organization* **74** 299–305

Rip M.R.,Keen C.S.,Kibel M.A. (1986) A medical geography of perinatal mortality in Metropolitan Cape Town, *South African Medical Journal* **27** 399–403

Robinson T.P. (2000) Spatial statistics and geographical information systems in epidemiology and public health, *Advances in Parasitology* **47** 81–128

Rogers D.J.,Randolph S.E.,Snow R.W.,Hay S.I. (2002) Satellite imagery in the study and forecast of malaria, *Nature* **415** 710–715

Rotondi R. (2002) On the influence of the proposal distribution on a reversible jump MCMC algorithm applied to the detection of multiple change-points, *Computational Statistics & Data Analysis* **40** 633–653

Rubin D.B. (1987) A noniterative Sampling/Importance Resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm, Comment to: Tanner, Wong: The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association* **82** 543–546

Rue H. (2000) Fast sampling of Gaussian Markov random fields, *Journal of the Royal Statistical Society, Series B* **63** 325–338

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

Schellenberg J.A.,Newell J.N.,Snow R.W.,Mung'ala V.,Marsh K.,Smith P.G.,Hayes R.J. (1998) An analysis of the geographical distribution of severe malaria in children in Kilifi District, Kenya, *International Journal of Epidemiology* **27** 323–329

Schultz T.P. (1979) Interpretation of relations among mortality, economics of the household and the health environment. In: *Proceedings of the Meeting on Socio-Economic Determinants and Consequences of Mortality* Mexico City, June 19–25 Geneva: World Health Organization

Smith A.F.M.,Roberts G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55** 3–23

Smith T.,Charlwood J.D.,Kihonda J.,Mwankusye S.,Billingsley P.,Meuwissen J.,Lyimo E., Takken W.,Teuscher T.,Tanner M. (1993) Absence of seasonal variation in malaria parasitaemia in an area of intense seasonal transmission, *Acta Tropica* **54** 55–72

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

Smith T.A.,Leuenberger R.,Lengeler C. (2001) Child mortality and malaria transmission intensity in Africa, *Trends in Parasitology* **17** 145–149

Snow J. (1855) *On the Mode of Communication of Cholera*, 2nd ed., The Commonwealth Fund, New York

Snow R.W.,Marsh K. (1995) Will reducing *Plasmodium falciparum* transmission alter malaria mortality among African children? *Parasitology Today* **11** 188–190

Snow R.W.,Omumbo J.A.,Lowe B.,Molineaux C.S.,Obiero J.O.,Palmer A.,Weber M.W., Pinder M.,Nahlen B.,Obonyo C.,Newbold C.,Gupta S.,Marsh K. (1997) Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa, *Lancet* **349** 1650–1654

Snow R.W.,Marsh K. (1998) New insights into the epidemiology of malaria relevant to disease control, *British Medical Journal* **54** 293–309

Snow R.W.,Gouws E.,Omumbo J.,Rapuoda B.,Craig M.H.,Tanser F.C.,le Sueur D.,Ouma J. (1998) Models to predict the intensity of *Plasmodium falciparum* transmission: applications to the burden of disease in Kenya, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92** 601–606

Snow R.W.,Craig M.H.,Deichman U.,Le Sueur D. (1999) A preliminary continental risk map for malaria mortality among African children, *Parasitology Today* **15** 99–104

Snow R.W.,Marsh K. (2002) The consequences of reducing transmission of *Plasmodium falciparum* in Africa, *Advances in Parasitology* **52** 235–264

Spiegelhalter D.J.,Best N.,Carlin B.P.,van der Linde A. (2002) Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B* **64** 583–639

Steketee R.W.,Nahlen B.L.,Parise M.E.,Menendez C. (2001) The burden of malaria in pregnancy in malaria-endemic areas, *American Journal of Tropical Medicine and Hygiene* **64** 28–35

Stein M.L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer

Steinwand D.R.,Hutchinson J.A.,Snyder J.P. (1992) Map projections from global and continental data sets, and an analysis of distortion caused by reprojection, *USGS/EDC Contract Report*, EROS Data Center, Sioux Falls

Sun D.C.,Tsutakawa R.K.,Kim H.,He Z. (2000) Spatio-temporal interaction with disease mapping, *Statistics in Medicine* **19** 2015–2035

Tanser F.C.,Sharp B.,Le Sueur D. (2000) Malaria seasonality and population exposure in Africa: a high resolution climatic model, In: *The Application of Geographical Information Systems to Infectious Diseases and Health System in Africa*, Tanser F.C., PhD thesis, University of Natal

Tanser F.C.,Sharp B.L.,le Sueur D. (2002) Malaria seasonality and climate change: implications for Africa's disease burden, Submitted

Thomas C.J.,Lindsay S.W. (2000) Local-scale variation in malaria infection amongst rural Gambian children estimated by satellite remote sensing, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94** 159–163

Thomson M.C.,Connor S.J.,Milligan P.J.M.,Flasse S.P. (1996) The ecology of malaria - As seen from Earth-observation satellites, *Annals of Tropical Medicine and Parasitology* **90** 243-264

Thomson M.C.,Connor S.J.,Milligan P.,Flasse S.P. (1997) Mapping malaria risk in Africa: What can satellite data contribute? *Parasitology Today* **13** 313–318

Thomson M.C.,Connor S.J.,D'Alessandro U.,Rowlingson B.,Diggle P.,Cresswell M.,Greenwood B. (1999) Predicting malaria infection in Gambian children from satellite data and bednet use surveys: The importance of spatial correlation in the interpretation of results, *American Journal of Tropical Medicine and Hygiene* **61** 2–8

Tubilla A. (1975) Error convergence rates for estimates of multidimensional integrals of random functions, *Technical Report* **72**, Department of Statistics, Stanford University, Stanford, CA

Uchudi J.M. (2001) Covariates of child mortality in Mali: Does the health-seeking behavior of the mother matter? *Journal of Biosocial Science* **33** 33–54

van der Hoek W.,Konradsen F.,Amerasinghe P.H.,Perera D.,Piyaratne M.K.,Amerasinghe F.P. (2003) Towards a risk map of malaria for Sri Lanka: the importance of house location relative to vector breeding sites, *International Journal of Epidemiology* **32** 280–285

Vounatsou P.,Smith T.,Gelfand A.E. (2000) Modeling of Multinomial data with latent structure: application to geographical mapping of human gene and haplotype frequencies, *Biostatistics* **1** 177–189

Waller L.A.,Carlin B.P.,Xia H.,Gelfand A.E. (1997) Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical Association* **92** 607–617

Wells D. (1988) *Guide to GPS Positioning*, University of Brunswick: Graphic Services, Canada

Whittle P. (1954) On stationary process in the plane, *Biometrika* **41** 434–449

Wolfinger R.,O'Connell M. (1993) Generalized linear mixed models: A pseudo-likelihood approach, *Journal of Statistical Computation and Simulation* **48** 233–243

World Resources Institute (1995) *African Data Sampler*, (CD-ROM) Edition I

Zeger S.L.,Liang K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes, *Biometrika* **42** 121–130

Zimmerman D.L.,Zimmerman M.B. (1991) A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors, *Technometrics* **33** 77–91

Zimmerman D.L.,Cressie N. (1992) Mean squared prediction error in the spatial linear model with estimated covariance parameters, *Annals of the Institute of Statistical Mathematics* **44** 27–43

Curriculum Vitae

Armin Gemperli

Date of Birth:	May 09, 1973
Nationality:	Swiss

Education

2000–2003	PhD thesis "Development of Spatial Statistical Methods for Modelling Point-Referenced Spatial Data in Malaria Epidemiology" (2003) under the supervision of Dr. P. Vounatsou und Prof. T. Smith at the Swiss Tropical Institute, University of Basel. Degree with summa cum laude honors.
1999	SAS-Masterclass-Certification, SAS-Institute Switzerland.
1998	Diplomawork "Nonparametric Methods for Changepoint Problems" (in german) under the supervision of Prof. J. Hüsler.
1993–1998	Studies in mathematical statistics (Prof. J. Hüsler) at the University of Bern, with subsidiary subjects Mathematics and Theory of Sciences.

Professional Activities and Teaching

1999	Statistician in the Business Intelligence Unit at SYSTOR AG in Basel.
1998–1999	Statistics Lecturer at the School for Nutrition at the Insel Hospital Bern.
1996–1999	Statistical Consulting at the Institute of Mathematical Statistics at the University of Bern. Lecturing (probability, applied stochastic and statistics for sports scientists).
1995	Lecturing by deputy in Mathematics: - Central Swiss Traffic School, Luzern. - Teachers Seminar Luzern.

Membership

Swiss Society of Statistics; International Biometric Society, with sections Austria-Swiss (ROeS) and Basel Biometric Section (BBS); Institute of Mathematical Statistics; American Statistical Association; Data Warehousing Institute.

Reviewer: American Journal of Epidemiology.