Geostatistical modelling of schistosomiasis transmission in Africa

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Nadine Schur
aus Deutschland

Basel, November 2011

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. J. Utzinger, PD Dr. P. Vounatsou, and Prof. Dr. M. Schumacher.

Basel, den 26. April 2011

Prof. Dr. Martin Spiess
Dekan

Dedicated to my family
and my beloved husband, Mike

Summary

Schistosomiasis is a parasitic disease that is currently endemic in more than 70 countries with the bulk of infections concentrated in Africa. The interest in schistosomiasis has recently grown, after many years of neglect, due to the commitment of substantial amounts of funding to the control of the so-called neglected tropical diseases. Targeting of interventions and allocation of financial resources should be driven by evidenced-based information on the spatial distribution and schistosomiasis burden estimates in order to increase cost-effectiveness and to meet local needs. Currently, decisions are mainly made based on crude schistosomiasis risk estimates which are largely obsolete due to ongoing control efforts, ecological transformations, demographic changes and improved hygiene, among other reasons.

Schistosomiasis transmission depends on the distribution of intermediate host snail species. This distribution is determined by climatic and other environmental conditions, such as temperature, precipitation or water flow velocity. Statistical models can be used to establish the relation between the aforementioned factors and schistosomiasis risk and to predict the risk at unobserved locations. Empirical risk mapping requires observed prevalence data distributed within the area of interest, however contemporary large-scale surveys are not available. To address this issue, the European Union (EU)-funded CONTRAST project initiated the development of the Global Neglected Tropical Disease (GNTD) database, an open-access source of survey data extracted from peer-reviewed publications, Health Ministry reports and other unpublished literature. To-date, the GNTD database is the most comprehensive schistosomiasis database in Africa.

Schistosomiasis prevalence data are spatially correlated, because locations in close proximity share common spatial exposures, which similarly influence transmission. Standard statistical models are not appropriate because they assume independence between locations, leading to imprecise parameter estimates and risk predictions. Geostatistical models

assume take into account potential spatial correlation by introducing location-specific random effects. These additional factors are considered as observations of latent Gaussian spatial processes. Geostatistical models typically contain large number of parameters. Bayesian model formulations, implemented via Markov chain Monte Carlo (MCMC) simulations methods, enable model fit overcoming the computational problems of the likelihood-based methods.

Geostatistical model fit requires the repeated inversion of the correlation matrix of the spatial process. The size of this matrix increases with the number of locations and, for the very large number of locations present in the GNTD database, matrix inversion is infeasible. This is known as the 'large N problem'. An important aspect of geostatistical model fit is the choice of predictors driving schistosomiasis transmission. There are a number of environmental factors which are correlated, complicating model fit. Rigorous geostatistical variable selection has not yet been applied in spatial schistosomiasis epidemiology. Data compilations contain heterogeneous surveys across locations in terms of age groups involved and diagnostic methods used. The lack of prevalence data reported in standard age groups complicates the estimation of age-adjusted schistosomiasis risk. A common assumption of geostatistical models is that of isotropy implying that spatial correlation is a function of distance between locations irrespective of direction or location. However in schistosomiasis risk mapping, spatial correlation is likely to be related to the direction of river flow due to water-dependent intermediate host snail species. This might introduce directional dependency (anisotropy). Schistosomiasis tends to be present in areas with other neglected diseases. Cost-effective interventions call for an integrated disease control, which requires estimating of the geographical distribution of high co-endemicity. The co-endemic diseases might be correlated, however surveys screening for multiple diseases are not available over large geographical areas. There rather exist data from independent surveys screening for single diseases on different sets of individuals.

The aim of this PhD thesis was (i) to develop Bayesian geostatistical models for the analysis of schistosomiasis survey data taking into account inherent data characteristics; and (ii) to validate and implement these models in order to produce spatially-explicit schistosomiasis risk estimates and number of infected individuals on regional scale in Africa.

In Chapter 3 the construction and development of the open-access GNTD database is described. As of January 2011, the database contained more than 12,000 geo-referenced schistosomiasis survey locations in more than 30 African countries. It consists of historical and recent data from various sources, including unpublished data. Therefore, the GNTD

database became a unique tool for schistosomiasis mapping purposes and analyses over time.

In *Chapter 4* Bayesian geostatistical models were developed to estimate the spatial distribution of *Schistoma haematobium* and *S. mansoni* at high spatial scales in West Africa. The 'large N problem' was addressed by estimating the spatial process from a subset of locations. Country-specific estimates on the number of infected individuals ≤ 20 years were derived resulting in more than 50 million infections. The analysis revealed that previous burden estimates were likely to be outdated for some countries.

In *Chapter 5* models were developed to take into account age-heterogeneity across surveys and to produce age-adjusted *S. haematobium* and *S. mansoni* risk estimates. The models related surveys on individuals aged ≤ 20 years with those on individuals aged >20 years and entire communities. This methodology was applied on survey data distributed over 11 eastern African countries extracted from the GNTD database. Model validation showed that regional age-alignment factor models were superior those assuming country-specific factors.

Chapter 6 employed the alignment factors models obtained from Chapter 5 for spatial risk prediction of schistosomiasis in eastern Africa. In addition, Bayesian geostatistical variable selection and Gaussian spatial process approximations were applied to reduce complexity of the model and to enable model fit, respectively. Our results indicated that schistosomiasis accounts for more than 120 million infections in eastern Africa, which is considerably higher than the previously reported (58 million).

In *Chapter 7* Bayesian geostatistical models were developed that incorporate anisotropic effects using simulated and real data obtained from a national school survey on urinary schistosomiasis in Senegal. Models assumed a global direction of anisotropy and locally-dependent directions fixed at environmental features. The results showed that anisotropic models improve model-based predictions and parameter estimation, even if anisotropy was not very prominent in the real dataset. Locally-dependent directional effects had not improved model predictive performance.

In *Chapter 8* geostatistical shared component models were proposed to model the geographical distribution and burden of co-infection risk from independent single disease surveys using simulated and real data. The data of the application were obtained from a survey screening for *Schistosoma mansoni*-hookworm co-infections in the region of Man, Côte d'Ivoire, however they were treated as if they were collected from independent surveys. The ability of the models to capture co-infection risk was assessed and compared

to multinomial models which can directly incorporate co-infection data. Model validation revealed that, for correlated diseases, joint risk modelling estimates obtained via shared component models have better predictive ability than commonly-used independent modeling approaches.

The main contributions of this thesis were (i) the development of Bayesian isotropic and anisotropic geostatistical models for high spatial resolution schistosomiasis risk mapping and prediction based on age-heterogeneous historical survey data collected over very large number of locations; (ii) the development of statistical methodology for assessing the geographical distribution of co-infection risk from independent single-disease surveys when diseases are correlated; and (iii) the estimation of location-specific schistosomiasis risk and number of infected people in 29 countries across West and eastern Africa. Hence, for first time, empirical model-based evidence of schistosomiasis risk and burden in those regions is provided. These estimates are of considerable importance for schistosomiasis control programmes, as they indicate high-risk areas requiring interventions, allow calculations of the number of praziquantel tablets required based on WHO guidelines at the appropriate administrative level, and provide baseline maps to assess effectiveness of interventions on the roadmap towards schistosomiasis elimination.

Zusammenfassung

Schistosomiasis ist eine parasitäre Erkrankung, welche zur Zeit in mehr als 70 Ländern vorkommt, mit einem Großteil der Infektionen in Afrika. Das Interesse an dieser Erkrankung ist nach vielen Jahren der Vernachlässigung endlich gewachsen, aufgrund erheblicher finanzieller Unterstützung zur Kontrolle der vernachlässigten tropischen Krankheiten. Das Planen von Interventionen und die Zuweisung von finanziellen Mitteln sollten durch wissenschaftliche Belege zur geographischen Verteilung und zuverlässigen Schätzungen der Belastung gesteuert werden, um die Kosteneffizienz zu steigern und die lokalen Bedürfnisse zu berücksichtigen. Bisherige Entscheidungen wurden hauptsächlich auf Grundlage von grob überschlagenen Werten ermittelt und sind weitestgehend veraltet, durch Maßnahmen zur Eindämmung der Schistosomiasis, Veränderungen der Umwelt, demografischem Wandel und verbesserter Hygiene.

Die Verbreitung der Schistosomiasis hängt stark von der räumlichen Verteilung des Zwischenwirtes (diverse Schneckenarten) ab. Dessen Verteilung ist wiederum durch bestimmte Umweltbedingungen und klimatische Faktoren, wie Temperatur, Niederschlag oder Strömungs-geschwindigkeit von Flüssen, bestimmt. Statistische Modelle können verwendet werden, um die Beziehung zwischen den zuvor genannten Faktoren und dem Schistosomiasis Risiko zu untersuchen und um das Risiko an unbekannten Standorten vorherzusagen. Empirische Kartierungen erfordern Feldstudien über die Häufigkeit der Erkrankung in dem zu erschließenden Gebiet, jedoch existieren keine aktuellen Einzelstudien über große Gebiete. Um dieses Problem zu beheben initiierte das von der Europäischen Union finanzierte CONTRAST Projekt die Entwicklung der globalen Datenbank zu vernachlässigten tropischen Erkrankungen (GNTD). Diese Datenbank besteht aus unzähligen Schistosomiasis Daten aus wissenschaftlichen Publikation, Berichten von Gesundheitsministerien und unveröffentlichtem Material, und ist die bisher umfangreichste frei verfügbare Sammlung von Studien in Afrika.

Daten zur Verbreitung von Schistosomiasis sind räumlich korreliert, da Standorte in

unmittelbarer Nähe durch ähnliche räumlichen Faktoren beeinflusst werden. Die gängigen statistischen Modelle gehen allerdings von Unabhängigkeit zwischen Studienorten aus, was zu ungenauen Parameterabschätzungen und Risikovorhersagen führt, und sind daher nicht zur Analyse geeignet. Geostatistische Modelle berücksichtigen dagegen mögliche räumliche Korrelationen durch die Einführung von ortsabhängigen Effektparametern. Diese zusätzlichen Parameter werden als Beobachtungen von verborgenen Normalverteilten Prozessen angesehen. Geostatistische Modelle haben in der Regel eine große Anzahl von Parametern. Modelformulierungen nach Bayes, die durch Markov Chain Monte Carlo (MCMC) Simulationen durchgeführt werden, ermöglichen die Modellanpassung und überwinden die rechnerischen Probleme von Maximum-Likelihood-basierten Methoden.

Geostatistische Modellanpassungen erfordern die wiederholte Invertierung der Kovarianzmatrix des räumlichen Prozesses. Die Größe dieser Matrix wird durch die Anzahl der Studienorte bestimmt und Matrixinvertierung wird mit einer großen Anzahl von Ortschaften, wie sie in der GNTD Datenbank auftreten, unmöglich. Dieses Problem ist als "Grosses N Problem"bekannt. Bestimmt Umfaltfaktoren sind miteinander korreliert, was die Modellanpassung erschwert. Die geostatistische Auswahl von Einflussfaktoren zur räumlichen Ausbreitung der Schistosomiasis wurde bisher noch nicht angewandt. Das Zusammentragen von verschiedenen Studien führt zu einer Vielzahl von Daten aus verschiedenen Altersgruppen. Der Mangel an Studien mit den selben Altersgruppen erschwert die Abschätzung des altersabhängigen Risikos and Schistosomiasis zu erkranken.

Eine häufige Annahme geostatistischer Modelle ist Isotropie, welche impliziert dass die räumliche Korrelation zwischen zwei Ortschaften auf deren Entfernung beruht und unabhängig von der Richtung und des Standortes ist. Allerdings, scheint es plausibler, dass die räumliche Korrelation im Falle der Schistosomiasis durch die Fliessrichtung von Flüssen bestimmt wird, da der Zwischenwirt (einige Schneckenarten) in Gewässern lebt. Dies führt unter Umständen zu einer Richtungsabhängigkeit (Anisotropie). Schistosomiasis tritt häufig in Gebieten mit anderen vernachlässigten Krankheiten auf. Möglichst Kosten sparende Interventionen sollten gleichzeitig mit anderen Krankheitsinterventionen durchgeführt werden. Dies erfordert die Bestimmung von Gebieten mit besonders hohem gleichzeitigem Auftreten im Bestimmungsgebiet. Manche Krankheiten treten abhängig voneinander auf, allerdings gibt es keine einzelnen Studien über große Bestimmungsgebiete, die sich mit dem Auftreten mehreren Erkrankungen gleichzeitig beschäftigen. Was es dagegen gibt, sind Daten von unabhängigen Studien die einzelne Krankheiten untersuchen.

Das Ziel dieser Dissertation war es, (i) Bayes'sche geostatistische Modelle für die Analyse von Schistosomiasis-Daten zu entwickeln und dabei die zu Grunde liegenden Besonderheiten der Daten zu berücksichtigen, und (ii) die entwickelten Methoden anzuwenden und zu überprüfen, um räumliche Risikobewertungen zu erstellen und die Anzahl der mit Schistosomiasis infizierten Personen in Afrika abzuschätzen.

In *Kapitel 3* wird die Konstruktion und Entwicklung der frei verfügbaren GNTD Datenbank beschrieben. Diese Datenbank enthielt im Januar 2011 bereits mehr als 12.000 georeferenzierte Studienorte zur Schistosomiasis verteilt in über mehr als 30 afrikanischen Ländern. Sie besteht aus historischen und aktuellen Daten aus verschiedenen Quellen, darunter auch bisher unveröffentlichte Studien. Dadurch stellt die GNTD Datenbank ein einzigartiges Werkzeug für die räumliche Kartierung der Schistosomiasis dar und erlaubt zudem eine zeitliche Analyse der Daten.

In *Kapitel 4* wurden Bayes'sche geostatistische Modelle entwickelt um die geographische Verteilung von *Schistosoma haematobium* und *S. mansoni* in Afrika, mit hoher räumlicher Auflösung, zu ermitteln. Dabei wurde der räumlichen Prozess mittels einer Zufallsauswahl von Ortschaften abgeschätzt, um das „Große N Problem" zu bewältigen. Länderspezifische Schätzungen über die Zahl der Infizierten, die unter 20 Jahre alt waren, resultierten in mehr als 50 Millionen Infektionen. Die Analyse ergab außerdem, dass die bestehenden Schätzungen für bestimmte Länder sehr wahrscheinlich veraltet waren.

In *Kapitel 5* wurden Modelle, die die Altersheterogenität zwischen Studien berücksichtigen, entwickelt und altersspezifische *S. haematobium* und *S. mansoni* Risikobewertungen erstellt. Die Modelle basierten auf den folgenden Altersgruppen: Personen jünger als 20 Jahre, Personen über 20 Jahre und die Gesamtbevölkerung. Die Methode wurde auf Erhebungsdaten der GNTD Datenbank aus 11 ostafrikanischen Ländern angewendet. Modelüberprüfungen zeigten, dass Modelle mit regionalen Faktoren zur Altersangleichung anderen Modellen mit länderspezifische Faktoren überlegen waren.

Kapitel 6 verwendet die Altersangleichungsfaktoren aus Kapitel 4 für die räumliche Risikovorhersage von Schistosomiasis in Ostafrika. Dabei wurden das Variablenauswahlverfahren nach Gibbs und die Abschätzung des räumlichen Prozesses genutzt, um die Komplexität des Modells zu reduzieren und die Modellanpassung zu ermöglichen. Unsere Ergebnisse zeigten, dass das Infektionsrisiko in Ostafrika mit mehr als 120 Millionen Infizierten erheblich höher ist als zuvor angenommen (58 Millionen).

In *Kapitel 7* wurden Bayes'sche geostatistische Modelle entwickelt, die anisotrope Effekte berücksichtigen. Die Analyse beruhte hierbei auf simulierten Daten und einer nationalen

Erhebung zur Schistosomiasis an Schulen in Senegal. Die Modelle basierten dabei auf einer globale Richtung des anisotropen Effektes und ortsabhängigen Richtungen, welche an Umweltfaktoren gekoppelt waren. Unsere Ergebnisse zeigten, dass anisotrope Modelle die Risikovorhersagen und die Abschätzung der Modellparameter verbessern, sogar wenn die direktionalen Effekte in den zu Grunde liegenden Daten nicht sehr ausgeprägt sind. Allerdings war die Fixierung dieser Effekte nachteilig für die Vorhersagekraft der Modelle.

In *Kapitel 8* wurden geostatistische Modelle mit gemeinsamen Komponenten genutzt, um die räumliche Verteilung und Belastung von Co-Infektionen mit Hilfe unabhängiger Studien zu modellieren. Dabei wurden simulierte und reale Daten verwendet. Die realen Daten wurden einer Studien über *S. mansoni*-Hakenwurm-Co-Infektionen in der Region um Man an der Elfenbeinküste entnommen, allerdings wurden die Daten so betrachtet als stammen sie von unabhängigen Studien. Die Fähigkeit von Modellen das Co-Infektionsrisiko zu ermitteln, wurde im Vergleich zu multinomialverteilten Modellen geprüft, die das Co-Infektionsrisiko direkt erfassen können. Die Überprüfung ergab, dass das Krankheitsrisikos, wenn die Krankheiten räumlich korreliert waren, mittels der häufig angewendeten getrennten Analysen weniger genau modelliert werden konnte, als mit gemeinsamen Komponenten.

Die wichtigsten Beiträge dieser Dissertation waren (i) die Entwicklung von Bayes'schen isotropen und anisotropen geostatistischen Modellen zur genauen Kartierung des Schistosomiasis-Risikos und der Vorhersage an unbekannten Orten basierend auf einer großen Anzahl von historischen Daten mit heterogen Altersgruppen; (ii) die Entwicklung von statistischen Methoden zur Bewertung der geographischen Verteilung des Co-Infektionsrisikos mittels unabhängigen Studien bei korrelierten Erkrankungen; und (iii) die Abschätzung des ortsspezifischen Schistosomiasis-Risikos und die Anzahl infizierter Personen in 29 Ländern in West- und Ostafrika. Zum ersten Mal wurden dabei empirische Daten zum Erkrankungsrisiko und der Belastung in diesen Regionen, die auf wissenschaftlichen Methoden beruhen, gewonnen. Diese Abschätzungen sind von erheblicher Bedeutung für Schistosomiasis-Kontrollprogramme, da sie besonders gefährdete Gebieten aufzeigen, welche Interventionen benötigen, die Berechnung von den erforderlichen Praziquantel-Tabletten auf den WHO-Richtlinien erlauben, und die Wirksamkeit von Interventionen auf dem Weg zur Schistosomiasis-Elimination beurteilen können.

Acknowledgements

The present PhD thesis was carried out within the EU-funded CONTRAST program under the joint supervision of PD Dr. Penelope Vounatsou and Prof. Jürg Utzinger and I am grateful to many people for help, both direct and indirect, in writing this thesis.

First and foremost, I am deeply indebted to Penelope for all her excellent scientific support. I could profit tremendously from her experience and passion and I am very grateful for her outstanding supervision, her patience and her encouragements in any aspects. Thank you very much for going together with me through this journey, not just as a supervisor, but also as a friend.

I also wish to express my sincerest gratitude to Jürg for all his constructive inputs and his expertise on schistosomiasis throughout this work. I greatly appreciated his valuably assistance, the inspiring discussions and his enthusiasm. Words cannot express my appreciation for his motivation to bring out the best in me.

Many thanks go to Prof. Martin Schumacher, who was willing to act as co-referee for this thesis, and Dr. David Rollinson, who kindly agreed to be the expert of the committee. My gratitude is expressed to my dear friends and colleagues of the CONTRAST project, especially Prof. Thomas Kristensen and Dr. Anna-Sofie Stensgaard, for their valuable comments and ideas, and their work related to the GNTD database. Special thanks are addressed to Prof. Thomas Smith for his kind help during my time as Master student and Dr. Giovanna Raso for very stimulating discussions and her collaboration. I would also like to thank Prof. Marcel Tanner for providing the institutional framework which was indispensable for the realization of this work, and for a cheerful evening in Atlanta.

While working at the Swiss TPH, I enjoyed the friendship and support of many colleagues, without which I wouldn't have managed to complete this thesis. I profited tremendously from the scientific knowledge and experience, which Laura Goşoniou was always willing to share. Furthermore, I am in indebted to her for her backing, encouragements and friendship, which I will forever value. Mulţumesc mult! The enjoyable spirit in the office

wouldn't have been possible without Susan Rumisha, who shared moments and thoughts with me I will never forget. Asante sana mpenzi wangu! I also enjoyed the wonderful support from Eveline Hürlimann, who provided invaluable work on the GNTD database, on which this thesis is based, and who is exceptionally gifted in tracing the coordinates of the tiniest villages. Many thanks in this regard also to Katrin Ziegelbauer, Maiti Laserna de Himpsl and Benjamin Speich. I am indebted to the whole 'Bayesian group' and my office mates who were always patient with me and wouldn't get tired of listening to my stories. A big 'thank you' goes to Simon Kasasa, Amek Ombek, Dominic Goşoniu, Verena Jürgens, Federica Giardina, Frederique Chammartin, Ronaldo Scholte and Amina Msengwa. Special thanks are addressed to Bianca Plüss and Thomas Fürst for their kind friendship and for teaching me Swiss German. I apologize to all my fellow students and Swiss TPH members, there is simply no space to mention them all.

My work wouldn't have been possible without all the supportive efforts and assistance. Foremost, I want to mention Margrit Slaoui and Zsuzsanna Györffi for all their administrative work and their kind words whenever needed. I also enjoyed the support and nice lunch breaks with the IT team. Merci for this to Dominique Forster, Marco Clementi and all the others. Additional thanks to the team of the library and the main secretariat who made my life smooth and pleasant.

I also owe a great deal to Emile Tchicaya, Kigbafori Silué, Dr. Benjamin Koudou, Prof. Eliézer N'Goran at CSRS and Dominik Glinz for their eagerness to support me with schistosomiasis data and a wonderful time in Côte d'Ivoire. In addition, many thanks to all our collaborators around the world who provided further data for the GNTD database.

This thesis wouldn't have come together without the kind support of my friends and family back home and I would like to acknowledge my brother Matti, my father Uwe, my grandparents, cousins and in laws and Nadine Köhler. My deepest gratitude goes to my mother Conny for the sacrifices she made to ensure that I had an excellent education, for her understanding and her endless patience ("Du bist die Beste!"). Not least, I also thank my (now) husband, Mike, for his patience and forbearance whilst I have spent hundreds of hours working on this thesis!

My apologies if I have inadvertently omitted anyone to whom acknowledgement is due. Thank you all!

In addition, I would like to acknowledge the finanical support I received from the Stiftungsrat of the 'Basler Studienstiftung' for printing this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Schistosomiasis

Schistosomiasis is one of the most prevalent parasitic diseases in tropical and subtropical countries. After many years of general neglect, there are growing interest and financial resources to control schistosomiasis. Yet, despite successful control programmes in different countries, schistosomiasis still affects over 200 million individuals with an estimated global burden that might exceed 4.5 million disability-adjusted life years (DALYs) lost annually (Utzinger et al., 2009).

1.1.1 Biology and life cycle

Schistosomiasis is a parasitic disease caused by trematode blood flukes of the genus *Schistosoma.* There are five species parasitizing humans, namely *S. haematobium*, *S. intercalatum*, *S. japonicum*, *S. mansoni* and *S. mekongi.* The life cycle of the schistosomes (depicted in Figure 1.1) includes different snail species that act as intermediate hosts. The intermediate hosts for *S. mansoni* are aquatic snails of genus *Biomphalaria,* while *S. haematobium* and *S. intercalatum* are transmitted by aquatic snails of the genus *Bulinus.* The am-

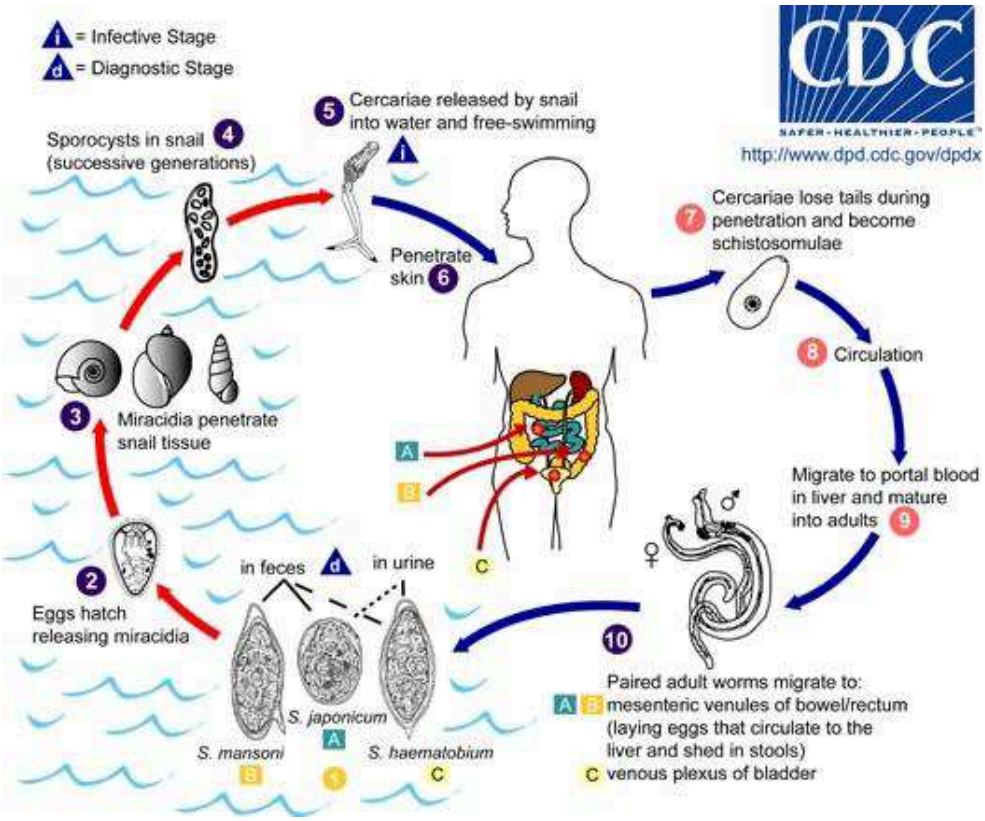

Figure 1.1: Life cycle of schistosomiasis (source: CDC).

phibian *Oncomelania* and aquatic *Tricula aperta* snails act as intermediate hosts for *S. japonicum* and *S. mekongi*, respectively. All intermediate host species have the ability to aestivate, enabling the parasite to survive during dry seasons. Infected snails release cercariae in suitable freshwater reservoirs like streams, ponds, lakes or man-made reservoirs and marshlands. The released cercariae actively targets the definitive human host (or animal reservoir hosts in the case of *S. japonicum*), penetrates the exposed skin on contact and becomes a schistosomula. Via the blood stream, the parasites reach the liver where they mate and mature. Adult worm pairs further migrate to their final peri-vesical (*S. haematobium*) or peri-intestinal (the other species) destination. There, they constantly produce eggs for an average of 3 to 5 years. Roughly half of the eggs are trapped in the tissues, while the remaining eggs are released with excreta via the urinary (*S. haematobium*) or intestinal tract (other species). On contact with water, the eggs hatch into miracidia and infect appropriate snail intermediate host species where they mature and multiply asexually into cercarial larvae. Re-infection of individuals only results from contact with infested water because schistosomes do not replicate in the human body (Gryseels et al., 2006; Hotez et al., 2006a; Davis, 2009).

1.1.2 Clinical conditions

Most schistosome infections are rather asymptomatic. An acute stage associated with fever and lymphadenopathy, known as Katayama syndrome, is triggered by an acute hypersensitivity reaction against migrating schistosomula (Ross et al., 2007). Chronic conditions are mainly resulting from blood vessel perforation and schistosome eggs trapped in the tissues. The eggs provoke granuloma formations and inflammations due to various proteolytic enzymes. *S. haematobium* eggs in the peri-vesical tissue can cause bladder wall pathologies and local ulcerations related with haematuria and dysuria. Severe conditions of the so-called urinary schistosomiasis are bladder calcification, genital tract lesions, renal failure and hydronephrosis (Hatz, 2001; van der Werf et al., 2003; Davis, 2009). The disease is also associated with an increased risk of bladder cancer. Intestinal schistosomiasis caused by eggs in the peri-intestinal tissues is characterized by intestinal bleeding and bloody diarrhoea. It leads to portal hypertension, splenomegaly, hepatosplenomegaly, periportal fibrosis and ascites, and can result in extensive liver pathology and life-threatening bleeding from gastro-oesophageal varices (Hotez et al., 2006a; Davis, 2009). Schistosmiasis in general also causes chronic growth faltering and contributes to anaemia (Gryseels et al., 2006; Hotez et al., 2006a; Davis, 2009).

1.1.3 Diagnosis and treatment

Techniques for the diagnosis of schistosomiasis can be divided into two categories based on direct and indirect detection of the parasite. Direct methods aim to detect either the parasite or their eggs, while indirect methods mainly rely on the detection of human antibodies or parasite antigens in response of an infection. In endemic settings, schistosome egg detection in urine or stool specimens via light microscopy is most common, because these methods are relatively rapid and inexpensive. Faecal samples are most frequently analysed by the Kato-Katz technique (Katz et al., 1972), but direct smears and formalin-based techniques are also common (Katz and Miura, 1954; Marti and Escher, 1990). Urine sedimentation, filtration and centrifugation are used to detect *S. haematobium* eggs. An indirect method screening for traces of blood and proteins in urine, a common condition of urinary schistosomiasis, are reagent strips. It has been shown that the number of eggs is correlated with the amount of blood and proteins found in urine (Wilkins et al., 1979). Simple questionnaires have been developed and self-reported blood in urine was shown to be a useful indicator for rapid identification of high-risk communities of urinary schistosomiasis (Lengeler et al., 2002).

Praziquantel is the current drug of choice for morbidity control and treatment of schistosomiasis (WHO, 2002), because it is generally well tolerated, easy to administer and relatively cheap (Utzinger and Keiser, 2004; Doenhoff et al., 2008). Furthermore, praziquantel is effective against the whole spectrum of human schistosome species, but not against the young developing stages of the parasite (e.g. schistosomula). The latter might be the underlying cause for treatment failures in some areas (Cioli and Pica-Mattoccia, 2003). Nevertheless, there is considerable concern of arising or already existing resistance to praziquantel (Cioli et al., 2004; Davis, 2009; Melman et al., 2009). The drugs that have been widely used against schistosomiasis are oxamniquine and metrifonate, the former being active against *S. mansoni* only, and the latter against *S. haematobium*. Oxamniquine has shown emerging resistance in Brazil (Conceio et al., 2000). Both drugs have been replaced in favour of praziquantel (Utzinger et al., 2011). Further drugs are currently being developed as alternatives to praziquantel in case of developing clinical relevant resistance (Utzinger et al., 2011).

1.1.4 Global schistosomiasis distribution and disease burden

Schistosomiasis occurs in Africa, Asia and the Americas and is currently endemic in 76 countries. *S. haematobium* and *S. mansoni* are mainly found in Africa and the Arabian

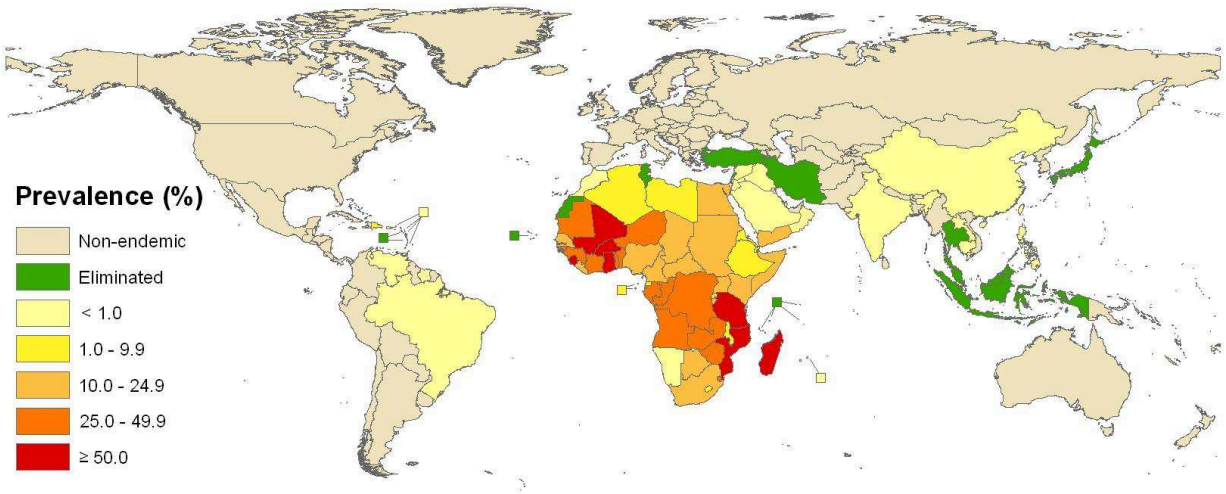

Figure 1.2: Global schistosomiasis prevalence map (source: Utzinger et al. (2011)).

Peninsula with *S. mansoni* being also present in South America and the Caribbean. *S. japonicum* is restricted to China, the Philippines and Indonesia. *S. mekongi* and *S. intercalatum* are only of regional importance in Cambodia and Lao People's Democratic Republic (*S. mekongi*) and Central Africa (*S. intercalatum*) (Gryseels et al., 2006; Hotez et al., 2006a; Davis, 2009).

Globally, almost 800 million people are believed to be at risk of schistosomiasis with more than 200 million infections. This relates to approximately 120 million individuals estimated to suffer from clinical manifestations and about 20 million severe morbidity cases (Chitsulo et al., 2000; Steinmann et al., 2006). Persons of particular risk are school-aged children (WHO, 2002). The global schistosomiasis burden has been estimated at 1.7-4.5 million DALYs lost (WHO, 2002; Utzinger and Keiser, 2004) but might actually be several times higher than the upper estimate due to discrepancies in schistosomiasis disability weight assignments (King et al., 2005; Hotez, 2009). The main disease burden is concentrated in Africa with approximately 97% of infections and 200,000 deaths per year (van der Werf et al., 2003; Steinmann et al., 2006). Country-specific schistosomiasis prevalence estimates and previously eliminated areas as of mid-2003 are shown in Figure 1.2. However, these statistics are largely based on population-adjusted data originally published by Utroska and colleagues in 1989 (Utroska et al., 1989). These estimates are likely to be outdated due to ecological transformations, socio-economic development and control interventions. Recent national wide schistosomiasis surveys are sparse and only exist for Ghana (Biritwum et al., unpublished data) and Sierra Leone (Koroma et al., 2010), while Mali, Burkina Faso and Niger are covered by a sub-national survey (Clements

et al., 2009b). Therefore, accurate estimates of the number of infected individuals and high-risk areas are essential tools for planning, coordination and evaluation of control activities.

1.1.5 Determinants of transmission

The geographical distribution of schistosomiasis is focal and results from a complex interaction between various factors acting on the human definitive and the snail intermediate host. Transmission depends upon the presence of freshwater sources and host snail species. Key determinants of snail species distribution are environmental factors, such as climate (precipitation, temperature), vegetation and geographical conditions (water flow velocity, soil-related parameters) which are often interrelated. Human exposure to contaminated water is mainly influenced by sanitary facilities, behavioural factors, land use, irrigation methods, population movement (urbanization, migration), socio-economic and health system-related factors. Additionally, ecological transformations due to human alteration (e.g. construction of dams and water management activities) have been shown to be important determinants of the distribution of the major intermediate host snail species and disease transmission (Steinmann et al., 2006; Stensgaard et al., 2011).

Infection intensity depends on the frequency and duration of body surface exposure during water-based activities like farming, washing or swimming (Davis, 2009). Highest prevalence and intensities of infection are observed in school-aged children because of frequent exposure (WHO, 2002). Behavioural differences in the population result in a skewed distribution of infection intensity. Few infected individuals excrete a large proportion of schistosome eggs, while the majority of individuals is only responsible for small amounts of eggs (Bradley, 1972; Polderman, 1979; Anderson and May, 1985). However, the amount of detected schistosome eggs is subject to individual day-to-day and intra-stool variation (Engels et al., 1997; Utzinger et al., 2001; Booth et al., 2003).

1.1.6 Prevention and control

Schistosomiasis prevention and control activities are aiming to achieve one or more of the following goals: reduction of (i) the amount of schistosome eggs reaching freshwater sources harboring intermediate host snail species; (ii) miracidia-snail contacts, (iii) reduction of cercariae densities; (iv) cercariae-human contacts; and (v) worm burden (Davis, 2009). The recommended and most common strategy is large-scale preventive chemotherapy with praziquantel. This approach seeks to reduce morbidity by lowering the worm

burden, and hence reducing the amount of excreted eggs, but re-infection is common. In high transmission settings, indicated by prevalence levels of at least 50%, preventive chemotherapy is recommended to be administered once a year to school-aged children. At median prevalence levels (at least 10%) and at low prevalence levels (below 10%) bi-annual treatment and treatment on primary school entry and leave, respectively, is recommended (WHO, 2002).

Before preventive chemotherapy became the key component in schistosomiasis control, transmission control with molluscicides was common. Importance decreased because large-scale application was expensive and of limited success due to rapid re-population of habitats. However, molluscicides are still a control tool but restricted to specific epidemiological settings (Davis, 2009).

Construction of adequate sanitary facilities and health education aim to increase population awareness of schistosomiasis transmission and health consequences in order to avoid contamination of freshwater sources. Provision of safe water supply systems and environmental transformations can also lead to significant long-term improvements but require large amounts of resources (Davis, 2009).

None of the presented tools alone will achieve sustainable control, because schistosomiasis transmission is a dynamic process dependent on various factors, such as the social-ecological context and the epidemiological situation. For cost-effective planning of control activities, it is essential to have reliable maps of the geographical distribution of areas with high morbidity and estimates of the number of infected individuals.

1.2 Mapping schistosomiasis transmission

Although lots of resources and efforts have been allocated to schistosomiasis control programmes, reliable burden estimates are rare and restricted to rather small areas. However, accurate spatially explicit maps of at-risk areas are valuable tools for various stakeholders and disease control managers to support decision-making on interventions and to assess the amount of treatment needed. In addition, disease risk maps are needed to monitor and evaluate effectiveness of control programmes. Schistosomiasis is an environmentally-driven disease because transmission is linked to the presence of freshwater sources and intermediate host snail species, which are sensitive to environmental conditions. Hence, schistosomiasis risk mapping is based on the availability of environmental and survey data as well as appropriate statistical methods.

1.2.1 Existing mapping efforts

Early schistosomiasis mapping efforts were not reliable because they were mainly based on climatic suitability thresholds lacking disease data (Bavia et al., 2001; Malone et al., 2001). Most empirical mapping efforts cover small geographical areas, e.g. single villages (Pinot de Moira et al., 2007) or health districts (Raso et al., 2005), or countries (Brooker et al., 2001; Clements et al., 2006a, 2009a). Empirical large-scale mapping efforts covering multiply countries currently do not exist besides few exceptions (Clements et al., 2006b, 2008, 2010). The lack of large-scale maps on the distribution of schistosomiasis is mainly due to a paucity of contemporary large-scale survey data. Therefore, existing survey data have to be compiled within a database and analysed together in order to cover large-scale areas for disease mapping.

1.2.2 Databases on schistosomiasis surveys

The first attempt to create a comprehensive compilation of historical schistosomiasis prevalence surveys at a global scale was carried out in the mid-1980s by (Doumenge et al., 1987). Recently, Brooker and colleagues (2010) collected data on soil-transmitted helminths and schistosomiasis within the global atlas of helminth infections (GAHI; `http://www.thiswormyworld.org`) project, but data access is limited. An up-to-date, open-access database of historical and contemporary schistosomiasis prevalence surveys was initiated as part of the European Union (EU)-funded CONTRAST project. A key objective of CONTRAST project was to assess the distribution of schistosomiasis risk and burden at high geographical scale for the spatial refinement of control interventions and the cost-effective allocation of scarce resources in sub-Saharan Africa. To fulfil this objective, the global database on neglected tropical diseases (GNTD database in short; `http://www.gntd.org`) was developed (Hürlimann et al., 2011).

1.2.3 Description of the GNTD database

The GNTD database is based on a systematic literature review employing various databases without restriction in time or language. Multiple sources of unpublished data, such as reports, doctoral theses, African university libraries, health research institutions or personal contacts were screened to maximise survey coverage. It assembles all available information on schistosomiasis prevalence studies, such as (i) publication-specific information about

the type/source of publication, authors and publication year; (ii) study-specific information about survey population, survey period, *Schistosoma* species and diagnostic test employed; and (iii) survey location-specific information about the number of infected individuals among those examined (stratified by age and sex if available). Post-intervention studies and studies on displaced populations (such as nomads, travelers, military personnel, expatriates) or non-representative population samples (such as HIV positives, hospital patients) were excluded, but in case baseline prevalence data for post-intervention studies were reported these data were included.

As of mid-January 2011, the GNTD database contained more than 12,000 ge-referenced survey locations for schistosomiasis in over 35 African countries. Countries with large amounts of survey locations (>500) were Mali, Cameroon, Niger, Senegal, Tanzania, Ethiopia and Nigeria (sorted in descending order). The majority of the data are surveys on *S. haematobium* (54.6%) and *S. mansoni* (41.8%). Further observed species were *S. intercalatum* (3.5%) and animal schistosomiasis species (*S. bovis*, *S. matthei* and *S. margrebowiei*) (Hürlimann et al., 2011).

Compilation of surveys increases the number of locations within the area of interest, but certain constraints need to be considered. The major drawback of data compilations is the lack of homogeneity and comparability between surveys. The broader the inclusion criteria the more diverse the studies in terms of survey population (variation in prevalence risk), survey period (variation in time), or diagnostic test (variation in diagnostic sensitivity and specificity). For instance, the GNTD database contains community and school-based surveys involving overlapping age-groups, and hence different risk groups. Simple joining of all these data, ignoring heterogeneity, introduces additional bias in the analyses and is likely to result in incorrect disease risk estimates.

1.2.4 Mapping tools

Schistosomiasis transmission is strongly linked to ecological factors, such as temperature, precipitation or soil-related parameters. Therefore, disease risk mapping requires knowledge on environment-outcome relations and the geographical distribution of environmental factors. Remote sensing (RS) is a technology for sampling (electromagnetic) radiation emitted or reflected from distant objects to extract information on the surface or atmosphere. Radiation is often detected by means of artificial satellites. RS is often used to determine spatially rich information on environmental predictors of schistosomiasis transmission. Even though various RS data types have already been implemented to study

disease transmission, further research is needed to validate and improve the proxies on environmental conditions based on ground data.

Geographical information systems (GIS) are computerized database management systems capable for the collection, storage, handling, analysis and display of all forms of geo-referenced data and are often employed to process RS data. Development and application of GIS in public health and disease risk mapping has made considerable progress over the past two decades. GIS arrange spatially defined information about a region as a set of maps (called layers) with each layer possessing information on a characteristic of the region. The global positioning system (GPS) is often used to geo-locate survey data in order to be linked with GIS.

In the mid-1980s, RS data were applied for the first time to map the occurrence of schistosomiasis in the Philippines and the Caribbean (Cross and Bailey, 1984; Cross et al., 1984). First GIS-based predictions on schistosomiasis transmission employing remotely-sensed temperature estimates were published by Malone et al. (1994) for the Nile delta in Egypt. The interest on the application of GIS and RS in the spatial epidemiology of schistosomiasis has been growing considerably (Brooker, 2002; Yang et al., 2005b; Simoonga et al., 2009) and several reviews have highlighted the potential of such techniques for disease control especially in combination with spatial statistics (Malone, 2005; Yang et al., 2005b; Brooker, 2007). Some GIS software have integrated statistical technologies that are well suited for explanatory analyses of disease epidemiology. However, these methods are inadequate to model the relationship between disease risk and its predictors, and to perform model-based predictions.

1.2.5 Statistical modeling

Statistical models identify the significant predictors of schistosomiasis transmission, give a mathematical description of the outcome-predictor relationship and provide estimates of disease risk at unsampled locations based on this relation. Locations in close proximity are characterised by similar infection risks due to shared spatial exposures. Unobserved spatially distributed exposures introduce spatial correlation to the data. Standard statistical modelling approaches are not appropriate for analysing spatially clustered data because they assume independence between locations. Ignoring potential spatial correlation in neighbouring areas could result in incorrect model estimates (Ver Hoef et al., 2001).

The type of the geographical information is influencing the choice of the spatial statistical method used to analyse data correlated in space. Three kinds of spatial data exist,

namely (i) point-level (or geostatistical) data; (ii) areal (or lattice) data; and (iii) point patterns. Spatial models introduce additional random effect parameters at each observed location or region which takes into account potential spatial correlation. Areal data are individual-level or aggregated data usually consisting of counts or rates with geographical information available over a set of regions with common boarders. Spatial correlation between areas is implemented based on a neighbouring structure. Analysis of areal data aims to identify trends and spatial patterns and to assess large-scale associations between schistosomiasis risk and environmental predictors. Point pattern data are sets of locations in a study region where a particular event occurred. Hence, the locations are not fixed but random quantities. The analysis focus on the detection of clusters in the spatial occurrence of events and associated risk factors. Geostatistical data represent observations obtained at specific locations over a continuous study area. Spatial proximity is defined by a function of distance between pairs of locations. Analysis of this type of data aims to identify the effect of covariates that determine schistosomiasis risk and to predict the outcome at new locations of the study area (referred to as kriging) (Banerjee et al., 2003).

Geostatistical models were implemented by Diggle et al. (1998) at the end of the twentieth century. These models introduce location-specific random effect parameters which take into account the underlying spatial process. The random effect parameters are assumed to be multivariate normally distributed with the covariance matrix defined as function of distance between locations. Such models typically contain large numbers of parameters and cannot be estimated by the commonly used maximum likelihood approaches (Kleinschmidt et al., 2000). Bayesian model formulation fitted via Markov chain Monte Carlo (MCMC) simulations methods is able to simultaneously estimate outcome-predictor relations as well as spatial correlation and avoid the computational problems of likelihood-based methods (Gosoniu et al., 2006). The most common MCMC methods are Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), Gibbs sampler algorithm (Gelfand and Smith, 1990) and reversible Jump MCMC (Green, 1995). These methods derive empirical approximations of the posterior distributions of the model parameters.

Bayesian geostatistical models have been applied in schistosomiasis risk mapping in various setting, for example by Raso et al. (2005), Beck-Wörner et al. (2007) and Vounatsou et al. (2009) in the region of Man, western Côte d'Ivoire; Clements et al. (2006a) in Tanzania; Clements et al. (2008) in Mali, Niger, and Burkina Faso; Clements et al. (2010) for Burundi, Uganda and parts of Kenya and Tanzania; Wang et al. (2008) in Dangtu county (Peoples Republic of China); or Yang et al. (2005a) in Jiangsu province, China).

1.2.6 Methodological issues

Various methodological problems are related to schistosomiasis risk mapping. Some of the main issues are: (i) analysis of non-stationarity and anisotropy; (ii) modelling large geostatistical data; (iii) variable selection; (iv) analysis of age-heterogeneous data; and (v) modelling of dependent diseases.

In disease risk mapping, the majority of geostatistical models are based on the assumption that spatial correlation is solely a function of distance between locations and stable throughout the study area, that is the spatial process is isotropic and stationary. This assumption might be inappropriate in the field of schistosomiasis because dry regions or other local characteristics might be less suitable for schistosomiasis transmission. Therefore, the amount of spatial correlation potentially varies between areas of the study region, referred to as non-stationarity. Some possibilities to model non-stationarity were developed by Kim et al. (2005) who partitioned the study area into random tiles assuming independent tile-specific stationary spatial processes, or by Gosoniu and Vounatsou (2011b) who assumed correlated tiles. Furthermore, the intermediate host snails of schistosomiasis are water-dependent species and spread along rivers, ponds and lake shores. Therefore, it is likely that spatial correlation is related to the direction of river flow and shores. Anisotropy is arising when association depends not only on distance but also on direction between pairs of locations. Geometric anisotropy is a special case of anisotropy defined by spatial decay parameters varying with direction between locations (Zimmerman, 1993). Geostatistical models can be easily expanded to incorporate additional model parameters accounting for the angle, range and ratio of anisotropy via a positive-definite correlation matrix (Banerjee et al., 2003). While isotropy stationary processes are well studied, non-stationary and anisotropic processes have been so far neglected in schistosomiasis risk mapping.

Very large number of survey locations (N) are often observed when dealing with data compilations. The spatial analysis of such data is computationally challenging because geostatistical model fit requires the repeated inversion of the covariance matrix (of size NxN). A number of approaches on handling large spatial dataset exist (Rue and Tjelmeland, 2002; Gemperli and Vounatsou, 2006; Paciorek, 2007). Recently, an approximation of the spatial process was proposed by Banerjee et al. (2008) based on a subset of survey locations (M, M<<N). This has the advantage that the large size of the covariance matrix is reduced to much smaller dimensions. This approach was further developed by Gosoniu et al. (2011a) and (Rumisha et al., 2011).

Schistosomiasis transmission is based on a complex relationship between numerous

(inter-related) factors. Geostatistical modelling including all relevant covariates might result in convergence problems due to the correlation between factors and large amount of modelling parameters. Expert opinions and evidence from previous publications could be used to reduce the set of parameters. However, the resulting set of covariates might not lead to the best possible results and miss some locally important factors. Widely used variable selection approaches in epidemiological applications are fitting of stepwise non-spatial regressions or univariate non-spatial models and selection of covariates based on thresholds of significance (Gosoniu et al., 2006; Kazembe et al., 2006; Raso et al., 2006a; Schur et al., 2011b). An alternative to reduce the complexity of the model are geostatistical variable selection approaches, for instance using Gibbs variable selection (George and McCulloch, 1993). The best fitting set of covariates is determined based on the posterior predictive probability of indicator variables linked to the regression coefficients that indicate presence or absence of the corresponding covariate. To our knowledge, variable selection taking into account spatial correlation has only been employed in geostatistical risk modeling by Giardina et al. (2011) and Gosoniu and Vounatsou (2011a).

A drawback of data compilations is the lack of homogeneity and comparability between surveys. Age-heterogeneity of compiled survey data complicates geostatistical disease risk estimation, because unadjusted joining of age-heterogeneous studies is likely to result in imprecise risk estimates. Studies for malaria addressed this issue by dropping surveys on the most heterogeneous age-groups (Kleinschmidt et al., 2000; Gosoniu et al., 2009). This contributes to smaller numbers of survey locations included in the analyses, and hence lower model accuracy, especially in regions with sparse data. Gemperli et al. (2006b) used mathematical transmission models to convert age-heterogeneous prevalence data to a common age-independent malaria transmission measure. This approach was further developed by Gosoniu et al. (2008) and Hay et al. (2009). In schistosomiasis risk mapping, the age-heterogeneity problem has not yet been addressed and incompatible surveys were often excluded from the analysis.

A host of communicable diseases show spatially overlapping distributions, which is called co-endemicity, leading to individuals being simultaneously infected with more than one disease (co-infection). *A priori* knowledge of areas with high risk of co-infections will enhance cost-effectiveness of integrated control programmes (Bundy et al., 1991; Brady et al., 2006). The reasons for co-infections vary depending on the diseases investigated. While some simultaneous occurring infections simply arise by chance, others are due to

shared risk factors (e.g. behavioural, environmental, demographic and socio-economic conditions), genetic predispositions or a combination of factors. Indeed, diseases are rarely independent and estimating co-endemicity by separately modeling each disease (Brooker et al., 2006) fails to account for inter-relations and might give imprecise estimates of the geographical distribution of risk. Raso et al. (2006b) and Brooker and Clements (2009) have implemented multinomial spatial models for predicting the risk of co-infection. Multinomial models depend on observed co-infection data on individuals, but these data rarely exist because most surveys screen for single infections. Joint risk analyses via shared component models have been proposed by Knorr-Held and Best (2001) and others (Tzala and Best, 2008; Kazembe et al., 2009) to detect shared and divergent patterns in the risk surface of multiple diseases, while separating the random effects into disease-specific and shared components. So far, joint risk analyses on combined survey data have not yet been implemented in schistosomiasis co-infection risk mapping.

Chapter 2

Goal and objectives

2.1 Goal

The overarching goals of the thesis are: (i) to develop Bayesian geostatistical models for the analysis of schistosomiasis survey data taking into account inherent data characteristics; and (ii) to validate and implement these models in the field of schistosomiasis to produce spatially-explicit risk estimates and number of infected individuals on regional scale in Africa.

2.2 Specific objectives

There are four specific objectives linked to these goals:

(i) development of geostatistical models for Binomial data which allow application to large data sets and estimation of the number of infected individuals in the field of schistosomiasis (Chapters 4 and 6);

(ii) development and validation of geostatistical models taking into account age-heterogeneity by incorporating alignment factors for large-scale schistosomiasis mapping (Chapters 5 and 6);

(iii) development and validation of methods for anisotropic prevalence data for mapping schistosomiasis transmission (Chapter 7); and

(iv) development and validation of shared component models to improve co-infection risk predictions from single disease surveys (Chapter 8).

The above mentioned models were applied on data extracted from the GNTD database and national survey data from Senegal:

(i) to produce smooth large-scale schistosomiasis risk maps and to estimate the number of infected individuals in West and eastern Africa;

(ii) to identify environmental predictors and to assess the affect of soil-related factors on schistosomiasis transmission in Senegal and eastern Africa;

(iii) to obtain spatially explicit schistosomiasis risk estimates in Senegal taking into account directional effects; and

(iv) to evaluate the shared and disease-specific spatial effects on *S. mansoni*-hookworm co-infection in the region of Man, Côte d'Ivoire.

Chapter 3

Toward an open-access global database for mapping, control, and surveillance of neglected tropical diseases

Hürlimann E.[1,2], Schur N.[1,2], Boutsika K.[1,2], Stensgaard AS.[3,4], de Himpsl ML.[1,2], Ziegelbauer K.[1,2], Laizer N.[5,6], Camenzind L.[5], Di Pasquale A.[1,2], Ekpo UF.[7], Simoonga C.[8,9], Mushinge G.[8], Saarnak CFL.[4], Utzinger J.[1,2], Kristensen TK.[4], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Center for Macroecology, University of Copenhagen, Copenhagen, Denmark
[4] DBL, University of Copenhagen, Frederiksberg, Denmark
[5] Informatics, Swiss Tropical and Public Health Institute, Basel, Switzerland
[6] The Open University of Tanzania, Dar es Salaam, United Republic of Tanzania
[7] Department of Biological Sciences, University of Agriculture, Abeokuta, Nigeria
[8] Department of Community Medicine, University of Zambia, Lusaka, Zambia
[9] Ministry of Health, Lusaka, Zambia

This paper has been accepted for publication in *PLoS Neglected Tropical Diseases* (DOI: 10.1371/journal.pntd.0001404).

Abstract

Background: After many years of general neglect, interest has grown and efforts came under way for the mapping, control, surveillance, and eventual elimination of neglected tropical diseases (NTDs). Disease risk estimates are a key feature to target control interventions, and serve as a benchmark for monitoring and evaluation. What is currently missing is a georeferenced global database for NTDs providing open-access to the available survey data that is constantly updated and can be utilized by researchers and disease control managers to support other relevant stakeholders. We describe the steps taken toward the development of such a database that can be employed for spatial disease risk modeling and control of NTDs.

Methodology: With an emphasis on schistosomiasis in Africa, we systematically searched the literature (peer-reviewed journals and 'grey literature'), contacted Ministries of Health and research institutions in schistosomiasis-endemic countries for location-specific prevalence data and survey details (e.g., study population, year of survey and diagnostic techniques). The data were extracted, georeferenced, and stored in a MySQL database with a web interface allowing free database access and data management.

Principal Findings: At the beginning of 2011, our database contained more than 12,000 georeferenced schistosomiasis survey locations from 35 African countries available under `http://www.gntd.org`. Currently, the database is expanded to a global repository, including a host of other NTDs, e.g. soil-transmitted helminthiasis and leishmaniasis.

Conclusions: An open-access, spatially explicit NTD database offers unique opportunities for disease risk modeling, targeting control interventions, disease monitoring, and surveillance. Moreover, it allows for detailed geostatistical analyses of disease distribution in space and time. With an initial focus on schistosomiasis in Africa, we demonstrate the proof-of-concept that the establishment and running of a global NTD database is feasible and should be expanded without delay.

3.1 Introduction

More than half of the world's population is at risk of neglected tropical diseases (NTDs), and over 1 billion people are currently infected with one or several NTDs concurrently, with helminth infections showing the highest prevalence rates (Hotez et al., 2006b; Hotez, 2008). Despite the life-long disabilities the NTDs might cause, they are less visible and receive lower priorities compared to, for example, the 'big three', that is malaria, tuberculosis, and HIV/AIDS (WHO, 2006a; Utzinger et al., 2010), because NTDs mainly affect the poorest and marginalized populations in the developing world (Hotez, 2008; King, 2010; Utzinger et al., 2010). Efforts are under way to control or even eliminate some of the NTDs of which the regular administration of anthelmintic drugs to at-risk populations - a strategy phrased 'preventive chemotherapy' - is a central feature (Fenwick, 2006; Hotez, 2009; Lammie et al., 2006; Molyneux, 2006; Smits, 2009).

There is a paucity of empirical estimates regarding the distribution of infection risk and burden of NTDs at the national, district, or sub-district level in most parts of the developing world (Brooker et al., 2000, 2009a,b; Brooker, 2010; Simoonga et al., 2009). Such information, however, is vital to plan and implement cost-effective and sustainable control interventions where no or only sketchy knowledge on the geographical disease distribution is available. There is a risk of missing high endemicity areas and distributing drugs to places which are not at highest priority, hence wasting human and financial resources. Consequently, integrated control efforts should be tailored to a given epidemiological setting (Brooker et al., 2009a).

The establishment of georeferenced databases is important to identify areas with no information on disease burden, to foster geographical modeling over time and space, and to control and monitor NTDs. In 1987 the bilingual (English and French) 'Atlas of the Global Distribution of Schistosomiasis' was published, which entailed country-specific maps of schistosomiasis distribution based on historical records, published reports, hospital-based data, and unpublished Ministry of Health (MoH) data (Doumenge et al., 1987). While recent projects like the Global Atlas of Helminth Infections (GAHI; `http://www.thiswormyworld.org`) (Brooker et al., 2010) and the Global Atlas of Trachoma (`http://trachomaatlas.org`) (Smith et al., 2011) offer maps on the estimated spatial distribution of soil-transmitted helminthiasis, schistosomiasis, and trachoma prevalence, they do not provide the underlying data for further in-depth analyses conducted by different research groups. An open-access global parasitological database for NTDs, which provides the actual data, is not available.

The Swiss Tropical and Public Health Institute (Swiss TPH) in Basel, Switzerland, together with partners from the University of Copenhagen, Denmark, and the University of Zambia (UNZA) in Lusaka, Zambia, were working together in a multidisciplinary project to enhance our understanding of schistosomiasis transmission (the CONTRAST project) (Stothard et al., 2009; Kristensen, 2008). One of the CONTRAST goals was to create a data repository on location-specific schistosomiasis prevalence surveys in sub-Saharan Africa. In this manuscript, we describe the steps taken toward the development of such an open-access schistosomiasis database which is currently expanded to a global scale and to include other NTDs (e.g., soil-transmitted helminthiasis and leishmaniasis) and that can be constantly updated based on new publications and reports, as well as field data provided by contributors.

3.2 Materials and Methods

3.2.1 Guiding framework

We selected schistosomiasis as the first disease to establish a proof-of-concept and populate our global NTD database. Indeed, schistosomiasis affects over 200 million people worldwide, with more than 95% concentrated in Africa. Both urinary schistosomiasis (caused by the blood fluke *Schistosoma haematobium*) and intestinal schistosomiasis (causative agents: *S. mansoni* and *S. intercalatum*) are endemic in Africa (Gryseels et al., 2006; Steinmann et al., 2006).

In order to obtain a large number of geographical locations to which prevalence data can be attached to our database, we conducted a systematic review. The specific steps of the process from identification of relevant surveys to data entry in the database, including various data sources, search criteria, data extraction and entry procedures, and quality control measures, are visualized in Figure 3.1, and will be described in more detail in the following sections.

3.2.2 Data sources

We systematically searched the following electronic databases with no restriction to date and language of publication: PubMed (`http://www.pubmed.gov`), ISI - Web of Knowledge (`http://www.isiwebofknowledge.com`), and African Journal Online (AJOL; `http://www.ajol.info/`). Using specific search terms, we retrieved relevant peer-reviewed publications with an emphasis on schistosomiasis prevalence data in Africa.

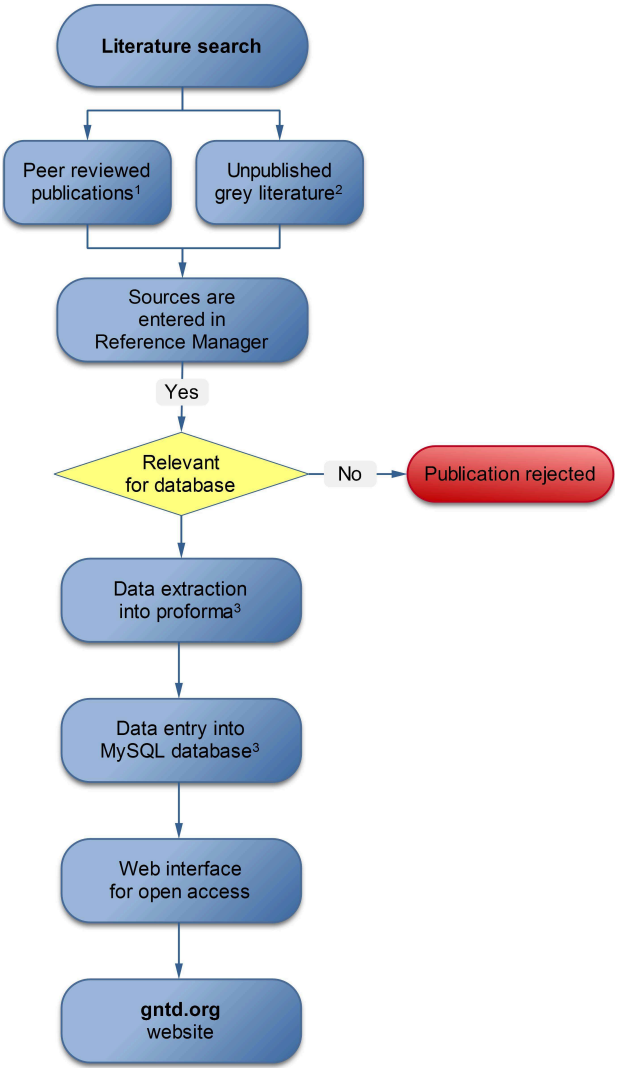

Figure 3.1: Flow-chart showing the steps used to assemble the GNTD database.
1. PubMed (`http://www.pubmed.gov`), ISI Web of Knowledge (`http://www.isiwebofknowledge.com`), African Journal Online (AJOL; `http://www.ajol.info/`), Institut de Recherche pour le Dveloppement (IRD)-resources documentaries (`http://horizon.documentation.ird.fr`), WHO library archive (`http://www.who.int/publications/en/`), Doumenge et al. (1987);
2. Dissertations and theses in local universities or public health departments, ministry of health reports, other reports and personal communication.
3. Proforma and MySQL database include: (i) data source (authors); (ii) document type; (iii) location of the survey; (iv) area information (rural or urban); (v) coordinates (lat-long in decimal degrees); (vi) method of the sample recruitment and diagnostic technique; (vii) description of survey (community-, school- or hospital-based); (viii) date of survey (month/year); and (ix) prevalence information (number of subjects examined and positive by age group and parasite species).

The keywords applied for our literature search on schistosomiasis in the electronic databases, as well as the terms for the future search strategy on other NTDs, usually consists of species names and disease expressions often abbreviated and supplied with an asterisk in order not to miss out any results due to the variety of different spellings. The search strategy can be generalized as follows: *country name OR continent AND disease* (alternative spellings were included). These keywords were combined with names of African countries, whereas also alternative or former country names were considered to have our search strategy as broad as possible. This approach enabled us to save literature search results on a country-by-country basis.

Along with articles from peer-reviewed journals, reports from health institutions (e.g., World Health Organization (WHO) and the Office de la Recherche Scientifique et Technique d'Outre-Mer (ORSTOM)/Organisation de Coordination et de Coopration pour la Lutte contre les Grandes Endmies (OCCGE)) and doctoral theses (so-called 'grey literature') compose an important literature source for schistosomiasis data. Grey literature is often restricted to internal use or is not available in an electronic format. Publication databases available from WHO (`http://www.who.int/publications/en/`) and the Institute de la Recherche pour le Dveloppement (IRD, former ORSTOM; `http://horizon.documentation.ird.fr`) offer at least partial access to such documents. Additional grey literature included was obtained directly via site visits by team members to African universities and health research and development institutions. Another important source for survey data is the direct communication with local contacts, i.e., collaborators and partners from different African countries, individual researchers, and staff from ministries of health. The majority of entries that can be retrieved in the database were extracted from peer-reviewed journals (46%), however 30.5% of the data was obtained from personal communication with authors and 23.5% from grey literature. The latter was usually more extensive in terms of survey locations than the former sources. Since the key terms we used for our systematic review were mainly species and abbreviated disease names (e.g., 'schisto*' and 'bilharz*') that are not language specific, we also extracted and included reports written in languages other than English, including French (especially for West African countries), Portuguese, Italian, Dutch, Scandinavian and few in Russian and Chinese. Sources from literature and from personal communication were stored, labeled and managed with Reference Manager 11 (Thomson ISI Research Soft).

Often, geographical information of the survey location was not given in the retrieved publications and reports (94%). Hence, we retrospectively georeferenced the locations. The

majority of the coordinates was retrieved using the GEOnet Names Server (55%) (GNS; `http://earth-info.nga.mil/gns/html/index.html`), topographic or sketch maps (23%), and GoogleMaps (14%) (`http://maps.google.com`). Personal communication with authors and local collaborators contributed another 5% of the retrospective geolocations, and only 3% were derived from other gazetteers and sources. Irrespective of the source of retrospective geolocation, we always mapped the coordinates in Google Maps to ensure that they are located in the study area and pointing to a human settlement. In general, we tried to adhere to the guidelines for georeferencing put forward by the MaNIS/HerpNEt/Ornis network to approach georeferecing in standardized manner (`http://manisnet.org/GeorefGuide.html`).

3.2.3 Data extraction

All data sources obtained (literature, data from personal communication, and field visits) were screened for relevance by applying defined inclusion and exclusion criteria. Studies were included if they comprised prevalence data of schistosomiasis, identified either by school-based or community-based surveys. We accepted different study designs (cluster sampling, random sampling, stratified sampling, systematic sampling, etc.) as long as the reported findings could be considered as representative for the population or a specific sub-group of the population (e.g., school children, women, fishermen) in a given area. Along with schistosome prevalence data, a minimal set of information was collected, such as survey location (school, village, and administrative unit), date of survey, and number of individuals examined and found schistosome-positive (irrespective of sample size). In case additional survey-specific data were available, such as infection status according to age and sex, or intermediate host snail species (i.e., *Bulinus spp.* for *S. haematobium* and *Biomphalaria spp.* for *S. mansoni*), such information was tagged, as it might be of relevance for subsequent data extraction.

Hospital-based investigations, case-control studies, drug efficacy studies, and clinical trials, as well as reports on disease infection among travelers, military personnel, expatriates, nomads, and other displaced or migrating populations were excluded from the database in order to avoid non-representative samples (e.g., individuals with symptoms or disease-related morbidity) were excluded. Thus, the data in the database reflect the actual spatial distribution of the disease at a given time point. In case baseline prevalence data were reported in the aforementioned study types, or if former migrant populations settled down and the given survey location was clearly defined, data were included. Although

having taken these precautionary steps, the database might still include prevalence data influenced by migration, since mobility and migration patterns of the rural population in sub-Saharan Africa are quite common (Shears and Lusty, 1987; Watts, 1987). Based on our exclusion criteria, we rejected more than 70% of the articles retrieved from the literature search. The rejection rate varied from country to country with a minimum rejection of 52% in Niger and a maximum of 95% in Guinea.

Once a source was identified as relevant, the data were extracted following a standard protocol with emphasize on to (i) the source of disease data such as authorship, journal, publication date, etc.; (ii) description of the parasitological survey specifying the country, the survey date (year, month, season), and the type of survey (community- or school-based); (iii) survey location reported at the highest spatial resolution available; and (iv) parasitological survey data. If relevant source included malacological data, details on snail survey methods used, snail species collected, and infection rate of the Planorbidae were also extracted.

The Kato-Katz technique for *S. mansoni* and urine filtration for *S. haematobium* diagnosis are often considered as 'gold' standard methods (WHO, 2002). If prevalence data were reported by different diagnostic methods, we only recorded in the database the results of the test with highest sensitivity and specificity. We applied the following ranking of diagnostic methods: (i) 'gold' standard; (ii) direct methods such as detection of eggs in urine/stool; and (iii) any other method such as antigen detection.

3.2.4 Database system

The data are stored and managed in a MySQL (MySQL, 1995) relational database with a web-interface built in hypertext preprocessor (PHP) (Arntzen et al., 2001). The process from prevalence data extraction to database entry is schematically depicted in Figure 3.1.

The database consists of six tables corresponding to the sections of data extraction. The system architecture supports two types of users: the administrators and the end-users (Widenius and Axmark, 2002). Registered administrators can enter new data, edit or delete existing entries under their username and password. In addition, administrators can temporarily mask confidential data as requested by authors contributing specific data. Then a summary measure is presented instead with the contact details of the data owner to enable direct communication between researchers. Users can search all records using different selection criteria, e.g., country, document category, disease, and journal. The user part was designed to fulfill the most common queries, e.g., all recorded data for a specific

parasite species in a given country or region within a specified period. The user will be able to download all information stored in the database matching different search criteria in an Excel file through an export function.

3.2.5 Data quality

To guarantee and improve data quality, the following measures have been taken. A first quality check is performed after data entry in the electronic database. For example, data extracted by assistants are always double-checked against the original source of information before becoming open-access, while data entries of senior staff are checked randomly. Data sent by contributors are inspected for completeness (e.g., in terms of study year and diagnostic technique), precise calculations (e.g., prevalence) and for correctness of coordinate information if provided. Additionally, we routinely screen the database for specific errors, i.e., by mapping survey locations and counterchecking whether the points are plotted in the expected area, by summarizing prevalence data per location and survey date to check for duplicate records, by testing for entry completeness.

Together with correctness of data extracted and entered, we also aim at the integrity of survey data. To further improve completeness of our database (e.g., date of surveys, disaggregated data) corresponding authors are contacted by e-mail asking for missing information. Approximately half of all reports had missing information, and so far we were able to get in touch with more than a third of the authors. Finally, missing coordinates for specific survey locations were obtained by re-checking additional maps and gazetteer sources, by communication with authors, and by employing global positioning system (GPS) databases created by collaborators during field visits for specific countries (i.e., Uganda, Zambia).

3.3 Results

On 10 January 2011, our database contained 12,388 survey locations for schistosomiasis that are georeferenced from 35 African countries and 568 data points on intermediate host snails for 20 African countries, giving information on 25 different mollusk species. The database is constantly updated and subjected to quality control as the project moves along. Surveys are dated as early as 1900 and the historical references that are part of the Doumenge et al. (1987) global schistosomiasis atlas are included by extracting data from the original source files. Since our main focus was on sub-Saharan Africa, the data currently included in the database covers all Western, Eastern, Middle and Southern African

countries, according to UN Population Division classification. Data extraction for Northern African countries is currently in progress. Survey coverage between countries shows considerable variation. Typically, larger numbers of survey locations were found in higher populated countries, but the amount of surveys also depends on existing national control or monitoring programs. In addition, temporal and spatial gaps in the survey distribution (as observed in Liberia, Rwanda, and Sierra Leone) might have occurred due to political instability and financial problems. The most widely used method for the diagnosis of intestinal schistosomiasis in the surveys that were fed into our repository is the Kato-Katz technique (76.7%, as single method or in combination). Stool concentration techniques accounted for a total of 13.3% (e.g., Ritchie/modified Ritchie technique (6.0%), concentration in ether solution (5.0%), merthiolate-iodine-formaline (MIF) concentration method (2.3%)) (Bergquist et al., 2009). With regard to *S. haematobium* diagnosis, microscopic examination of urine after concentration (82.0%) such as urine filtration, urine centrifugation and urine sedimentation, as well as reagent strip testing (12.8%) for the detection of blood in urine (i.e., microhematuria) or a combination of both approaches (2.3%) are most commonly employed.

Most of the surveys currently included in our database focus on school-aged children (70.1%), whereas less than a third (29.9%) of the surveys include all age groups. Furthermore, among the prevalence data of schistosomiasis collected, *S. haematobium* (54.4%) and *S. mansoni* (40.8%) were the most prevalent species. The third schistosome species parasitizing humans in Africa, *S. intercalatum* (4.8%), was only reported in surveys carried out in Cameroon and Nigeria, confirming that this species is restricted to some parts of West and Central Africa (Figure 3.2). Additionally, two zoonotic Schistosoma species were reported, namely *S. bovis* (0.02%) and *S. matthei* (0.01%), in the first cattle being the reservoir, while the latter is naturally affecting different antelope species (Table 3.1). Co-occurrence of multiple species was reported in 20.4% of the surveys, the majority of which (97.6%) was *S. mansoni-S. haematobium* co-occurrence. Currently, two schistosomiasis datasets in the GNTD database are confidential and about 100 datasets still await quality control. Hence, these data were masked and cannot yet be accessed by the database users.

The distributions of *S. mansoni* and *S. haematobium* are shown in Figure 3.3 and Figure 3.4, respectively. The applied prevalence cut-offs of 10% and 50% were chosen based on WHO recommendations to distinguish between low (<10%), moderate (between 10 and 50%) and high (>50%) endemicity communities (WHO, 2002). The compiled

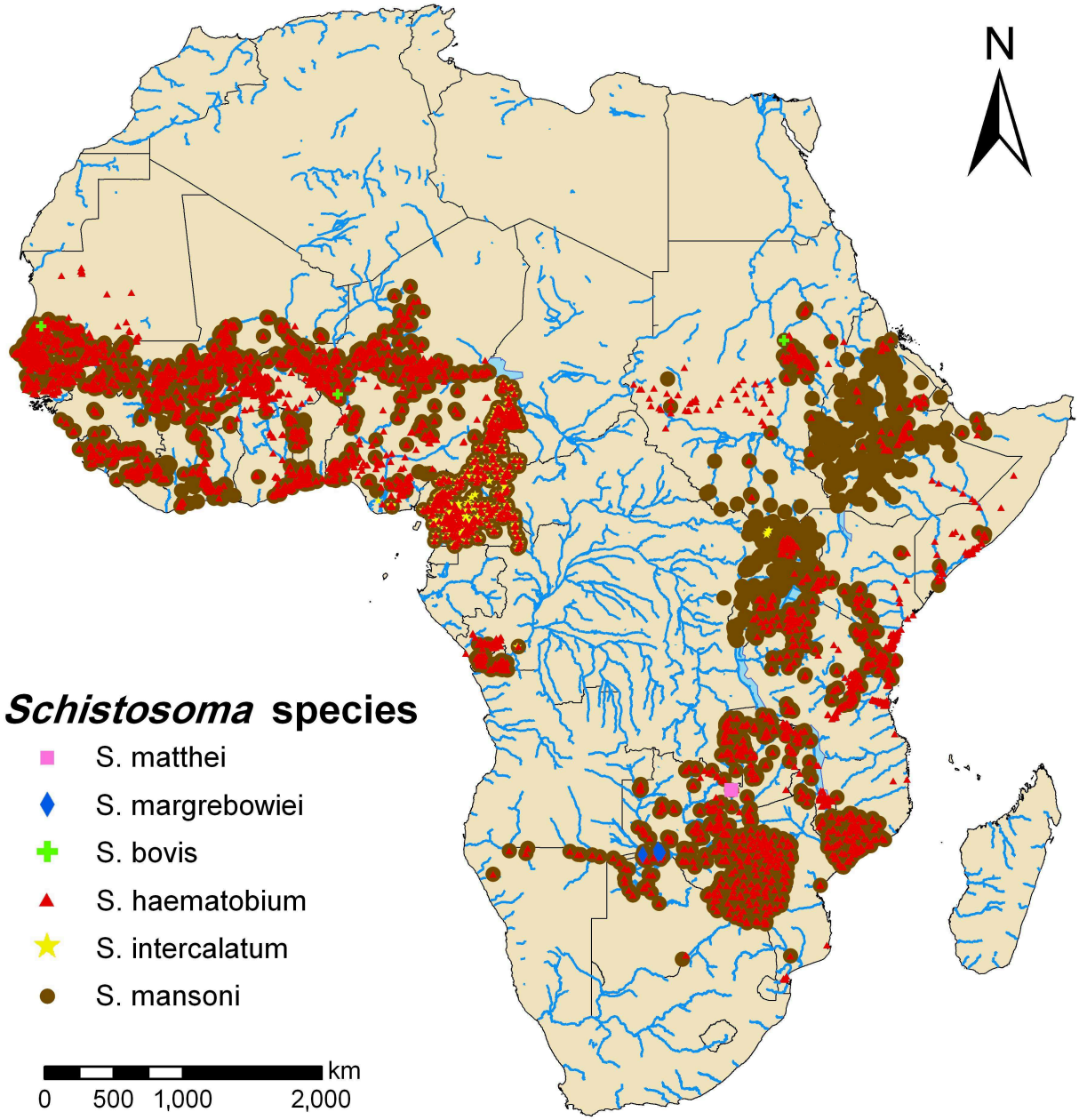

Figure 3.2: African map of schistosomiasis survey locations based on current progress of the GNTD database. Survey locations are represented by pink squares for *S. matthei*, blue diamonds for *S. margrebowiei*, yellow stars for *S. intercalatum*, green crosses for *S. bovis*, brown dots for *S. mansoni* and red triangles for *S. haematobium*. Surveys where subjects were screened for co-occurrence of multiple species are indicated with overlapping symbols.

Table 3.1: Number of *Schistosoma spp.* survey locations in the GNTD database in Africa stratified by country. Number of survey locations of *S. mansoni* (A), *S. haematobium* (B), *S. intercalatum* (C), *S. bovis* (D), *S. matthei* (E) and *S. margrebowiei* (F) as of 10 January 2011.

Countries	A	B	C	D	E	F	Total
Angola	1	1	0	0	0	0	2
Benin	15	11	0	0	0	0	26
Botswana	34	26	0	0	0	0	60
Burkina Faso	55	257	0	0	0	0	312
Burundi	87	0	0	0	0	0	87
Cameroon	467	528	415	0	0	0	1410
Congo	2	86	0	0	0	0	88
Congo DRC	129	117	1	0	0	0	247
Côte d'Ivoire	229	225	0	0	0	0	454
Djibouti	1	0	0	0	0	0	1
Eritrea	10	8	0	0	0	0	18
Ethiopia	671	107	0	0	0	0	778
Gambia	5	56	0	0	0	0	61
Ghana	22	112	0	0	0	0	134
Guinea	37	38	0	0	0	0	75
Guinea-Bissau	0	38	0	0	0	0	38
Kenya	208	193	0	0	0	0	401
Liberia	93	120	0	0	0	0	213
Malawi	23	87	0	0	0	0	110
Mali	935	1007	0	0	0	0	1942
Mauritania	51	95	0	0	0	0	146
Mozambique	96	105	0	0	0	0	201
Namibia	32	32	0	0	0	4	68
Niger	237	858	0	1	0	0	1096
Nigeria	111	406	17	0	0	0	534
Rwanda	4	0	0	0	0	0	4
Senegal	238	699	0	1	0	0	938
Sierra Leone	37	64	0	0	0	0	101
Somalia	10	69	0	0	0	0	79
Sudan	202	179	0	1	0	0	382
Tanzania	292	576	0	0	0	0	868
Togo	80	77	0	0	0	0	157
Uganda	414	57	3	0	0	0	474
Zambia	183	311	0	0	1	0	495
Zimbabwe	169	219	0	0	0	0	388
Total	5180	6764	436	3	1	4	12388

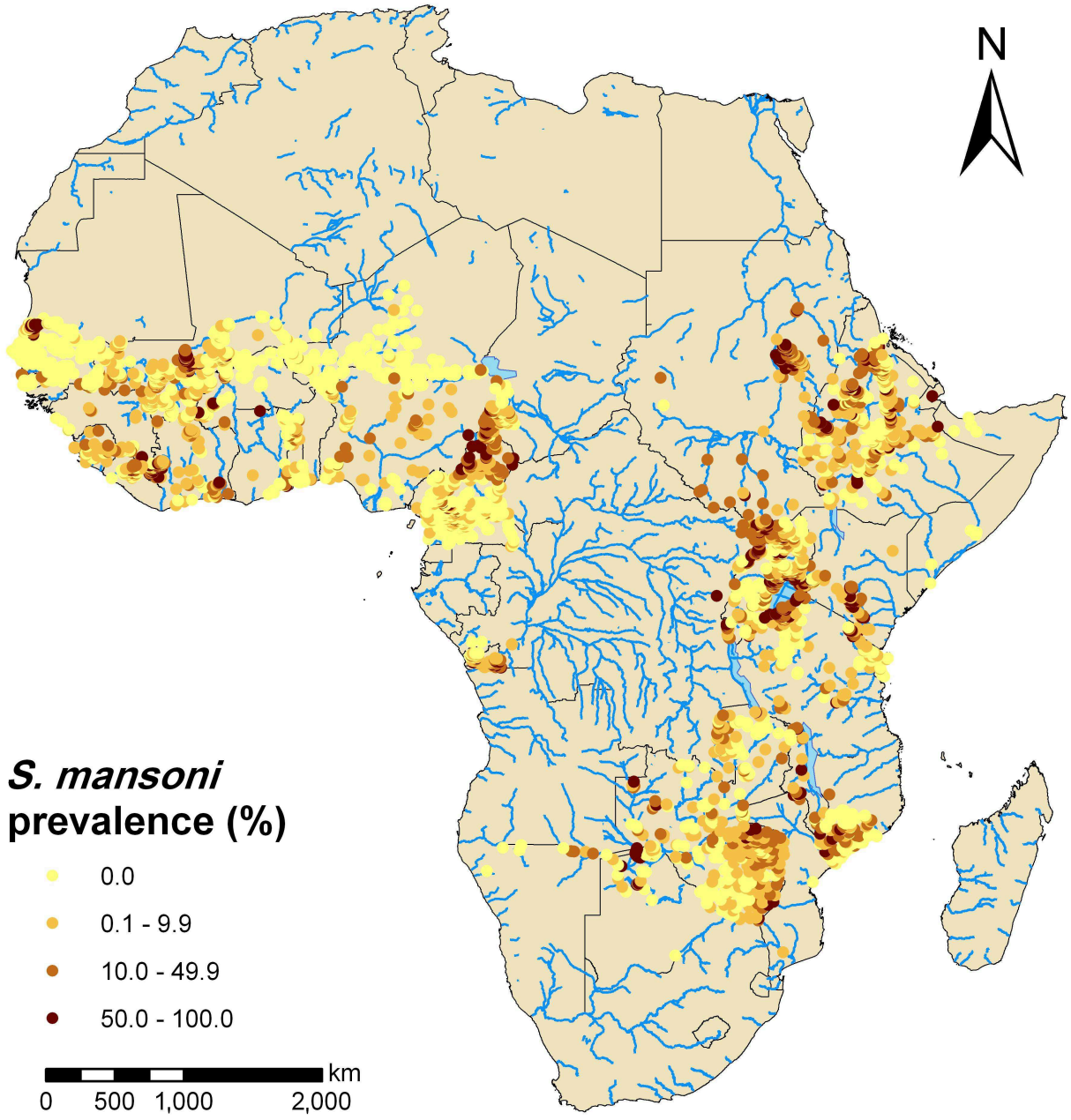

Figure 3.3: Observed prevalence of *S. mansoni* based on current progress of the GNTD database in Africa. The data included 4604 georeferenced survey locations. Prevalence equal to 0% in yellow dots, low infection rates (0.1-9.9%) in orange dots, moderate infection rates (10.0-49.9%) in light brown dots and high infection rates (>50%) in brown dots. Cut-offs follow WHO recommendations.

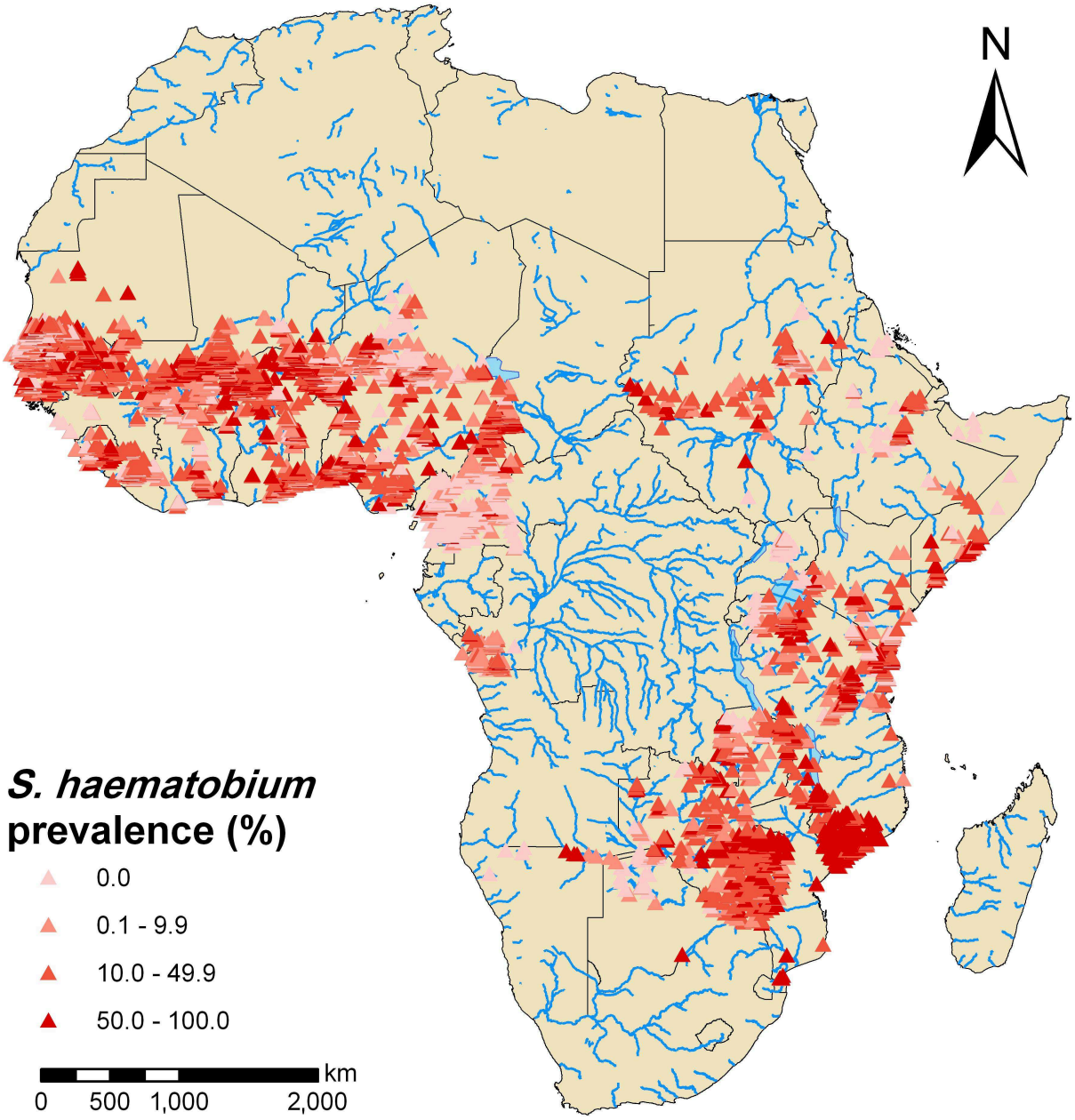

Figure 3.4: Observed prevalence of *S. haematobium* based on current progress of the GNTD database in Africa. The data included 5807 georeferenced survey locations. Prevalence equal to 0%, low infection rates (0.1- 9.9%), moderate infection rates (10.0-49.9%) and high infection rates (>50%) indicated by a red scale from light red to dark red. Cut-offs follow WHO recommendations.

survey data in the database suggest that *S. mansoni* predominates in East Africa, whereas *S. haematobium* prevalence is higher than *S. mansoni* in many African countries.

3.4 Discussion

Data repositories are important tools for the development and validation of data-driven models to estimate the distribution and burden of NTDs, such as for malaria (Le Sueur et al., 1997; Hay et al., 2009). Model-based predictions based on the compiled survey data will facilitate mapping of disease endemicity in areas without data and spatially explicit targeting of control interventions and long-term surveillance. With regard to NTDs, progress has been made for helminthic diseases (Brooker et al., 2010) and trachoma (Smith et al., 2011). The information included in a database helps to identify where current information is missing, request feedback from endemic countries, and initiate the collection of new data at those areas. Here, we described our efforts toward the establishment of an open-access database for NTDs. The database (`http://www.gntd.org`) allows for subsequent mapping of the observed survey data in order to identify high risk areas and to produce smooth risk maps, as exemplified by Schur et al. (2011b).

3.4.1 Open-access

The work presented here and the issue of open-access in relation to data, information sharing, and services, is not a new one. Indeed, we are following the successful implementations in different fields, e.g., open-access publishing (e.g., Public Library of Sciences (PLoS) and BioMed Central (BMC) journals), PubMed, genomic data (Emmert et al., 1994; Lawson et al., 2009; Ramana and Gupta, 2009), biodiversity (GBIF; `http://data.gbif.org/welcome.htm`), drug trial results (Lee et al., 2008; Pan et al., 2009), and entertainment technologies (Cohen, 2008).

With regard to epidemiological research, mapping disability, mortality, and disease burden due to infectious diseases, two recent open-access georeferenced epidemiological databases include include the Mapping Malaria Risk in Africa (MARA), which is reporting malaria survey data in Africa dating back to 1900 (Le Sueur et al., 1997), and the Malaria Atlas Project (MAP) (Hay et al., 2009), which provides maps of raw and model-based estimates of malaria risk at a global scale and country level. Other examples are the WorldWide Antimalarial Resistance Network (WWARN; `http://www.wwarn.org`) (Sibley et al., 2007), the MosquitoMap, a geospatially referenced clearinghouse for mosquito species collection records and distribution models (`http://wrbu.si.edu/mosqMap/index.htm`),

and the Disease Vectors Database (Moffett et al., 2009), which is a georeferenced database on the presence of vector species of Chagas disease, dengue, leishmaniasis, and malaria. The GAHI project created a database of schistosomiasis and soil-transmitted helminthiasis survey data (Brooker et al., 2009b, 2010), similar to our GNTD database, with the goal to provide open-access information on the global disease distribution and to highlight areas requiring mass drug administration. While the GAHI project focuses on mapping country-specific disease risk estimates, the GNTD database provides open-access to the mainly location-specific survey data. Free access to the data enables the users to conduct analyses for their own purposes. The existence of both databases offers the opportunity to join forces and to move forward in a unified way. As a first step it would be interesting to validate the two existing databases, align and harmonize them into a single comprehensive data repository, and discuss ways of harnessing synergies. Involvement of partners at WHO and other organizations will be essential.

3.4.2 Limitations

Despite the benefits of free and public data repositories, data sharing is a challenge. Data owners may hesitate to provide their data, especially when they have not yet been published. However, confidential data can be masked through a special database feature as explained in the *Methods* section. As more and more data are included into the GNTD database, the current lack in the geographical extent of location-specific survey data across countries and regions will become less critical. Undoubtedly, a host of valuable information exists within countries, in the form of unpublished local archived sources. Efforts are ongoing to access this information with the help of our in-country scientific partners in ministries of health and research institutions by visiting the countries of interest to strengthen and further expand our global network of collaborators. Nevertheless, it is likely that there will remain significant areas with scarce data because no surveys have been conducted or data are not readily accessible or have been lost in the face of civil war, political unrest, or inappropriate archiving procedures. Such geographical lacks of survey data might be only known to local experts while the international community might not be aware.

Data from systematic literature searches or unpublished reports may contain different levels of reliability. For instance, snail identification is complex and without the guidance of experienced morphologists incorrect results may be reported. The quality of diagnostic methods must also be improved, for example through repeated stool and urine sampling

over several consecutive days, since schistosome egg-output varies from one day to another. Unfortunately, only few surveys adopted such adopted such rigorous diagnosis due to generally limited financial and human resources (Utzinger et al., 2001; WHO, 2008). Furthermore, historical surveys differ in study design and are heterogeneous in terms of the age groups considered, the diagnostic methods applied, and the survey dates. Heterogeneity is also present in the way data are reported. For example in the past, numerous studies often aggregated their results at province or district level (Ouma and Waithaka, 1978; Wenlock, 1977), while currently information are frequently provided or shared at village or even individual level (Rudge et al., 2008; Yapi et al., 2005). All these points form important limitations of database compilations of epidemiological data. However, data are as limited as the sources from which they were derived. Developing standard NTD survey protocols, will enhance data comparability in the future (Gray et al., 2009).

Georeferencing historical surveys are not a straightforward undertaking. We have used a number of different sources to geolocate surveyed locations, the most common ones were described in the *Data sources* section. However, several villages may have the same name within a single country. In such cases, information regarding the administrative boundaries of the village or its distance from nearby rivers, lakes, or towns is essential. A further complication is that administrative boundaries as well as region and district names may change over time. For instance, in Uganda, 23 new districts have been created in 2005 and 2006 (Green, 2008).

In order to maintain high quality of the database, the entries are checked continually using systematic screening approaches as described in more detail in the *Methods* section. Additionally, we aim to further complement gaps (on date of survey, geographical coordinates, age group, number of people examined, etc.) and to obtain disaggregated survey data by contacting authors or collaborators, and by cross-checking new sources (maps, databases, and grey literature).

3.5 Summary and outlook

Our database is a global, freely-available, public, online resource, which hosts information pertaining to the distribution of NTDs. At present, the database contains more than 12,000 survey locations with emphasis on schistosomiasis prevalence data across Africa. It is currently expanded with information on soil-transmitted helminthiasis from Latin American and Southeast Asian countries. Our short-term goal is to extend the database

from schistosomiasis to include other NTDs (i.e., ascariasis, hookworm disease, trichuriasis, lymphatic filariasis, onchocerciasis, and trachoma). Future versions of the database will supplement prevalence information from other NTDs (Buruli ulcer, Chagas disease, cysticercosis, dracunculiasis, leishmaniasis, leprosy, and human African trypanosomiasis). The approach for inclusion of further NTDs, as well as the search strategy that is going to be applied, will be the same as described in this article. We are aware that data on soil-transmitted helminthiasis is often given alongside intestinal schistosomiasis data and could have been extracted simultaneously. However, the database evolved from the CONTRAST project that focused on schistosomiasis. While screening for schistosomiasis, we labeled relevant references on other NTDs in our reference database, which will speed up future work steps, such as literature review and data extraction of relevant sources.

The structure of the database allows entering not only parasitological data, but also other attributes, like geospatially referenced data on the disease vectors. At present, our database has limited malacological survey information, and it does not include historical collections, however, we plan to add the georeferenced historical collection compiled by the Mandahl-Barth Centre for Biodiversity and Health in Copenhagen, Denmark, which holds information on about 7,000 georeferenced snail samples.

Our hope is to provide to scientists and policy makers, a user-friendly and useful platform which is continuously updated in order to facilitate data sharing, and retrieval of disease surveillance and epidemiological data. We welcome contributions from other researchers in possession of prevalence data from various NTDs. Users may contribute by download the template offered after registration and providing the required information. An administrator checks the data for quality and sends a confirmation e-mail before including the data in the database. Researchers who may not wish to share their data may only provide limited information about the data they possess (survey location, year, and amount of data) so that the database becomes a library of potential data sources. Furthermore, we plan to add an option for the GNTD database users to contact and interact with the contributors by providing a 'send e-mail to contributor' function.

Another immediate goal is to develop a web-based interface, which will combine raw disease data and spatial model-based estimates of disease burden at different geographical levels with country boundaries and geophysical information. The results will be accessed in geo-referenced kml format, which is displayed automatically on a Google Earth interface on the website (Stensgaard et al., 2009). This will allow users to obtain estimates of disease burden at different spatial resolutions (village, district, region, country, etc.) and to display

model predictions including prediction uncertainties and raw data on the map.

A more distant option is to allow end-users to upload their own data, for instance regional and community-based health practitioners could directly upload disease prevalence to the MySQL database using hand-held smart phones with GPS functionality (Aanensen et al., 2009). Success of the project will depend on active collaboration and contribution of researchers and disease control managers from around the world. We hope that our efforts will be recognized as a helpful tool contributing to the control and eventual elimination of NTDs.

Acknowledgements

This work was carried out under the CONTRAST project (`http://www.eu-contrast.eu`) funded by European Union grant, FP6-STREP-2004-INCO-DEV Project no. 032203. JU (project no. PPOOB-102883 and PPOOB-119129) and PV (project no. 325200-118379) are grateful to the Swiss National Science Foundation. AS is supported by a PhD studentship partly funded by DHI Denmark and we thank the Danish National Research Foundation for its support of the Center for Macroecology, Evolution and Climate. We are thankful to Nadine Köhler and Marco Clementi who assisted with data extraction and software development, respectively.

Chapter 4

Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged $\leq$20 years in West Africa

Schur N.[1,2], Hürlimann E.[1,2], Garba A.[1,2,3], Traoré MS.[4], Ndir O.[5], Ratard RC.[6], Tchuem Tchuenté LA.[7,8,9], Kristensen TK.[10], Utzinger J.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Ministère de la Santé Publique et de la Lutte Contre les Endémies, Niamey, Niger
[4] Institut National de Recherche en Sant Publique, Ministre de la Sant, Bamako, Mali
[5] Faculté de Médecine, Pharmacie et Odontologie, Université Cheikh Anta Diop, Dakar, Sénégal
[6] Office of Public Health, New Orleans, Louisiana, United States of America
[7] Ministry of Public Health, Yaoundé, Cameroon
[8] Université de Yaoundé I, Yaoundé, Cameroon
[9] Centre for Schistosomiasis and Parasitology, Yaoundé, Cameroon
[10] DBL, University of Copenhagen, Frederiksberg, Denmark

This paper has been published in *PLoS Neglected Tropical Diseases* 2011, 5(6):e1194.

Abstract

Background: Schistosomiasis is a water-based disease that is believed to affect over 200 million people with an estimated 97% of the infections concentrated in Africa. However, these statistics are largely based on population re-adjusted data originally published by Utroska and colleagues more than 20 years ago. Hence, these estimates are outdated due to large-scale preventive chemotherapy programs, improved sanitation, water resources development and management, among other reasons. For planning, coordination, and evaluation of control activities, it is essential to possess reliable schistosomiasis prevalence maps.

Methodology: We analyzed survey data compiled on a newly established open-access global neglected tropical diseases database (i) to create smooth empirical prevalence maps for *Schistosoma mansoni* and *S. haematobium* for individuals aged $\leq$20 years in West Africa, including Cameroon, and (ii) to derive country-specific prevalence estimates. We used Bayesian geostatistical models based on environmental predictors to take into account potential clustering due to common spatially structured exposures. Prediction at unobserved locations was facilitated by joint kriging.

Principal Findings: Our models revealed that 50.8 million individuals aged $\leq$20 years in West Africa are infected with either *S. mansoni*, or *S. haematobium*, or both species concurrently. The country prevalence estimates ranged between 0.5% (The Gambia) and 37.1% (Liberia) for *S. mansoni*, and between 17.6% (The Gambia) and 51.6% (Sierra Leone) for *S. haematobium*. We observed that the combined prevalence for both schistosome species is two-fold lower in Gambia than previously reported, while we found an almost two-fold higher estimate for Liberia (58.3%) than reported before (30.0%). Our predictions are likely to overestimate overall country prevalence, since modeling was based on children and adolescents up to the age of 20 years who are at highest risk of infection.

Conclusion/Significance: We present the first empirical estimates for *S. mansoni* and *S. haematobium* prevalence at high spatial resolution throughout West Africa. Our prediction maps allow prioritizing of interventions in a spatially explicit manner, and will be useful for monitoring and evaluation of schistosomiasis control programs.

4.1 Introduction

Schistosomiasis is a water-based disease caused by trematodes of the genus *Schistosoma*. The five schistosome species that are known to infect humans are *Schistosoma mansoni*, *S. haematobium*, *S. intercalatum*, *S. mekongi*, and *S. japonicum*. School-aged children are at highest risk of infection and are the main target group for interventions (WHO, 2002).

Despite successful efforts to control schistosomiasis in different parts of the world, more than 200 million individuals are still estimated to be infected and the annual global burden due to schistosomiasis might exceed 4.5 million disability-adjusted life years (DALYs) lost (WHO, 2002; Steinmann et al., 2006; King et al., 2005). A substantial amount of this burden is concentrated in West Africa, including Cameroon. Indeed, 72 million infections are thought to occur in this part of the world (Chitsulo et al., 2000). However, the current statistics, as presented by Chitsulo et al. (2000), Steinmann et al. (2006), and Utzinger et al. (2009), are largely based on population re-adjusted data originally published by Utroska and colleagues in the late 1980s (Utroska et al., 1989). Hence, the estimates are likely to be outdated due to, among other reasons, large-scale preventive chemotherapy campaigns, improved sanitation, water resources development and management, and socio-economic development.

Recently, donors have provided new funds to control the so-called neglected tropical diseases (NTDs), including schistosomiasis. For cost-effective planning and evaluation of control activities, it is essential to have reliable baseline maps of the geographical distribution of at-risk population and disease burden. Early schistosomiasis mapping efforts have been based on climatic suitability thresholds (Malone et al., 2001; Bavia et al., 2001). These maps are not reliable because they are not based on disease data. Apart from a few studies (Clements et al., 2009b,a; Brooker et al., 2000, 2001), empirical maps of disease distribution over large areas are not available since there is a paucity of contemporary large-scale survey data.

The first comprehensive compilation of historical schistosomiasis prevalence surveys at a global scale was carried out by Doumenge et al. in the mid-1980s (Doumenge et al., 1987). More recent collections are available by Brooker et al. (2010) for soil-transmitted helminthiasis and schistosomiasis, but data access is limited. The European Union (EU)-funded CONTRAST project initiated the development of an open-access global NTD database, which is updated in real time (GNTD database; `http://www.gntd.org`) (Hürlimann et al., 2011). A key objective of CONTRAST is to employ this database for large-scale schistosomiasis prevalence mapping and prediction in sub-Saharan Africa for the spatial

refinement of control interventions and the cost-effective allocation of resources.

Geographical locations in close proximity share common exposures which influence the disease outcome similarly. The geographical information of the survey locations in the GNTD database allows taking into account the potential spatial correlation and therefore creation of more realistic models. Standard statistical modeling approaches assume independence between locations (Diggle et al., 1998). Ignoring potential spatial correlation in neighboring areas due to common exposures could result in incorrect model estimates (Gosoniu et al., 2006). Geostatistical models take into account spatial clustering by introducing location-specific random effect parameters in the covariance matrix by a function of distance between locations (Diggle et al., 1998). Such models typically contain large numbers of parameters and cannot be estimated by the commonly used maximum likelihood approaches (Kleinschmidt et al., 2000). Bayesian model formulations enable model fit via Markov chain Monte Carlo (MCMC) simulations (Diggle et al., 1998).

Bayesian geostatistical models have been applied in mapping schistosomiasis at different spatial scales, for example by Raso et al. (2005) in the region of Man, western Côte d'Ivoire, and Clements et al. (2008) in Mali, Niger, and Burkina Faso. Brooker et al. (2010) developed a global predictive map highlighting those areas where preventive chemotherapy against schistosomiasis and soil-transmitted helminthiasis are warrant. However, to our knowledge, there is neither a model-based *S. haematobium* nor a *S. mansoni* large-scale prevalence map and spatially explicit burden estimates for the whole West African region.

In this paper, we developed Bayesian geostatistical models based on environmental and climatic risk factors to obtain reliable empirical schistosomiasis prevalence maps for individuals aged $\leq$20 years by analyzing the GNTD data for West Africa, including Cameroon. Prediction was based on joint kriging in order to summarize the results as population-adjusted country prevalence estimates. Emphasis was placed on the distribution of *S. haematobium* and *S. mansoni*. We neglected *S. intercalatum* due to low infection risks, especially outside Cameroon.

4.2 Methods

4.2.1 Disease data

The GNTD database was used to obtain prevalence data on schistosomiasis. This database assembles general information about the type of publication, authors, and publication

year, as well as study-specific information about survey population, survey period, *Schistosoma* species, diagnostic test employed, and the number of infected individuals among those examined, stratified by age and sex (if available). Hospital studies, data on specific susceptible groups (such as HIV positives), and post-intervention studies were not included in the database (Hürlimann et al., 2011). For this study, we analyzed all point-level data on settled populations in West Africa on either *S. haematobium* or *S. mansoni*: 4550 and 2611 survey locations, respectively. We excluded (i) surveys with missing geographical coordinates; (ii) missing numbers of individuals screened; (iii) surveys carried out before 1980; (iv) individuals aged >20 years; and (v) entries based on certain diagnostic techniques. With regard to the latter exclusion criteria, we rejected all non-direct diagnostic examination techniques, such as immunofluorescence tests, antigen detections or questionnaire data, and direct fecal smears that have very low diagnostic sensitivities (overall, 4% of the data for *S. mansoni* and 0.1% for *S. haematobium* were excluded). Hence, the surveys included were mainly based on the Kato-Katz thick smear method (*S. mansoni*) and urine filtration or sedimentation (*S. haematobium*). Sensitivity and specificity of the diagnostic techniques were not incorporated in the model due to usually unknown sampling effort (e.g., number of stool samples, number of slides examined under a microscope, etc.), which affect diagnostic accuracy.

We assumed that the proportion of rejected diagnostic techniques among the data with missing information on the technique (*S. mansoni*: 33.5% missing, *S. haematobium*: 20.6% missing) is similar. Therefore, we considered the bias that would arise from ignoring the missing data as larger than the bias from potentially rejected diagnostic techniques among the missing data. A separate model validation on the reduced datasets confirmed that by including data with incomplete records the predictive ability increased compared to the model excluding this information (results not presented).

4.2.2 Climatic, environmental, and population data

Climatic, environmental, and population data were obtained from different freely accessible remote sensing data sources, as summarized in Table 4.1. Data on day and night temperature were extracted from land surface temperature (LST) data. The normalized difference vegetation index (NDVI) was used as a proxy for vegetation. Digitized maps on freshwater body sources (e.g., rivers, lakes, and wetlands) in West Africa were acquired with the characteristic of being either perennial or temporary.

Processing of the MODIS/Terra data was carried out using the ‘MODIS Reprojection

Table 4.1: Remote sensing data sources.

Data type	Source	Date	Temporal resolution	Spatial resolution
LST	MODIS/Terra[1]	2000-2008	8-days	1 km
NDVI	MODIS/Terra[1]	2000-2008	16-days	1 km
Land cover	MODIS/Terra[1]	2001-2004	Yearly	1 km
Rainfall	ADDS[2]	2000-2008	10-days	8 km
Altitude	DEM[3]	-	-	1 km
Freshwater bodies	HealthMapper[4]	-	-	Not known
Population counts	LandScan[5]	2008	-	1 km

[1] Moderate Resolution Imaging Spectroradiometer (MODIS). Available at: `https://lpdaac.usgs.gov/lpdaac/products/modis_products_table` (accessed: 5 January 2009)
[2] African Data Dissemination Service (ADDS). Available at: `http://earlywarning.usgs.gov/adds/` (accessed: 5 January 2009)
[3] Digital elevation model (DEM). Available at: `http://eros.usgs.gov/` (accessed: 4 January 2009)
[4] HealthMapper database. Available at: `http://www.who.int/health_mapping/tools/healthmapper/en/index.html` (accessed: 4 March 2009)
[5] LandScan^{TM} Global Population Database. Available at: `http://www.ornl.gov/landscan/` (accessed: 20 January 2011)

Tool' (U.S. Geological Survey, USGS, `http://lpdaac.usgs.gov`) and code implemented in Fortran 90 (DIGITAL Equipment Corporation, `http://www.fortran.com`) to summarize the temporal changes by an overall yearly average based either on the mean (NDVI, day and night LST) or the mode (land cover). Furthermore, the land cover categories, as defined by the International Geosphere-Biosphere Programme, were re-grouped into six categories as follows: (i) sparsely vegetated; (ii) deciduous forest and savanna; (iii) evergreen forest; (iv) cropland; (v) urban; and (vi) wet areas. Rainfall estimates were processed via the software IDIRSI 32 (Clarks Labs, Worcester, MA, USA). Yearly averaged rainfall was calculated as summary measure. Distance calculations to the nearest freshwater body source were done in ArcMap version 9.2 of the Environmental Systems Research Institute (ESRI; Redlands, CA, USA).

A classification scheme of West Africa into ecological zones was obtained using a demo version of the Earth Resources Data Analysis System Imagine 9.3 software (ERDAS; ERDAS Inc., Atlanta, USA). The datasets were subjected to an unsupervised classification, via the 'Iterative Self-Organizing Data Analysis Technique' (ISODATA), to map areas of environmental clustering which were further summarized into five main classes based on between-class similarities. The resulting map matched existing classifications (Global Agro-Ecological Zones, Food and Agriculture Organization) and the classes can be interpreted as (i) desert/semi-desert; (ii) sahelian zone; (iii) savannah; (iv) forest; and (v)

tropical rainforest.

Population count data obtained from LandScan for 2008 were converted to 5 x 5 km spatial resolution and adjusted to 2010 using country-specific average annual rates of change for 2005-2010 provided by the United Nations (UN) (United Nations, 2007). Estimates for the percentage of individuals aged ≤20 years among the total population per country were extracted from the U.S. Census Bureau International Database (IDB; U.S. Census Bureau, `http://www.census.gov`) for the year 2010. Population counts were linked to the percentage of children. The estimated number of infected individuals ≤20 years was calculated by combining a sample of the joint predictive posterior distribution of the disease prevalence predicted at pixel level with the population size of that age group within the pixel. The predictive posterior distribution of the number of infected individuals per country was estimated by summing up the pixel-samples and calculating summary statistics. The combined schistosomiasis prevalence (infection with *S. mansoni* or *S. haematobium* or both) was calculated on the assumption that the two infections are independent from each other, as *Schistosoma spp.* = *S. mansoni* + *S. haematobium* - (*S. mansoni* * *S. haematobium*).

Extraction of the remotely sensed data at the survey locations and at the prediction locations for the two databases was performed via a self written Fortran 90 code. The prediction surface for West Africa was built in ArcMap with a spatial resolution of 0.05 x 0.05 (approximately 5 x 5 km) resulting in approximately 220,000 pixels covering the study region. The data were displayed in ArcMap.

4.2.3 Statistical analysis

For each *Schistosoma* species, bivariate logistic regressions were performed in STATA/IC 10.1 (StataCorp LP, `http://www.stata.com`) in order to assess potential covariates in relation to the outcome (the number of infected individuals over the number of individuals screened per location). Continuous covariates were categorized into four groups based on quartiles to account for potential non-linearity in the outcome-predictor relationship on the logit. The Bayesian information criterion (BIC) was employed to detect whether linear or categorized covariates on the logits have smaller BIC and therefore predict the outcome more accurately. We used the following covariates in both linear and categorical scales: altitude, day LST, night LST, rainfall, NDVI, and distance to the nearest freshwater body. The type of freshwater body, ecological zone, and land cover were measured in categorical dimensions.

The study year was also included as linear and categorical covariate in order to account for possible temporal trends. The categories were defined on decades as follows: 1980-1989, 1990-1999, and from 2000 onwards. For *S. mansoni*, half of the data were from the 1980s (49.7%), 24.1% from the 1990s, whereas 26.2% were obtained in the new millennium. For *S. haematobium*, 37.8% of the data stem from the 1980s, 35.7% from the 1990s, and 26.5% from 2000 onwards.

Relevance of continuous or categorized covariates to predict the outcome was assessed based on p-values resulting from likelihood ratio tests (LRTs) at significance levels of 0.15. All significant covariates were included in the Bayesian analysis.

Bayesian geostatistical logistic regression models were fitted with location-specific random effects. Spatial correlation was modeled assuming that the random effects follow a multivariate normal distribution with variance-covariance matrix related to an exponential correlation function between any pair of locations. Model fit requires the inversion of this matrix. Due to the large number of survey locations in our datasets, parameter estimation becomes unfeasible. An approximation of the spatial process by a subset of m survey locations ($m < n$) proposed by Banerjee et al. (2008) and further developed by Gosoniu et al. (2011a) and Rumisha et al. (2011) was implemented instead. We employed MCMC simulation to estimate the model parameters. Prevalence of infection at 220,000 locations was predicted for the most recent decade (from the year 2000 onwards) via Bayesian kriging using joint predictive posterior distributions (Diggle et al., 1998). Due to computational issues, we modeled the multivariate Gaussian spatial process separately for each country. The performance of the models was assessed using model validation via different approaches: mean predictive errors (ME), mean absolute predictive errors (MAE), discriminatory performance on a 50% prevalence cut-off, and Bayesian credible interval (BCI) comparisons (Gosoniu et al., 2006). Further details pertaining to the Bayesian geostatistical model, sub-sampling, and model validation approaches are given in the Appendix.

4.3 Results

4.3.1 Final datasets and preliminary statistics

A schematic overview of the study profile on obtaining prevalence data on schistosomiasis from the GNTD is given in Figure 4.1. The final datasets consisted of 1993 and 1179 survey locations for *S. haematobium* and *S. mansoni*, respectively, out of which 1722 and 1094 locations were unique. Observed prevalence of the survey locations ranged from 0% to

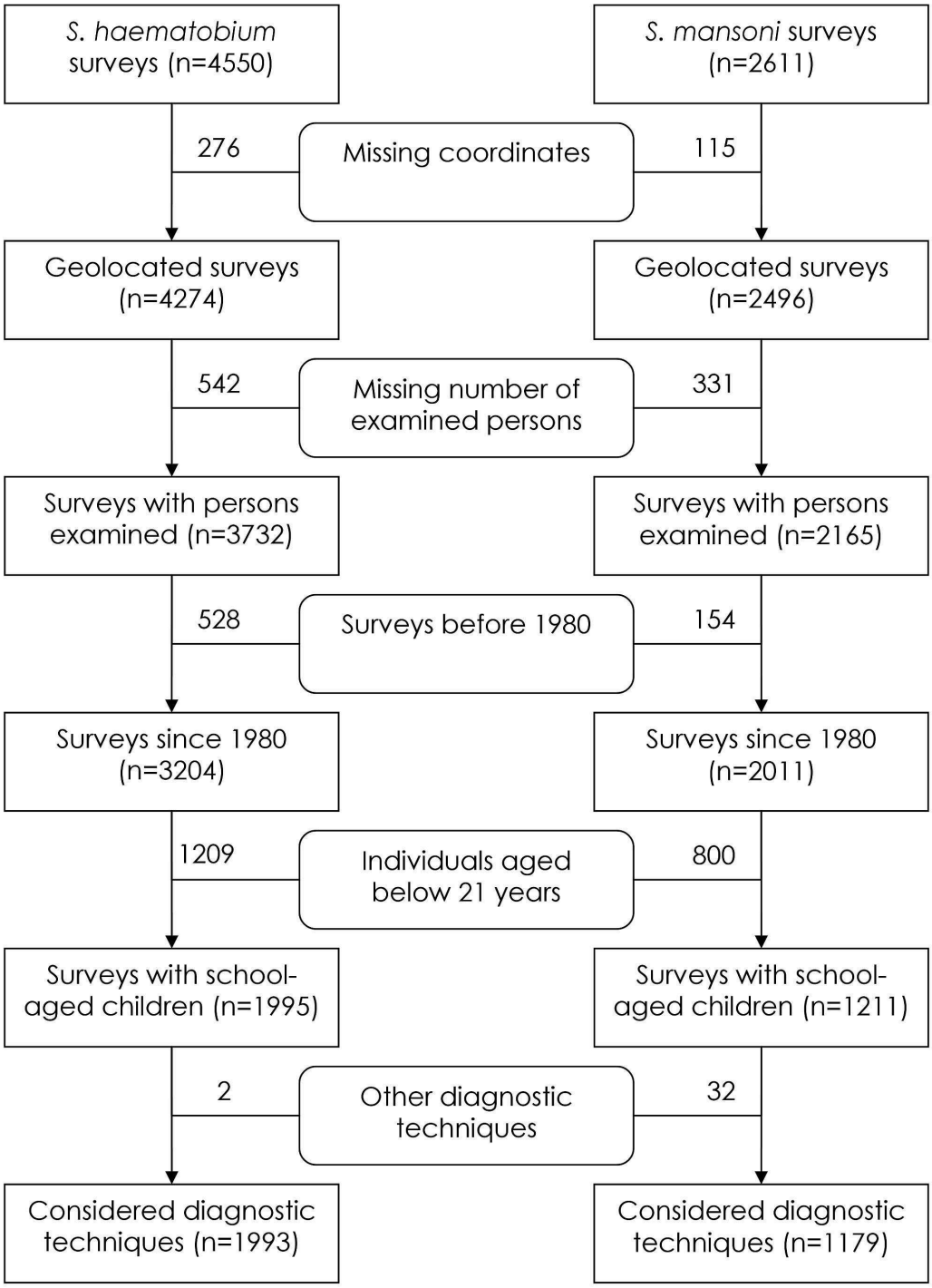

Figure 4.1: Study profile. Schematic overview of the study profiling process. The numbers in brackets in the acute-angled boxes represent the number of survey locations (which may not be unique) included in the current GNTD dataset, while the numbers outside the boxes represent the amount of survey dropped due to the reason given in the boxes with rounded corners.

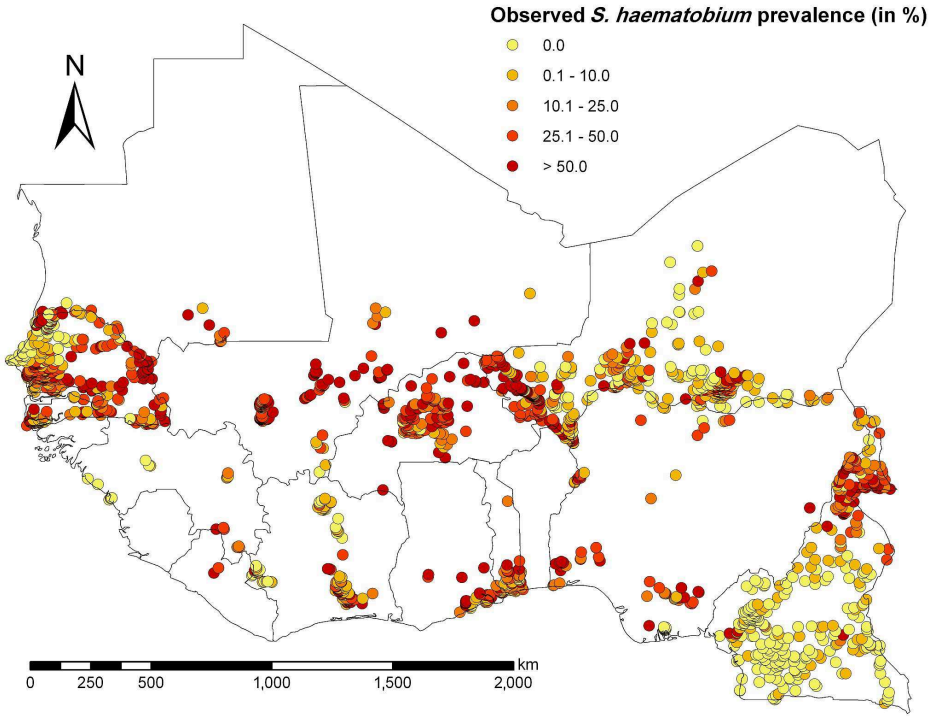

Figure 4.2: Observed prevalence of *S. haematobium* among individuals aged ≤20 years across West Africa, including Cameroon.

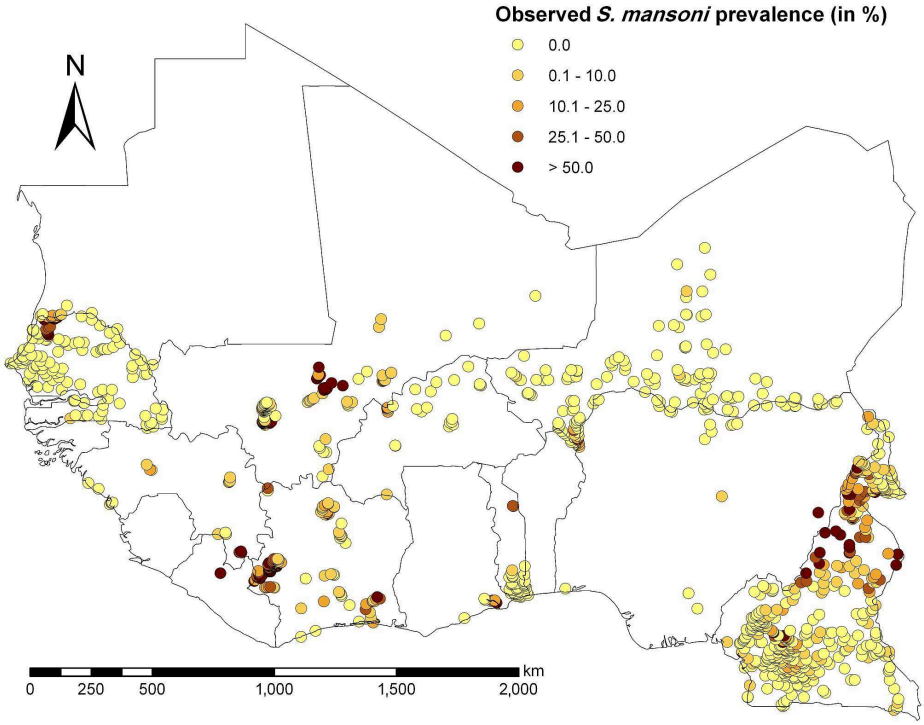

Figure 4.3: Observed prevalence of *S. mansoni* among individuals aged ≤20 years across West Africa, including Cameroon.

Table 4.2: Overview on the survey data included in the analysis stratified by country.

	Locations		Survey year			Diagnostic technique*		Survey type	Prevalence
	Total	Unique	1980s	1990s	2000+	UT	RS	School	Mean
S. haematobium									
Benin	5	5	0	5	0	5	0	5	18.2
Burkina Faso	123	119	92	8	23	35	88	117	46.4
Cameroon	349	342	335	4	10	18	0	342	22.2
Côte d'Ivoire	183	108	1	178	4	63	120	137	19.5
The Gambia	1	1	1	0	0	0	0	0	100
Ghana	47	47	22	8	17	47	0	36	38.5
Guinea	24	20	0	24	0	23	0	21	10.6
Guinea-Bissau	0	0	0	0	0	0	0	0	-
Liberia	3	2	3	0	0	3	0	0	51.3
Mali	139	130	83	23	33	137	0	33	45.4
Mauritania	27	25	8	11	8	27	0	19	34.8
Niger	544	442	104	304	136	473	0	455	32.7
Nigeria	86	71	36	21	29	80	1	48	38.3
Senegal	423	374	29	125	269	205	218	263	25.1
Sierra Leone	0	0	0	0	0	0	0	0	-
Togo	39	37	39	0	0	39	0	8	25.3
TOTAL	1993	1723	753	711	529	1155	427	1484	31
S. mansoni						KK	Other		
Benin	0	0	0	0	0	0	0	0	-
Burkina Faso	28	24	0	5	23	23	5	28	11.7
Cameroon	416	412	403	1	12	13	0	415	9.7
Côte d'Ivoire	201	157	12	118	71	200	0	141	33.3
The Gambia	0	0	0	0	0	0	0	0	-
Ghana	8	8	7	0	1	1	7	7	8.8
Guinea	22	20	0	22	0	22	0	20	12.7
Guinea-Bissau	0	0	0	0	0	0	0	0	-
Liberia	2	1	2	0	0	1	1	0	72.8
Mali	132	124	80	22	30	131	0	32	19.9
Mauritania	19	17	0	11	8	19	0	19	9.4
Niger	170	159	36	0	134	130	40	155	2.2
Nigeria	7	7	5	1	1	4	3	3	5.5
Senegal	133	126	0	104	29	132	0	27	18.2
Sierra Leone	0	0	0	0	0	0	0	0	-
Togo	41	39	41	0	0	38	3	8	4.4
TOTAL	1179	1094	586	284	309	714	59	855	17.7

Details given on the number of surveys per survey year, diagnostic technique, survey type, and observed mean prevalence given per country and Schistosoma species.
* UT = urine test, RS = reagent strip, KK = Kato Katz thick smear

100% for each *Schistosoma* species with mean prevalence of 31.0% (median 15.0%, standard deviation (SD) 29.0%) for *S. haematobium*, and 17.7% (median 0.0%, SD 24.4%) for *S. mansoni*. The distribution and the prevalence level of the survey locations are shown in Figure 4.2 and Figure 4.3 for *S. haematobium* and *S. mansoni*, respectively. An overview of the number of surveys with details given regarding sampling period, diagnostic technique, survey type, and mean prevalence, stratified by country, is given in Table 4.2.

Spatial distributions of potential covariates influencing the distribution of schistosomiasis are presented in Figure 4.4. Bivariate logistic regressions of the continuous factors in relation to the disease outcomes showed that categorical variables predicted better based on BIC values than linear variables for both *Schistosoma* species (results not presented). Each potential covariate considered for the analyses had a p-value of <0.001 based on LRTs and was therefore included in the multivariate analyses. Backwards logistic regressions demonstrated the importance of the whole set of covariates for each species. The resulting odds ratios (ORs) of bivariate and multivariate non-spatial logistic regressions are summarized in Table 4.3 for *S. haematobium*, and Table 4.4 for *S. mansoni*. The only non-significant outcome-predictor relations in a multivariate framework for the former species were yearly averaged precipitation between 300 mm and 399 mm, and NDVI levels between 0.33 and 0.52. For the latter species, only altitude levels of at least 500 m above sea level and night LSTs between 20.0 C and 20.7 C were non-significant.

4.3.2 Spatial modeling outcomes

Model parameter estimates for *S. haematobium* and *S. mansoni* are presented in Table 4.3 and Table 4.4, respectively. Introduction of spatial correlation led to changes in the significance of covariates and the direction of outcome-predictor relations compared to the corresponding non-spatial multivariate logistic regression models. For example, the influence of rainfall for *S. mansoni* became more important while the effect of the survey period and non-perennial freshwater bodies was reduced. The spatial range was estimated to be 398 km (95% BCI: 384-412 km) and 387 km (95% BCI: 375-402 km) for *S. haematobium* and *S. mansoni*, respectively. These estimates suggest strong spatial correlation for both species. The spatial variation was similar for the two species (4.02 for *S. haematobium* vs. 4.05 for *S. mansoni*).

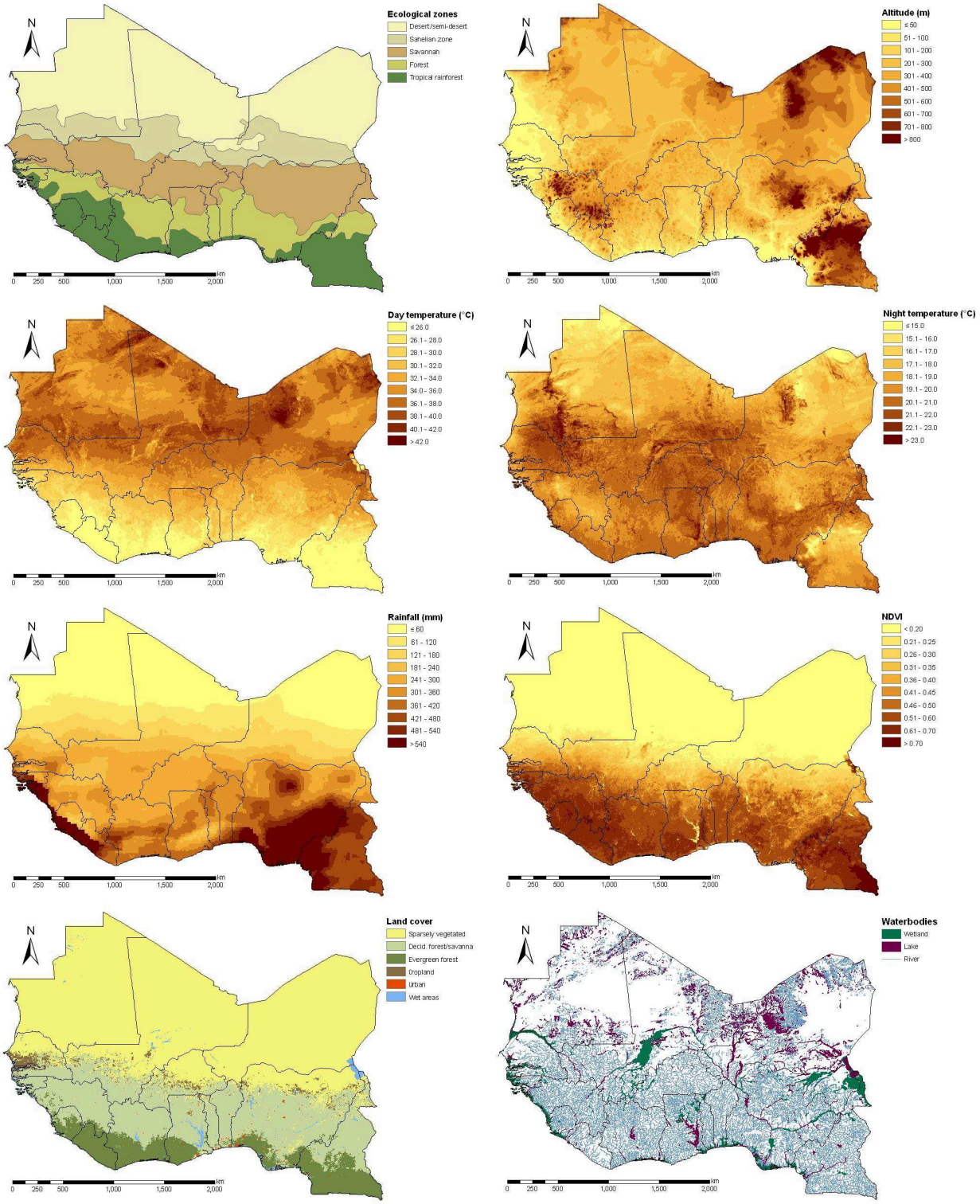

Figure 4.4: Spatial distribution of remotely sensed covariates for West Africa, including Cameroon. Climatic covariates were summarized via yearly averages.

Table 4.3: Logistic regression parameter estimates for *S. haematobium*.

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Decade			
1980-1989	1	1	1
1990-1999	1.09 (1.07, 1.12)*	1.22 (1.18, 1.25)*	1.26 (1.22, 1.30)*
2000 onwards	1.16 (1.13, 1.19)*	1.26 (1.22, 1.29)*	1.14 (1.09, 1.20)*
Ecological zone			
Tropical rainforest	1	1	1
Forest	1.61 (1.56, 1.67)*	1.45 (1.40, 1.51)*	1.70 (1.63, 1.77)*
Savannah	2.05 (1.99, 2.12)*	2.33 (2.21, 2.46)*	1.28 (1.21, 1.36)*
Sahelian	1.97 (1.91, 2.03)*	2.05 (1.92, 2.19)*	1.01 (0.90, 1.14)
Desert/semi-desert	1.09 (0.99, 1.19)	1.35 (1.20, 1.52)*	0.57 (0.51, 0.65)*
Altitude (m)			
≤55	1	1	1
56-224	1.83 (1.78, 1.88)*	1.59 (1.55, 1.65)*	1.51 (1.45, 1.57)*
225-408	1.28 (1.25, 1.32)*	1.11 (1.07, 1.14)*	0.91 (0.86, 0.96)*
>408	0.81 (0.78, 0.83)*	1.32 (1.27, 1.37)*	0.93 (0.86, 1.00)
Day LST (C)			
≤28.3	1	1	1
28.4-34.8	1.43 (1.39, 1.47)*	0.78 (0.75, 0.82)*	0.72 (0.68, 0.77)*
34.9-36.4	1.49 (1.45, 1.54)*	0.76 (0.71, 0.81)*	0.57 (0.53, 0.61)*
>36.4	1.19 (1.15, 1.22)*	0.63 (0.59, 0.67)*	0.49 (0.45, 0.53)*
Night LST (C)			
≤19.2	1	1	1
19.3-20.4	2.15 (2.08, 2.23)*	1.86 (1.79, 1.93)*	1.70 (1.62, 1.79)*
20.5-21.1	2.84 (2.75, 2.94)*	2.52 (2.43, 2.62)*	1.99 (1.92, 2.05)*
>21.1	3.30 (3.20, 3.42)*	3.11 (2.99, 3.23)*	2.18 (2.12, 2.25)*
Rainfall (mm)			
0-249	1	1	1
250-299	1.45 (1.41, 1.49)*	1.13 (1.09, 1.18)*	1.16 (1.13, 1.21)*
300-399	1.12 (1.08, 1.15)*	0.95 (0.91, 1.00)	0.96 (0.92, 0.99)*
≥400	0.81 (0.78, 0.83)*	0.94 (0.89, 0.99)*	0.56 (0.51, 0.61)*
NDVI			
≤0.22	1	1	1
0.23-0.32	0.97 (0.94, 1.00)	1.05 (1.02, 1.09)*	0.96 (0.93, 0.99)*
0.33-0.52	0.93 (0.91, 0.96)*	1.05 (1.00, 1.10)	1.16 (1.13, 1.21)*
>0.52	0.67 (0.65, 0.69)*	0.91 (0.85, 0.97)*	1.20 (1.15, 1.25)*
Land cover			
Sparsely vegetated	1	1	1
Deciduous forest/savanna	0.72 (0.70, 0.74)*	0.72 (0.69, 0.75)*	0.78 (0.76, 0.80)*

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Evergreen forest	0.75 (0.72, 0.77)*	1.13 (1.07, 1.20)*	1.36 (1.28, 1.42)*
Cropland	1.07 (1.04, 1.11)*	1.14 (1.10, 1.19)*	0.78 (0.75, 0.81)*
Urban	0.66 (0.64, 0.69)*	0.47 (0.45, 0.49)*	0.49 (0.46, 0.51)*
Wet areas	1.27 (1.18, 1.37)*	0.84 (0.77, 0.91)*	0.82 (0.75, 0.89)*
Distance to closest freshwater body (km)	0.95 (0.95, 0.95)*	0.98 (0.98, 0.99)*	0.98 (0.97, 0.98)*
Type of closest water body			
Perennial	1	1	1
Non-perennial	0.85 (0.83, 0.87)*	0.72 (0.70, 0.73)*	0.81 (0.78, 0.84)*

Logistic regression parameter estimates for *S. haematobium* summarized by odds ratios (OR), 95% confidence intervals (CI), and 95% Bayesian credible intervals (BCI).
*: Significant correlation based on 95% CI/BCI

Table 4.4: Logistic regression parameter estimates for *S. mansoni*.

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Decade			
1980-1989	1	1	1
1990-1999	3.17 (3.03, 3.31)*	2.70 (2.55, 2.86)*	1.60 (1.46, 1.73)*
2000 onwards	1.82 (1.73, 1.91)*	1.36 (1.28, 1.44)*	1.14 (1.02, 1.28)*
Ecological zone			
Tropical rainforest	1	1	1
Forest	0.45 (0.42, 0.49)*	0.69 (0.61, 0.77)*	1.16 (1.01, 1.34)*
Savannah	0.40 (0.39, 0.42)*	0.78 (0.68, 0.89)*	0.20 (0.18, 0.23)*
Sahelian	0.82 (0.78, 0.85)*	3.22 (2.73, 3.80)*	0.07 (0.06, 0.08)*
Desert/semi-desert	0.01 (0.01, 0.02)*	0.05 (0.01, 0.20)*	0.01 (0.01, 0.01)*
Altitude (m)			
≤185	1	1	1
186-326	2.70 (2.57, 2.83)*	4.25 (3.98, 4.53)*	2.51 (2.32, 2.69)*
327-499	1.59 (1.52, 1.68)*	2.45 (2.29, 2.63)*	1.95 (1.70, 2.25)*
>499	0.98 (0.92, 1.04)	1.06 (0.97, 1.16)	1.80 (1.58, 2.05)*
Day LST (C)			
≤25.0	1	1	1
25.1-31.2	0.83 (0.79, 0.87)*	1.45 (1.34, 1.56)*	1.34 (1.23, 1.45)*
31.3-35.6	0.78 (0.75, 0.82)*	1.90 (1.68, 2.15)*	2.05 (1.92, 2.18)*
>35.6	0.21 (0.19, 0.22)*	0.66 (0.57, 0.76)*	2.10 (1.88, 2.32)*

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Night LST (C)			
≤18.9	1	1	1
19.0-19.9	4.56 (4.30, 4.84)*	2.08 (1.94, 2.23)*	2.36 (2.18, 2.59)*
20.0-20.7	1.87 (1.76, 2.00)*	1.03 (0.95, 1.12)	0.97 (0.91, 1.03)
>20.7	0.92 (0.86, 0.99)*	0.47 (0.43, 0.51)*	0.46 (0.43, 0.50)*
Rainfall (mm)			
0-269	1	1	1
270-339	0.75 (0.71, 0.79)*	1.12 (1.03, 1.21)*	3.32 (2.89, 3.82)*
340-469	1.77 (1.69, 1.85)*	1.96 (1.77, 2.17)*	4.44 (3.97, 4.95)*
≥470	1.11 (1.05, 1.17)*	1.52 (1.36, 1.70)*	3.53 (3.16, 3.90)*
NDVI			
≤0.26	1	1	1
0.27-0.43	1.40 (1.33, 1.47)*	1.52 (1.39, 1.66)*	1.82 (1.62, 2.03)*
0.44-0.59	1.11 (1.05, 1.17)*	0.83 (0.73, 0.94)*	1.84 (1.52, 2.25)*
>0.59	2.97 (2.83, 3.12)*	1.45 (1.25, 1.67)*	0.94 (0.77, 1.15)
Land cover			
Sparsely vegetated	1	1	1
Deciduous forest/savanna	1.20 (1.14, 1.26)*	1.39 (1.28, 1.51)*	1.25 (1.17, 1.34)*
Evergreen forest	2.36 (2.26, 2.47)*	1.56 (1.40, 1.73)*	1.55 (1.45, 1.67)*
Cropland	1.46 (1.38, 1.55)*	1.51 (1.38, 1.66)*	0.82 (0.71, 0.94)*
Urban	1.41 (1.32, 1.50)*	1.27 (1.15, 1.41)*	1.72 (1.58, 1.88)*
Wet areas	0.47 (0.39, 0.57)*	0.62 (0.51, 0.76)*	0.60 (0.47, 0.77)*
Distance to closest water body (km)	0.92 (0.91, 0.92)*	0.91 (0.91, 0.92)*	0.94 (0.93, 0.94)*
Type of closest water body			
Perennial	1	1	1
Non-perennial	0.33 (0.32, 0.35)*	0.32 (0.31, 0.34)*	0.70 (0.64, 0.76)*

Logistic regression parameter estimates for *S. mansoni* summarized by odds ratios (OR), 95% confidence intervals (CI), and 95% Bayesian credible intervals (BCI).
*: Significant correlation based on 95% CI/BCI

4.3.3 Schistosomiasis prevalence maps

Figure 4.5A presents the prevalence map for *S. haematobium* based on the median of the predictions. Low-prevalence areas (predicted infection prevalence <10%) were primarily observed in the Sahara, Cameroon, north-west Côte d'Ivoire, and Senegal. Prevalence >50% are mainly spread along the Niger River, in Sierra Leone, east/central Senegal, and south Nigeria. The map of the SD of model predictions for this species (Figure 4.5B) demonstrates that small prediction errors were primarily found around the survey locations used for sub-sampling.

The median spatial *S. mansoni* prevalence map is shown in Figure 4.6A with the corresponding error presented in Figure 4.6B. High-prevalence areas (predicted prevalence >50%) were mainly found in north-east Liberia, east Côte d'Ivoire, west Ghana, north/central Benin, west Nigeria, north Cameroon, and central Mali in close proximity to Niger River. Very low prevalence areas (predicted prevalence <10%) were predominant in Senegal, The Gambia, Guinea-Bissau, Mauritania, and Niger. Furthermore, low prevalence areas were predicted for north Mali, south Togo, and parts of Cameroon. Areas of high prediction accuracy were found around the sub-sampled survey locations and in desert/semi-desert ecological zones.

4.3.4 At-risk population estimates

Table 4.5 shows population-adjusted country prevalence estimates. For *S. haematobium*, prevalence estimates range between 17.6% (The Gambia) and 51.6% (Sierra Leone), whereas for *S. mansoni* they range between 0.5% (The Gambia) and 37.1% (Liberia). *S. haematobium* was found to be the predominant species throughout West Africa with a difference compared to *S. mansoni* of up to 30% in Burkina Faso and a minimum difference of about 4% in Liberia. Combined *Schistosoma* prevalence estimates, assuming independence of the occurrence of the two species, varied from 18.1% (The Gambia) to 58.3% (Liberia) with high numbers of infected individuals aged ≤20 years (more than 5 million) in Ghana and Nigeria. Lower numbers (<1 million) of infected individuals aged ≤20 years were found in The Gambia, Guinea-Bissau, Liberia, and Mauritania. The overall number of infected individuals aged ≤20 years in West Africa is 50.8 million.

4.3.5 Model validation results

Model validation based on 80% of the survey locations resulted in MEs of -1.7 for *S. haematobium* and 0.0 for *S. mansoni*, and respective MAEs of 19.5 and 7.3. The percentage of test locations correctly predicted by 95% BCIs was 72.9% for *S. haematobium*, and 72.5% for *S. mansoni*. ME and MAE comparisons between spatial and exchangeable random effect models showed that spatial models result in better predictive ability (*S. haematobium*: ME=3.8, MAE=27.7; *S. mansoni*: ME=-0.8, MAE=14.9).

Discriminatory performance based on a 50% prevalence cut-off showed that the models correctly predicted 93.2% and 76.9% of the validation locations for *S. mansoni* and *S. haematobium*, respectively. False-high predictions were obtained for 5.5% (*S. mansoni*) and 18.8% (*S. haematobium*) of the test locations.

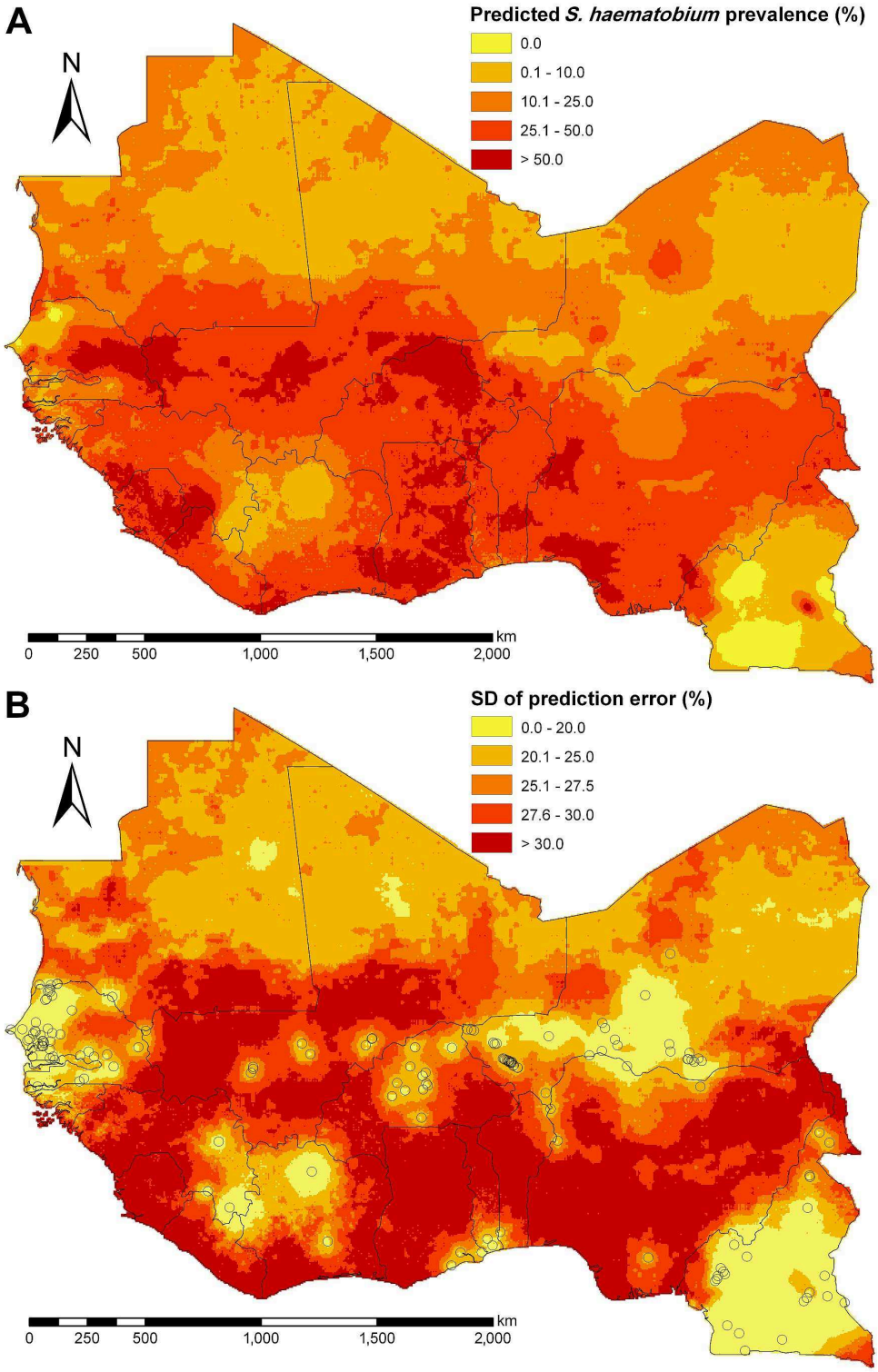

Figure 4.5: (A) Predicted median of prevalence for *S. haematobium* among individuals aged ≤20 years during the period of 2000-2009 based on Bayesian kriging, and (B) standard deviation (SD) of the prediction error with sub-sampled survey locations.

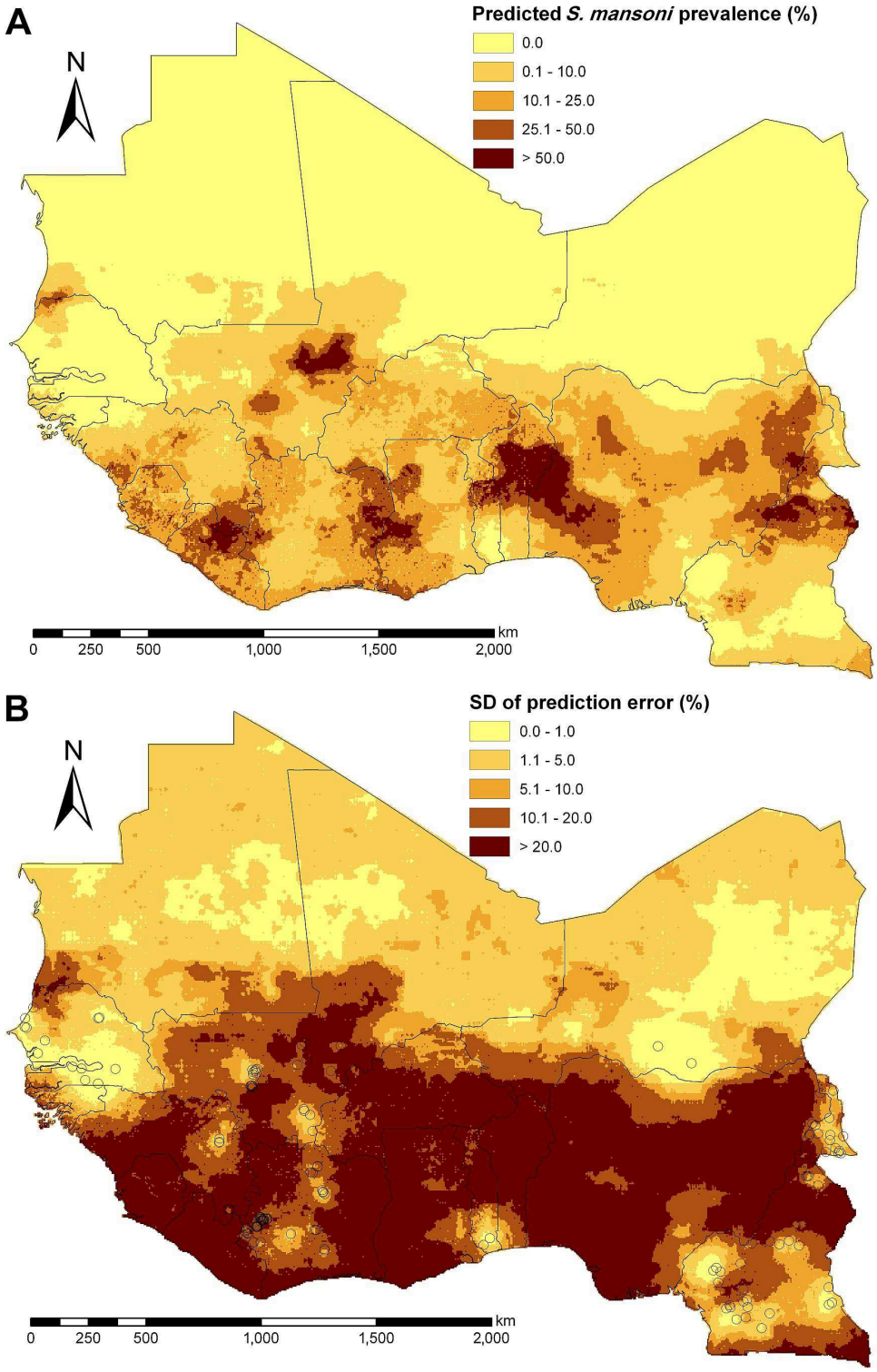

Figure 4.6: (A) Predicted median of prevalence for *S. mansoni* among individuals aged ≤20 years during the period of 2000-2009 based on Bayesian kriging, and (B) standard deviation (SD) of the prediction error with sub-sampled survey locations.

Table 4.5: Median prevalence and estimated number of infected individuals (aged ≤20 years) per country (predicted for the period 2000-2009) based on 2010 population estimates with 95% Bayesian credible interval (BCI).

Country	Population Children (x10^6)	*S. haematobium* Prevalence (%) 95% BCI	*S. haematobium* Infected (x10^6) 95% BCI	*S. mansoni* Prevalence (%) 95% BCI	*S. mansoni* Infected (x10^6) 95% BCI	*Schistosoma spp.* Prevalence (%) 95% BCI	*Schistosoma spp.* Infected (x10^6) 95% BCI	Prevalence (%)[a]	Infected (x10^6)[a]
Benin	4.62	38.8 (18.0, 63.1)	1.792 (0.830, 2.914)	20.3 (5.9, 36.5)	0.94 (0.271, 1.687)	46 (22.1, 71.1)	2.124 (1.020, 3.282)	35.5	1.95
Burkina Faso	9.434	45.4 (32.3, 59.4)	4.282 (3.043, 5.606)	15.3 (4.5, 38.2)	1.446 (0.427, 3.604)	50.2 (34.7, 67.8)	4.738 (3.274, 6.400)	60	6.24
Cameroon	10.3	20.4 (13.5, 29.0)	2.099 (1.389, 2.986)	9.2 (6.9, 12.5)	0.952 (0.715, 1.289)	25.9 (18.8, 34.7)	2.668 (1.934, 3.573)	26.5	3.02
Côte d'Ivoire	11	31.5 (16.4, 50.9)	3.229 (1.677, 5.213)	22.1 (12.6, 35.5)	2.262 (1.293, 3.642)	41.8 (25.4, 60.8)	4.286 (2.605, 6.235)	40	5.6
Gambia	4.872	17.6 (9.3, 36.9)	0.168 (0.088, 0.352)	0.5 (0.0, 5.5)	0.005 (0.000, 0.053)	18.1 (9.3, 38.7)	0.173 (0.089, 0.369)	37.5	0.33
Ghana	0.822	46.1 (26.5, 67.2)	5.077 (2.918, 7.396)	24.2 (9.8, 49.5)	2.659 (1.081, 5.452)	53.7 (31.0, 76.0)	5.912 (3.408, 8.365)	72.5	12.4
Guinea	10.3	37.4 (18.8, 57.0)	1.824 (0.914, 2.776)	20.5 (9.3, 35.9)	0.999 (0.455, 1.749)	46.4 (25.8, 66.0)	2.259 (1.255, 3.214)	25.8	1.7
Guinea-Bissau	0.953	24.7 (6.7, 59.6)	0.203 (0.055, 0.490)	2.9 (0.2, 21.3)	0.024 (0.002, 0.175)	26.5 (7.0, 63.0)	0.218 (0.057, 0.518)	30	0.33
Liberia	1.585	41.5 (14.7, 69.5)	0.658 (0.233, 1.102)	37.1 (14.1, 66.3)	0.588 (0.223, 1.051)	58.3 (24.6, 84.4)	0.924 (0.390, 1.338)	30	0.648
Mali	4.43	45.1 (27.9, 63.2)	1.997 (1.237, 2.801)	19.1 (13.0, 27.2)	0.845 (0.573, 1.204)	51.7 (35.5, 67.7)	2.291 (1.572, 3.000)	60	5.88
Mauritania	0.944	31.7 (19.0, 46.6)	0.299 (0.180, 0.44)	5.8 (2.7, 10.8)	0.055 (0.026, 0.101)	35.2 (22.0, 51.1)	0.333 (0.208, 0.483)	27.4	0.63

Continued on next page

Country	Population Children ($x10^6$)	*S. haematobium* Prevalence (%) 95% BCI	*S. haematobium* Infected ($x10^6$) 95% BCI	*S. mansoni* Prevalence (%) 95% BCI	*S. mansoni* Infected ($x10^6$) 95% BCI	*Schistosoma spp.* Prevalence (%) 95% BCI	*Schistosoma spp.* Infected ($x10^6$) 95% BCI	Prevalence (%)[a]	Infected ($x10^6$)[a]
Niger	5.16	25.6 (19.4, 33.2)	1.321 (1.001, 1.712)	3.5 (0.6, 12.1)	0.179 (0.031, 0.625)	27.1 (19.9, 35.7)	1.397 (1.028, 1.841)	26.7	2.4
Nigeria	39.9	39.4 (24.7, 55.7)	15.741 (9.866, 2.253)	23.2 (11.8, 38.0)	9.257 (4.717, 5.175)	47 (30.0, 63.9)	18.754 (11.976, 25.505)	25.2	25.83
Senegal	6.358	21 (16.7, 24.9)	1.338 (1.062, 1.581)	2.9 (1.5, 5.9)	0.183 (0.094, 0.372)	23 (18.1, 27.6)	1.464 (1.151, 1.755)	15.3	1.3
Sierra Leone	3.476	51.6 (15.4, 84.7)	1.792 (0.535, 2.944)	24.5 (4.4, 59.8)	0.853 (0.153, 2.080)	57.5 (17.6, 89.6)	1.999 (0.612, 3.113)	67.6	2.5
Togo	2.985	36.9 (18.1, 58.5)	1.102 (0.540, 1.745)	14 (3.6, 31.4)	0.419 (0.107, 0.938)	41.9 (21.0, 62.6)	1.251 (0.628, 1.869)	25.1	1.03

[a] Estimated country prevalence and number of infected individuals with schistosomiasis over all age groups in 1995 as presented by Chitsulo et al. (2000) for West Africa.

4.4 Discussion

To our knowledge, we provide the first model-based prevalence maps for both *S. haematobium* and *S. mansoni* for individuals aged ≤20 years in West Africa, including Cameroon. We used a readily available open-access database consisting of a large number of historical and contemporary geolocated and standardized survey data (Hürlimann et al., 2011), coupled with Bayesian-based geostatistical tools. Standard geostatistical methods are not able to handle large numbers of survey locations due to computational problems. Therefore, for the first time, an approximation of the spatial process was implemented in *Schistosoma* prevalence modeling.

In comparison to existing prevalence estimates, major shortcomings of previous studies have been addressed, and hence our prevalence maps show a higher spatial resolution and we believe that they are more accurate than heretofore. This claim is justified as follows. First, our estimates are based on the GNTD database that has gone live in July 2010, developed as part of the EU-funded CONTRAST project. As of February 2010, the GNTD contained more than 4500 and 2600 unique entries in West Africa for *S. haematobium* and *S. mansoni*, respectively. Second, data-tailored statistical methods based on Bayesian geostatistical modeling were used in order to incorporate spatial correlation between survey locations and to obtain more accurate estimates of the uncertainty of the predictions. Third, climatic and environmental covariates were employed in the models to evaluate the effect on the disease outcomes. The climatic and environmental factors were obtained at high spatial resolution to be able to predict small hotspots of risk, which could arise due to the focal distribution of schistosomiasis, which is an important epidemiological feature of the disease (Lengeler et al., 2002). An existing *S. haematobium* prevalence map for three West African countries (i.e., Burkina Faso, Mali, and Niger) using Bayesian geostatistical modeling was previously presented by Clements et al. (2008) based on data from 2004-2006. However, this map does not show the actual level of schistosomiasis prevalence but rather probabilities that the predicted prevalence is above a pre-defined cut-off, arbitrarily set at 50%. This cut-off has been proposed by the World Health Organization (WHO) (WHO, 2002) to distinguish between low and high risk areas, and hence such maps are useful to detect areas where preventive chemotherapy might be warranted on an annual basis. However, the maps do not provide detailed information for lower risk areas or the number of infected individuals and they cannot be used for monitoring and evaluation purposes following interventions. A more recent publication by Clements et al. (2009b) presented a *S. haematobium* prevalence map for the same three West African countries.

This map shows similar patterns to our map with the exception of north Burkina Faso. In this area, Clements and colleagues predicted prevalence levels of 10-20% for high and low egg-intensities, while our estimates suggest much higher prevalence (¿50%). These discrepancies are most likely due to differences in the underlying survey data. The Clements et al. data were only partially included in the GNTD database as we could not access them fully.

The estimated spatial correlation for both *Schistosoma* species was very strong with spatial ranges of approximately 400 km. Previously reported spatial ranges in parts of West Africa vary between 7.5 km (Raso et al., 2005) and approximately 180 km (Clements et al., 2008). However, these estimates were based on recent surveys, and hence influenced by recently established control programs. Interventions are likely to reduce the predictive power of environmental and climatic factors on the distribution of schistosomiasis and, thus, reduce spatial correlation. Similar effects were found for malaria, where historic data showed stronger spatial correlation (Gemperli et al., 2006a) than recent surveys (Gosoniu et al., 2009; Riedel et al., 2010).

We overlaid population data adjusted to 2010 on the predicted prevalence surfaces for the two *Schistosoma* species in order to obtain country-specific estimates of the number of infected individuals aged $\leq$20 years. Previous country estimates, for instance those presented by Chitsulo et al. (2000), Steinmann et al. (2006), or Utzinger et al. (2009), are interpolations of limited observations for a whole country, and hence lack empirical modeling. Chitsulo and colleagues reported a higher number of infected people for West Africa (71.8 million) compared to our estimate (50.8 million). Of note, the Chitsulo et al. estimates are based on the whole population, while our new estimates concern the age group $\leq$20 years. Moreover, the Chitsulo et al. estimates pertain to mid-1990s population estimates, compared to our adjusted estimates for the year 2010. In countries like Cameroon, The Gambia, Ghana, and Liberia, characterized by high rural-to-urban migration in the last decade, the Chitsulo et al. prevalence estimates should be treated with care due to rapid urbanization. Our study revealed that the combined prevalence of *S. haematobium* and *S. mansoni* in The Gambia, for example, is two-fold lower than previously reported by Chitsulo et al. (18.1% vs. 37.5%). However, in Benin, Guinea, Liberia, Nigeria, and Togo, we found prevalence estimates that are more than 10 percentage points higher than the previous estimates. On the one hand, differences might be related to sparse data, for example, in Benin, The Gambia, Guinea, Guinea-Bissau, Liberia, Mauritania, Nigeria, and Sierra Leone. Previous estimates failed to take into account model-based predictions on

the basis of climate, environment and disease data. Since we modeled disease prevalence on individuals aged ≤20 years (highest risk groups), the prevalence estimates correspond to the former risk group. Therefore they are likely to overestimate the prevalence in the whole population.

We estimated the country-specific overall schistosomiasis prevalence by assuming independence between the occurrence of *S. haematobium* and *S. mansoni* in each area. However, it is conceivable that simultaneous infections with both species is more frequent than expected by chance in areas where the species co-exist as infection pathways are similar and highly behavioral related. Hence, the combined prevalence estimates potentially underestimate the true schistosomiasis situation in West Africa. A modeling approach via joint spatial random effects (Schur et al., 2011a) could assess the effect of potential dependence between the species, but would increase the number of spatial parameters and is therefore computationally challenging.

We might also underestimate schistosomiasis prevalence in Cameroon, Mali, and Nigeria because of the presence of *S. intercalatum* (Chitsulo et al., 2000). We did not include this species in the analysis since the GNTD database currently only contains 17 survey locations outside Cameroon. However, it is assumed that S. intercalatum has a low prevalence (Chitsulo et al., 2000) and there are signs that this species is further declining in importance (Tchuem Tchuenté et al., 2003).

Model validation has shown that the *S. haematobium* predictions seem to overestimate the actual prevalence, while the *S. mansoni* model revealed no tendency to over- or underestimate the overall prevalence. The MAE for the *S. haematobium* model is nearly three times larger than the one for *S. mansoni*. This is expected because the mean prevalence for *S. haematobium* was about double than that for *S. mansoni*. Our models correctly predict about 72% of the survey locations when considering 95% BCIs. We are encouraged by these results, since perfect predictions are rather unlikely in reality due to the complexity of disease transmission.

However, our models are based on assumptions, which could influence model performance. We assumed that the diagnostic techniques employed have similar ability to detect an infection, but different diagnostic techniques show differences in sensitivity and specificity, which also depends on the overall prevalence and infection intensity (Bergquist et al., 2009). This might have led to an underestimation of prevalence due to the imperfect sensitivity of direct diagnostic techniques (Bergquist et al., 2009). Additional model parameters accounting for the performance of the different diagnostic techniques could be incorporated

in the models. However in the absence of detailed information regarding sampling effort, assumptions would be required which may be debatable and introduce additional biases. We are currently examining the effect of different approaches on addressing this issue on the model-based predictions.

We did not adjust the outcome according to age and sex even though the age groups differ and especially school surveys are likely to include more boys than girls due to prevailing cultural issues in many parts of West Africa. Therefore, our results are likely to be biased and potentially overestimate schistosome prevalence. However, many publications do not present stratified results by these subgroups. Age-adjustment models are feasible but difficult to implement because age-prevalence curves have to be fitted for different transmission settings (Gemperli et al., 2006b). Furthermore, disease data are often reported at wide age ranges (i.e., school-aged children) and individuals might not be well distributed within the range introducing bias even though an age-prevalence model is taken into account.

Surveys are typically conducted in endemic areas leading to high observed prevalence levels. This could result in an overestimation of prevalence in the present analysis. However, in the data we analyzed, 45% of the locations for *S. haematobium* and 73% for *S. mansoni* had an observed prevalence levels below 10%. We therefore assume that a location selection bias is unlikely. Another concern is the large amount of zero outcomes (i.e., none of the study participants found to be infected) especially for *S. mansoni* (*S. mansoni*: 54.1%; *S. haematobium*: 20.1%). To overcome this issue, zero-inflated models need to be incorporated, which modify the likelihood function and add an additional model parameter capturing the over-dispersion arising by the zeros (Vounatsou et al., 2009).

The models presented in this manuscript did only include spatial random errors, and hence we ignored potential measurement errors. Inclusion of location-specific non-spatial error terms might have improved model predictions. However, location-specific non-spatial error terms would have doubled the number of error terms leading to highly parameterized models.

We further assumed isotropic stationary models. Non-stationary models imply that the spatial random effect is varying from one region to another and is not stable throughout the study area (Gosoniu et al., 2009). This assumption has been confirmed by semi-variogram comparisons showing that the estimated spatial range parameters for *S. mansoni* differ between eco-zones. However, semi-variogram analyses did not indicate non-stationarity in the spatial distribution of *S. haematobium*. Isotropic models assume that the spatial correlation is the same within the same distance irrespective of direction (Ecker and Gelfand,

2003). This assumption might not be valid since intermediate host snails spread along rivers and lakeshores and, therefore, introduce correlation attributed to directions.

The choice and size of sub-sampled locations required to adequately approximate the spatial Gaussian process is a research area on its own in spatial statistics. Many different approaches are available to optimize selection. We implemented a method based on semi-variogram comparisons. This selection is aiming to preserve the spatial surface of the original dataset. However, it might fail to identify a sub-sample, which minimizes the prediction error. The spatially averaged predictive variance (SAPV) method proposed by Finley is trying to optimize the variance in the predictions, but implementation is computationally highly demanding (Gosoniu et al., 2011b).

Time-dependent covariates, such as the climatic factors, might have changed between the 1980s and the 2000s. However, our geographical covariates were solely based on recent remote sensing data (from 2000 onwards), because historical remote sensing data are, to our knowledge, not freely available at high spatial and temporal resolution. The long run averages of the recent data enable us to maintain high spatial resolution although they cannot capture variation in the observed outcome due to unusual climatic conditions or climate change that might have occurred since the 1980s and 1990s.

Preliminary residual analyses suggest that there is only weak temporal correlation in the data. We therefore only modeled a spatial rather than a spatio-temporal process. This led to a more parsimonious model and facilitated model fit. Nevertheless, we incorporated temporal trends in the prevalence estimation by including the survey year as covariate. Both *Schistosoma* species showed that the predicted prevalence was highest during the 1990s. This increase might be explained by water resources development and management activities (e.g., the construction of dams and irrigation systems), political unrests and civil restructuring. Water resources development and management projects might have improved the suitability of the environment for snail intermediate hosts that might have spread into previously snail-free zones together with the parasites. Since the beginning of the new millennium, a number of large-scale preventive chemotherapy programs are underway in parts of West Africa and it will be important to monitor how the prevalence of schistosomiasis changes in space and over time. The effectiveness of control interventions may vary across areas but, to our knowledge, a comprehensive database compiling this information with high spatio-temporal resolution has yet to be established.

Concluding, our country-specific *Schistosoma* prevalence estimates and numbers of individuals aged $\leq$20 years infected with either *S. mansoni*, or *S. haematobium*, or both

species concurrently presented here are useful tools for disease control managers and other stakeholders to support decision-making on interventions. Our maps can also serve as a benchmark to monitor the impact of control interventions and for long-term evaluation on transmission dynamics. Model-based estimates in areas with scarce data and high uncertainty could be improved by additional surveys to enhance our knowledge on the distribution of schistosomiasis and disease burden. We plan to further expand this work to other regions and address the issues of non-stationarity, diagnostic sensitivity, and age-heterogeneity across surveys. Finally, we will test the assumption of independence between the *Schistosoma* species to improve accuracy of the joint prevalence estimates.

Acknowledgements

Many thanks are addressed to Dr. Anna-Sofie Stensgaard for her work related to the GNTD database. Special thanks go to Mr. Dominic Gosoniu and Ms. Susan Rumisha for further development and implementation of the spatial process approximation to handle large datasets. We are also grateful to all our collaborators from Benin, Burkina Faso, Cameroon, Côte d'Ivoire, The Gambia, Ghana, Guinea, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, and Togo who contributed geolocated schistosomiasis survey data for the GNTD database.

4.5 Appendix

4.5.1 Geostatistical modelling

Let Y_i and N_i be the number of infected and screened individuals at location i ($i = 1, \ldots, n$) and p_i the probability of infection. We assume that Y_i arises from a Binomial distribution, i.e., $Y_i \sim Bin(p_i, N_i)$. The influence of covariates $\underline{X}_i$ and location-specific spatial random effects ω_i are modelled on the logit, as $logit(p_i) = \underline{X}_i^T \underline{\beta} + \omega_i$, where $\underline{\beta}$ is the vector of regression coefficients. Unobserved spatial variation is introduced on ω_i by assuming that $\underline{\omega} = (\omega_1, \ldots, \omega_n)^T$ follows a latent stationary Gaussian process over the study region, $\underline{\omega} \sim MVN(\underline{0}, \Sigma)$. Σ is a matrix with elements Σ_{ij} accounting for the covariance between any pair of locations i and j. Assuming an isotropic exponential correlation function, the matrix elements are defined by $\Sigma_{ij} = \sigma^2 exp(-\rho d_{ij})$ with spatial variance σ^2, rate of correlation decay ρ and the distance between locations d_{ij}. The data are spread over large areas and Euclidean distances are not appropriate any longer, since they are unable to account for the curvature of the surface of the Earth. Therefore, the great-circle distance was used (Vincenty, 1975). The minimum distance for which the spatial correlation is less than 5% is referred to as range and can be calculated by $3/\rho$ in the exponential correlation function setting.

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

The model was fitted using Markov chain Monte Carlo (MCMC) simulation implemented in Fortran 90 code written by the investigators using the standard numerical libraries (Numerical Algorithms Group Ltd, NAG). The code was run with two chains and a burn-in of 5000 iterations. Starting values for the chains were based on non-spatial model estimates from STATA/IC 10.1 and semi-variogram estimates for the spatial model parameters. Convergence was assessed by inspection of ergodic averages of selected model parameters during the sampling period of 50,000 iterations. The models converged after approximately 30,000 iterations. Samples of 500 iterations per chain were saved for each model.

Predictive posterior distributions at the 220,000 prediction locations were estimated via Bayesian kriging (Diggle et al., 1998) implemented in Fortran 90 using the standard numerical libraries. Our predictions are based on the period from 2000 onwards.

4.5.2 Spatial process approximation

Depending on the number of survey location, parameter estimation can be very slow or infeasible (computational costs are in the order of n^3), because the variance-covariance matrix of the spatial process $\Sigma_{n\times n}$ needs to be inverted at every iteration during the fitting and kriging process. Each of the processed datasets for *S. mansoni* and *S. haematobium* in West Africa includes more than 1000 unique survey locations, therefore it is not possible to include the full matrix to estimate spatial correlation between locations. We overcame the computational burden by an approximation of the spatial process via a subset of m survey locations ($m < n$) (Banerjee et al., 2008).

The subset was selected via balanced sampling (Deville and Tillé, 2004) with a modified inclusion probability based on the variability of the outcome (Gosoniu et al., 2011a; Rumisha et al., 2011). A grid of 15 equally sized tiles (A_i, $i = 1, \ldots, 15$) was created over the study area and each survey location was allocated to the tile surrounding it. The within-tile variability σ_a^2 and total variability σ_A^2 were assessed and the inclusion probability of a location within a specific tile a was calculated by σ_a^2/σ_A^2. Sampling of the locations based on the inclusion probability and upon the selected covariates was performed in R 2.10.0 via the 'samplecube' function of the 'sampling' library.

A semi-variogram analysis was performed to identify the minimum size of the sub-sample still preserving the spatial correlation surface of the original datasets of the two *Schistosoma* species. The location subset of choice was used in model fit as a proxy of the original locations to estimate the spatial variance and correlation decay. The model was implemented in Fortran 90 code developed by the authors.

Semi-variogram comparisons for this study suggested that samples of 150 locations preserve the original spatial correlation surface sufficiently, while smaller samples were unable to capture spatial range and variance simultaneously. We sampled different sets of locations between 50 and 300 locations before the selection of the final sub-sample. The semi-variogram of each sample was compared with the semi-variogram of the complete dataset fitted via exponential correlation functions in R 2.10.0. The results indicated that samples of at least 150 locations sufficiently preserve the original correlation structure while samples with 50 or 100 locations fail. The semi-variogram based on the sub-sample of the 150 selected locations compared to the original set for each Schistosoma species is shown in the Figure 4.7.

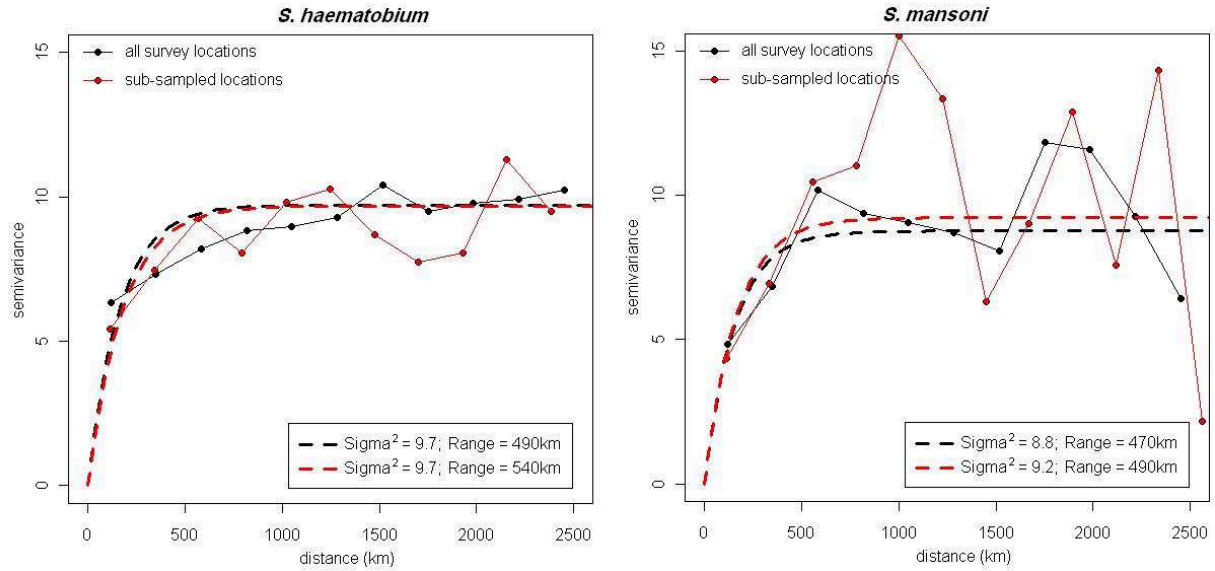

Figure 4.7: Semi-variogram comparision between sub-sampled and original *S. haematobium* and *S. mansoni* survey locations in West Africa.

4.5.3 Model validation

The performance of the models was assessed using model validation. A sample of 80% of the survey locations was employed as training set for model fit while the remaining 20% of the locations (test locations) were kept for model validation. The predicted outcomes at the test locations are compared to the observed outcomes via three different approaches: ME, MAE, and BCI comparisons (Gosoniu et al., 2006). The ME shows the overall tendency of a model to over- or underestimate prevalence and it is calculated by $ME = 1/k \sum_{i=1}^{k} p_i - \widehat{p}_i$, where p_i is the observed outcome and $\widehat{p}_i$ the median of the predictions at test location i. The MAE provides information about the accuracy of a model based on the absolute distances between predictions and observations, $MAE = 1/k \sum_{i=1}^{k} |p_i - \widehat{p}_i|$. The proportion of test locations being correctly predicted within the q-th BCI of the posterior predictive distribution (restricted by the lower centiles c_i^l and upper centiles c_i^u) is the outcome of the BCI approach, i.e., $BCI_q = \frac{1}{k} \sum_{i=1}^{k} min\left(I(c_i^l < p_i), I(c_i^u > p_i)\right)$.

Chapter 5

Modelling age-heterogeneous *Schistosoma haematobium* and *S. mansoni* survey data via alignment factors

Schur N.[1,2], Utzinger J.[1,2], Vounatsou P.[1,2]

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

This paper has been published in *Parasites & Vectors* 2011, 4(1):142.

Abstract

Background: Reliable maps of the geographical distribution, number of infected individuals and burden estimates of schistosomiasis are essential tools to plan, monitor and evaluate control programmes. Large-scale disease mapping and prediction efforts rely on compiled historical survey data obtained from the peer-reviewed literature and unpublished reports. Schistosomiasis surveys usually focus on school-aged children, whereas some surveys include entire communities. However, data are often reported for non-standard age groups or entire study populations. Existing geostatistical models ignore either the age-dependence of the disease risk or omit surveys considered too heterogeneous.

Methodology: We developed Bayesian geostatistical models and analysed existing schistosomiasis prevalence data by estimating alignment factors to relate surveys on individuals aged $\leq$20 years with surveys on individuals aged >20 years and entire communities. Schistosomiasis prevalence data for 11 countries in the eastern African region were extracted from an open-access global database pertaining to neglected tropical diseases. We assumed that alignment factors were constant for the whole region or a specific country.

Results: Regional alignment factors indicated that the risk of a *Schistosoma haematobium* infection in individuals aged >20 years and in entire communities is smaller than in individuals $\leq$20 years, 0.83 and 0.91, respectively. Country-specific alignment factors varied from 0.79 (Ethiopia) to 1.06 (Zambia) for community-based surveys. For *S. mansoni*, the regional alignment factor for entire communities was 0.96 with country-specific factors ranging from 0.84 (Burundi) to 1.13 (Uganda).

Conclusions: The proposed approach could be used to align inherent age-heterogeneity between school-based and community-based schistosomiasis surveys to render compiled data for risk mapping and prediction more accurate.

5.1 Background

An estimated 200 million individuals are infected with *Schistosoma spp.* in Africa, and yet schistosomiasis is often neglected (Utzinger et al., 2009). The global strategy to control schistosomiasis and several other neglected tropical diseases (NTDs) is the repeated large-scale administration of anthelminthic drugs to at-risk populations, an approach phrased 'preventive chemotherapy'(WHO, 2006b, 2010). The design, implementation, monitoring and evaluation of schistosomiasis control activities require knowledge of the geographical distribution, number of infected people and disease burden at high spatial resolution.

In the absence of contemporary surveys, large-scale empirical risk mapping heavily relies on analyses of historical survey data. For example, Brooker et al. (2010) compiled survey data and presented schistosomiasis (and soil-transmitted helminthiasis) risk maps within the global atlas of helminth infections (GAHI) project (`http://www.thiswormyworld.org/`). The GAHI database, however, is not fully open-access, and country-specific predictive risk maps only show probabilities of infection prevalence below and above pre-set thresholds where preventive chemotherapy is warranted (e.g. >50% of school-aged children infected, which demand annual deworming of all school-aged children and adults considered to be at risk) (WHO, 2006b). Starting in late 2006, the European Union (EU)-funded CONTRAST project developed a global database pertaining to NTDs, the GNTD database (`http://www.gntd.org`) (Hürlimann et al., 2011). This open-access database compiled raw survey data from published (i.e. peer-reviewed literature) and unpublished sources (e.g. Ministry of Health reports). It is continuously updated and data can be downloaded as soon as they are entered in the database. In early 2011, the GNTD database consisted of more than 12,000 survey locations for schistosomiasis in Africa (Hürlimann et al., 2011). The database has already been utilised for high-spatial resolution schistosomiasis risk mapping and prediction in West Africa (Schur et al., 2011b) and East/southern Africa.

An important drawback of data compilation is the lack of homogeneity and comparability between surveys, such as target population (different age groups), time of survey, diagnostic method employed, among other issues. The GNTD database is populated with schistosomiasis prevalence surveys conducted in schools, as well as in entire communities, involving different, sometimes overlapping age-groups (Hürlimann et al., 2011). However, each population sub-group carries a different risk of infection, with school-aged children and adolescence known to carry the highest risk of infection (Woolhouse, 1998; Jordan and Webbe, 1982). Simple pooling of this type of studies is likely to result in incorrect disease

risk estimates.

Schistosomiasis survey data are correlated in space because the disease transmission is driven by environmental factors (Raso et al., 2005; Clements et al., 2009b; Brooker et al., 2001). However, standard statistical modelling approaches assume independence between locations, which could result in inaccurate model estimates (Gosoniu et al., 2006). Geostatistical models take into account potential spatial clustering by introducing location-specific random effects and are estimated using Markov chain Monte Carlo (MCMC) simulations (Diggle et al., 1998). Geostatistical models have been applied on compiled survey data for disease risk prediction, for example in malaria (Gemperli et al., 2006a; Gosoniu et al., 2006; Hay et al., 2009) and helminth infections, including schistosomiasis (Schur et al., 2011b; Pullan et al., 2011).

Age-heterogeneity of survey data has been addressed in geostatistical modelling by omitting those surveys which consist of particularly heterogeneous age-groups (Schur et al., 2011b; Gosoniu et al., 2009). As a result, the number of survey locations included in the analysis is reduced, and hence model accuracy is lowered, especially in regions with sparse data. Gemperli et al. (2006b) used mathematical transmission models to convert age-heterogeneous malaria prevalence data to a common age-independent malaria transmission measure. This approach has been further developed by Gosoniu (2008) and Hay et al. (2009). To our knowledge, the age-heterogeneity problem has yet to be investigated in schistosomiasis.

In this paper, we developed Bayesian geostatistical models, which take into account age-heterogeneity by incorporating alignment factors to relate schistosomiasis prevalence data from surveys on individuals aged $\leq$20 years with surveys on individuals >20 years and entire communities. Different models were implemented assuming regional and country-specific alignment factors. The predictive performance of the models was assessed using a suite of model validation approaches. Our analysis is stratified for *Schistosoma haematobium* and *S. mansoni* with a geographical focus on eastern Africa.

5.2 Methods

5.2.1 Disease data

Prevalence data of *S. haematobium* and *S. mansoni* from 11 countries in eastern Africa were extracted from the GNTD database. We excluded non-direct diagnostic examination techniques, such as immunofluorescence tests, antigen detections or questionnaire data.

Table 5.1: Remote sensing data sources.[a]

Data type	Source	Date	Temporal resolution	Spatial resolution
LST	MODIS/Terra[1]	2000-2009	8-days	1 km
NDVI	MODIS/Terra[1]	2000-2009	16-days	1 km
Land cover	MODIS/Terra[1]	2001-2004	Yearly	1 km
Rainfall	ADDS[2]	2000-2009	10-days	8 km
Altitude	DEM[3]	-	-	1 km
Water bodies	HealthMapper[4]	-	-	Unknown

[a] All data accessed on 3 February 2011
[1] Moderate Resolution Imaging Spectroradiometer (MODIS). Available at: `https://lpdaac.usgs.gov/lpdaac/products/modis_products_table`
[2] African Data Dissemination Service (ADDS). Available at: `http://earlywarning.usgs.gov/adds/`
[3] Digital elevation model (DEM). Available at: `http://eros.usgs.gov/`
[4] HealthMapper database. Available at: `http://www.who.int/health_mapping/tools/healthmapper/en/index.html`
[5] LandScan^{TM} Global Population Database. Available at: `http://www.ornl.gov/landscan/`

Hospital-based studies and data on non-representative groups, such as HIV positives, are not part of the GNTD database (Hürlimann et al., 2011).

The remaining data were split into three groups and stratified for the two Schistosoma species according to study type. The three groups correspond to surveys on (i) individuals aged $\leq$20 years, (ii) individuals >20 years and (iii) entire community surveys. In case a survey contained prevalence data on multiple age groups, we separated the data according to groups (i) and (ii).

Preliminary analyses suggested only weak temporal correlation in the data for either *Schistosoma* species. Hence, spatial models instead of spatio-temporal models were fitted in the subsequent analyses employing the study year only as a covariate. We grouped the study years as follows: surveys conducted (i) before 1980; (ii) between 1980 and 1989; (iii) between 1990 and 1999; and (iv) from 2000 onwards.

5.2.2 Environmental data

Freely accessible remote sensing data on climatic and other environmental factors were obtained from different sources, as shown in Table 5.1. Data with temporal variation were obtained from launch until the end of 2009 and summarised as overall averages for the available period. Estimates for day and night temperature were extracted from land surface temperature (LST) data. The normalized difference vegetation index (NDVI) was used as a proxy for vegetation. Land cover categories were restructured into six categories: (i)

shrublands and savannah; (ii) forested areas; (iii) grasslands; (iv) croplands; (v) urbanized areas; and (vi) wet areas. Digitized maps of rivers and lakes were combined as a single freshwater map covering the study area. Characteristics on perennial and seasonal water bodies at each survey location were obtained using the spatial join function of ArcMap version 9.2. In addition, the minimum distance between the locations and the closest freshwater source was calculated with the same function.

All data were used as covariates for modelling. Continuous covariates were categorized based on quartiles in order to account for potential non-linear outcome-predictor relations. Processing and extraction of the climatic and environmental data at the survey locations was performed in ArcMap version 9.2, IDRISI 32 and the Modis Reprojection Tool.

5.2.3 Geostatistical model formulation and age-alignment

Let Y_i be the number of infected individuals and N_i the number of individuals screened at location i ($i = 1, \ldots, n$). We assumed that Y_i arises from a Binomial distribution, i.e. $Y_i \sim Bin(p_i, N_i)$, with probability of infection p_i. We introduced covariates $\underline{X}_i$ on the logit scale, such as $logit(p_i) = \underline{X}_i\underline{\beta}$, where $\underline{\beta}$ is the vector of regression coefficients. Unobserved spatial variation can be modelled via additional location-specific random effects, φ_i. We assumed that $\underline{\varphi} = (\varphi_1, \ldots, \varphi_n)^T$ arises from a latent stationary Gaussian spatial process, $\underline{\varphi} \sim MVN(\underline{0}, \sigma^2 R)$ with correlation matrix R modelling geographical dependence between any pairs of locations i and j via an isotropic exponential correlation function, defined by $R_{ij} = exp(-\rho d_{ij})$, where d_{ij} is the distance between i and j, ρ a correlation decay parameter and σ^2 the spatial variance. A measurement error can also be introduced via location-specific non-spatial random effects, ϵ_i, such as $\epsilon_i \sim N(0, \tau^2)$, with non-spatial variance τ^2.

We aligned the risk measured by the different types of studies by incorporating a factor α_s such that $Y_{i,s} \sim Bin(\alpha_s q_{i,s}, N_{i,s})$, with $q_{i,s} = \alpha_s p_i$ and $s = 1$ (surveys with individuals aged $\leq$20 years); $s = 2$ (surveys with individuals aged >20 years); and $s = 3$ (entire community surveys). School-aged children carry the highest risk of *Schistosoma* infection, and hence many studies focus on this age group. We set $\alpha_1 = 1$ in order to use the probability of infection for individuals aged $\leq$20 years as baseline and to align the other groups to this designated baseline.

To complete Bayesian model formulation, we assumed non-informative priors for all parameters. Normal prior distributions with mean 0 and large variance were used for the regression coefficients, $\underline{\beta}$. Non-informative Gamma distributions with mean 1 were

assumed for the variance parameters, σ^2, τ^2 and the alignment factors α_s, while a uniform distribution was implemented for the spatial decay parameter ρ.

Models were developed in OpenBUGS version 3.0.2 (OpenBUGS Foundation; London, UK) and run with two chains and a burn-in of 5000 iterations. Convergence was assessed by inspection of ergodic averages of selected model parameters and history plots. After convergence, samples of 500 iterations per chain with a thinning of 10 were extracted for each model resulting in a final sample of 1000 estimates per parameter.

5.2.4 Model types

We implemented four different models, separately for *S. haematobium* and *S. mansoni*. The models varied based on different features. The first feature was the underlying data. Model A only consisted of schistosomiasis prevalence data on individuals aged ≤20 years ($s = 1$), while models B-D included data on all three kinds of study types ($s = 1, 2, 3$). The second feature was the introduction of alignment factors for disease risk modelling. Model C assumed common alignment factors across the entire study region, while model D assumed country-specific alignment factors.

5.2.5 Model validation

Validation for each model was carried out to identify the model with the highest predictive ability for either *Schistosoma* species and to compare models with and without alignment factors. All models were fitted on a subset of the data (training set) and validated by comparing the posterior median of the predicted risk p_j^* with the observed risk p_j for the remaining set of the data (test set, $j = 1, \ldots, m$, $m < n$). The test set consisted of 20% of the locations from the dataset on individuals aged ≤20 years and was congruent over all models.

Comparisons of predicted vs. observed risk were based on three different validation approaches. Mean absolute errors (MAE) calculate the absolute difference between observed and predicted schistosomiasis risk by $MAE = 1/m \sum_{j=1}^{m} \left|p_j^* - p_j\right|$. An alternative way to quantify divergences in the predictions to the observed data is the χ^2 measure, defined as $\chi^2 = 1/m \sum_{j=1}^{m} \frac{(p_j^* - p_j)^2}{p_j}$. The best predicting model based on these two methods is the model with smallest MAE and χ^2 estimates and therefore with predictions closest to the observed values.

The proportion of the test data being correctly predicted within the q-th Bayesian credible interval (BCIq) of the posterior predictive distribution is calculated by $BCI_q =$

$\frac{1}{m}\sum_{j=1}^{m} min\left(I(ci^{l}_{j(q)} < p_j), I(ci^{u}_{j(q)} > p_j)\right)$, with $q = 50\%, 70\%, 90\%$ and 95%. For this approach, the best performing model contains most test locations within BCIs of smallest width.

5.3 Results

5.3.1 Schistosomiasis prevalence data

Figure 5.1 shows the distribution of the observed schistosomiasis prevalence data over the study region, stratified by study type. An overview of the amount of observed data and mean prevalence levels per country for either *Schistsoma* species, stratified by survey period and diagnostic methods, is given in Table 5.2. Some countries (e.g. Kenya and Tanzania), contain large numbers of survey locations, while other countries, such as Burundi, Eritrea, Rwanda, Somalia and Sudan, are not well covered. Burundi and Rwanda do not include any locations for *S. haematobium*, and Rwanda contains only four surveys on individuals aged >20 years for *S. mansoni*. As expected, there were more surveys carried out with individuals aged ≤20 years than surveys focussing on adult populations or entire communities.

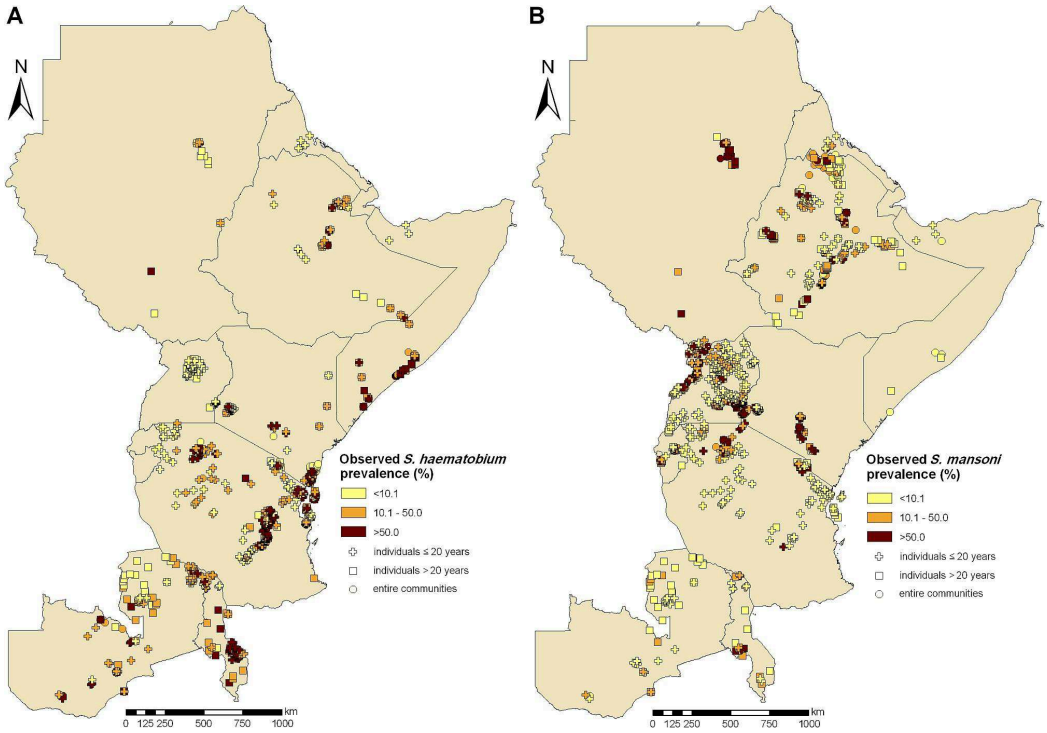

Figure 5.1: Compiled prevalence data of *Schistosoma haematobium* (A) and *S. mansoni* (B) across eastern Africa. Prevalence data are stratified by three different age groups.

Table 5.2: Overall and country-specific mean observed prevalence data (expressed in %) are given, along with number of survey locations (in brackets) for both *S. haematobium* and *S. mansoni* stratified by survey year, diagnostic method and age group.

	Age*	Survey year				Diagnostic**		TOTAL
		<1980	1980-1989	1990-1999	2000-2009			
S. haematobium						UC	RS	
Burundi	1	-	-	-	-	-	-	-
	2	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-
Eritrea	1	0.0 (4)	-	-	-	0.0 (4)	-	0.0 (4)
	2	0.0 (1)	-	-	-	0.0 (1)	-	0.0 (1)
	3	0.0 (2)	-	-	-	0.0 (2)	-	0.0 (2)
Ethiopia	1	17.0 (7)	30.5 (6)	25.2 (18)	-	22.7 (10)	22.2 (17)	24.4 (31)
	2	15.0 (6)	27.6 (5)	16.4 (12)	-	20.7 (11)	17.9 (6)	18.5 (23)
	3	35.0 (3)	29.5 (4)	74.1 (1)	-	38.4 (5)	-	37.1 (8)
Kenya	1	21.0 (65)	52.8 (15)	54.0 (30)	40.6 (2)	34.4 (109)	37.5 (3)	34.5 (112)
	2	22.7 (6)	49.6 (7)	30.3 (3)	-	34.9 (15)	50.7 (1)	35.9 (16)
	3	24.8 (7)	14.0 (25)	64.9 (2)	45.8 (6)	23.2 (40)	-	23.3 (40)
Malawi	1	22.6 (2)	-	55.5 (40)	-	21.5 (6)	59.3 (36)	53.9 (42)
	2	48.2 (7)	34.3 (8)	75.0 (1)	31.6 (5)	36.4 (17)	31.5 (1)	40.2 (21)
	3	31.4 (1)	-	-	-	31.4 (1)	-	31.4 (1)
Rwanda	1	-	-	-	-	-	-	-
	2	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-
Somalia	1	39.4 (11)	-	-	-	39.4 (11)	-	39.4 (11)
	2	55.5 (22)	87.1 (1)	-	-	56.9 (23)	-	56.9 (23)
	3	53.1 (21)	-	-	-	53.1 (21)	-	53.1 (21)
Sudan	1	5.2 (1)	45.3 (3)	-	-	31.2 (2)	-	35.2 (4)
	2	-	1.8 (7)	-	53.0 (3)	10.7 (8)	-	17.2 (10)

Continued on next page

	Age*	Survey year				Diagnostic**		TOTAL
		<1980	1980-1989	1990-1999	2000-2009			
	3	1.8 (2)	20.6 (2)	-	-	20.6 (2)	-	11.2 (4)
Tanzania	1	41.5 (48)	39.9 (173)	30.2 (12)	20.0 (84)	30.6 (173)	59.8 (73)	34.5 (317)
	2	45.5 (41)	29.2 (9)	20.0 (8)	10.0 (2)	39.1 (47)	-	38.5 (60)
	3	47.1 (16)	30.3 (3)	-	28.4 (3)	42.2 (20)	-	42.3 (22)
Uganda	1	42.5 (2)	-	-	0.6 (36)	2.8 (38)	-	2.8 (38)
	2	9.8 (14)	-	-	0.8 (3)	8.7 (16)	-	8.2 (17)
	3	33.3 (2)	-	-	3.0 (2)	18.1 (4)	-	18.1 (4)
Zambia	1	28.2 (22)	30.5 (11)	29.6 (32)	49.2 (4)	31.0 (38)	29.8 (31)	30.5 (69)
	2	13.7 (30)	35.1 (6)	22.9 (5)	13.3 (1)	17.9 (42)	-	17.9 (42)
	3	17.0 (3)	62.4 (3)	35.5 (1)	32.5 (3)	37.1 (10)	-	37.1 (10)
TOTAL	1	28.8 (162)	40.1 (208)	42.4 (132)	15.7 (126)	28.6 (391)	49.4 (160)	32.8 (628)
	2	33.1 (127)	31.0 (43)	22.0 (29)	25.2 (14)	30.5 (180)	26.7 (8)	30.6 (213)
	3	40.3 (57)	21.3 (37)	59.8 (4)	33.1 (14)	34.2 (105)	-	33.8 (112)
S. mansoni						KK	SC	
Burundi	1	-	16.4 (12)	38.3 (3)	-	20.8 (15)	-	20.8 (15)
	2	-	20.8 (19)	44.1 (2)	-	23.0 (21)	-	23.0 (21)
	3	-	19.8 (8)	50.5 (2)	-	25.9 (10)	-	25.9 (10)
Eritrea	1	12.5 (4)	-	-	-	-	12.5 (4)	12.5 (4)
	2	10.0 (1)	-	-	-	-	10.0 (1)	10.0 (1)
	3	41.7 (2)	-	-	-	-	41.7 (2)	41.7 (2)
Ethiopia	1	2.7 (27)	25.0 (23)	28.0 (51)	33.0 (4)	30.1 (69)	6.9 (36)	22.2 (105)
	2	25.7 (15)	18.6 (93)	18.4 (36)	41.8 (7)	18.8 (100)	23.2 (50)	20.3 (151)
	3	4.6 (9)	16.0 (8)	21.1 (62)	36.3 (4)	26.4 (28)	16.0 (55)	19.5 (83)
Kenya	1	18.6 (48)	72.8 (9)	74.3 (18)	53.8 (15)	68.4 (43)	8.7 (39)	41.0 (90)
	2	49.2 (7)	75.2 (7)	36.0 (5)	33.3 (1)	65.4 (14)	28.1 (6)	54.2 (20)
	3	30.7 (15)	71.7 (9)	23.6 (2)	-	69.8 (12)	22.5 (14)	44.3 (26)
Malawi	1	19.4 (1)	37.5 (2)	20.3 (6)	-	30.8 (7)	0.4 (2)	24.0 (9)

Continued on next page

	Age*	Survey year				Diagnostic**		TOTAL
		<1980	1980-1989	1990-1999	2000-2009			
	2	17.9 (6)	36.8 (6)	-	-	35.3 (8)	-	27.3 (12)
	3	-	-	-	-	-	-	25.9 (21)
Rwanda	1	-	-	-	-	-	-	-
	2	-	4.6 (4)	-	-	-	-	4.6 (4)
	3	-	-	-	-	-	-	-
Somalia	1	0.0 (3)	-	-	-	-	0.0 (3)	0.0 (3)
	2	0.0 (2)	-	-	-	-	0.0 (2)	0.0 (2)
	3	0.0 (3)	0.0 (2)	-	-	-	0.0 (5)	0.0 (5)
Sudan	1	61.7 (4)	61.5 (4)	-	-	61.9 (6)	-	61.6 (8)
	2	62.3 (4)	64.9 (8)	47.0 (1)	56.3 (3)	65.8 (15)	0.3 (1)	61.7 (16)
	3	41.0 (2)	52.3 (4)	-	-	50.5 (5)	-	48.5 (6)
Tanzania	1	20.3 (27)	25.3 (4)	30.8 (7)	8.9 (77)	39.6 (17)	21.7 (21)	13.5 (115)
	2	22.2 (26)	12.0 (1)	25.6 (3)	38.6 (5)	46.4 (6)	26.3 (18)	24.8 (35)
	3	27.3 (14)	11.6 (1)	44.1 (3)	44.4 (3)	49.8 (5)	27.3 (14)	31.4 (21)
Uganda	1	48.0 (5)	48.7 (3)	14.5 (6)	22.1 (263)	22.2 (272)	48.0 (5)	22.7 (277)
	2	24.2 (17)	56.3 (3)	66.7 (5)	40.8 (12)	49.6 (20)	20.9 (12)	37.9 (37)
	3	41.7 (7)	47.8 (4)	45.0 (5)	55.8 (12)	50.5 (21)	47.0 (6)	48.3 (28)
Zambia	1	2.9 (16)	5.7 (7)	71.0 (1)	33.2 (1)	36.1 (4)	4.3 (10)	7.7 (25)
	2	8.4 (30)	0.0 (1)	-	41.7 (1)	41.7 (1)	8.1 (31)	9.2 (32)
	3	5.0 (2)	9.5 (1)	60.1 (1)	33.5 (1)	34.4 (3)	5.0 (2)	22.6 (5)
TOTAL	1	16.6 (135)	31.8 (64)	36.7 (92)	20.7 (360)	29.5 (433)	11.5 (120)	23.2 (651)
	2	21.7 (108)	25.3 (142)	26.8 (52)	42.0 (29)	31.7 (185)	19.1 (121)	25.8 (331)
	3	25.0 (54)	28.6 (243)	24.9 (75)	49.1 (20)	41.7 (84)	19.9 (98)	29.7 (186)

*: 1, individuals aged ≤20 years; 2, individuals aged >20 years; 3, entire communities.
**: UC, urine concentration by sedimentation, filtration or centrifugation; RS, reagent strips; KK, Kato-Katz thick smear method; SC, stool concentration methods. Results for surveys with missing diagnostic methods were omitted.

The mean prevalence per country for surveys on individuals aged ≤20 years varies between 0% (Eritrea) and 53.9% (Malawi) for *S. haematobium* and between 0% (Somalia) and 61.6% (Sudan) for *S. mansoni*. We found an overall mean prevalence of *S. haematobium* and *S. mansoni* of 32.8% and 23.2%, respectively. Community surveys usually showed higher mean prevalence levels. However, the survey locations might not be the same among the different types of studies and therefore the observed prevalence levels are not directly comparable.

Two-third of the *S. haematobium* survey data were obtained before the 1990s (66.5%), while few surveys were compiled from 2000 onwards (16.2%). On the other hand, *S. mansoni* surveys were mainly conducted in the 1980s (32.7%) and from 2000 onwards (29.8%), whereas only 15.9% of the surveys were carried out in the 1990s. The distribution of surveys within the different time periods varies from country to country and between the two *Schistsoma* species. While some countries (e.g. Eritrea and Somalia) only have surveys for one or two periods, other countries (e.g. Kenya, Tanzania and Zambia) are well covered over time. The data also vary in the diagnostic methods. For example, even though 67.4% of the *S. mansoni* surveys with known diagnostic methods employed the Kato-Katz thick smear method, in Somalia and Eritrea only stool concentration methods (e.g. Ritchie technique or ether-concentration technique) were used.

5.3.2 Model validation

For *S. haematobium*, model validation based on the MAE measure (Table 5.3) showed no difference between disease risk modelling on individuals aged ≤20 years (model A) and unaligned modelling of all three survey types (model B), while the χ^2 measure led to improved predictions. The introduction of regional alignment factors in spatial modelling based on all survey types (model C) further enhanced model predictive ability based on the MAE and χ^2 measures. Model D, including country-specific alignment factors, showed similar predictive performance as model B. Validation based on different BCIs demonstrated that the proportion of correctly predicted test locations was similar among all models. Model A predicted most test locations correctly within the 95% BCI, while model C was superior for 50% BCIs and model D for 70% BCIs. Regardless of the model used, average BCI widths were comparable.

For *S. mansoni*, model predictive performance in terms of MAE and χ^2 measures was best for model C, followed by models B and D. The differences among the models for the BCI method were small and not consistent between the examined BCIs. For example,

Table 5.3: Model validation results based on MAE, χ^2 measure and BCIs.

Age groups Alignment	Model A ≤20 years -	Model B All -	Model C All Regional	Model D All Country
S. haematobium				
MAE	16.4	16.7	15.5	16.5
χ^2	126.4	95.7	72.6	96.6
50% BCI (width of BCI)	39.7 (24.8)	41.1 (24.3)	42.1 (25.2)	38.1 (24.5)
70% BCI (width of BCI)	57.1 (36.9)	57.1 (36.3)	61.1 (37.7)	61.9 (36.5)
90% BCI (width of BCI)	75.4 (54.1)	75.4 (53.5)	75.4 (54.7)	75.4 (53.9)
95% BCI (width of BCI)	84.9 (61.0)	81.0 (60.4)	79.4 (61.2)	80.2 (60.6)
S. mansoni				
MAE	11.5	11.3	11	11.5
χ^2	48.3	46.8	39.7	43.1
50% BCI (width of BCI)	41.5 (18.5)	34.6 (16.1)	36.9 (15.2)	40.0 (16.4)
70% BCI (width of BCI)	57.7 (29.0)	60.8 (25.4)	60.8 (24.7)	60.8 (26.0)
90% BCI (width of BCI)	80.0 (47.6)	79.2 (41.3)	80.0 (41.6)	81.5 (44.0)
95% BCI (width of BCI)	88.5 (56.5)	84.6 (49.6)	83.8 (50.2)	83.8 (52.7)

BCI, Bayesian credible interval; MAE, mean absolute error.

at 70% BCI, model A included least of the test locations, while at 95% BCI, this model correctly predicted most of the test locations but the averaged width of the BCI was widest.

5.3.3 Alignment factors

Regional and country-specific schistosomiasis risk alignment factors for *S. haematobium* and *S. mansoni* are presented in Table 5.4. Some countries had insufficient data, and hence country-wide alignment factors could not be estimated. A mean regional alignment factor of 0.83 (95% BCI: 0.81-0.85) confirmed that the risk of *S. haematobium* in individuals aged ≤20 years is greater than in individuals >20 years. *S. haematobium* risk estimation from entire community survey was related to the risk of individuals aged ≤20 years with 0.91 (95% BCI: 0.90-0.93). Mean country-specific alignment factors varied from 0.62 (Ethiopia) to 1.26 (Zambia) among individuals >20 years and from 0.79 (Ethiopia) to 1.06 (Zambia) in entire communities. In Ethiopia and Sudan, the country-specific alignment factors were significantly smaller than the overall alignment factor, whereas in Somalia and Zambia, country-specific factors were significantly larger.

For *S. mansoni*, the mean regional alignment factor among individuals aged >20 years was 0.94 (95% BCI: 0.92-0.96), while country-specific estimates varied from 0.64 (Zambia)

Table 5.4: Overall and country-specific number of survey locations (N), mean observed prevalence (p) and alignment factor results (with 95% BCI given in brackets) per age group and *Schistosoma* species.

		S. haematobium			*S. mansoni*		
	Age*	N	p	Alignment factor	N	p	Alignment factor
Burundi	1	0	-	1	15	20.8	1
	2	0	-	-	21	23	0.78 (0.71, 0.87)
	3	0	-	-	10	25.9	0.84 (0.76, 0.93)
Eritrea	1	4	0	1	4	12.5	1
	2	1	0	-	1	10	-
	3	2	0	-	2	41.7	-
Ethiopia	1	31	24.4	1	105	22.2	1
	2	23	18.5	0.62 (0.55, 0.68)	151	20.3	0.71 (0.70, 0.73)
	3	8	37.1	0.79 (0.72, 0.87)	83	19.5	0.85 (0.83, 0.88)
Kenya	1	112	34.5	1	90	41	1
	2	16	35.9	0.84 (0.79, 0.89)	20	54.2	1.09 (1.05, 1.13)
	3	40	23.2	0.86 (0.83, 0.89)	26	44.3	1.02 (0.99, 1.05)
Malawi	1	42	53.9	1	9	24	1
	2	21	40.2	0.86 (0.82, 0.92)	12	27.3	-
	3	1	31.4	-	0	-	-
Rwanda	1	0	-	1	0	-	1
	2	0	-	-	4	4.6	-
	3	0	-	-	0	-	-
Somalia	1	11	39.4	1	3	0	1
	2	23	56.9	1.05 (0.95, 1.18)	2	0	-
	3	21	53.1	1.02 (0.94, 1.12)	5	0	-
Sudan	1	4	35.2	1	8	61.6	1
	2	10	17.2	0.69 (0.64, 0.74)	16	61.7	1.02 (0.94, 1.10)
	3	4	11.2	-	6	48.5	1.00 (0.95, 1.06)
Tanzania	1	317	34.5	1	115	13.5	1
	2	60	38.5	0.84 (0.82, 0.86)	35	24.8	1.18 (1.12, 1.24)
	3	22	42.3	0.94 (0.91, 0.96)	21	31.4	1.13 (1.08, 1.17)
Uganda	1	38	2.8	1	277	22.7	1
	2	17	8.2	0.89 (0.77, 1.03)	37	37.9	1.06 (1.01, 1.11)
	3	4	18.1	-	28	48.3	1.01 (0.96, 1.04)
Zambia	1	69	30.5	1	25	7.7	1
	2	42	17.9	1.26 (1.08, 1.43)	32	9.2	0.64 (0.49, 0.89)
	3	10	37.1	1.06 (0.99, 1.13)	5	22.6	-
TOTAL	1	628	32.8	1	651	23.2	1
	2	213	30.6	0.83 (0.81, 0.85)	331	25.8	0.94 (0.92, 0.96)
	3	112	33.8	0.91 (0.90, 0.93)	186	29.7	0.96 (0.95, 0.98)

*: 1, individuals aged $\leq$20 years; 2, individuals aged >20 years; 3, entire communities.

to 1.18 (Tanzania). In community surveys, the regional alignment factor was 0.96 (95% BCI: 0.95-0.98) with country-specific alignment factors between 0.84 (Burundi) and 1.13 (Uganda). Significantly smaller country-specific alignment factors compared to the overall alignment factor were found in Burundi, Ethiopia and Zambia, while significantly larger factors were obtained for Kenya, Tanzania and Uganda.

The regional alignment factor estimates for *S. haematobium* compared to *S. mansoni* are much lower, e.g. 17% risk reduction for individuals aged >20 years vs. 6% risk reduction. This relation is also found in country-specific estimates, except for Zambia.

5.4 Discussion

In this study, we derived factors to align schistosomiasis prevalence estimates from age-heterogeneous surveys across an ensemble of 11 countries in eastern Africa. We found correction factors that are significantly different from 1. As a result, geostatistical model-based predictions from school-based and community-based surveys are further enhanced. The estimates of the regional alignment factors confirm that individuals aged ≤20 years are at a higher risk of a *Schistosoma* infection than adults (Woolhouse, 1998; Jordan and Webbe, 1982; Anderson and May, 1985). Interestingly, the alignment factor estimates for *S. haematobium* were slightly lower than those for *S. mansoni*. This finding might be explained by differences in the age-prevalence curves between the two species. *S. haematobium* prevalence usually peaks in the age group 10-15 years (Woolhouse et al., 1991), while the peak of *S. mansoni* prevalence occurs somewhat later, up to the age of 20 years (Fulford et al., 1992). Consequently, there is a larger difference in infection risk between children and adults for *S. haematobium* compared to *S. mansoni*. Additionally, the peak of *S. mansoni* prevalence might be further shifted towards older age groups due to the so-called peak shift. Indeed, it has been shown that the peak of infection prevalence is more flat and reaches its maximum in older age groups if transmission is low-to-moderate, while prevalence peaks are higher and they are observed at a younger mean age if transmission is high (Woolhouse, 1998). Several African countries have implemented large-scale preventive chemotherapy programmes against schistosomiasis (WHO, 2010; Fenwick et al., 2009). These programmes reduced schistosomiasis-related morbidity (Koukounari et al., 2007) and might have had some impact on transmission (King et al., 2006; French et al., 2010). It is therefore conceivable that the peak of *Schistosoma* infection might slightly shift to older age groups. It should also be noted that, disparities in the spatial risk distribution of the two *Schistosoma* species and in the implementation of control strategies in these

areas could have led to differences in the alignment factors.

Considerable differences between country-specific alignment factors and prevalence ratios based on the raw data were found for Ethiopia, Tanzania, Uganda and Zambia in *S. haematobium*, and for Burundi and Zambia in *S. mansoni*. These differences are mainly due to the spatial distribution of the survey locations, which vary between age groups. For example, surveys focussing on individuals aged $\leq$20 years are located in central and eastern Zambia, while surveys on individuals >20 years in Zambia are mainly located in the north of the country. The north is characterised by lower schistosomiasis transmission risk. Therefore, the crude prevalence ratio between the two groups is artificially small, while the alignment factor, which is based on the predicted prevalence risk in this area, is much higher.

Model validation showed that regional alignment factors improved predictive performance of the models for both *Schistosoma* species, however, country-specific alignment factors did not further improve the models. The predictive performance of the model with regional factors was good, as 79.4% and 83.8% of the test locations were correctly predicted within 95% BCIs for *S. haematobium* and *S. mansoni*, respectively. All models estimated relatively wide BCIs, indicating large variation in the data that could not be explained by the model covariates. Socioeconomic and health system factors might play a role in the spatial distribution of schistosomiasis, however these data do not exist at high spatial distribution for the entire study area, and hence could not be used for model fit and prediction. Part of the variation might have arisen by the model assumptions of stationarity and isotropy and the heterogeneity in the diagnostic methods.

The proposed alignment factor approach is scaling the predicted prevalence of schistosomiasis and leads to an easy interpretation of the parameters. In addition, it allows defining meaningful prior distributions, and hence resulting in better model convergence. An alternative way to include age in the models is to introduce age as a covariate. This approach is scaling the odds instead of the prevalence. Preliminary analyses preformed by the authors, on the same data using age as covariate, resulted in serious model convergence problems, leading to the implementation of age alignment factors as proposed in this manuscript.

A limitation of our work is the assumption of constant disease risk within each age group. This is not true especially for school-aged children for whom the schistosomiasis risk reaches a maximum at around 11-14 years. A more rigorous model formulation should take into account the age-prevalence curve and standardise the surveys using a

mathematical description of this curve. Raso et al. (2007a) derived a Bayesian formulation of the immigration-death model to obtain age-specific prevalence of *S. mansoni* from age-prevalence curves. We are currently exploring geostatistical models, coupled with mathematical immigration-death models, to fully consider the age-dependence of the schistosomiasis risk.

5.5 Conclusions

We have shown that age-alignment factors should be included to improve prevalence estimates of population-based risk of schistosomiasis, especially for large-scale modelling and prediction efforts. Indeed, large-scale modelling cannot be achieved without compilation of primarily historical survey data assembled over large study areas using different study designs and age groups. The proposed alignment factor approach can be used to relate the most frequent survey types, i.e. studies focussing on individuals aged $\leq$20 years (mainly school surveys) with studies on individuals aged >20 years and entire communities. Unaligned survey compilation leads to imprecise disease risk estimates and potentially wrong recommendations to decision makers for the implementation of control activities and subsequent monitoring and evaluation.

Acknowledgements and funding

NS is grateful for financial support of the EU-funded CONTRAST project. This investigation received further financial support from the Swiss National Science Foundation for JU (project no. IZ70Z0-123900) and PV (project no. 325200-118379) and UBS Optimus Foundation. Special thanks are addressed to Eveline Hürlimann and our partners at the University of Copenhagen, Denmark for their work related to the GNTD database.

Chapter 6

Spatially explicit *Schistosoma* infection risk in eastern Africa using Bayesian geostatistical modelling

Schur N.[1,2], Hürlimann E.[1,2,3], Stensgaard AS.[4,5], Chimfwembe K.[6], Mushinge G.[6], Simoonga C.[7], Kabatereine NB.[8], Kristensen TK.[5], Utzinger J.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Centre Suisse de Recherches Scientifique en Côte d'Ivoire, Abidjan, Côte d'Ivoire
[4] Center for Macroecology and Evolution, University of Copenhagen, Copenhagen, Denmark
[5] DBL, University of Copenhagen, Frederiksberg, Denmark
[6] Department of Community Medicine, University of Zambia, Lusaka, Zambia
[7] Ministry of Health, Lusaka, Zambia
[8] Vector Control Division, Ministry of Health, Kampala, Uganda

This paper has been accepted for publication in *Acta Tropica*.

Abstract

Background: Schistosomiasis remains one of the most prevalent parasitic diseases in the tropics and subtropics, but current statistics are outdated due to demographic and ecological transformations and ongoing control efforts. Reliable risk estimates are important to plan and evaluate interventions in a spatially explicit and cost-effective manner.

Methodology: We analysed a large ensemble of georeferenced survey data derived from an open-access neglected tropical diseases database to create smooth empirical prevalence maps for *Schistosoma mansoni* and *S. haematobium* for a total of 13 countries of eastern Africa. Bayesian geostatistical models based on climatic and other environmental data were used to account for potential spatial clustering in spatially structured exposures. Geostatistical variable selection was employed to reduce the set of covariates. Alignment factors were implemented to combine surveys on different age-groups and to acquire separate estimates for individuals aged $\leq$20 years and entire communities. Prevalence estimates were combined with population statistics to obtain country-specific numbers of *Schistosoma* infections.

Principal Findings: We estimate that 122 million individuals in eastern Africa are currently infected with either *S. mansoni*, or *S. haematobium*, or both species concurrently. Country-specific population-adjusted prevalence estimates range between 12.9% (Uganda) and 34.5% (Mozambique) for *S. mansoni* and between 11.9% (Djibouti) and 40.9% (Mozambique) for *S. haematobium*. Our models revealed that infection risk in Burundi, Eritrea, Ethiopia, Kenya, Rwanda, Somalia and Sudan might be considerably higher than previously reported, while in Mozambique and Tanzania, the risk might be lower than current estimates suggest.

Conclusion/Significance: Our empirical, large-scale, high-resolution infection risk estimates for *S. mansoni* and *S. haematobium* in eastern Africa can guide future control interventions and provide a benchmark for subsequent monitoring and evaluation activities.

6.1 Introduction

Schistosomiasis remains one of the most prevalent parasitic diseases in tropical and subtropical areas, particularly in sub-Saharan Africa (Steinmann et al., 2006; Utzinger et al., 2009). After many years of neglect, there is growing interest in the control of schistosomiasis and other neglected tropical diseases (Hotez et al., 2007; Fenwick et al., 2009; Utzinger et al., 2009). Reliable baseline maps of the geographical distribution of at-risk areas and estimates of the number of infected individuals are important tools to plan and evaluate control interventions in a cost-effective manner.

Most empirical mapping efforts for schistosomiasis only cover small geographical areas, e.g. a single village (Pinot de Moira et al., 2007), a district (Raso et al., 2005) or an entire country (Clements et al., 2006a). Indeed, besides a few exceptions (Clements et al., 2008, 2010; Schur et al., 2011b), there is a paucity of large-scale mapping efforts. As part of the European Union (EU)-funded CONTRAST project, an up-to-date, open-access database of historical and contemporary prevalence surveys on schistosomiasis in Africa was developed (`http://www.gntd.org`) (Hürlimann et al., 2011; Schur et al., 2011d; Stensgaard et al., 2011). Recently, we presented the first empirical schistosomiasis prevalence estimates for West Africa, based on the aforementioned database and a Bayesian-based geostatistical modelling approach using climatic and other environmental predictors (Schur et al., 2011b). We also presented population-adjusted risk estimates at country level and noted considerable differences from the widely cited statistics put forth by Chitsulo and colleagues for the mid-1990s (Chitsulo et al., 2000) and extrapolated estimates for mid-2003 (Steinmann et al., 2006). These previous estimates were based on population-adjusted statistics originally published by Utroska et al. (1989) and lack empirical modelling. Moreover, the estimates are likely outdated due to demographic and ecological transformations (e.g. water resources development and management), socio-economic development (e.g. improved access to clean water and sanitation) and implementation of large-scale control interventions, most notably regular deworming of school-aged children (Fenwick, 2006; Steinmann et al., 2006; Fenwick et al., 2009; Utzinger et al., 2009; WHO, 2010).

Bayesian geostatistical models fitted by Markov chain Monte Carlo (MCMC) simulation methods are increasingly utilised in disease risk mapping and prediction (Diggle et al., 1998; Banerjee et al., 2003). Such models have been employed for schistosomiasis risk profiling, i.e. mapping the distribution of *Schistosoma mansoni* in Burundi, Uganda and parts of Kenya and Tanzania (Clements et al., 2010). However, to our knowledge, large-scale model-based high-resolution *S. haematobium* and *S. mansoni* infection risk maps, including the

number of infected individuals for the entire eastern African region, do not exist.

To fill this gap, we developed Bayesian geostatistical models based on climatic and other environmental risk factors, including different soil characteristics, to obtain empirical schistosomiasis risk maps and population-adjusted country prevalence estimates for an ensemble of 13 countries in eastern Africa. We analysed readily available survey data and implemented alignment factor models to account for the age-heterogeneity in the compiled survey data (Schur et al., 2011d). Geostatistical variable selection was applied to reduce the set of covariates to the most important predictors (George and McCulloch, 1993). Here, we report prevalence maps at 5 x 5 km spatial resolution for *S. haematobium* and *S. mansoni* and estimated numbers of infected individuals at country level.

6.2 Data and methods

6.2.1 Disease data

Prevalence data on schistosomiasis for eastern Africa were extracted from the 'Global Neglected Tropical Disease' (GNTD) database (version: 5 October 2010) for all available survey years. The database assembles general information from the included publications, as well as study-specific information on survey population, time of the study, *Schistosoma* species, diagnostic test employed, and the number of infected individuals among those examined, stratified by age and sex (if available) (Hürlimann et al., 2011). Data currently lacking geographical reference information were excluded. We also excluded entries based on non-direct diagnostic tests (e.g. immunofluorescence tests, antigen detections or questionnaire data) due to lower diagnostic sensitivities compared to direct diagnostic tests (e.g. schistosome egg detection in urine or stool). The proportion of rejected diagnostic techniques was low: 2.5% for *S. mansoni* and 0.6% for *S. haematobium*. Entries with missing information on the diagnostic technique (*S. mansoni*: 4.6% missing, *S. haematobium*: 8.4% missing) were assumed to be also largely based on direct examination techniques. We considered the bias that would arise from ignoring the missing data, as larger than the bias from potentially rejected diagnostic techniques among the data with missing information on the examination technique.

6.2.2 Climatic, demographic and environmental data

Climatic, demographic and environmental data were obtained from different freely accessible remote sensing data sources, as summarised in Table 6.1. Land surface temperature

(LST) data were used as a proxy for day and night temperature, the normalized difference vegetation index (NDVI) as a proxy for moisture (Huete et al., 2002) and the human influence index (HII) for changes in the environment due to anthropometric activities (Sanderson et al., 2002). The land cover categories were re-grouped into six categories, as follows: (i) savannah and shrublands; (ii) forests; (iii) grasslands and sparsely vegetated areas; (iv) croplands; (v) urban areas; and (vi) wet areas. The soil parameters used were the following: bulk density (in kg/dm3), available water capacity (in cm/m), pH and texture class (fine, medium and coarse). Furthermore, digitised maps on water body sources

Table 6.1: Data sources and properties of the climatic and other environmental covariates used to model schistosomiasis prevalence in eastern Africa.[a]

Source	Data type	Data period	Temporal resolution	Spatial resolution
Moderate Resolution Imaging Spectroradiometer (MODIS)/Terra[1]	Land surface temperature (LST) for day and night	2000-2009	8 days	1 km
	Normalized difference vegetation index (NDVI)	2000-2009	16 days	1 km
	Land cover	2001-2004	Yearly	1 km
African Data Dissemination Service (ADDS)[2]	Rainfall	2000-2009	10 days	8 km
Earth Resources Observation (EROS) Center[3]	Altitude, slope and aspect	-	-	1 km
International Soil Reference and Information Centre (ISRIC)[4]	Soil parameters	-	-	8 km
HealthMapper database[5]	Water bodies	-	-	Unknown
Socioeconomic Data and Applications Center (SEDAC)[6]	Human influence index (HII)	-	-	1 km
LandScanTM Global Population Database[7]	Population counts	2008	-	1 km

[a] All data accessed on 03. February 2011
[1] Available at: `https://lpdaac.usgs.gov/lpdaac/products/modis_products_table`
[2] Available at: `http://earlywarning.usgs.gov/fews/africa/index.php`
[3] Available at: `http://edc.usgs.gov/\#/Find_Data/Products_and_Data_Available/gtopo30/hydro/`
[4] Available at: `http://www.isric.org/data/isric-wise-derived-soil-properties-5-5-arc-minutes-global-grid-version-11`
[5] Available at: `http://gis.emro.who.int/PublicHealthMappingGIS/HealthMapper.aspx`
[6] Available at: `http://sedac.ciesin.columbia.edu/wildareas/`
[7] Available at: `http://www.ornl.gov/landscan/`

(rivers and lakes) in eastern Africa were combined and distance to the nearest water body source was calculated. LST, NDVI and rainfall data were summarised as overall averages over the period of 2000-2009 based on the mean. Land cover data from 2001-2004 were combined based on the most frequent category. Estimates for the percentage of individuals aged $\leq$20 years among the total population, stratified by country, were extracted from the U.S. Census Bureau International Database for the year 2010.

The MODIS/Terra data were processed using the 'MODIS Reprojection Tool' (Land Processes DAAC, USGS EROS). Rainfall estimates were converted in IDRISI 32 (Worcester, Clark University). Processing of the remaining data, distance calculations, and displaying of data and results were performed in ArcMap version 9.2 (ESRI). Further data processing was performed in Fortran 90 codes written by the authors. Remote sensing data were aligned to a common resolution of 5 x 5 km. In particular, for data with high initial resolution (1 x 1 km), the 'Aggregate' function of ArcMap was used to calculate the mean of all valid input cells that encompass the output cells (5 x 5 km resolution). For data with low initial resolution (8 x 8 km), the value of the input cell with the centroid closest to the centroid of the output cell was taken.

6.2.3 Statistical analysis

Bivariate logistic regressions were carried out to determine the relationship between the risk of *Schistosoma* infection and the potential covariates for each *Schistosoma* species separately. Covariates with non-linear outcome-predictor relations were treated as categorical.

Bayesian geostatistical logistic regression models with location-specific random effects and age-alignment factors (for details, see Appendix) were fitted to identify the most significant predictors and to obtain spatially explicit schistosomiasis risk estimates. The random effects were considered as latent observations of a Gaussian process with variance-covariance matrix related to an exponential correlation function between any pair of locations. The variance-covariance is a matrix of $n \times n$, where n is the number of survey locations. Model fit requires the inversion of this matrix. The datasets used for this study contain large numbers of survey locations and parameter estimation becomes unfeasible. Therefore, an approximation of the spatial process was used in the current application (see Appendix).

The best set of covariates was determined using Gibbs variable selection (George and McCulloch, 1993). Indicator variables were linked to the regression coefficients to specify presence or absence of the corresponding covariate. In this study, variable selection was

based on the estimation of the posterior inclusion probability with prior probability of 0.25. All covariates with a posterior inclusion probability larger than 0.5 were employed in the final model.

We employed MCMC simulation to estimate model parameters. Infection risk at unobserved locations was predicted via joint Bayesian kriging. A grid of prediction locations with a spatial resolution of 0.05° x 0.05° (approximately 5 x 5 km) was used, resulting in approximately 260,000 pixels. Population count estimates were linked to the grid to calculate the number of individuals aged ≤20 years and above per pixel. The number of individuals was merged with the model-based schistosomiasis risk predictions at the same locations to estimate the averaged number of infected individuals. A combined estimate for schistosomiasis risk of *S. mansoni* and *S. haematobium* was calculated on the assumption of independence between the two species, i.e. prevalence of *Schistosoma spp.* = prevalence of *S. mansoni* + prevalence of *S. haematobium* - (prevalence of *S. mansoni* * prevalence of *S. haematobium*).

6.2.4 Model validation

The performance of the models was assessed using a suite of model validations. In a first step, a sample of 80% of the survey locations was employed as training set for model fit, while the remaining 20% of the locations (test locations) were kept for model validation. Second, the predicted outcomes at the k test locations were compared to the observed outcomes via three different approaches: mean errors (ME), mean absolute errors (MAE) and Bayesian credible interval (BCI) comparisons (Gosoniu et al., 2006). The ME shows the overall tendency of a model to over- or underestimate the risk, and it is calculated by $ME = 1/k \sum_{i=1}^{k} p_i - \widehat{p}_i$, where p_i is the observed outcome and $\widehat{p}_i$ the median of the predictions at test location i. The MAE provides information on the accuracy of a model based on the absolute distances between predictions and observations, $MAE = 1/k \sum_{i=1}^{k} |p_i - \widehat{p}_i|$. The proportion of test locations being correctly predicted within the q-th BCI of the posterior predictive distribution (restricted by the lower centiles c_i^l and upper centiles c_i^u) is the outcome of the BCI approach, i.e. $BCI_q = \frac{1}{k} \sum_{i=1}^{k} min\left(I(c_i^l < p_i), I(c_i^u > p_i)\right)$.

6.3 Results

6.3.1 Final datasets and preliminary statistics

The final datasets consisted of 1406 and 1851 survey locations for *S. haematobium* and *S. mansoni*, respectively. Among these, there were 1208 and 1558 unique locations, respec-

Table 6.2: Overview of schistosomiasis data, stratified by survey year and age group given by country.

	Total	Survey year				Age group		
	(unique locations)	<1980	1980-1989	1990-1999	2000-2009	≤20 years	>20 years	All
S. haematobium								
Burundi	0 (0)	0	0	0	0	0	0	0
Djibouti	0 (0)	0	0	0	0	0	0	0
Eritrea	7 (7)	7	0	0	0	4	1	2
Ethiopia	82 (56)	27	24	31	0	59	13	10
Kenya	172 (136)	76	50	33	13	123	18	31
Malawi	67 (62)	9	8	43	7	58	9	0
Mozambique	105 (103)	93	0	4	8	103	2	0
Rwanda	0 (0)	0	0	0	0	0	0	0
Somalia	73 (60)	48	25	0	0	40	18	15
Sudan	152 (135)	7	142	0	3	124	27	1
Tanzania	421 (351)	93	185	29	114	375	39	7
Uganda	57 (50)	16	2	0	39	50	7	0
Zambia	270 (248)	56	25	45	144	187	78	5
Total	1406 (1208)	432	461	185	328	1123	212	71
S. mansoni								
Burundi	85 (35)	0	67	18	0	47	38	0
Djibouti	0 (0)	0	0	0	0	0	0	0
Eritrea	11 (10)	8	3	0	0	7	1	3
Ethiopia	528 (438)	94	249	137	48	373	132	23
Kenya	142 (109)	68	31	22	21	119	15	8
Malawi	21 (21)	7	8	6	0	14	7	0
Mozambique	101 (96)	93	0	6	2	96	5	0
Rwanda	4 (4)	0	4	0	0	0	4	0
Somalia	10 (9)	8	2	0	0	4	1	5
Sudan	183 (156)	47	128	3	5	149	30	4
Tanzania	151 (129)	52	5	12	82	125	23	3
Uganda	432 (383)	28	21	21	362	381	38	13
Zambia	183 (168)	47	15	9	112	118	63	2
Total	1851 (1558)	452	533	234	632	1433	357	61

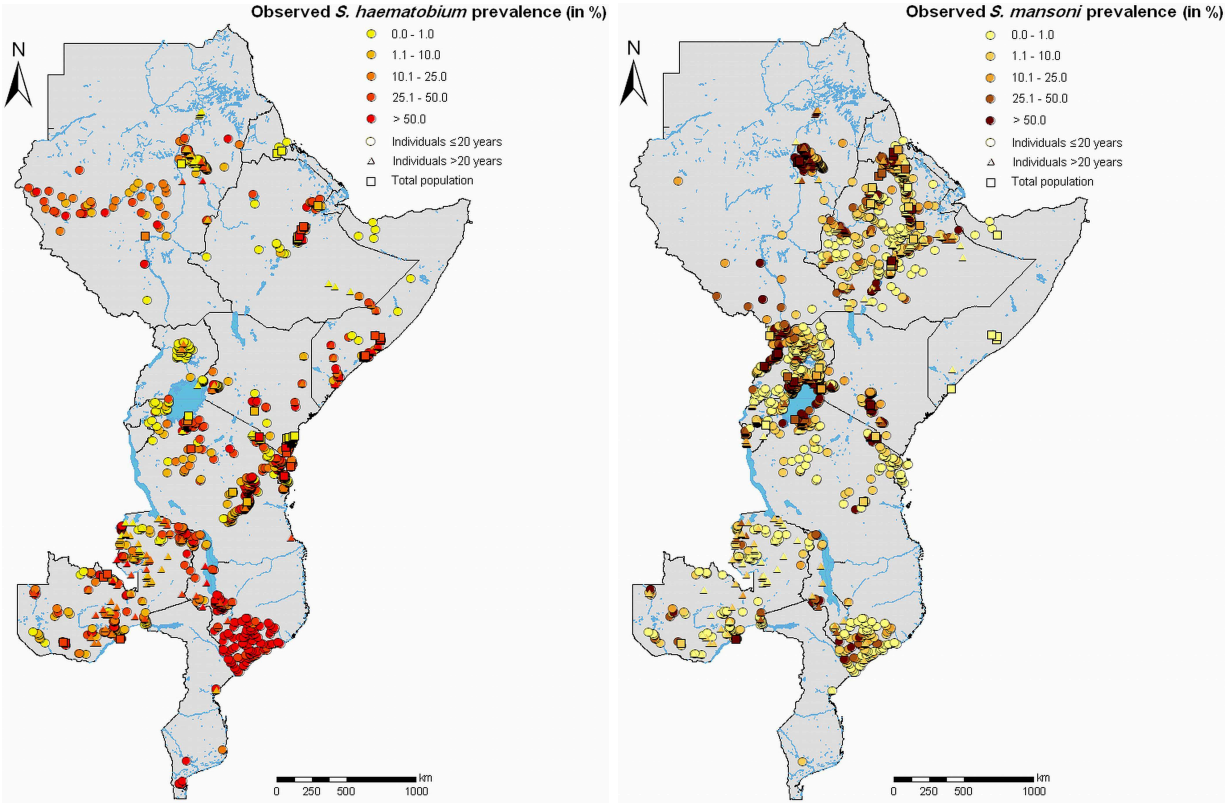

Figure 6.1: Observed prevalence of *S. haematobium* (left) and *S. mansoni* (right) across eastern Africa obtained from the GNTD database, including data until October 2010.

tively. The number of surveys per country, year and age group are listed in Table 6.2. Prevalence in individuals aged ≤20 years ranged from 0% to 100% for both *Schistosoma* species with mean prevalence of 34.0% (median 26.7%, standard deviation (SD) 34.0%) for *S. haematobium* and 21.5% (median 8.6%, SD 27.4%) for *S. mansoni*. The distribution and the observed prevalence of the survey locations are shown in Figure 6.1.

Data on the spatial distribution of the potential covariates influencing the distribution of schistosomiasis are presented in the Appendix. All considered covariates were significant in the bivariate logistic regressions. Typically, categorised covariates showed better predictive ability based on BIC than linear covariates, except for day temperature and altitude for *S. haematobium* and HII for *S. mansoni*. The implementation of the geostatistical variable selection approach led to a reduction in the final set of covariates. Distance to the closest freshwater body was excluded from the *S. haematobium* model, while NDVI was removed in the final *S. mansoni* model.

6.3.2 Spatial modelling results

Parameter estimates of the geostatistical model for *S. haematobium* are shown in Table 6.3. The results indicate that there was a significant decrease in *S. haematobium* infection risk from the beginning of the 1990s onwards. Latitude and longitude effects suggested an elevated risk around Lake Victoria and the countries south of Tanzania. Higher day LST averages and yearly rainfall estimates above 216 mm were associated with low risk of infection. There was a non-linear relation between *S. haematobium* with night LST and NDVI indicating a positive relation and a decrease in risk at the lowest and the highest values of those two covariates. Areas covered with croplands showed a higher schistosomiasis risk than those with savannahs and shrublands. Altitude showed a negative relation, while topological parameters showed high and low risk in East and West direction, respectively. Most of our locations were in low flow velocity flat areas, and therefore the observed positive effect of high slope regions is rather misleading. Soil-related parameters showed that schistosomiasis is more common in areas where the soils have a pH that is neutral to acid, high water capacity and coarse texture. The influence of anthropometric activities to the environment appeared to be positively associated with schistosomiasis at intermediate levels of the covariate.

Table 6.3: Logistic regression parameter estimates for *S. haematobium* summarised by odds ratios (OR), 95% confidence intervals (CI) and 95% Bayesian credible intervals (BCI).

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Survey year			
<1980	1	1	1
1980-1989	0.58 (0.57, 0.59)*	1.07 (1.04, 1.10)*	1.02 (0.99, 1.08)
1990-1999	0.81 (0.79, 0.84)*	0.94 (0.90, 0.97)*	0.70 (0.67, 0.74)*
2000-2009	0.52 (0.51, 0.54)*	0.59 (0.57, 0.61)*	0.47 (0.45, 0.50)*
Latitude (in °)			
<-11.1	1	1	1
-5.6	0.97 (0.94, 0.99)*	0.37 (0.35, 0.39)*	0.74 (0.64, 0.83)*
-7.3	1.49 (1.45, 1.53)*	0.39 (0.37, 0.41)*	1.37 (1.22, 1.57)*
>1.7	0.60 (0.59, 0.62)*	0.34 (0.31, 0.35)*	0.61 (0.57, 0.65)*
Longitude (in °)			
<32.7	1	1	1
32.7-35.5	1.18 (1.14, 1.21)*	0.90 (0.86, 0.94)*	0.59 (0.54, 0.63)*
35.6-39.1	1.69 (1.64, 1.75)*	1.26 (1.18, 1.34)*	1.00 (0.96, 1.05)
>39.1	2.56 (2.49, 2.63)*	1.57 (1.47, 1.68)*	0.67 (0.63, 0.72)*

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Altitude (m)			
<200	1	1	1
200-559	0.27 (0.26, 0.28)*	0.48 (0.46, 0.51)*	1.00 (0.96, 1.05)
560-1102	0.49 (0.47, 0.50)*	0.66 (0.62, 0.70)*	0.57 (0.54, 0.60)*
>1102	0.54 (0.52, 0.55)*	0.57 (0.53, 0.61)*	0.50 (0.47, 0.53)*
Day LST (°C)	0.92 (0.92, 0.92)*	0.97 (0.97, 0.98)*	0.92 (0.92, 0.92)*
Night LST (°C)			
<17.7	1	1	1
17.7 - 19.8	2.34 (2.27, 2.41)*	2.27 (2.16, 2.38)*	2.20 (2.09, 2.34)*
19.9 - 21.7	1.53 (1.49, 1.58)*	1.91 (1.81, 2.03)*	2.26 (2.09, 2.43)*
>21.7	2.09 (2.03, 2.15)*	1.84 (1.73, 1.97)*	1.85 (1.70, 2.00)*
Rainfall (mm)			
<216	1	1	1
216-276	1.71 (1.68, 1.75)*	1.49 (1.44, 1.54)*	0.63 (0.60, 0.67)*
>276	0.85 (0.83, 0.87)*	1.12 (1.08, 1.17)*	0.77 (0.72, 0.83)*
NDVI			
<0.40	1	1	1
0.40-0.49	2.74 (2.67, 2.80)*	1.23 (1.19, 1.28)*	1.67 (1.61, 1.76)*
0.50-0.59	2.08 (2.03, 2.14)*	1.15 (1.10, 1.20)*	1.53 (1.45, 1.64)*
>0.59	2.32 (2.26, 2.38)*	0.89 (0.85, 0.94)*	1.23 (1.15, 1.33)*
Land cover			
Savannah/ shrublands	1	1	1
Forests	0.98 (0.96, 1.01)	0.94 (0.90, 0.97)*	1.00 (0.99, 1.03)
Grasslands/ sparsely vegetated	0.79 (0.77, 0.81)*	0.81 (0.78, 0.84)*	0.67 (0.64, 0.70)*
Croplands	1.25 (1.22, 1.28)*	1.24 (1.20, 1.28)*	1.07 (1.01, 1.12)*
Urban	0.86 (0.83, 0.88)*	0.63 (0.60, 0.66)*	1.00 (0.97, 1.02)
Wet areas	0.87 (0.83, 0.92)*	0.65 (0.60, 0.70)*	0.72 (0.66, 0.78)*
Slope (in °)			
<0.16	1	1	1
0.16-0.43	0.62 (0.60, 0.64)*	0.74 (0.71, 0.76)*	0.87 (0.83, 0.90)*
0.44-1.15	0.52 (0.51, 0.54)*	0.78 (0.75, 0.81)*	0.71 (0.67, 0.75)*
>1.15	0.74 (0.72, 0.76)*	1.08 (1.04, 1.12)*	1.17 (1.12, 1.23)*
Aspect (in °)			
<48.8	1	1	1
48.8-105.0	1.40 (1.36, 1.43)*	1.40 (1.35, 1.45)*	1.43 (1.39, 1.48)*
105.1-202.4	1.05 (1.03, 1.08)*	1.04 (1.01, 1.07)*	1.00 (0.99, 1.02)
>202.4	0.60 (0.58, 0.61)*	0.95 (0.92, 0.99)*	0.93 (0.89, 0.97)*
Human influence index			

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
<17	1	1	1
17-19	1.19 (1.15, 1.23)*	1.15 (1.10, 1.19)*	1.00 (0.97, 1.04)
20-24	1.48 (1.43, 1.52)*	1.40 (1.35, 1.45)*	1.07 (1.01, 1.13)*
>24	1.20 (1.17, 1.24)*	1.31 (1.26, 1.36)*	0.93 (0.89, 0.98)*
Bulk density (in kg/dm3)			
<1.32	1	1	1
1.32-1.34	1.25 (1.21, 1.29)*	0.65 (0.62, 0.68)*	0.81 (0.75, 0.91)*
1.35-1.50	1.35 (1.31, 1.39)*	0.49 (0.47, 0.52)*	0.55 (0.50, 0.61)*
>1.50	1.98 (1.93, 2.03)*	1.43 (1.35, 1.51)*	1.02 (0.99, 1.04)
Available water capacity (in cm/m)			
<8	1	1	1
8-9	0.78 (0.76, 0.80)*	2.43 (2.25, 2.63)*	1.69 (1.55, 1.84)*
10-11	0.31 (0.30, 0.32)*	0.95 (0.88, 1.02)	2.06 (1.90, 2.22)*
>12	0.56 (0.55, 0.57)*	1.04 (0.95, 1.12)	1.43 (1.33, 1.52)*
pH in water			
<5.9	1	1	1
5.9-6.8	1.11 (1.09, 1.13)*	0.85 (0.82, 0.89)*	0.76 (0.72, 0.81)*
>6.8	0.67 (0.66, 0.69)*	1.08 (1.01, 1.15)*	1.17 (1.07, 1.26)*
Texture class			
Medium	1	1	1
Coarse	1.76 (1.72, 1.80)*	0.49 (0.44, 0.53)*	1.30 (1.10, 1.49)*
Fine	0.79 (0.77, 0.80)*	0.43 (0.41, 0.45)*	0.90 (0.80, 1.00)
			Mean (95% BCI)
$Sigma^2$	-	-	3.83 (3.15, 4.49)
Range (km)	-	-	355.3 (341.0, 372.0)

*: Significant correlation based on 95% CI or 95% BC

The estimated odds ratios (ORs) of the predictors for *S. mansoni* infection risk are given in Table 6.4. The risk increased during the 1980s and 1990s, but decreased slightly from 2000 onwards. A South-to-North and West-to-East trend of increasing prevalence is indicated by latitude and longitude. Low risk of infection is associated with high average day LST and low night LST values and rainfall showed a significant positive relation. Forested areas were found to be related to highest risk, while low risk estimates were found in built up environments, such as urbanised settings. An elevated risk was suggested at very low and very high levels of altitude in flat regions and areas pointing towards West, as indicated by the aspect parameter. Locations with distances of more than 1.6 km to the nearest freshwater body did not appear to be different from locations in close

proximity (within 0.6 km) to freshwater bodies. However, intermediate distances showed a positive relation to *S. mansoni* infections. Higher risks were associated with soils of medium texture, high basicity, low bulk density and low water capacity. The HII showed no relation with *S. mansoni*.

Table 6.4: Logistic regression parameter estimates for *S. mansoni* summarised by odds ratios (OR), 95% confidence intervals (CI) and 95% Bayesian credible intervals (BCI).

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Study year			
<1980	1	1	1
1980-1989	1.30 (1.26, 1.33)*	1.21 (1.17, 1.25)*	1.86 (1.79, 1.91)*
1990-1999	1.84 (1.79, 1.89)*	1.98 (1.92, 2.05)*	1.99 (1.89, 2.08)*
2000-2009	1.11 (1.08, 1.14)*	1.13 (1.09, 1.17)*	1.48 (1.41, 1.55)*
Latitude (in °)			
<-2.8	1	1	1
-3.8	2.21 (2.15, 2.27)*	3.67 (3.53, 3.81)*	1.80 (1.57, 2.08)*
1.1-9.8	2.42 (2.34, 2.49)*	4.21 (4.04, 4.38)*	1.41 (1.29, 1.55)*
>9.8	2.22 (2.16, 2.29)*	5.39 (5.12, 5.66)*	2.44 (2.28, 2.61)*
Longitude (in °)			
<32.9	1	1	1
32.9-37.2	1.30 (1.28, 1.33)*	1.00 (0.97, 1.03)	2.21 (2.03, 2.39)*
>37.2	1.45 (1.42, 1.48)*	1.63 (1.55, 1.70)*	2.82 (2.62, 2.97)*
Altitude (m)			
<801	1	1	1
801-1154	0.80 (0.78, 0.82)*	0.57 (0.54, 0.60)*	0.53 (0.50, 0.57)*
1155-1513	0.88 (0.86, 0.90)*	0.59 (0.56, 0.63)*	0.50 (0.46, 0.55)*
>1513	0.59 (0.58, 0.60)*	0.59 (0.55, 0.63)*	1.04 (0.99, 1.18)
Day LST (°C)			
<27.8	1	1	1
27.8-30.1	1.16 (1.13, 1.19)*	0.96 (0.93, 1.00)	1.23 (1.19, 1.27)*
30.2-33.1	0.88 (0.86, 0.90)*	0.75 (0.72, 0.78)*	1.00 (0.99, 1.02)
>33.1	1.18 (1.15, 1.20)*	0.40 (0.38, 0.42)*	0.75 (0.72, 0.79)*
Night LST (°C)			
<15.7	1	1	1
15.7-18.0	2.68 (2.60, 2.76)*	3.82 (3.67, 3.98)*	1.67 (1.59, 1.75)*
18.1-20.1	2.35 (2.28, 2.42)*	5.10 (4.85, 5.37)*	2.27 (2.09, 2.46)*
>20.1	2.57 (2.50, 2.64)*	4.01 (3.74, 4.30)*	1.61 (1.50, 1.80)*
Rainfall (mm)			
<216	1	1	1
216-312	0.49 (0.48, 0.50)*	0.77 (0.74, 0.80)*	1.56 (1.48, 1.66)*
>312	0.80 (0.79, 0.82)*	0.91 (0.87, 0.95)*	3.19 (3.02, 3.46)*

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Land cover			
Savannah/ shrublands	1	1	1
Forests	0.87 (0.84, 0.89)*	0.78 (0.75, 0.81)*	1.15 (1.10, 1.19)*
Grasslands/ sparsely vegetated	1.15 (1.12, 1.18)*	1.03 (1.00, 1.07)	1.00 (0.99, 1.02)
Croplands	0.82 (0.80, 0.84)*	0.94 (0.91, 0.97)*	1.00 (0.97, 1.02)
Urban	0.39 (0.38, 0.40)*	0.74 (0.71, 0.78)*	0.58 (0.54, 0.61)*
Wet areas	1.13 (1.09, 1.17)*	0.79 (0.75, 0.84)*	0.97 (0.90, 1.00)
Slope (in °)			
<0.28	1	1	1
0.28-0.92	1.35 (1.32, 1.38)*	1.29 (1.26, 1.33)*	1.01 (0.99, 1.05)
>0.92	1.10 (1.08, 0.13)*	1.03 (1.01, 1.06)*	0.96 (0.92, 0.99)*
Aspect (in °)			
<73.1	1	1	1
73.1-182.2	1.06 (1.03, 1.08)*	1.28 (1.24, 1.31)*	1.03 (1.00, 1.08)
>182.2	1.27 (1.24, 1.29)*	1.34 (1.30, 1.38)*	1.13 (1.09, 1.18)*
Human influence index	0.96 (0.96, 0.96)*	0.96 (0.96, 0.97)*	1.00 (0.99, 1.00)
Distance to closest water body (km)			
<0.59	1	1	1
0.59-1.56	1.23 (1.20, 1.26)*	1.14 (1.11, 1.18)*	1.16 (1.11, 1.21)*
1.57-3.70	1.04 (1.02, 1.07)*	0.85 (0.83, 0.88)*	0.99 (0.96, 1.01)
>3.70	0.72 (0.71, 0.74)*	0.72 (0.70, 0.75)*	0.99 (0.96, 1.05)
Bulk density (in kg/dm3)			
<1.30	1	1	1
1.30-1.31	1.49 (1.45, 1.53)*	1.69 (1.61, 1.78)*	1.31 (1.21, 1.39)*
1.32-1.40	0.73 (0.71, 0.75)*	0.57 (0.54, 0.59)*	0.59 (0.56, 0.62)*
>1.40	0.57 (0.56, 0.59)*	0.39 (0.37, 0.41)*	0.84 (0.80, 0.88)*
Available water capacity (in cm/m)			
<8	1	1	1
8-9	0.72 (0.70, 0.75)*	0.47 (0.44, 0.50)*	0.32 (0.28, 0.36)*
10-12	1.00 (0.96, 1.04)	0.35 (0.33, 0.38)*	0.59 (0.54, 0.65)*
>12	1.05 (1.02, 1.09)*	0.23 (0.21, 0.25)*	0.65 (0.60, 0.72)*
pH in water			
<5.2	1	1	1
5.2-6.7	0.79 (0.77, 0.81)*	1.16 (1.10, 1.21)*	0.77 (0.72, 0.85)*
>6.7	1.28 (1.25, 1.31)*	1.36 (1.29, 1.43)*	0.47 (0.44, 0.49)*

Continued on next page

	Bivariate non-spatial OR (95% CI)	Multivariate non-spatial OR (95% CI)	Multivariate spatial OR (95% BCI)
Texture class			
Medium	1	1	1
Coarse	0.56 (0.53, 0.59)*	0.36 (0.33, 0.39)*	0.48 (0.44, 0.54)*
Fine	0.92 (0.91, 0.94)*	0.58 (0.56, 0.60)*	0.69 (0.67, 0.72)*
			Mean (95% BCI)
Sigma2	-	-	3.79 (3.16, 4.50)
Range (km)	-	-	356.9 (336.9, 379.0)

*: Significant correlation based on 95% CI or 95% BC

The estimated spatial parameters were similar for both *Schistosoma* species. Spatial ranges of 355 km (95% BCI: 341-372 km) and 357 km (95% BCI: 337-379 km) were observed for *S. haematobium* and *S. mansoni*, respectively, and respective spatial variation of 3.83 (95% BCI: 3.15-4.49) and 3.79 (95% BCI: 3.16-4.50).

6.3.3 *Schistosoma* infection risk maps

The spatial distribution of *S. haematobium* risk throughout eastern Africa is shown in Figure 6.2A. Large areas of high infection risk (>50%) were predicted for central Mozambique, the south of Lake Victoria and around the Sudanese and Eritrean border. Low risk areas with predicted infection risks <10% were mainly located in mid/northern Zambia, around Mount Kilimanjaro, in the north of Lake Victoria, northern Sudan and in Ethiopia. The map of the SD of the prediction error for *S. haematobium* (Figure 6.2B) demonstrates that areas of relatively high uncertainty (above 30%) are mainly found in areas of high infection risk and far away from sampled survey locations. We found a mean SD of about 23%, varying from 0% to 32.5%, with areas of low uncertainty typically in close proximity to sub-sampled locations.

Figure 6.3A displays the *S. mansoni* infection risk map, and Figure 6.3B shows the corresponding map of the SD of the prediction error. Low risk areas (predicted infection risk <10%) occur in large parts of Zambia, central Tanzania, around the Ugandan and Kenyan border and the Ethiopian highlands. High risk areas (predicted infection risk >50%) are located in northern Mozambique, southern Tanzania, around Lake Victoria, south-eastern Kenya and small areas in Ethiopia and Sudan. Main areas of high uncertainty are found in Mozambique, south-western Sudan, northern Eritrea, and parts of Somalia and Kenya, while areas of low uncertainty are located around the sampled survey locations and low-risk settings. We calculated a mean SD of 23.5%, varying from 0% to 32.5%.

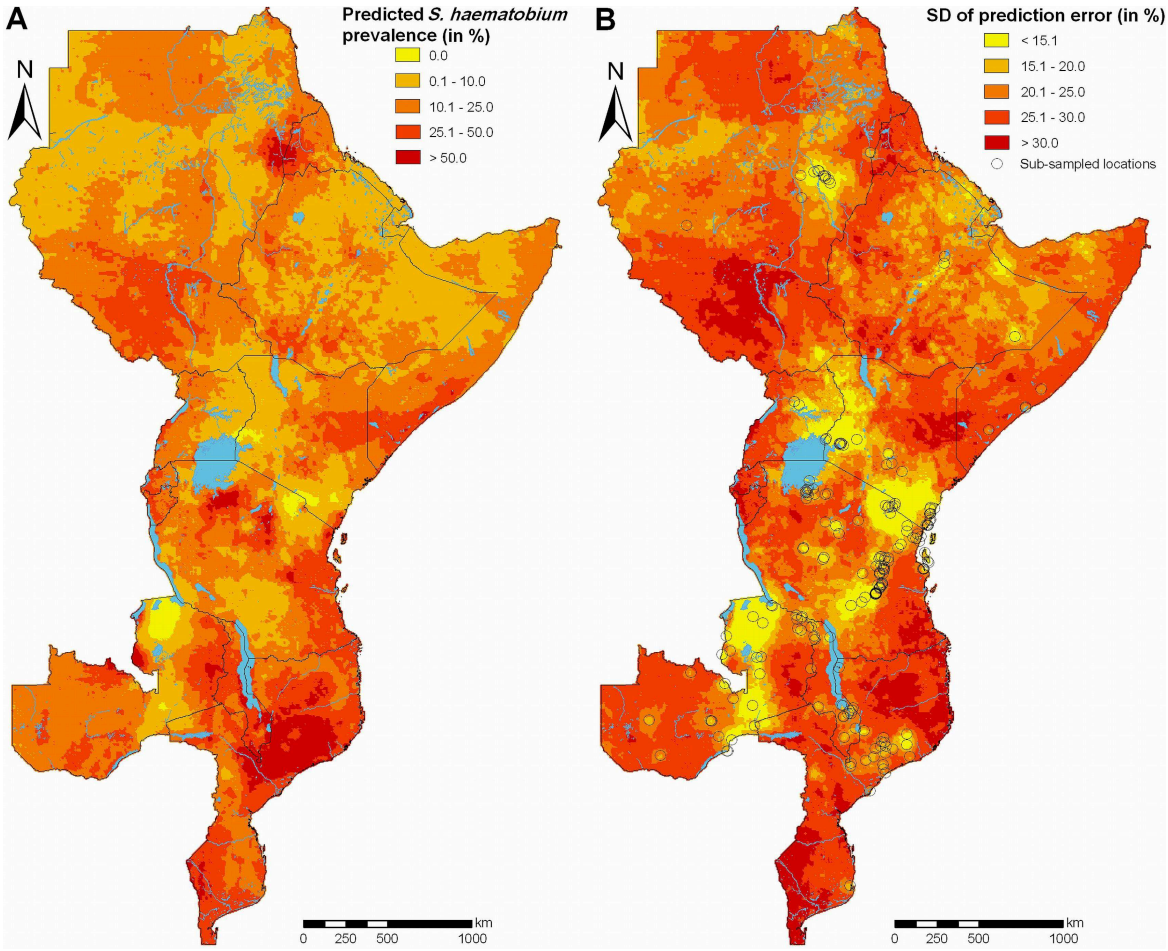

Figure 6.2: Predicted median of infection risk for individuals aged ≤20 years for *S. haematobium* during the period of 2000-2009 based on joint Bayesian kriging (A) and standard deviation (SD) of the prediction error with sub-sampled survey locations (B).

6.3.4 Country prevalence estimates and numbers of infected individuals

Population-adjusted country prevalence estimates are summarised in Table 6.5 for individuals aged ≤20 years and the total population. For *S. haematobium*, prevalence estimates for the total population vary from 11.9% (Djibouti) to 40.9% (Mozambique), whereas for *S. mansoni* they vary between 12.9% (Uganda) and 34.5% (Mozambique). In Burundi, Malawi, Mozambique, Rwanda and Zambia, *S. haematobium* was the predominant species, while in Djibouti, Eritrea, Kenya, Somalia and Sudan, *S. mansoni* was the primary *Schistosoma* species. Both species were estimated to have similar country prevalence in Ethiopia, Tanzania and Uganda. Combined schistosomiasis prevalence estimates, assuming independence in the occurrence of the two species, ranged between 25.3% (Uganda)

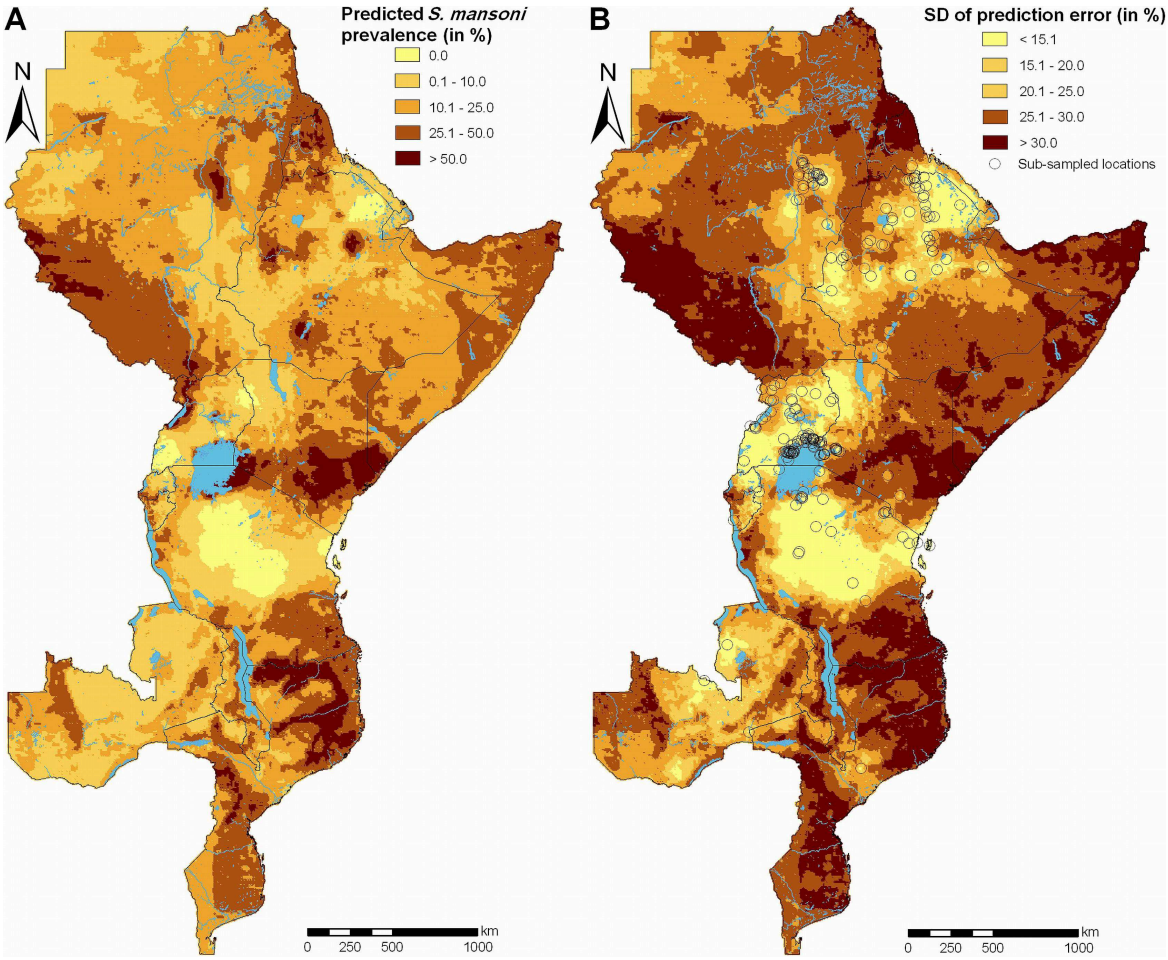

Figure 6.3: Predicted median of infection risk for individuals aged ≤20 years for *S. mansoni* during the period of 2000-2009 based on joint Bayesian kriging (A) and standard deviation (SD) of the prediction error with sub-sampled survey locations (B).

and 55.6% (Mozambique) for the total population.

The number of infected individuals per country, stratified by individuals aged ≤20 years and the total population, is given in Table 6.6. High numbers of infected individuals among the total population (>5 million) were predicted for Ethiopia, Malawi, Mozambique, Sudan, Tanzania and Uganda for *S. haematobium*, and in Ethiopia, Kenya, Mozambique, Sudan and Tanzania for *S. mansoni*. Low numbers (<1 million) were only observed in Djibouti for *S. haematobium* and *S. mansoni*. The combined number of infected individuals vary from 147,000 (Djibouti) to 29.1 million (Ethiopia) with a total of approximately 122 million infections in the 13 countries considered here in eastern Africa.

Table 6.5: Population-adjusted prevalence of *S. haematobium* and *S. mansoni* in individuals (≤20 years) and in the total population, stratified by country in eastern Africa (predicted for the period 2000-2009) based on 2010 population estimates with 95% Bayesian credible interval (BCI).

Country	*S. haematobium* prevalence (%)		*S. mansoni* prevalence (%)		Schistosomiasis prevalence (%)		
	≤20 years 95% BCI	Entire pop. 95% BCI	≤20 years 95% BCI	Entire pop. 95% BCI	≤20 years[a] 95% BCI	Entire pop.[a] 95% BCI	Entire pop.[b]
Burundi	32.2 (5.1, 76.9)	29.9 (4.7, 71.2)	21.0 (6.3, 49.5)	20.2 (6.1, 47.7)	42.3 (11.7, 83.9)	40.3 (11.1, 80.7)	13.3
Djibouti	12.8 (0.9, 69.0)	11.9 (0.8, 63.9)	21.6 (2.0, 77.6)	20.8 (1.9, 74.9)	30.0 (2.7, 87.8)	28.8 (2.6, 86.0)	-
Eritrea	24.0 (6.6, 64.3)	22.2 (6.1, 59.6)	32.4 (11.3, 61.6)	31.2 (10.9, 59.3)	44.1 (16.1, 79.4)	42.5 (15.4, 77.6)	7.2
Ethiopia	19.5 (9.8, 31.0)	18.1 (9.1, 28.7)	22.9 (15.9, 30.9)	22.1 (15.4, 29.8)	33.8 (22.4, 45.0)	32.5 (21.6, 43.4)	7.1
Kenya	16.6 (10.1, 25.5)	15.4 (9.4, 23.7)	35.6 (21.6, 51.6)	34.3 (20.8, 49.8)	42.9 (27.8, 58.6)	41.4 (26.7, 56.8)	30.0
Malawi	37.8 (27.0, 50.5)	35.0 (25.1, 46.8)	27.1 (9.3, 56.6)	26.1 (9.0, 54.6)	50.0 (32.5, 72.2)	47.7 (30.8, 70.1)	42.9
Mozambique	44.2 (32.4, 55.2)	40.9 (30.0, 51.2)	35.8 (21.4, 49.4)	34.5 (20.6, 47.6)	57.7 (42.7, 69.7)	55.6 (40.9, 67.5)	69.8
Rwanda	31.3 (5.8, 75.0)	29.0 (5.3, 69.6)	15.8 (3.8, 41.6)	15.3 (3.6, 40.1)	38.6 (9.2, 81.4)	36.6 (8.8, 78.1)	5.9
Somalia	26.3 (17.5, 37.2)	24.4 (16.3, 34.5)	32.3 (18.6, 50.1)	31.2 (17.9, 48.3)	44.0 (29.4, 59.2)	42.5 (28.2, 57.8)	18.0
Sudan	23.4 (16.2, 33.8)	21.7 (15.1, 31.3)	29.5 (21.8, 39.0)	28.4 (21.1, 37.6)	40.5 (30.7, 52.2)	39.0 (29.5, 50.5)	18.2
Tanzania	24.8 (19.2, 32.0)	23.0 (17.8, 29.7)	20.0 (13.9, 28.7)	19.3 (13.4, 27.6)	38.2 (30.6, 48.0)	36.4 (29.1, 45.8)	51.5
Uganda	17.5 (7.7, 33.5)	16.2 (7.1, 31.1)	13.4 (10.0, 18.0)	12.9 (9.6, 17.3)	26.6 (16.8, 42.2)	25.3 (16.1, 39.9)	32.0
Zambia	26.1 (17.9, 34.3)	24.2 (16.6, 31.8)	16.2 (8.0, 28.0)	15.6 (7.7, 27.0)	34.4 (23.4, 46.1)	32.8 (22.1, 44.2)	26.6

[a] Both *S. haematobium* and *S. mansoni* combined, assuming independence between the two species.

[b] Estimated country prevalence of infected individuals with schistosomiasis over all age groups in 1995, as presented by Chitsulo et al. (2000).

Table 6.6: Estimated number of infected in individuals (≤20 years) and in the total population, stratified by country in eastern Africa (predicted for the period 2000-2009) based on 2010 population estimates with 95% Bayesian credible interval (BCI).

Country	Total		*S. haematobium* infected (mill)		*S. mansoni* infected (mill)		Schistosomiasis infected (mill)		
	≤20 years (mill)	Entire pop. (mill)	≤20 years 95% BCI	Entire pop. 95% BCI	≤20 years 95% BCI	Entire pop. 95% BCI	≤20 years[a] 95% BCI	Entire pop.[a] 95% BCI	Entire pop.[b]
Burundi	5.528	9.445	1.78 (0.281, 4.248)	2.820 (0.445, 6.728)	1.159 (0.347, 2.736)	1.908 (0.571, 4.506)	2.340 (0.645, 4.640)	3.806 (1.046, 7.620)	0.84
Djibouti	0.251	0.512	0.032 (0.002, 0.173)	0.061 (0.004, 0.327)	0.054 (0.005, 0.195)	0.107 (0.010, 0.383)	0.076 (0.007, 0.221)	0.147 (0.013, 0.440)	-
Eritrea	3.006	5.477	0.721 (0.197, 1.932)	1.218 (0.333, 3.262)	0.974 (0.340, 1.850)	1.710 (0.597, 3.249)	1.324 (0.484, 2.387)	2.329 (0.843, 4.252)	0.260
Ethiopia	52.200	89.500	10.165 (5.096, 16.180)	16.157 (8.100, 25.718)	11.946 (8.313, 16.132)	19.746 (13.740, 26.665)	17.656 (11.712, 23.500)	29.095 (19.320, 38.803)	4.0
Kenya	21.900	40.300	3.647 (2.217, 5.606)	6.209 (3.774, 9.543)	7.813 (4.730, 11.326)	13.833 (8.373, 20.051)	9.420 (6.105, 12.867)	16.693 (10.775, 22.899)	6.14
Malawi	8.390	14.400	3.168 (2.267, 4.233)	5.047 (3.612, 6.745)	2.272 (0.782, 4.753)	3.764 (1.295, 7.875)	4.194 (2.726, 6.055)	6.883 (4.435, 10.114)	4.2
Mozambique	11.900	20.200	5.273 (3.867, 6.594)	8.263 (6.060, 10.334)	4.271 (2.556, 5.899)	6.96 (4.166, 9.613)	6.895 (5.100, 8.317)	11.224 (8.253, 13.624)	11.3
Rwanda	5.868	10.700	1.834 (0.337, 4.403)	3.113 (0.573, 7.473)	0.928 (0.221, 2.439)	1.639 (0.390, 4.306)	2.263 (0.541, 4.776)	3.930 (0.943, 8.387)	0.38
Somalia	5.149	9.150	1.354 (0.902, 1.917)	2.23 (1.486, 3.158)	1.664 (0.957, 2.578)	2.851 (1.639, 4.415)	2.266 (1.513, 3.047)	3.890 (2.575, 5.289)	1.71
Sudan	23.400	42.100	5.482 (3.801, 7.900)	9.148 (6.343, 13.183)	6.902 (5.110, 9.119)	11.976 (8.867, 15.825)	9.465 (7.180, 12.218)	16.416 (12.421, 21.260)	4.85
Tanzania	23.600	42.100	5.84 (4.537, 7.544)	9.666 (7.510, 12.487)	4.717 (3.282, 6.759)	8.119 (5.650, 11.634)	8.998 (7.219, 11.309)	15.304 (12.229, 19.273)	15.24
Uganda	21.300	33.600	3.73 (1.634, 7.150)	5.45 (2.388, 10.447)	2.858 (2.131, 3.831)	4.343 (3.238, 5.821)	5.674 (3.589, 8.996)	8.511 (5.402, 13.434)	6.14
Zambia	6.473	10.900	1.688 (1.160, 2.221)	2.635 (1.810, 3.467)	1.051 (0.518, 1.809)	1.706 (0.841, 2.938)	2.229 (1.513, 2.983)	3.578 (2.408, 4.815)	2.39
TOTAL	188.965	328.384	44.714	72.017	46.609	78.662	72.800	121.806	57.45

[a] Both *S. haematobium* and *S. mansoni* combined, assuming independence between the two species.

[b] Estimated country prevalence of infected individuals with schistosomiasis over all age groups in 1995, as presented by Chitsulo et al. (2000).

6.3.5 Model validation results

Model validation based on 80% of the survey locations resulted in MEs of 0.6 for *S. haematobium* and -1.7 for *S. mansoni*, and MAEs of 15.3 and 14.2, respectively. The percentage of test locations correctly predicted by 95% BCIs is 78.3% for *S. haematobium* and 71.1% for *S. mansoni*.

6.4 Discussion

To our knowledge, we present the first smooth empirical schistosomiasis prevalence maps at a spatial resolution of 5 x 5 km for an ensemble of 13 countries in eastern Africa. The maps are stratified by the two main *Schistosoma* species, *S. haematobium* and *S. mansoni*. Bayesian geostatistical models with an approximation of the spatial process were employed to handle the large amount of unique survey locations extracted from a readily available open-access GNTD database (Hürlimann et al., 2011; Schur et al., 2011b,d; Stensgaard et al., 2011). Our prevalence maps are accompanied by contemporary population-adjusted prevalence estimates and number of infected individuals on a country-by-country basis. An attempt was made to employ factors to align surveys arising from different risk groups, namely individuals aged $\leq$20 years and entire communities. This enabled us to obtain age-adjusted risk estimates for individuals aged $\leq$20 years, who are known to carry the highest schistosomiasis risk (WHO, 2002), as well as entire populations. The spatial resolution of 5 x 5 km is a compromise between computational burden and estimation accuracy. A map of 1 x 1 km resolution would result in a total of more than 6 million pixels in the study area, as compared to 260,000 pixels when a 5 x 5 km resolution is chosen (a 25-fold difference). In addition, schistosomiasis risk is influenced by local factors (e.g. people's movements, behaviour, socio-economic factors), which are unknown and therefore prediction at very high spatial resolution may not be rational.

The GNTD database represents an important output of the EU-funded CONTRAST project. As of early October 2010, the database contained over 4000 survey locations across eastern Africa that have been obtained through a systematic search of published and unpublished sources. Importantly, various remotely sensed parameters were incorporated into our models to evaluate the effect of climate and other environmental factors on *Schistosoma* infection risk. For the first time in large-scale schistosomiasis risk profiling, different soil characteristics were also included, such as pH and available water capacity, which might have an effect on the intermediate host snails, and hence potentially influence disease risk. Another initiative on mapping helminthic infections has been taken by the

Global Atlas of Helminth Infections (GAHI; `http://www.thiswormyworld.org`) project (Brooker et al., 2010). It would be interesting to compare our estimates with the ones from the GAHI project once they become available for schistosomiasis.

Clements et al. (2010) previously presented a *S. mansoni* risk map using Bayesian geostatistical modelling for Burundi, Uganda and parts of Kenya and Tanzania based on geo-referenced school surveys carried out between 1998 and 2007. Their map shows similar schistosomiasis risk patterns than the one presented here, yet discrepancies are evident in the area north of Lake Albert and areas in proximity to the Kyoga, Edward and George lakes. Importantly though, the prevalence estimates in both maps are similar. One decade ago, another risk map using non-spatial logistic regression was published for Tanzania, focussing on *S. haematobium* (Brooker et al., 2001). However, this map does not show the actual level of schistosomiasis risk, but rather probabilities that the predicted risk is above a certain cut-off fixed at 50%. This cut-off has been proposed by the World Health Organization (WHO); areas where >50% of school-aged children are infected warrant yearly preventive chemotherapy to entire communities (WHO, 2002, 2006b). However, such maps do not provide detailed information for lower risk areas or the number of infected individuals, which is important for operational and programmatic reasons. Additionally, such maps cannot be used for monitoring and evaluation of interventions. A smaller Bayesian geostatistical risk map covering areas of north-western Tanzania for *S. haematobium* (Clements et al., 2006a) revealed similar patterns of risk as predicted in our map. Nonetheless, differences, especially in the estimated prevalence level, can be found in areas further away from Lake Victoria where we predicted higher risk of infections than Clements and colleagues. In the 1950s and 1970s, higher prevalence of *S. haematobium* compared to *S. mansoni* was found in the Lango region of Central Northern Uganda (Schwetz, 1951; Bradley et al., 1967), while recent investigations in the same area only detect few *S. haematobium* cases. The underlying reasons for this decline remain to be determined (Adriko et al., 2011). Therefore, we are likely to overestimate the current *S. haematobium* and schistosomiasis risk in this region.

We obtained estimates of the number of infected individuals aged ≤20 years, and for all age groups, on a country basis by overlaying population data adjusted for 2010 on the predicted risk surfaces for the two *Schistosoma* species. These estimates are empirical model-based, while previous country estimates presented by Chitsulo et al. (2000), Steinmann et al. (2006) and Utzinger et al. (2009), are interpolations of limited survey data for a whole country. Chitsulo and colleagues reported 57.5 million infected individuals in eastern

Africa, which is less than half of our combined schistosomiasis prevalence estimate (122 million). We observed at least three-fold more infected individuals in Burundi, Eritrea, Ethiopia, Rwanda and Sudan and similar numbers in Mozambique and Tanzania.

How can these differences be explained? First and foremost, populations have grown. The Chitsulo et al. estimates are calculated for the mid-1990s (estimated population in the 13 countries: 219.4 million) compared to our estimates for the year 2010 (328.4 million). Second, individuals might have moved into areas in close proximity to freshwater bodies or newly established irrigation systems. These areas are likely to be linked to higher *Schistosoma* infection risks (Steinmann et al., 2006), even though no significant effect for the distance to the nearest freshwater body was observed for *S. haematobium*. Third, large-scale preventive chemotherapy programmes (Fenwick et al., 2009; WHO, 2010), improved sanitation (WHO and UNICEF, 2010), water resources development and management (Fenwick, 2006; Steinmann et al., 2006), urban-rural movements and socio-economic development are important underlying determinants of changing schistosomiasis risk patterns. Fourth, discrepancies might be related to interpolations of few data points over large areas without taking into account model-based predictions on the basis of climate, environment and disease data. Fifth, we might also underestimate country-specific schistosomiasis prevalence due to the assumption of independence between the occurrence of *S. haematobium* and *S. mansoni*. Simultaneous infections with both species in areas where the species co-exist might be more frequent (e.g. due to similar and highly behavioural infection pathways) or less frequent (e.g. due to protective factors) than expected by chance.

Model validation at 20% of the original survey locations included in our models showed that we are able to correctly predict more than 70% of the locations when considering 95% BCIs. In general, our predictions are approximately 14-15% away from the observed prevalence with a small tendency for *S. mansoni* to overestimate the risk. We are encouraged by these results due to the complexity of schistosomiasis disease transmission in reality (Stensgaard et al., 2011). Nevertheless, certain modelling assumptions might have influenced model performance. For example, overall prevalence and infection intensity depend on the sensitivity and specificity of the diagnostic technique (Bergquist et al., 2009). However, we assumed that the different diagnostic techniques in our dataset have similar ability to detect a *Schistosoma* infection, which might bias the results. Spatial models accounting for sensitivity and specificity could be incorporated in the models, as demonstrated by Wang et al. (2008). However, due to a large number of missing or incomplete information in our underlying data, assumptions on the diagnostic techniques and

the sampling effort would be required, which may introduce considerable bias. Another concern is the amount of zero outcomes (i.e. none of the study participants found to be infected), especially for *S. mansoni* (*S. mansoni*: 30.0%; *S. haematobium*: 14.3%). Zero-inflated models could be implemented instead. Such models modify the likelihood function and add an additional model parameter capturing the over-dispersion arising by the zeros (Vounatsou et al., 2009). Furthermore, our models are assumed to be isotropic stationary, which implies that the spatial random effect is stable throughout the study area (Gosoniu et al., 2009) and that the spatial correlation is the same within the same distance irrespective of direction (Ecker and Gelfand, 2003). This is a potentially inappropriate assumption because dry regions might be less suitable for the disease to spread than humid region and therefore the spatial range could be smaller. Additionally, intermediate host snails spread along rivers and lakeshores, and hence correlation is likely to be attributed to directions.

School-aged children are known to carry the highest risk of *Schistosoma* infection, and hence are the key target group for preventive chemotherapy (WHO, 2002, 2006b). However, large amounts of surveys included in the database are either community-based or involved adults only. It follows that these surveys are related to lower risks of infection. Hence, unadjusted combination of all surveys in one model, irrespective of age-group, would result in inaccurate community risk estimates. We incorporated age-alignment factors to merge studies based on the three main age groups (individuals aged $\leq$20 years, individuals aged >20 years, entire communities) present in our data. These factors are expected to increase model performance compared to models considering only one age group or models without any alignment between the different age groups (Schur et al., 2011d). However, age adjustment could be refined and adopted to different disease transmission settings in order to further enhance model performance.

Temporal trends included in the risk estimation highlighted the differences in schistosomiasis prevalence levels between the 1980s, 1990s and the 2000s. The risk of infection for *S. haematobium* has been lowered during the past two decades, while *S. mansoni* risk increased during the 1980s and 1990s and dropped slightly during the present decade. Major water resources development and management activities might explain the observed increase in *S. mansoni* risk over the past decades. Human-altered habitats has been shown to be an important determinant of the distribution of the major intermediate host snail species at the African continent (Stensgaard et al., 2011). Indeed, there is evidence that urinary schistosomiasis is replaced by intestinal schistosomiasis in face of irrigation schemes and large dams (Abdel-Wahab et al., 1979; Steinmann et al., 2006). This phenomenon

is also referred to as 'Nile shift', as it has been documented first in the Nile delta of Egypt after the completion of the Aswan dam. On the one hand, several African countries have (re-)established national schistosomiasis control programmes emphasizing preventive chemotherapy to school-aged children (Fenwick et al., 2009; WHO, 2010), which reduced the community prevalence of both *S. haematobium* and *S. mansoni*.

In comparison to a similar analysis done for West Africa, including Cameroon (Schur et al., 2011b), we implemented further potential covariates in our models such as latitude, longitude, slope, aspect and soil parameters. To our knowledge, we have now implemented soil parameters for the first time in large-scale geostatistical schistosomiasis risk mapping. Importantly, soil parameters were indeed related to the risk of schistosomiasis transmission, and hence improved outcome predictions. While, pH was a predictor for both schistosome species, available water capacity was associated with *S. haematobium*, whereas bulk density, and texture class showed an association with *S. mansoni*. These soil factors directly influence snail habitats and larval survival in the environment (Madsen, 1985b,a; Bavia et al., 1999), and hence are important predictors of schistosomiasis.

6.5 Conclusions and outlook

Our country-specific estimates on the number of schistosome-infected individuals in eastern Africa revealed considerable differences to previous and widely cited statistics. Our new estimates, together with the *Schistosoma* infection prevalence maps, are useful decision tools for disease control managers to efficiently guide interventions to high-risk areas, to plan the frequency of deworming campaigns, to estimate the required drug supplies at the operational unit of drug deployment (e.g. district) in order to reduce the burden of schistosomiasis, and to monitor progress of interventions to ultimately interrupt transmission. Regions of high model uncertainty need to be studied in greater detail to further validate our results and to deepen our knowledge on the spatial distribution of *Schistosoma* infection. In the future, we plan to further expand this work to obtain Africa-wide prevalence estimates and to study temporal trends. In addition, we will include the geographical distribution of key intermediate host snail species in our spatial models to improve model-based predictions and to study the importance of climatic and other environmental covariates on schistosomiasis risk, while accounting for the presence and absence of intermediate hosts. Finally, we will probe the assumption of independence between *S. haematobium* and *S. mansoni* by jointly modelling both species to enhance our combined risk estimates.

Acknowledgements

We thank the many collaborators who contributed geo-referenced schistosomiasis survey data to our GNTD database. NS and EH are grateful for the financial support of the EU-funded CONTRAST project (FP6-STREP-2004-INCO-DEV project no. 032203). This investigation received further financial support from the Swiss National Science Foundation granted to PV (project no. 325200-118379). ASS was supported by a PhD fellowship at the University of Copenhagen, partly funded by DHI Denmark, and thanks the Danish National Research Foundation for its support of the Center for Macroecology, Evolution and Climate.

6.6 Appendix

6.6.1 Geostatistical modelling

Let Y_i and N_i be the number of Schistosoma-infected and screened individuals at location i $(i = 1, \ldots, n)$ and p_i the probability of infection. We assume that Y_i arises from a Binomial distribution, i.e. $Y_i \sim Bin(p_i, N_i)$. The influence of covariates $\underline{X}_i$ and location-specific spatial random effects ω_i are modelled on the logit, as $logit(p_i) = \underline{X}_i^T \underline{\beta} + \omega_i$, where $\underline{\beta}$ is the vector of regression coefficients. Unobserved spatial variation is introduced on ω_i by assuming that $\underline{\omega} = (\omega_1, \ldots, \omega_n)^T$ follows a latent stationary Gaussian process over the study region, $\underline{\omega} \sim MVN(\underline{0}, \Sigma)$. Σ is a matrix with elements Σ_{ij} accounting for the covariance between any pair of locations i and j. The datasets used for this study contain large numbers of survey locations render parameter estimation infeasible. Hence, an approximation of the spatial process was implemented using a subset of m survey locations $(m < n)$. This approach was proposed by Banerjee et al. (2008) and further developed by Gosoniu et al. (2011a) and Rumisha et al. (2011).

Assuming an isotropic exponential correlation function, the matrix elements Σ_{ij} are defined by $\Sigma_{ij} = \sigma^2 exp(-\rho d_{ij})$ with spatial variance σ^2, rate of correlation decay ρ and the distance between locations d_{ij}. The data are spread over large areas and Euclidean distances are not appropriate any longer, since they are unable to account for the curvature of the surface of the Earth. Therefore, the great-circle distance was used (Vincenty, 1975). The minimum distance for which the spatial correlation is less than 5% is referred to as range and can be calculated by $3/\rho$ in the exponential correlation function setting.

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

The model was fitted using MCMC simulation implemented in Fortran 90 code written by the investigators using the standard numerical libraries. Predictive posterior distributions at the prediction locations were estimated via joint Bayesian kriging (Diggle et al., 1998) implemented in Fortran 90 using the standard numerical libraries. Our predictions are based on the period from 2000-2009.

Chapter 7

Bayesian modeling of anisotropic geostatistical data: An application in mapping urinary schistosomiasis in Senegal

Schur N.[1,2], Ndir O.[3], Utzinger J.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Faculté de Médecine, Pharmacie et Odontologie, Université Cheikh Anta Diop, Dakar, Sénégal

This paper has been submitted for publication to *Statistics in Medicine*.

Abstract

Background: A common assumption of geostatistical models is isotropy, that is spatial correlation is a function of distance between locations, irrespective of direction. Anisotropy, in contrast, is characterized by direction-dependent spatial ranges, the most common form of which is geometric range anisotropy with directions defined by an ellipse. For some diseases, geometric range anisotropy might be present due to the transmission process. For instance, freshwater snails act as intermediate hosts in the transmission of schistosomiasis, and hence direction of river flow might be important.

Methodology: We developed Bayesian geostatistical models that explicitly incorporate anisotropic effects based on simulated and real data obtained from a national survey on urinary schistosomiasis in Senegal. Two anisotropic models were developed assuming a global direction of anisotropy and locally dependent directions fixed at the geographical aspect. Model outcomes were compared to those produced by isotropic models and an empirical risk map was obtained from the model with the best predictive ability.

Principal Findings: Model validation results showed that an anisotropic model with a global direction of anisotropy predicted urinary schistosomiasis risk in Senegal more accurately than other isotropic or locally-dependent anisotropic models. Directional effects were pointing toward the main direction of river flow with maximum spatial range of 65 km and a ratio of anisotropy of 4:3.

Conclusion/Significance: Relaxing the isotropic assumption toward geometric range anisotropy leads to improved model-based schistosomiasis risk mapping and more precise parameter estimates.

7.1 Introduction

Urinary schistosomiasis is caused by a chronic infection with the blood fluke *Schistosoma haematobium*. Aquatic snails of the genus *Bulinus* act as intermediate host (Gryseels et al., 2006). The disease affects the urinary track and can cause severe bladder wall pathology and major hydronephrosis (Hatz, 2001; van der Werf et al., 2003). A common condition is blood in urine (hematuria) which is useful indicator for identification of high-risk communities of *S. haematobium* (Lengeler et al., 2002; Robinson et al., 2009). An important epidemiological feature of schistosomiasis is its focal distribution (Utzinger et al., 2010). Recently, there is high interest in controlling the disease and therefore in obtaining high-resolution estimates of the disease burden (Hotez et al., 2007; Utzinger et al., 2009; Magalhães et al., 2011).

Empirical maps of disease distribution can be obtained via regression-based approaches using environmental predictors. Standard regression approaches assume independence between locations. However, unobserved spatially distributed exposures similarly affect locations in close proximity, which in turn introduce spatial correlation to the data. Geostatistical methods take into account the underlying spatial process via location-specific random effect parameters following a zero-mean Gaussian process with a covariance matrix based on a function of distance between locations (Diggle et al., 1998). Such models typically contain large numbers of parameters and cannot be estimated by the commonly used maximum likelihood approaches (Kleinschmidt et al., 2000). Bayesian model formulations fitted via Markov chain Monte Carlo (MCMC) simulations methods are able to simultaneously estimate model parameters and remedy the computational problems of likelihood-based methods (Diggle et al., 1998). Geostatistical models are well established in disease risk mapping, such as for malaria (Gemperli et al., 2006a; Gosoniu et al., 2006; Hay et al., 2009; Gosoniu et al., 2010; Riedel et al., 2010), schistosomiasis (Magalhães et al., 2011; Raso et al., 2005; Clements et al., 2006b, 2008, 2009a; Beck-Wörner et al., 2007; Schur et al., 2011b) and other helminthiases (Raso et al., 2006a; Brooker and Clements, 2009; Clements et al., 2010).

A common assumption of geostatistical models is that of isotropy, that is spatial correlation acts as a function of distance between locations, irrespective of direction (see Glossary). However, the intermediate host snails of schistosomiasis are freshwater-dependent species and spread along rivers, ponds and lake shores. This biological feature is likely to introduce spatial correlation in the data related to the direction of river flow and currents. Association upon distance and direction is referred to as anisotropy (see Glossary). Three

types of anisotropy can be defined; namely (i) sill, (ii) nugget and (iii) range anisotropy (Zimmerman, 1993) (see Glossary). In practice, range anisotropy is the most common type, characterized by direction-dependent spatial ranges. In case the spatial range is defined by directions related to an ellipse, then the spatial process is said to be geometrically anisotropic (see Glossary). In order to describe geometric range anisotropy, the maximum and minimum spatial range given by the major and minor axes of the associated ellipse and the angle of the direction related to the maximum spatial range need to be defined (see Glossary). The ratio of maximum to minimum spatial range is the magnitude of anisotropy.

Geometric range anisotropy has been described in the statistical literature (Banerjee et al., 2003; Diggle and Ribeiro, 2007). Applications to model scallop catches (Ecker and Gelfand, 1999) and fish abundance (Schmidt and Rodriguez, 2010) have been described. With regard to disease risk mapping, to our knowledge, the only application of anisotropic spatial models pertains to predicting malaria risk at global scale (Hay et al., 2009). However, the authors did neither provide estimates on the magnitude and direction of the association, nor compare model outcomes of malaria risk with isotropic model specifications.

Glossary: Definition of selected terms used in this article

Isotropy: spatial process that is depending only on the distance between locations irrespective of direction.

Anisotropy: spatial process that is depending on the distance and direction between locations.

Sill: total amount of variation within a dataset, or the finite limiting value of the semi-variogram.

Nugget: measurement error and/or microscale effect within the data, or the intercept of the semi-variogram.

Range: distance at which locations are effectively uncorrelated (given a specific direction), or the distance at which spatial correlations falls below 5%.

Geometric anisotropy: property of the spatial process assuming that the direction of spatial correlation is determined by an ellipse.

Angle of anisotropy: angle of the associated ellipse that is pointing towards the direction of maximum spatial range.

Aspect: the geographical aspect refers to the horizontal direction of steepest decrease within a given terrain.

A common way of addressing geometric range anisotropy is to expand the covariance structure of isotropic geostatistical models via an additional positive-definite matrix that accounts for the angle, maximum range and ratio of anisotropy (Ecker and Gelfand, 2003). Ecker and Gelfand (2003) implemented the Wishart prior distribution for this matrix to ensure positive-definiteness during model fit. This approach has the limitation that none of the anisotropy parameters can be updated separately or fixed at a certain value defined by environmental features. Following an idea proposed by Johnson (2005), the matrix can be decomposed into three positive-definite matrices, allowing prior distributions to be specified on the parameters of anisotropy. This might be advantageous when modeling urinary schistosomiasis because the direction of anisotropy is potentially linked to the geographical aspect (i.e. the direction of the greatest decrease of the terrain, see Glossary), which is directly related to the direction of river flow. Instead of estimating a global direction of anisotropy, the direction might be fixed locally at the geographical aspect. This might improve model performance and predictive ability, especially in areas with various directions of river flow, which cannot be precisely estimated assuming an anisotropic process based on a global direction.

In this study, we critically determined the assumption of isotropy and developed validated modeling approaches for schistosomiasis risk mapping by explicitly incorporating anisotropy based on simulated and real data from a national survey carried out among schoolchildren in Senegal. For the applied data, we implemented two different anisotropic models assuming (i) global direction of anisotropy estimated by the model; and (ii) locally dependent directions, fixed at the geographical aspect of the survey locations, respectively. Model predictive ability for each model was assessed via several validation methods and compared to isotropic spatial and non-spatial models. The best fitting model was used to obtain a smooth empirical risk map throughout the study area.

7.2 Data

7.2.1 Schistosomiasis survey data

A national survey pertaining to *S. haematobium* infection in Senegal was carried out during May and June 2003 by the national schistosomiasis control program (Ndir, 2003). Overall, 229 schools were selected in all regions of Senegal with the exception of Dakar and Thiés regions. In each school, 50 children aged 6-14 were randomly chosen from grades 4 and 5 (CE2 and CM1) and invited to provide a single urine sample. In case fewer than 50

children were present, additional children from other grades were randomly selected in order to achieve the desired sample size of 50 children per school. Reagent strips were used to detect microhematuria, a proxy for *S. haematobium* infection, in 187 schools. In the remaining 42 schools, eggs of *S. haematobium* were determined under a microscope, using a urine filtration method. Although results from urine filtration and reagent strip testing correlate well at the population level (Lengeler et al., 2002) the diagnostic accuracy of those two methods for determining *S. haematobium* vary, especially in settings outside East Africa (Lengeler et al., 1993; Brooker et al., 2009a). Hence, urine filtration surveys were excluded to remain with a homogeneous data set in terms of diagnostic approach. Schools that could not be retrospectively geo-located were also excluded from subsequent analysis. Overall, our sample consisted of 143 unique locations.

7.2.2 Environmental data

The environmental data included as covariates in the geostatistical models were obtained from freely accessible data sources, as summarized in Table 7.1. In brief, land surface temperature (LST) data were used as a proxy for day and night temperature, the normalized difference vegetation index (NDVI) as proxy for vegetation, rainfall estimate (RFE) for precipitation, and the human influence index (HII) for changes in the environment due to anthropometric activities. NDVI, RFE, day and night LST were summarized as averages 1 year prior to the survey with a lag of 1 month.

Topographical conditions of the area were described by altitude, aspect, and slope. Digitized maps on water body sources (rivers and lakes) in Senegal were acquired and the distance between surveyed schools and the nearest freshwater source estimated. Additionally, the following soil parameters were employed: amount of coarse fragments >2 mm (expressed in %), available water capacity (in cm/m), gypsum content (in g/kg), organic carbon content (in g/kg), pH and soil drainage class (extremely well, well, moderately, poorly drained). Land cover characteristics were re-grouped into four categories, as follows: (i) savannah and shrublands; (ii) forests; (iii) grasslands and sparsely vegetated areas; and (iv) croplands.

The MODIS/Terra data were processed using the 'MODIS Reprojection Tool' (Land Processes DAAC, USGS EROS). RFE were converted in IDRISI 32 (Worcester, Clark University). Processing of the remaining data, distance calculations and displaying of data and results were performed in ArcMap version 9.2 (ESRI). Further data processing was performed, using in-house developed Fortran 90 codes. All environmental data were

Table 7.1: Data sources and properties of the climatic and other environmental covariates used to model schistosomiasis prevalence in eastern Africa.[a]

Source	Data type	Data period	Temporal period	Spatial resolution
Moderate Resolution Imaging Spectroradiometer (MODIS)/Terra[1]	Land surface temperature (LST) for day and night	Apr-2002 - Mar-2003	8 days	1 km
	Normalized difference vegetation index (NDVI)	Apr-2002 - Mar-2003	16 days	1 km
	Land cover	2003	Yearly	1 km
African Data Dissemination Service (ADDS)[2]	Rainfall	Apr-2002 - Mar-2003	10 days	8 km
Earth Resources Observation (EROS) Center[3]	Altitude, slope and aspect	-	-	1 km
International Soil Reference and Information Centre (ISRIC)[4]	Soil parameters	-	-	8 km
HealthMapper database[5]	Water bodies	-	-	Unknown
Socioeconomic Data and Applications Center (SEDAC)[6]	Human influence index (HII)	-	-	1 km

[a] All data accessed on 03. February 2011
[1] Available at: `https://lpdaac.usgs.gov/lpdaac/products/modis_products_table`
[2] Available at: `http://earlywarning.usgs.gov/fews/africa/index.php`
[3] Available at: `http://edc.usgs.gov/\#/Find_Data/Products_and_Data_Available/gtopo30/hydro/`
[4] Available at: `http://www.isric.org/data/isric-wise-derived-soil-properties-5-5-arc-minutes-global-grid-version-11`
[5] Available at: `http://gis.emro.who.int/PublicHealthMappingGIS/HealthMapper.aspx`
[6] Available at: `http://sedac.ciesin.columbia.edu/wildareas/`
[7] Available at: `http://www.ornl.gov/landscan/`

extracted at the survey locations and for the grid of prediction locations with a spatial resolution of 0.01° x 0.01° (approximately 1 x 1 km) resulting in approximately 165,000 pixels covering Senegal.

7.3 Methods

7.3.1 Isotropic model specifications

Let N_i and Y_i be the number of children examined and those with microhematuria, respectively, at location s_i $(i = 1, \ldots, n)$ of the study region $A \subset \Re^2$ with coordinates x_i and y_i. We assumed that Y_i are binomially distributed, that is $Y_i \sim Bin(N_i, p_i)$, with p_i

measuring microhematuria risk. We modeled the association between a set of m covariates $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{im})^T$ and microhematuria via logistic regression $logit(p_i) = \mathbf{X}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ is the vector of regression coefficients.

The aforementioned model is based on the assumption of independence between survey locations, ignoring potential correlation between neighboring sites. However, different factors such as climate, topology or cultural practices are expected to be more similar in neighboring locations introducing spatial variation to the data. We accounted for unobserved spatial correlation in the model via location-specific random effect parameters ω_i implemented on the logit, $logit(p_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \omega_i + \epsilon_i$, assuming that $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ follows a latent stationary Gaussian process with $MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and that ϵ_i are exchangeable random effect parameters with $\epsilon_i \sim N(0, \tau^2)$, where τ^2 is the nugget. The elements Σ_{ij} of the variance-covariance matrix $\boldsymbol{\Sigma}$ are related to an exponential correlation function defined by $\Sigma_{ij} = \sigma^2 exp(-\sqrt{\mathbf{d}_{ij}^T \mathbf{B} \mathbf{d}_{ij}})$ with spatial variance σ^2 and distance vector $\mathbf{d}_{ij} = (x_i - x_j, y_i - y_j)^T$ between any pair of locations s_i and s_j. $\mathbf{B} = \{b_{ij}\}_{i,j=1,2}$ is a symmetric and positive definite matrix in $\Re^2$. In the case of isotropy, $\mathbf{B}$ is reduced to a diagonal matrix with positive elements $b_{11} = b_{22} = b$, where b^2 is a spatial decay parameter (Banerjee et al., 2003).

7.3.2 Geometric range anisotropy

Anisotropy arises when spatial dependence is not only a function of distance but also depends on the direction between pairs of locations. Geometric range anisotropy is a special case of anisotropy with an elliptical form of the spatial correlation and direction-dependent spatial decay parameter (Zimmerman, 1993). The three parameters describing anisotropy are the angle of anisotropy indicating the direction associated with the maximum spatial range of the related ellipse, the spatial range that is the distance at which spatial correlation becomes less than 5%, and the magnitude of anisotropy which is defined by the ratio of maximum to minimum spatial range. The angle of anisotropy α can be calculated by $cot(2\alpha) = (b_{11} - b_{22})/(2b_{12})$. The spatial range r in any direction ϕ is given by $r_\phi = -ln(0.05)/\sqrt{\mathbf{h}_\phi^T \mathbf{B} \mathbf{h}_\phi}$ with $\mathbf{h}_\phi^T = (cos\phi, sin\phi)$. The magnitude of anisotropy λ is calculated by $\lambda = r_\alpha / r_{\alpha + 0.5\pi}$ for a given $\phi = \alpha$.

Geometric range anisotropy can also be interpreted as an isotropic spatial process on a transformed coordinate system based on stretching and rotation of the axes, such as $(x', y') = (x, y) \begin{pmatrix} cos\alpha & -sin\alpha \\ sin\alpha & cos\alpha \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ (Diggle and Ribeiro, 2007). Hence, the above model could be formulated assuming an isotropic form of $\mathbf{B}$ and $\mathbf{d}_{ij} = (x_i' - x_j', y_i' - y_j')^T$.

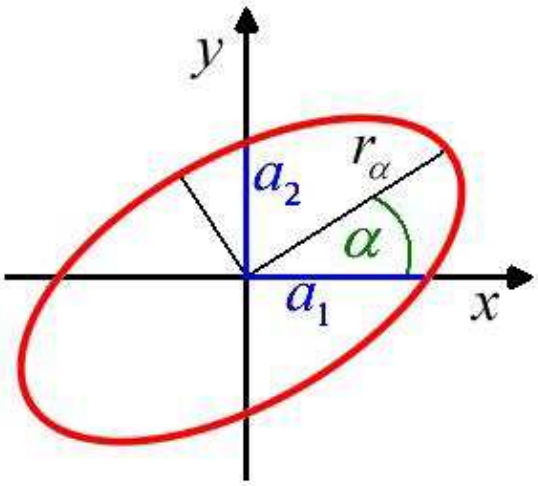

Figure 7.1: Illustration of the associated ellipse of geometric range anisotropy with related parameters.

7.3.3 Prior distributions

A Bayesian model formulation requires the specification of prior distributions of all model parameters. We adopted non-informative uniform prior distributions for the regression coefficients $\boldsymbol{\beta}$ with bounds $-\infty$ and ∞ and a gamma distribution with mean 1 and large variance for the spatial variance σ^2. To ensure that $\mathbf{B}$ remains positive definite, the Wishart prior distribution has been proposed (Banerjee et al., 2003; Ecker and Gelfand, 1999, 2003). This specification is rather inflexible because it does not allow prior distributions directly on the anisotropy parameters. In this study, we re-parameterized $\mathbf{B}$ such as $\mathbf{B} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}$, where $\mathbf{A}$ is a diagonal matrix with positive elements a_1 and a_2. The 2×2 matrix $\boldsymbol{\Psi}$ is parameterized as $\boldsymbol{\Psi} = \begin{pmatrix} 1 & \psi \\ \psi & 1 \end{pmatrix}$ with ψ defined between -1 and $+1$ to ensure positive definiteness of the matrix. Given a_1, a_2 and ψ, the parameters r_α, $r_{\alpha+0.5\pi}$, α and the ratio of anisotropy λ can be calculated (see Appendix). A uniform prior distribution with bounds 0° and 180° was employed for the angle of anisotropy . Uniform prior distributions with bounds at the minimum and maximum distance between locations were adopted for the spatial range parameters r_α and $r_{\alpha+0.5\pi}$. Figure 7.1 depicts the associated ellipse of anisotropy together with the most relevant parameters.

7.4 Implementation details and model validation

7.4.1 Variable selection

Gibbs variable selection was employed to determine a parsimonious set of covariates that show the best fit with the observed data of the application (George and McCulloch, 1993). Inclusion or exclusion of covariates was specified using indicator variables linked to the regression coefficients. In this study, variable selection was based on the estimation of the posterior inclusion probability with prior probability of 0.80. The final model consists of all covariates with a posterior inclusion probability larger than 0.5.

7.4.2 Directional semi-variogram plots

Directional semi-variogram plots were employed to detect evidence of anisotropy in the urinary schistosomiasis survey data from Senegal. The plots were created in R 2.10.0, using the variog and variofit command of the geoR library for angles every 30 with a tolerance of 15°.

7.4.3 Model fit and convergence

The model was fitted using MCMC simulation implemented in Fortran 90 code written by the investigators using standard numerical libraries. Predictive posterior distributions at the prediction locations were estimated via Bayesian kriging (Diggle et al., 1998).

Models were run for two chains with a thinning of 10 and a burn-in of 1000 iterations. Convergence was assessed every 10,000 iterations by inspection of ergodic averages of selected model parameters. After convergence, samples of 500 iterations per chain using a thinning of 10 iterations were extracted for each model resulting in a final sample of 1000 estimates per parameter.

7.4.4 Model validation

Model validation was performed on four different models to obtain the best fitting model for risk mapping of urinary schistosomiasis risk in Senegal. Model A is a non-spatial model, model B is an isotropic spatial model, model C is an anisotropic spatial model with a global angle of anisotropy estimated by the model itself and, finally, model D is an anisotropic spatial model with local angles of anisotropy fixed at the geographical aspect of the locations.

Model performance was assessed by using a training set of 80% of the survey locations for model fit. The remaining 20% (test locations) were kept for model validation.

The predicted outcomes at the test locations were compared to the observed outcomes employing mean absolute errors (MAEs) and χ^2 divergence measures. MAEs provide estimates on the accuracy of a model based on absolute distances between observed outcomes p_i and the median of the predictions $\widehat{p}_i$ at the ith $(i = 1, \ldots, k)$ test location, as $MAE = 1/k \sum_{i=1}^{k} |p_i - \widehat{p}_i|$. The χ^2 measure is calculated by $\chi^2 = \sum_{i=1}^{k} (p_i - \widehat{p}_i)^2 / \widehat{p}_i$.

Table 7.2: Implemented simulation parameters (bold) and results of model fit using an anisotropic model formulation with global angle of anisotropy. Model parameter estimates are based on the median and 95% CIs given in brackets.

Simulation	Coefficient (β_0)	Spatial variance (σ^2)	Angle (α)	Maximum range (r_α)	Minimum range ($r_{\alpha+0.5\pi}$)	Ratio (λ)
1	0.1	1	-	0.5	0.5	1
	0.11	1.12	125.9*	0.7	0.52	1.28
	(-0.20, 0.21)	(0.85, 1.69)	(5.9, 175.0)*	(0.41, 1.28)	(0.35, 0.88)	(1.00, 2.55)
2	0.1	0.2	-	1	1	1
	0.09	0.27	63.3*	1.47	1.14	1.26
	(-0.19, 0.23)	(0.18, 0.46)	(3.0, 173.6)*	(0.74, 3.04)	(0.60, 2.17)	(1.00, 2.14)
3	0.1	0.5	-	2	2	1
	0.29	0.53	75.5*	2.19	1.78	1.19
	(0.15, 0.50)	(0.36, 0.84)	(4.3, 176.1)*	(1.22, 3.99)	(1.04, 3.14)	(1.00, 1.75)
4	0.1	1	0°	1	0.5	2
	0	1.22	-6.7	1.23	0.72	1.7
	(-0.23, 0.18)	(0.88, 1.90)	(-49.6, 35.1)	(0.60, 2.13)	(0.45, 1.21)	(1.04, 2.89)
5	0.1	1	0°	1	0.25	4
	0.04	1.17	-0.3	1.3	0.36	3.6
	(-0.24, 0.30)	(0.85, 1.78)	(-12.1, 9.9)	(0.72, 2.32)	(0.21, 0.63)	(2.04, 6.76)
6	0.1	1	7.5°	1.2	0.25	4.8
	0.06	1.12	5.9	1.48	0.33	4.56
	(-0.07, 0.29)	(0.83, 1.63)	(-1.5, 13.9)	(0.90, 2.65)	(0.20, 0.53)	(2.54, 8.29)
7	0.1	0.5	10°	4	0.5	8
	0.03	0.52	9.6	3.95	0.61	6.18
	(-0.07, 0.15)	(0.39, 0.75)	(4.5, 14.5)	(2.45, 4.92)	(0.40, 0.96)	(3.72, 9.57)
8	0.1	1	25	2.5	0.5	5
	0.03	1.21	26.1	3.46	0.67	4.99
	(-0.32, 0.38)	(0.85, 1.84)	(18.4, 33.2)	(1.92, 4.86)	(0.41, 1.11)	(2.92, 8.26)
9	0.1	0.5	45°	3	1	3
	0.38	0.63	36.5	3.83	1.29	2.85
	(0.05, 0.52)	(0.42, 0.95)	(25.7, 51.8)	(2.25, 4.93)	(0.72, 2.24)	(1.59, 5.05)
10	0.1	1	172.5°	1.2	0.25	4.8
	-0.09	1.24	171.1	1.51	0.37	3.96
	(-0.29, 0.38)	(0.91, 2.00)	(158.4, 180.2)	(0.80, 2.98)	(0.22, 0.71)	(2.27, 7.32)

7.5 Results

7.5.1 Simulation study

Three and seven different data sets under the assumption of isotropy and anisotropy, respectively, at the exact same set of the study locations of our application in Senegal. The data were simulated from binomial distributions, assuming 100 individuals examined for *S. haematobium* using reagent strips per location. Anisotropic data had different magnitudes and angles of anisotropy and zero nugget. Parameter specifications are detailed in Table 7.2.

The results showed that for isotropic data (data sets 1, 2 and 3), the magnitude of anisotropy was estimated close to 1, while the angle could not be estimated correctly, as indicated by the uniform histogram plots of the posterior distribution and the very wide credible intervals (CIs) (see Table 7.2 and Figure 7.2). For anisotropic data (data sets

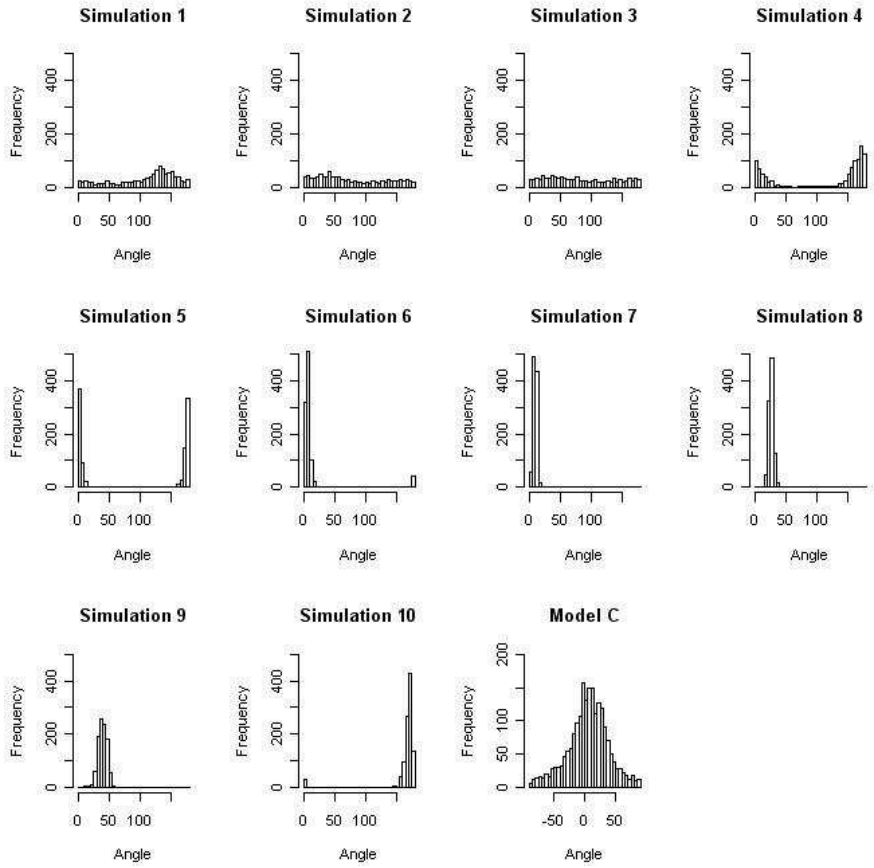

Figure 7.2: Predicted angles of anisotropy for different simulated datasets and model C (anisotropic model with global angle of anisotropy) based on a national school survey on urinary schistosomiasis carried out 2003 in Senegal using reagent strips.

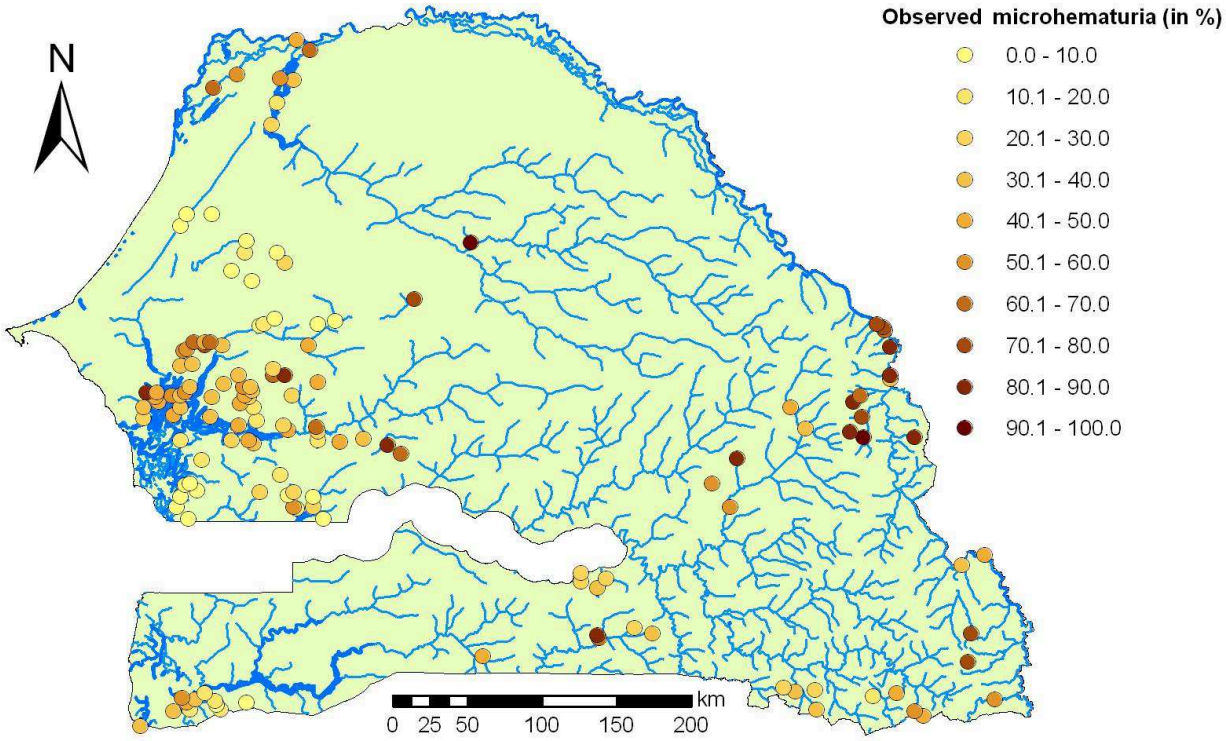

Figure 7.3: Observed prevalence of microhematuria across Senegal obtained from a national schistosomiasis school survey carried out in 2003 using reagent strips.

4-10), the 95% CI for the angle is more narrow and the histogram plot peaks at a value close to the simulated parameter, especially for large ratios of anisotropy. The remaining model parameters are always included in their corresponding 95% CIs highlighting good model performance.

7.5.2 Schistosomiasis data

The final data set from our application consisted of 143 georeferenced locations across Senegal, as shown in Figure 7.3. The prevalence of microhematuria at the unit of the school ranged between 0% and 100% with a mean prevalence of 39.3%. Directional semi-variogram analyses implied evidence of anisotropy at an angle of approximately 0° with maximum and minimum spatial range at approximately 200 km and 50 km, respectively (results depicted in Figure 7.4 and Table 7.3).

Exploratory analyses were carried out to assess linearity of covariates. Bivariate logistic regressions on the potential environmental predictors suggested categorizing of the following covariates: altitude, day temperature, NDVI, RFE, aspect, pH and content of

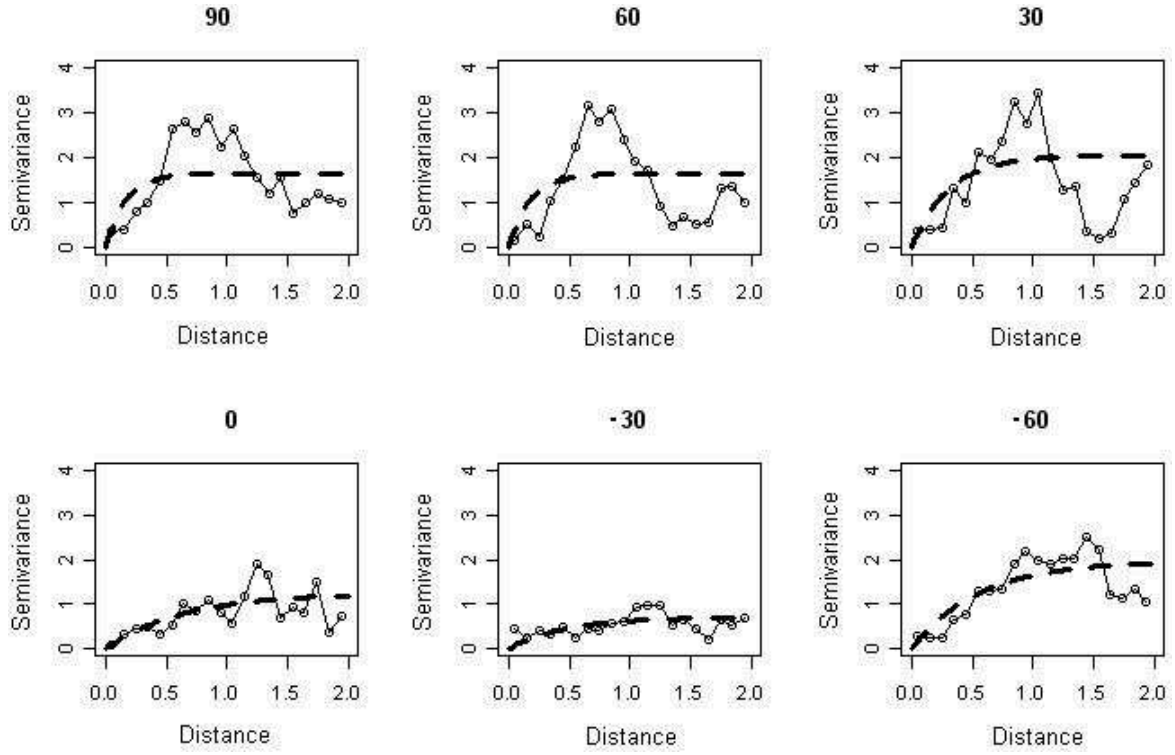

Figure 7.4: Directional semi-variograms performed in R 2.10.0, using the variog and variofit command of the geoR library, based on a national school survey on urinary schistosomiasis carried out 2003 in Senegal using reagent strips.

Table 7.3: Results of the directional semi-variogram analysis performed in R 2.10.0, using the variog and variofit command of the geoR library, based on national school survey data on urinary schistosomiasis in Senegal carried out in 2003.

Direction	Spatial variance	Spatial range (in km)
90°	1.64	49.7
60°	1.65	53.4
30°	2.05	94.5
0°	1.26	196.4
-30°	0.71	140.9
-60°	1.98	168.3
-90°	1.64	49.7

organic carbon. All considered environmental predictors were highly significant in bivariate logistic regressions. However, geostatistical variable selection identified a reduced set of covariates consisting of night temperature, slope, amount of coarse fragments >2 mm, gypsum content and available water capacity to best predict microhematuria risk. The spatial distribution of these covariates is shown in Figure 7.5.

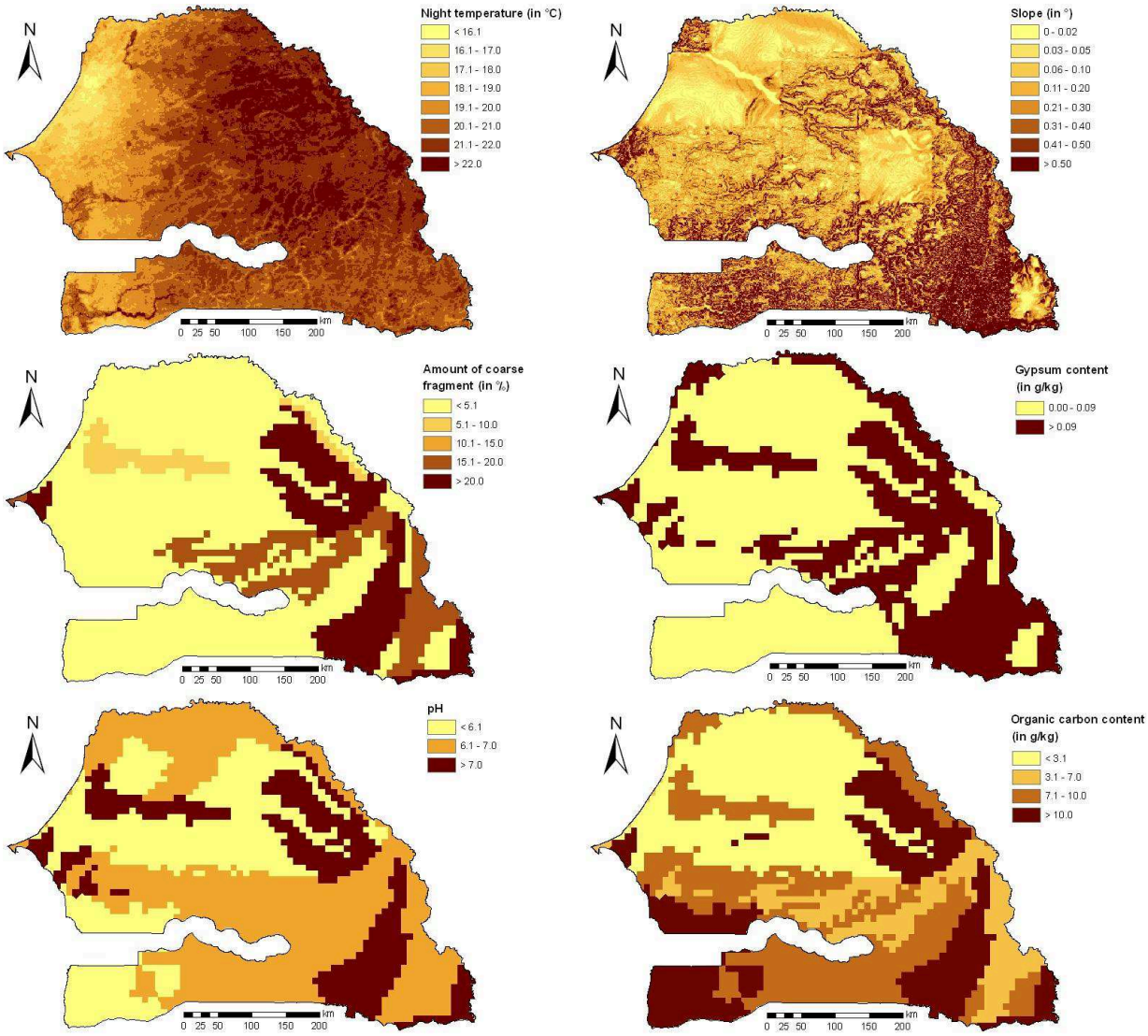

Figure 7.5: Spatial distribution of the environmental predictors of the final set of covariates used for model fit and prediction.

7.5.3 Model validation results

Model validation based on MAE and χ^2 measures showed that the anisotropic model with global angle of anisotropy (model C) was best predicting haematuria in Senegal (see Table 7.4). The isotropic spatial model (B) and the anisotropic one with local angles (D) were ranked second and third, respectively. Therefore, we considered the anisotropic model with a global angle estimated by the model as the best predicting model used for kriging.

Table 7.4: Model validation results based on mean absolute errors (MAE) and χ^2 measures of the 4 implemented models. Model A, non-spatial; model B, isotropic spatial; model C, anisotropic spatial with a global angle of anisotropy; and model D, anisotropic spatial with local angles of anisotropy fixed at the geographical aspect of the locations.

Model	MAE	χ^2 measure
A	21.26 (15.93, 27.31)	32.60 (15.05, 106.25)
B	20.19 (15.20, 25.44)	30.39 (14.50, 77.40)
C	19.69 (14.55, 25.48)	28.33 (13.30, 79.14)
D	20.98 (15.54, 27.71)	31.93 (14.86, 101.45)

7.5.4 Model parameter results

Model parameter estimates of all four models implemented on the applied data set are summarized in Table 7.5. The introduction of spatial random effects resulted in a reduced influence of slope and pH on microhematuria risk and a loss of significance of this association compared to model A, while the effect of night temperature on the outcome remained relatively stable and significant throughout the models. Models accounting for directional effects led to significant associations of the organic carbon content and to opposite (but non-significant) effects of the gypsum content to microhematuria risk. Associations with the amount of coarse fragments remained at the same level, but were significant in model C, while the effect of the slope on the outcome was almost negligible.

In the anisotropic model with a global angle (best predicting model based on our model validation approaches), night temperature and the amount of coarse fragments >2 mm were significantly positively associated with the risk of microhematuria at the school level, while organic carbon content of at least 10 g/kg was significantly negatively correlated. Slope, gypsum content and pH showed no significant association with the risk of microhematuria, whereas acid soils (pH above 7) had a slightly higher risk of microhematuria among surveyed children. The amount of total variation within the data was estimated to be around 1.1 with 85% of the variation associated to spatial effects. Directional effects were pointing

Table 7.5: Model parameter estimates of the 4 implemented models based on a national school survey on urinary schistosomiasis carried out 2003 in Senegal using reagent strips. Model A, non-spatial; model B, isotropic spatial; model C, anisotropic spatial with global angle of anisotropy; and model D, anisotropic spatial with local angles of anisotropy fixed at the geographical aspect of the locations.

Parameter	Model A OR (95% CI)	Model B OR (95% CI)	Model C OR (95% CI)	Model D OR (95% CI)
Night temperature	1.53 (1.50, 1.55)*	1.42 (1.32, 1.52)*	1.40 (1.27, 1.53)*	1.47 (1.43, 1.56)*
Slope	0.39 (0.25, 0.60)*	0.85 (0.50, 1.07)	1.02 (0.42, 1.34)	0.80 (0.54, 1.10)
Amount of coarse fragments >2mm	1.02 (1.00, 1.06)	1.03 (0.98, 1.06)	1.03 (1.01, 1.06)*	1.04 (1.00, 1.08)
Gypsum content (g/kg)				
<0.1	1	1	1	1
≥0.1	1.09 (0.89, 1.96)	1.04 (0.80, 2.48)	0.93 (0.66, 1.25)	0.97 (0.84, 1.28)
pH				
<6.1	1	1	1	1
6.1-7.0	2.60 (1.53, 3.97)*	1.15 (0.78, 3.01)	1.02 (0.79, 2.83)	1.15 (0.88, 1.59)
≥7.1	1.50 (0.96, 2.99)	1.08 (0.77, 2.09)	1.17 (0.98, 2.56)	1.09 (0.80, 2.04)
Organic carbon content (g/kg)				
<3.1	1	1	1	1
3.1-7.0	0.97 (0.77, 1.38)	0.95 (0.72, 1.16)	1.00 (0.80, 1.19)	1.02 (0.78, 1.31)
7.1-10.0	0.97 (0.81, 1.22)	1.29 (0.86, 1.94)	1.02 (0.83, 1.81)	1.24 (1.03, 2.76)
≥10.1	0.97 (0.63, 1.26)	0.98 (0.74, 1.39)	0.45 (0.24, 0.89)*	0.50 (0.32, 0.92)*
	Median (95% CI)	Median (95% CI)	Median (95% CI)	Median (95% CI)
τ^2	1.02 (0.77, 1.31)	0.20 (0.12, 0.32)	0.24 (0.15, 0.34)	0.15 (0.10, 0.31)
σ^2	-	1.00 (0.70, 1.51)	0.93 (0.65, 1.34)	0.91 (0.64, 1.32)
Rhomax (km)	-	65 (42, 105)	65 (28, 149)	39 (19, 60)
Rhomin (km)	-	65 (42, 105)	43 (22, 77)	33 (16, 50)
Ratio	-	1	1.33 (1.00, 3.37)	1.13 (1.00, 1.66)
Angle	-	-	6.7° (-67.2°,69.7°)	-

CI = Credible interval; OR = Odds ratio.
* Significant based on 95% CI

towards 7° (see Figure 7.2) and the ratio of anisotropy was 4:3, including isotropy in a 95% CI. Maximum and minimum ranges were estimated at 65 km and 43 km, respectively.

7.5.5 Risk map of microhematuria for Senegal

Figure 7.6 depicts the predicted microhematuria risk throughout Senegal based on the median of the predictive posterior distribution. Low-risk areas (predicted risk of microhematuria $\leq$10%) are found within approximately 100 km proximity to the seashore. Extremely high risks of microhematuria (>70%) were predicted for the eastern part of Senegal and some hotspots in the central part of the country. The standard deviation (SD) of the posterior predictive distribution is also shown in Figure 7.6. Areas of comparably high uncertainty (SD >0.225) were found in the north-eastern part of the country and at large distances to the survey locations. Figure 7.7 shows the distribution of the posterior predictive distribution of the spatial and non-spatial random effect parameters. Large areas of negative and positive random errors were associated with locations of observed microhematuria <30% and >60%, respectively.

7.6 Discussion

For the first time, we have shown that relaxing the isotropic assumption toward geometric range anisotropy leads to improved model-based risk predictions of urinary schistosomiasis. Interestingly, anisotropy was not highly prominent in the data, and hence it is conceivable that schistosomiasis risk predictions in other settings might be further enhanced. Bayesian geostatistical models were employed to account for potential directional effects of microhematuria risk across Senegal. Microhematuria prevalence data at the unit of the schools were obtained from a national survey carried out in 2003 (Ndir, 2003). Environmental factors (e.g. climatic, topographic and soil parameters) were obtained at high spatial resolution and incorporated in our models to determine outcome-predictor relations and to predict the outcome at a spatial resolution of approximately 1x1 km in order to enhance risk prediction of this highly focal disease. Predictive accuracy of anisotropic models was compared to spatial and non-spatial models with a suite of model validation methods.

Our findings are important for schistosomiasis risk mapping and control, particularly for Africa where more than 95% of the estimated cases of schistosomiasis are concentrated (Utzinger et al., 2009; Steinmann et al., 2006). To enable applied statisticians to employ the geometric range anisotropy models presented within the manuscript, an example OpenBUGS code (OpenBUGS Foundation, London, UK) is provided in the Appendix. Although

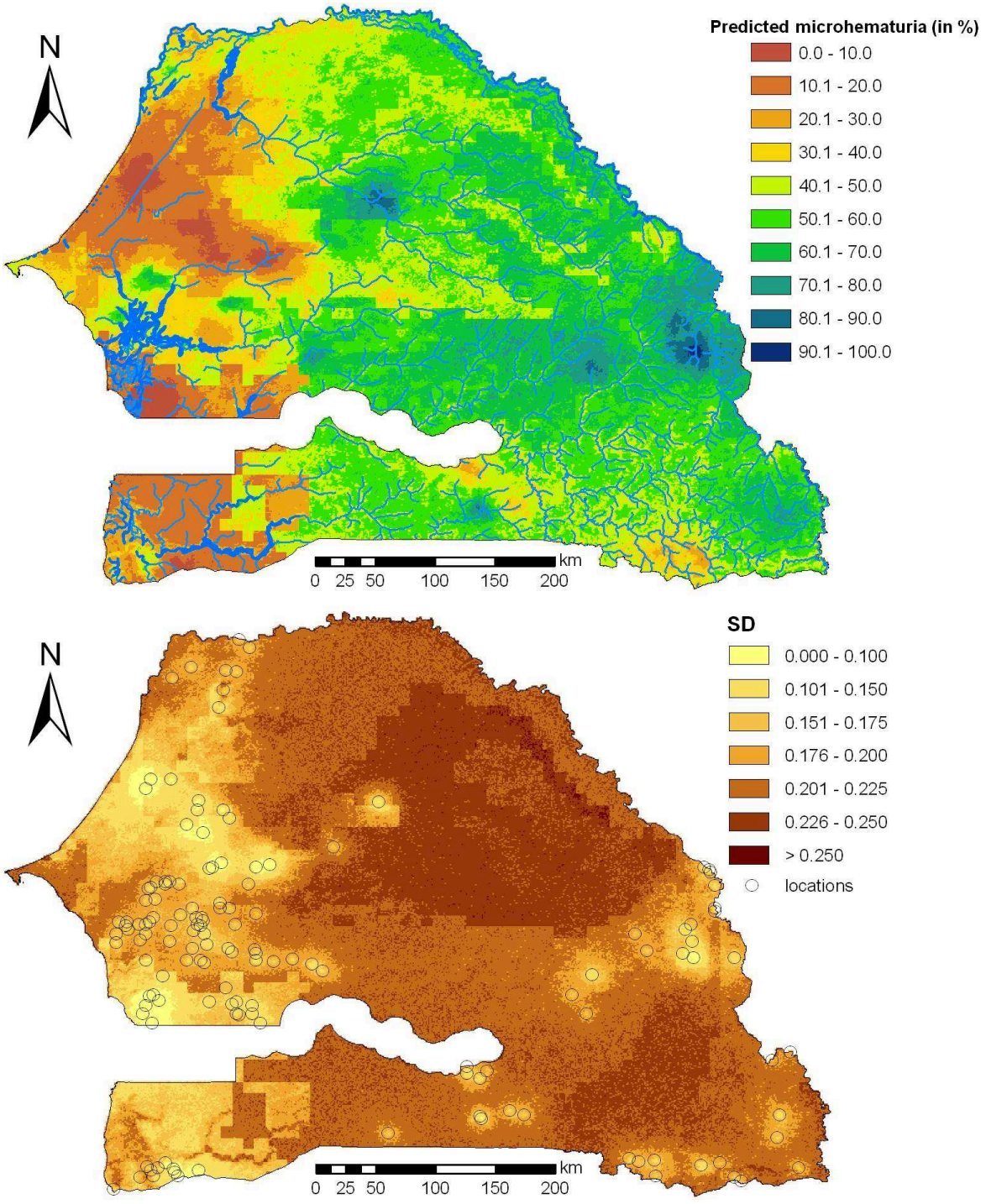

Figure 7.6: Predicted microhematuria risk and standard deviation (SD) of the prediction of the anisotropic spatial model with global angle of anisotropy (model C) based on a national school survey on urinary schistosomiasis with 143 survey locations carried out 2003 in Senegal using reagent strips.

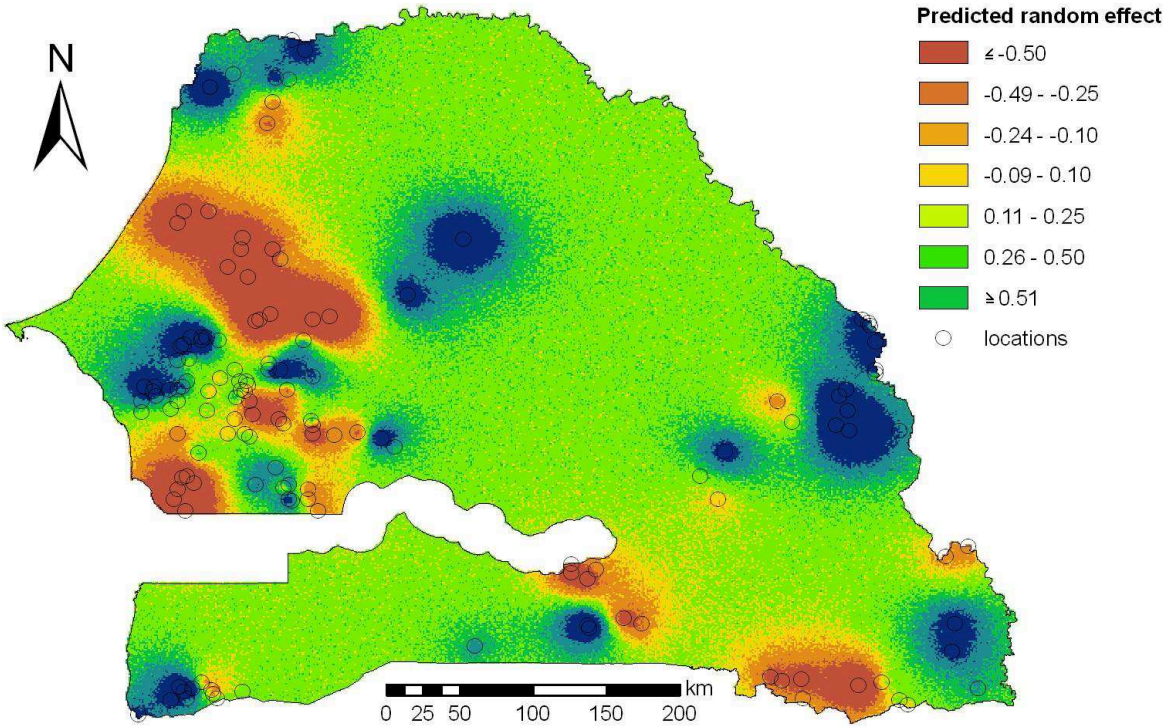

Figure 7.7: Predicted random effect of the microhematuria risk prediction using the anisotropic spatial model with global angle of anisotropy (model C) based on a national school survey on urinary schistosomiasis with 143 survey locations carried out 2003 in Senegal using reagent strips.

model compilation of this code might take several hours or even few days, depending on the number of survey locations and machine power, implementation in Fortran reduced the running time by about half.

Our model validation approach showed that the anisotropic model with global angle of anisotropy (model C) performed most accurately. Furthermore, it correctly predicted 93.0% of all test locations within a 95% CI and discriminatory performance based on 50% and 10% cut-offs resulted in 72.4% and 86.2% correct predictions, respectively.

The final prediction map shows that high-risk areas of urinary schistosomiasis (prevalence of microhematuria >50%) were present in central/eastern Senegal and low-risk areas ($\leq$10%) are located within 100 km from the seashore. The risk patterns on the final prediction map are similar to the patterns of our earlier work pertaining to *S. haematobium* risk mapping across West Africa (Schur et al., 2011b). Discrepancies in the maps can be observed relatively small microhematuria hotspots in northern and southern Senegal. However, in our previous work, we used isotropic spatial models employing a large ensemble of historical survey data pertaining to *S. haematobium* infections. These surveys used

different diagnostic approaches, age ranges of study participants were heterogeneous and different covariates were included. In our previous modelling focussing on West Africa, we estimated a spatial range for *S. haematobium* of approximately 400 km. In the current application focusing on a single West African country, we estimated ranges of spatial correlation between 43 and 65 km, depending on the direction between locations. Introduction of directional effects might reduce spatial range due to an enhanced ability of identifying the spatial surface of underlying data. This claim is supported by the results of the corresponding isotropic spatial model rating spatial range around the maximum range of the anisotropic model (i.e. 65 km). In addition, the estimates presented here are based on a single and quite recent national survey with the key outcome measure (i.e. prevalence of microhematuria at the unit of the school) influenced by schistosomiasis control interventions that might have reduced the effect of environmental predictors and decreased spatial correlation. Discrepancies in the predicted maps, especially in the magnitude of risk, may not only reflect the effect of the relaxation of the isotropic assumptions, but also differences in the data. The earlier work on *S. haematobium* risk in West Africa was based on historical data using different diagnostic techniques (e.g. urine filtration and centrifugation), while the current map is based on a national survey on haematuria (using reagent strips) and might overestimate *S. haematobium* risk.

The angle of anisotropy was estimated around 0° which relates to east-west directional effects. This result is in line with our expectations that schistosomiasis transmission is highly linked to the main direction of river flow (Beck-Wörner et al., 2007). We further assumed that the angle could be fixed at the aspect of the locations representing direction of slope correlated with river flow. However, anisotropic models with fixed angle did not improve microhematuria risk estimates. This observation might be valid for other settings also characterized by a main direction of river flow. In areas of various directions, anisotropic models with global angle of anisotropy might fail to capture directional effects and anisotropic models with local angles fixed at the geographical aspect might be superior in terms of model predictive ability. Furthermore, the study area could be partitioned into sub-regions with similar directions of river flow based on catchments and models could be developed to allow sub-region-specific anisotropic spatial processes (Beck-Wörner et al., 2007). Such an association upon distance and location is known as non-stationarity. The stationary spatial processes in each sub-regions could be either assumed to be independent (Kim et al., 2005), or spatially correlated (Gosoniu and Vounatsou, 2011b).

In conclusion, geometric range anisotropy takes into account spatial correlation related

to distance and direction between locations, and hence might play an important role in risk mapping of diseases that are governed by vectors and intermediate hosts upon which the environment introduces directional effects (e.g. direction of river flow or predominant wind directions). Ignoring anisotropic effects is likely to influence the strength of association and significance of model coefficients and the estimates of the spatial range parameter. These effects might reduce model predictive ability, particularly in the presence of strong anisotropy.

Acknowledgements

NS is grateful for the financial support of the EU-funded CONTRAST project (FP6-STREP-2004-INCO-DEV Project no. 032203). This investigation received further financial support from the Swiss National Science Foundation for PV (project no. 32003B-135769 andPDFMP3-137156).

7.7 Appendix

7.7.1 Formulas

The positive definite matrix $\mathbf{B} = \{b_{ij}\}_{i,j=1,2}$ can be rewritten by $\mathbf{B} = \mathbf{A\Psi A}$, such as $b_{11} = a_1^2$, $b_{22} = a_2^2$ and $b_{12} = b_{21} = \psi a_1 a_2$. Given that $cot(2\alpha) = (b_{11} - b_{22})/(2b_{12})$, the parameter ψ can be calculated by $\psi = tan(2\alpha(a_1^2 - a_2^2)/(2a_1 a_2))$. Furthermore, it can be shown that:

$$a_1^2 = \frac{t_\alpha(cos^2\alpha + z) - t_\omega(sin^2\alpha - z)}{(cos^2\alpha + z)^2(sin^2\alpha - z)^2}$$

and

$$a_2^2 = \frac{t_\omega - a_1^2(sin^2\alpha - z)}{(cos^2\alpha + z)}$$

with $t_\alpha = (-ln(0.05)/r_\alpha)^2$, $t_\omega = (-ln(0.05)/r_{\alpha+0.5\pi}^2$ and $z = tan(2\alpha)sin(\alpha)cos(\alpha)$ by using $r_\phi = -ln(0.05)/\sqrt{\mathbf{h}_\phi^T \mathbf{B} \mathbf{h}_\phi}$, for $\phi = \alpha$ and $\phi = \alpha + 0.5\pi$, and the above formula for ψ.

7.7.2 Code

The following code can be implemented in an OpenBUGS environment, a software that is available free of charge by the OpenBUGS Foundation (`http://www.openbugs.info/w/`). This code sets prior distributions on the anisotropy parameters and can be used to assess the effect of anisotropy in any dataset. Of note, the compilation can take several hours or even a few days, depending on the amount of data locations and machine power. We implemented our models in Fortran because computation was faster.

```
model{
# N = number of survey locations
for (i in 1:N) {
  # p = estimated disease prevalence
  positives[i] ~ dbin(p[i],examined[i])
  # v = spatial random effect, w = exchangable random effect
  logit(p[i]) <- q[i] +v[i] +w[i]
  q[i] <- b[1] +b[2]*covariate.1[i] +b[3]*covariate.2[i] +...
}
for (j in 1:n) { b[j]~dnorm(0,0.01) }

# Omega = inverse covariance matrix, H = covariance matrix
v[1:N] ~ dmnorm(mu[],Omega[,])
Omega[1:N,1:N] <- inverse(H[1:N,1:N])
```

```
# calculating the covariance matrix
for (i in 1:N) {
  for (j in 1:N) {
    x[i,j] <- longitude[i]-longitude[j]
    y[i,j] <- latitude[i]-latitude[j]
    d[i,j] <- x[i,j]*x[i,j]*b11 +y[i,j]*y[i,j]*b22+2*x[i,j]*y[i,j]*b12
    H[i,j] <- sigma.v*exp(-sqrt(d[i,j]))
}}

for (i in 1:N) {
  mu[i] <- 0.0
  w[i] ~ dnorm(0,tau.w) }

tau.w ~ dgamma(2.01,1.01)
sigma.w <- 1./tau.w
tau.v ~ dgamma(2.01,1.01)
sigma.v <- 1./tau.v

# anisotropy parameters
alpha ~ dunif(0,1.570796)
range.max ~ dunif(min,max)
range.min ~ dunif(min,range.max)
ratio <- range.max/range.min

# B matrix elements
b11 <- a1*a1
b22 <- a2*a2
b12 <- psi*a1*a2

# auxiliary variables
psi <- (pow(a1,2)-pow(a2,2))*tan(2*alpha)/(2*a1*a2)
a1 <- sqrt((ta*(pow(cos(alpha),2)+z)-to*(pow(sin(alpha),2)-z))/
  (pow(pow(cos(alpha),2)+z,2)-pow(pow(sin(alpha),2)-z,2)))
```

```
a2 <- sqrt((to-pow(a1,2)*(pow(sin(alpha),2)-z))/
  (pow(cos(alpha),2)+z))

ta <- pow(-log(0.05)/range.max,2)
to <- pow(-log(0.05)/range.min,2)
z <- tan(2*alpha)*sin(alpha)*cos(alpha)
}
```

Chapter 8

Modelling the geographical distribution of co-infection risk from single-disease surveys

Schur N.[1,2], Gosoniu L.[1,2], Raso G.[1,2,3], Utzinger J.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Centre Suisse de Recherches Scientifiques, Abidjan, Côte d'Ivoire

This paper has been published in *Statistics in Medicine* 2011, 30(14):1761-1776.

Abstract

Background: The need to deliver interventions targeting multiple diseases in a cost-effective manner calls for integrated disease control efforts. Consequently maps are required that show where the risk of co-infection is particularly high. Co-infection risk is preferably estimated via Bayesian geostatistical multinomial modelling, using data from surveys screening for multiple infections simultaneously. However, only few surveys have collected this type of data.

Methodology: Bayesian geostatistical shared component models (allowing for covariates, disease-specific and shared spatial and non-spatial random effects) are proposed to model the geographical distribution and burden of co-infection risk from single disease surveys. The ability of the models to capture co-infection risk is assessed on simulated datasets based on multinomial distributions assuming light- and heavily-dependent diseases, and a real dataset of *Schistosoma mansoni*-hookworm co-infection in the region of Man, Côte d'Ivoire. The data were restructured as if obtained from single disease surveys. Estimated results of co-infection risk, together with independent and multinomial model results, were compared via different validation techniques.

Principal Findings: The results showed that shared component models result in more accurate estimates of co-infection risk than models assuming independence in settings of heavily-dependent diseases. The shared spatial random effects are similar to the spatial co-infection random effects of the multinomial model for heavily-dependent data.

Conclusion/Significance: In the absence of true co-infection data geostatistical shared component models are able to estimate the spatial patterns and burden of co-infection risk from single disease survey data, especially in settings of heavily-dependent diseases.

8.1 Introduction

Communicable disease control programmes, especially in poorly-resourced settings such as sub-Saharan Africa, require the estimation of the geographical distribution of the underlying disease risk profile in order to identify areas of high burden, so that interventions can be targeted in a cost-effective manner. Risk mapping based on Bayesian geostatistical spatial modelling with remotely-sensed environmental covariates has become the state-of-the-art (Diggle et al., 1998; Raso et al., 2005; Gosoniu et al., 2006; Kazembe et al., 2006; Clements et al., 2008). A host of communicable diseases show spatially overlapping distributions which is called co-endemicity, leading to individuals being simultaneously infected with more than one disease (co-infection/multiple infection) (for definitions, see Glossary). In terms of clinical presentation, multiple infections can go far beyond the combination of the single-disease symptoms. They can lead to a significantly aggravated morbidity and even mortality, as for example in the case of tuberculosis and HIV. However, surprisingly few studies have addressed the issues of co-endemicity (e.g. Brooker et al. (2006) and Raso et al. (2007b)) and co-infection (e.g. Raso et al. (2006b), Kazembe and Namangale (2007) and Brooker and Clements (2009)). *A priori* knowledge of areas with high risk of multiple infections might enhance cost-effectiveness of interventions and disease control programmes if integrated/combined approaches are feasible (Bundy et al., 1991; Brady et al., 2006) and will improve the understanding of the underlying causes.

The reasons for co-infections vary depending on the diseases investigated. While some simultaneous occurring infections simply arise by chance, others are due to shared risk factors (e.g. behavioural, climatic, environmental, ecological, demographic and socio-economic conditions), genetic predispositions or a combination of factors. Besides being already infected with one contagious agent might have an effect on the chances of becoming infected with another one. As a result of the latter, true co-infection risks are likely to be different from co-infection risks due to chance. Indeed, diseases are rarely independent and

Glossary: Definition of selected terms used in this article

Co-endemicity: An area where the investigated diseases are simultaneously endemic

Co-infection: Also known as multiple infection. This applies to individuals harbouring infections simultaneously.

Mono-infection: Individuals harbour only one infection with regard to other investigated endemic infections.

Single infection: When individuals harbour an infection and other infections that might be present are not considered.

estimating co-endemicity while modelling each disease separate (Brooker et al., 2006) fails to account for these relations and might give imprecise estimates of the geographical distribution of risk. Raso et al. (2006b) and Brooker and Clements (2009) have implemented multinomial spatial models for predicting the risk of co-infection with multiple parasitic worms (helminths). Such models depend on observed co-infection data arising from a single survey on the same individuals. However, there is lack of this type of data since most surveys consider single infections. Therefore, we need validated approaches in order to estimate co-infection risk by combining data from multiple surveys screening for single infections.

In case of independent diseases, separate model-based risk predictions based on independent models can be multiplied/overlaid with each other in order to obtain estimates of co-infection risk. Diseases may not be independent because they share common risk factors. To account for between-disease correlations a joint analysis needs to be carried out, which simultaneously models disease risks. Such an approach holds promise to enhance our understanding of the epidemiology of multiple diseases, since it identifies common risk factors and geographical pattern. Shared component models (SCM) have been proposed by Knorr-Held and Best (2001) and others (Tzala and Best, 2008; Kazembe et al., 2009) in order to detect shared and divergent patterns in the risk surface, while separating the random effects into disease-specific and shared (common) components. Common factors are assumed "to be responsible for inter-correlations between the observed variables" (Tzala and Best, 2008). An alternative approach proposed by Zellner (1962) is known as seemingly unrelated regressions (SUR). SUR models consist of a set of multiple regression equations with random effects to capture between-location correlations. The models allow across-regression correlations in case outcomes are correlated. In contrast to SCM, SUR models do not estimate a common random effect across the regression equations which, in the case of co-infection would model an underlying common spatial pattern.

The aim of this paper is to assess the performance of SCM's to estimate the geographical patterns and the amount of co-infection based on disease data obtained from two independent disease surveys each carried out at the same locations and screening for single infections. For this purpose, we analyse simulated as well as real data arising from a single-disease survey screening for *Schistosoma mansoni* and hookworm infections simultaneously among more than 3500 individuals in Côte d'Ivoire. However, we treat these data as if they were obtained from separate single disease surveys in order to compare the results to those from multinomial models (MNM) as well as models assuming disease independence.

The MNM is considered as 'gold standard', since it makes use of the true co-infection data. Model validation was carried out to identify the model with the best predictive ability. Models were fitted on a subset of the observed locations (training locations) and validated by comparing their predictions with data observed at the remaining locations (test locations). Comparison between observed and model based predictions was based on different model validation approaches. The model with the best predictive ability was used for final predictions.

The paper is organized as follows. Section 8.2 describes the datasets underlying our models and contains the methodology of the models used and the validation approaches. In section 8.3, results for the different simulated datasets are given, while section 8.4 details the application with real data. Section 8.5 summarizes and discusses the main findings.

8.2 Data and Methods

8.2.1 Real data

The data which motivated this work have been collected in 2001 and 2002 in the region of Man, western Côte d'Ivoire. In the frame of a cross-sectional epidemiological survey, schoolchildren's infection status with *S. mansoni* and hookworm were assessed. Overall 3578 schoolchildren aged 6-16 years of 56 schools within an area of around 40x60 km were included. Demographic data were available from education registries while information pertaining to socio-economic status were extracted from questionnaires via an asset-based approach (Filmer and Pritchett, 2001). The infection status was assessed by a single stool sample of each child and processed with two diagnostic approaches (single Kato-Katz (Katz et al., 1972) and ether-concentration method (Allen and Ridley, 1970)). More details about the study have been described by Raso et al. (2005).

The geographic coordinates for all schools were obtained by using a hand-held global positioning system. In the Man region, land-cover and altitude have been identified by Raso et al. (2006b) as the key environmental covariates for the joint analysis of the two helminth infections of interest in the current application. The data on land-cover were downloaded from the U.S. Geological Survey (USGS) Earth Resources Observation System (EROS) Data Center at 1x1 km spatial resolution for the period September 2001 to August 2002. Land-cover was categorized into the following 5 categories: woody savannah (used as baseline category), tropical forest, deforested savannah/crops and tropical rainforest. Data on altitude were obtained from an interpolated digital elevation model (DEM) from

the USGS EROS Data Center and split into altitude levels below or above 400m.

The data are based on a single survey screening for multiple infections where co-infection as well as mono-infection risks are measured simultaneously. Out of the total number of schoolchildren screened for *S. mansoni* and hookworm infections, 680 were co-infected (co-infection risk among the schools ranged between 0 and 60%), 862 *S. mansoni* mono-infected, 869 hookworm mono-infected, whereas the remaining 1167 children showed no infection with either parasite. For this kind of data joint disease modeling via MNM is feasible. However, for the purpose of this study we restructured the data as if they arose from single disease surveys screening for *S. mansoni* and hookworm infections carried out at the same locations.

8.2.2 Multinomial model (MNM) formulation

Let $Y_{ijk}^{(m)}$ be the binary infection outcome of individual j at location i ($i = 1, \ldots, n$) and $p_{ijk}^{(m)}$ the corresponding probability of infection for this model. The infection status k indicates either disease 1 mono-infection ($k = 1$), disease 2 mono-infection ($k = 2$), co-infection of the two considered disease ($k = 3$), or neither of these infections ($k = 4$). We assume that $Y_{ijk}^{(m)}$ arises from a multinomial distribution, i.e.

$$(Y_{ij1}^{(m)}, Y_{ij2}^{(m)}, Y_{ij3}^{(m)}, Y_{ij4}^{(m)}) \sim MN(1, p_{ij1}^{(m)}, p_{ij2}^{(m)}, p_{ij3}^{(m)}, p_{ij4}^{(m)}),$$

and we model the influence of covariates $\underline{X}_{ij}$, location-specific random effects $\phi_{il}^{(m)}$ and $\epsilon_{il}^{(m)}$ on the $log(p_{ijl}^{(m)}/p_{ij4}^{(m)})$ with $l = 1, 2, 3$ that is

$$MNM : log\left(p_{ijl}^{(m)}/p_{ij4}^{(m)}\right) = \underline{X}_{ij}^{T}\underline{\beta}_{l}^{(m)} + \epsilon_{il}^{(m)} + \phi_{il}^{(m)},$$

where $\underline{\beta}_{l}^{(m)}$ is the vector of regression coefficients for multinomial category l and $p_{ijl}^{(m)}/p_{ij4}^{(m)}$ is the risk ratio (RR) of the infection status with regard to no infection. We introduce spatial correlation on the $\phi_{il}^{(m)}$'s by assuming that $\underline{\phi}_{l}^{(m)} = (\phi_{1l}^{(m)}, \ldots, \phi_{nl}^{(m)})^{T}$ follow a latent stationary Gaussian process, $\underline{\phi}_{l}^{(m)} \sim MVN(\underline{0}, \Sigma_{l}^{(m)})$, with variance-covariance matrix $\Sigma_{l}^{(m)}$ (Diggle et al., 1998). An isotropic exponential correlation function is used to model geographical correlation as a function of distance, i.e. $\Sigma_{irl}^{(m)} = \sigma_{l}^{2\,(m)} exp(-\rho_{l}^{(m)} d_{ir})$, where d_{ir} is the Euclidean distance between locations i and r, $\sigma_{l}^{2\,(m)}$ is the spatial variability known as partial sill, and $\rho_{l}^{(m)}$ is a smoothing parameter controlling the rate of correlation decay. The geographic dependency (referred to as range) is defined as the minimum distance at which spatial correlation between locations is less than 5% for each multinomial category

and is calculated by $3/\rho_l^{(m)}$. Non-spatial correlation is captured by the $\epsilon_{il}^{(m)}$'s assuming normal distributions with fixed variance $\tau_l^{2\,(m)}$ (known as nugget), as $\epsilon_{il}^{(m)} \sim N(0, \tau_l^{2\,(m)})$.

8.2.3 Independent model (IND) formulation

Suppose Y_{ijk} and p_{ijk} are the single-infection status and probability of infection k ($k = 1, 2$), respectively. In comparison to MNM, we assume that Y_{ijk} arises from a Bernoulli distribution, $Y_{ijk} \sim Be(p_{ijk})$, and we model covariates $\underline{X}_{ij}$ and random effects on the logit scale, as

$$\begin{aligned} IND\,1: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} \\ IND\,2: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \phi_{ik}. \end{aligned}$$

Spatial and non-spatial correlation are introduced as in subsection 8.2.2, however IND 1 only includes non-spatial random effects ϵ_{ik}, $\epsilon_{ik} \sim N(0, \tau_k^2)$, while IND 2 additionally includes disease-specific spatial random errors ϕ_{ik} with partial sill σ_k^2 and decay parameter ρ_k.

8.2.4 Shared component model (SCM) formulation

The basis of the SCM is the independent model. We introduce additional location-specific shared components to the logit scale to incorporate possible dependencies between the diseases at locations i, such as

$$\begin{aligned} A1: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \delta_k E_i \\ A2: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \phi_{ik} + \delta_k E_i \\ B1: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \lambda_k \Phi_i \\ B2: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \phi_{ik} + \lambda_k \Phi_i \\ C1: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \delta_k E_i + \lambda_k \Phi_i \\ C2: &\quad logit\,(p_{ijk}) = \underline{X}_{ij}^T \underline{\beta}_k + \epsilon_{ik} + \phi_{ik} + \delta_k E_i + \lambda_k \Phi_i \end{aligned}$$

where the E_i's and Φ_i's are the common non-spatial and spatial random effects, respectively. Similar to the MNM we suppose that $\underline{\Phi} = (\Phi_1, \ldots, \Phi_n)^T$ follows a latent stationary Gaussian process, i.e. $\underline{\Phi} \sim MVN(\underline{0}, \Sigma)$, with the shared variance-covariance matrix defined as $\Sigma_{ir} = \sigma^2 exp(-\rho d_{ir})$, and with $E_i \sim N(0, \tau^2)$.

The coefficients δ_k and λ_k are referred to as factor loadings and can be interpreted as

disease-specific weighting factors for the shared processes since the diseases might have unobserved common non-spatial and spatial effects but at different levels. Models of type A include shared exchangeable random effects, of type B shared spatial random effects and of type C both forms. Models A1 and A2 (B1 and B2, C1 and C2, respectively) vary in the presence of non-common spatial random effects.

8.2.5 Model fit

Model fit was implemented within a Bayesian framework of inference. Vague normal prior distributions were assigned to the regression coefficients, $\underline{\beta}_l^{(m)}, \underline{\beta}_k \sim N(0, 100)$, inverse gamma distributions to the partial sills, $\sigma_l^{(m)}, \sigma_k \sim IG(1, 100)$, and uniform priors to the spatial decay parameters $\rho_l^{(m)}, \rho_k, \rho$. The factor loadings were considered to have vague normal distributions with mean 0 and variance of 100. In order to remove model non-identifiability and to guarantee unique solutions for SCM's, one factor loading of δ_k and λ_k had to be restricted to be positive to avoid the problem of 'flipping states' (another possible solution with all factor loading changing their signs). Additionally, the common variances τ and σ needed to be constrained to be equal to 1 (Tzala and Best, 2008).

Markov chain Monte Carlo (MCMC) simulations were used to estimate the model parameters. We ran all models with two chain-samplers and a burn-in of 5000 iterations. We assessed convergence by the inspection of ergodic averages of selected model parameters. After convergence, a sample of 500 estimates of each model parameter was used for model validation and to generate smooth risk maps of co-infection risk via Bayesian kriging.

The analyses were performed in WinBUGS v1.4.3 (Imperial College and Medical Research Council, UK) and kriging was carried out in Fortran 95 (Compaq Visual Fortran Professional 6.6.0) using standard numerical libraries (NAG, The Numerical Algorithms Group Ltd.).

8.2.6 Validation methods

In total we fitted 9 models per dataset (one multinomial model, two independent models and six shared component models). For each model and dataset, a sample of 80% of the survey locations was used as training set for model fit, while the remaining 20% of the locations (test locations) were used for validation. The goodness-of-fit of each model was assessed using the deviance information criterion (DIC) (Spiegelhalter et al., 2002). This measure considers the fit of the data and penalizes models that are very complex. Accuracy and bias of the predictions were determined by comparing observed and predicted

co-infection risk, p_i and p_i^* respectively, at the test locations i ($i = 1, \ldots, m$, $m < n$) using different approaches such as mean absolute errors (MAE), χ^2 measures, Kullback-Leibler divergences (KL) and credible intervals (CI) plots. Under the IND's and SCM's p_i^* was calculated as a product of the single-disease risk estimates.

The MAE is a measure of model accuracy calculated by

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |p_i^* - p_i| .$$

A χ^2 measure is an alternative way of comparing model accuracy weighting differences between observed and predicted values with large weights for small observed values, i.e.

$$\chi^2 = \frac{1}{m}\sum_{i=1}^{m} \frac{(p_i^* - p_i)^2}{p_i}.$$

The KL is a weighted measure of model bias giving higher weights to large observed values as

$$KL = \frac{1}{m}\sum_{i=1}^{m} p_i log\left(\frac{p_i}{p_i^*}\right).$$

For the above measures p_i^* was replaced by the posterior median instead of the whole posterior predictive distribution because it gave more coherent model validation measures. The measures are equal to 0 if predictions are perfect.

The outcome of the CI approach is the proportion of test locations being correctly predicted within the qth credible interval (restricted by the lower and upper quantiles $ci^l_{i(q)}$ and $ci^u_{i(q)}$ respectively) of the posterior predictive distribution, i.e.

$$CI_q = \frac{1}{m}\sum_{i=1}^{m} min(I(ci^l_{i(q)} < p_i), I(ci^u_{i(q)} > p_i))).$$

The best model is the one including most test locations within the smallest CI. If two models include same proportions of test locations in the same CI, the model with the smallest width of the interval, defined by $ci^u_{i(q)} - ci^l_{i(q)}$, is considered to be superior.

8.3 Simulation

8.3.1 Data simulation

We simulated six different datasets for modelling co-infection risk in Fortran 95 using standard numerical libraries (NAG, The Numerical Algorithm Group Ltd.). Simulations were done for 200 locations equally distributed within a rectangular grid of [0,0.1]x[0,0.2] distance units. The data have been generated from multinomial distributions with 4 categories $(Y_{ij1}^{(m)}, Y_{ij2}^{(m)}, Y_{ij3}^{(m)}, Y_{ij4}^{(m)}) \sim MN(1, p_{ij1}^{(m)}, p_{ij2}^{(m)}, p_{ij3}^{(m)}, p_{ij4}^{(m)})$ assuming 100 individuals screened at each location $(i = 1, \ldots, 200)$, resulting in a sample size of 20,000 individuals. The total sample size of 20,000 individuals is unnecessary high and was only chosen to reduce sampling variation. The infection risks were estimated from the respective log's assuming only location-specific spatial and non-spatial random effects without covariates and considering non-infected as baseline (as in subsection 8.2.2).

The model parameters used to simulate the datasets (summarized in Table 8.1) vary to obtain six sets with either heavy (datasets 1,3,5) or light (datasets 2,4,6) dependency between the infections. The comparison between the two types of dependency was based on the difference between the simulated co-infection risks and co-infection risks arising by

Table 8.1: Simulation parameters used to create heavy and light dependent datasets for two infections

	"heavy dependent"			"light dependent"		
Parameter	co-inf	mono1	mono2	co-inf	mono1	mono2
	Dataset 1			Dataset 2		
$\beta^{(m)}$	1.0	-0.5	-0.5	0.0	-0.1	-0.2
$\rho^{(m)}$	30	150	300	300	60	60
$\sigma^{2\,(m)}$	0.10	0.01	0.01	0.05	0.01	0.01
$\tau^{2\,(m)}$	0.10	0.01	0.01	0.05	0.01	0.01
	Dataset 3			Dataset 4		
$\beta^{(m)}$	2.0	0.1	0.5	0.1	0.0	0.0
$\rho^{(m)}$	50	400	200	50	400	200
$\sigma^{2\,(m)}$	0.05	0.05	0.05	0.05	0.05	0.05
$\tau^{2\,(m)}$	0.05	0.05	0.05	0.05	0.05	0.05
	Dataset 5			Dataset 6		
$\beta^{(m)}$	2.0	0.1	0.5	0.1	-0.1	0.3
$\rho^{(m)}$	400	50	200	400	50	200
$\sigma^{2\,(m)}$	0.05	0.05	0.05	0.01	0.01	0.01
$\tau^{2\,(m)}$	0.05	0.05	0.05	0.01	0.01	0.01

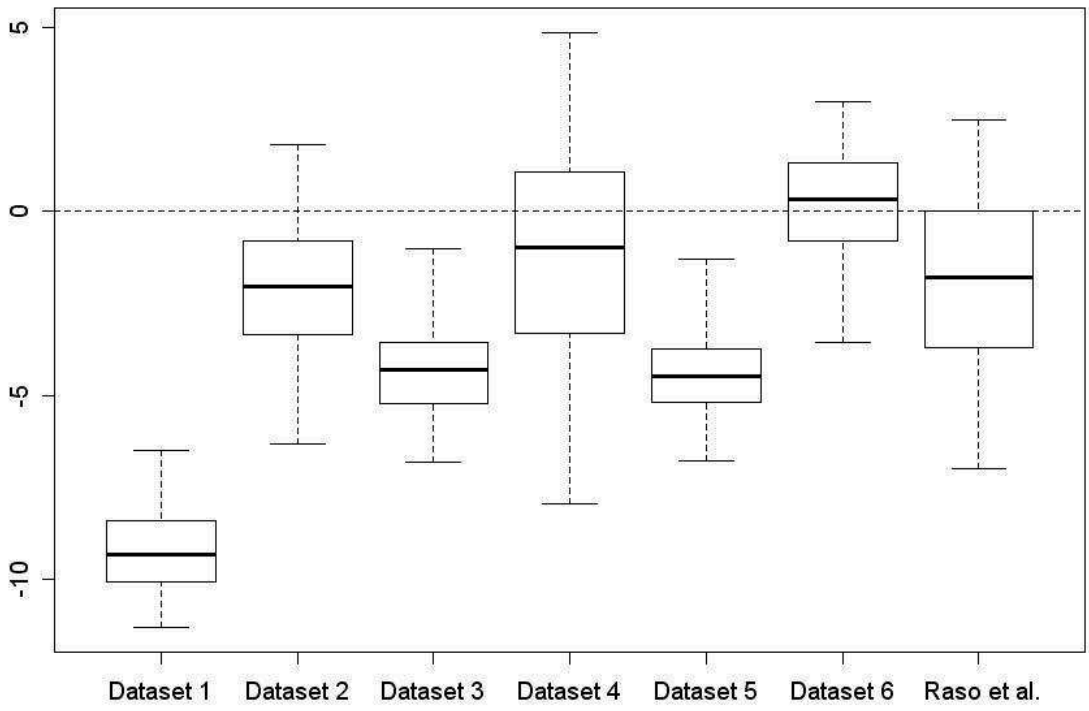

Figure 8.1: Differences in co-infection due to chance and observed co-infection risk for each dataset summarized as boxplots.

chance, that is $p_{i3}^{(m)} - \widetilde{p}_{i3}$ for each location i. The latter risk is computed by the product of the single infection risks (calculated by the sum of mono-infection and co-infection risks) per location, as $\widetilde{p}_{i3} = (p_{i1}^{(m)} + p_{i3}^{(m)}) \times (p_{i2}^{(m)} + p_{i3}^{(m)})$. The differences for each dataset are summarized in the boxplots of Figure 8.1. Dataset 1 shows a median difference in the co-infection risks of around -9.3%, dataset 3 of 4.3% and dataset 5 of 4.5%. The 95% quantiles for these datasets do not include 0 which indicates strong dependence between the infections. In comparison, dataset 2 has a median difference of -2.0%, dataset 4 of 1.0% and dataset 6 of 0.3% (including 0 within the 95% quantiles), therefore the infections are considered to be light dependent.

The inspection of ergodic averages of selected model parameters of the IND's and SCM's for the simulated datasets showed convergence after 30,000 iterations at most. Convergence of MNM's was observed after about 50,000 iterations. After convergence, we collected a sample of 500 from the posterior distribution with a thinning of 50 iterations. Autocorrelation between samples was generally low with exception the constant term of some models.

Table 8.2: Model validation results for heavy dependent (1,3,5), light dependent (2,4,6) and real datasets (7) and significance of factor loadings indicated by + (significant) or - (non-significant)

Model	DIC	MAE	χ^2	KL	δ	λ	DIC	MAE	χ^2	KL	δ	λ
			Dataset 1						Dataset 2			
MNM	3972	7.10	1.39	0.80			4168	4.34	1.09	0.25		
IND 1	2916	12.53	3.36	12.27			2954	4.38	1.07	2.17		
IND 2	2913	11.14	2.53	11.24			2953	4.43	1.06	2.08		
SCM A1	2891	12.39	3.31	11.67	+		2952	4.46	1.10	2.24	-	
SCM A2	2905	12.26	2.58	11.24	+		2942	4.47	1.08	2.14	+	
SCM B1	2842	10.77	2.53	10.95		+	2940	4.51	1.09	2.11		-
SCM B2	2412	10.93	2.58	11.21		-	2939	4.49	1.08	2.10		-
SCM C1	2825	10.79	2.44	10.89	+	+	2944	4.39	1.07	2.13	-	+
SCM C2	2751	11.21	2.56	11.31	+	+	2959	4.42	1.08	2.14	-	+
			Dataset 3						Dataset 4			
MNM	3939	5.20	0.67	0.59			4216	4.85	1.31	1.37		
IND 1	2833	7.71	1.21	5.53			2973	6.21	2.01	2.49		
IND 2	2832	6.46	0.87	5.20			2974	5.10	1.40	1.76		
SCM A1	2828	7.72	1.22	5.43	+		2954	6.26	2.03	2.48	-	
SCM A2	2819	6.54	0.89	5.40	-		2961	5.13	1.45	1.82	-	
SCM B1	2825	6.41	0.89	5.15		+	2959	6.00	1.83	2.33		-
SCM B2	2824	6.44	0.90	5.21		+	2963	5.08	1.45	1.80		-
SCM C1	2642	6.33	0.85	5.06	-	+	2928	5.86	1.73	2.21	-	-
SCM C2	2723	6.31	0.85	5.11	-	+	2947	5.08	1.40	1.94	-	-
			Dataset 5						Dataset 6			
MNM	3943	5.57	0.75	0.33			4175	2.74	0.45	0.17		
IND 1	2837	6.89	0.98	4.98			2941	2.59	0.44	-0.35		
IND 2	2837	6.77	0.97	4.96			2952	2.74	0.48	-0.40		
SCM A1	2830	6.79	0.99	4.83	+		2938	2.56	0.44	-0.35	-	
SCM A2	2828	6.90	1.00	5.10	-		2952	2.82	0.51	-0.37	-	
SCM B1	2801	6.71	0.98	4.79		+	2937	2.55	0.44	-0.34		-
SCM B2	2818	6.88	0.99	4.90		+	2952	2.80	0.50	-0.40		-
SCM C1	2746	6.80	0.96	4.78	-	+	2935	2.54	0.44	-0.38	-	-
SCM C2	2575	6.60	0.93	4.70	+	-	2951	2.71	0.48	-0.30	-	-
			Raso et al. dataset									
MNM	743	10.33	7.41	5.64								
IND 1	550	11.01	6.61	11.87								
IND 2	552	14.22	12.46	7.73								
SCM A1	540	11.57	7.40	12.65	-							
SCM A2	544	12.44	8.75	8.62	-							
SCM B1	549	13.85	10.91	9.59		-						
SCM B2	543	14.12	11.83	8.89		-						
SCM C1	528	13.91	11.71	8.61	-	-						
SCM C2	523	13.28	10.39	9.12	-	-						

8.3.2 Model comparison and validation

Heavy disease dependency: The MNM's of all heavily-dependent datasets (datasets 1,3,5) had the best predictive ability based on all validation methods except for DIC (Table 8.2 and Figure 8.2, left handed plots) while models including shared components typically outperformed the independent models. However, independent surveys screening for single infections do not generate multinomial data and therefore multinomial modelling is not applicable.

For dataset 1, the introduction of shared spatial random effects (SCM's B and C) enhanced the predictive ability of the models, while the results with an additional shared exchangeable random effect (SCM's C) did not much improve the predictive ability. Among SCM's B and C, the models without disease-specific spatial random effects (SCM B1 and C1) showed better validation results but were nearly indistinguishable from each other. However, SCM B1 is the preferred model because it contains fewer parameters. The DIC goodness of fit measure suggested that SCM B2 fitted better than the other models followed by SCM C2. Since we were mainly interested in the predictive ability of a model rather than its goodness of fit, we presented the results on DIC to indicate that the best fitting models do not necessarily predict well. For dataset 3, the validation methods showed that the best models included spatial and non-spatial shared components. However, SCM C1 model contains the fewest parameters and therefore was chosen as the best predictive model. Results of the simulated dataset 5 suggest that the SCM C2 was the model with the best predictive ability.

The MNM was able to correctly estimate the parameter values used to simulate all datesets (Table 8.1, left). The parameters estimated from models assuming single disease surveys are not comparable to those of the MNM's, since the latter fit multiple survey data based on mono-infections and co-infections instead of single infections. However, we found that the single disease estimates of β's and ρ's are closer to the ones of the co-infection category of MNM which might underline the importance of co-infection for these specific datasets.

The factor loadings λ and δ show positive contributions to the shared random effects. Nearly all factor loadings are significantly different from zero indicating the need to include shared components in the joint analysis of infection risk under heavy disease dependency. The only non-significant factor loadings were δ_2 of dataset 3 based on SCM C1 and λ_2 of dataset 5 based on SCM C2. Dataset 5 was simulated to have very weak spatial correlation in the co-infection category while dataset 1 and 3 were simulated to have strong spatial

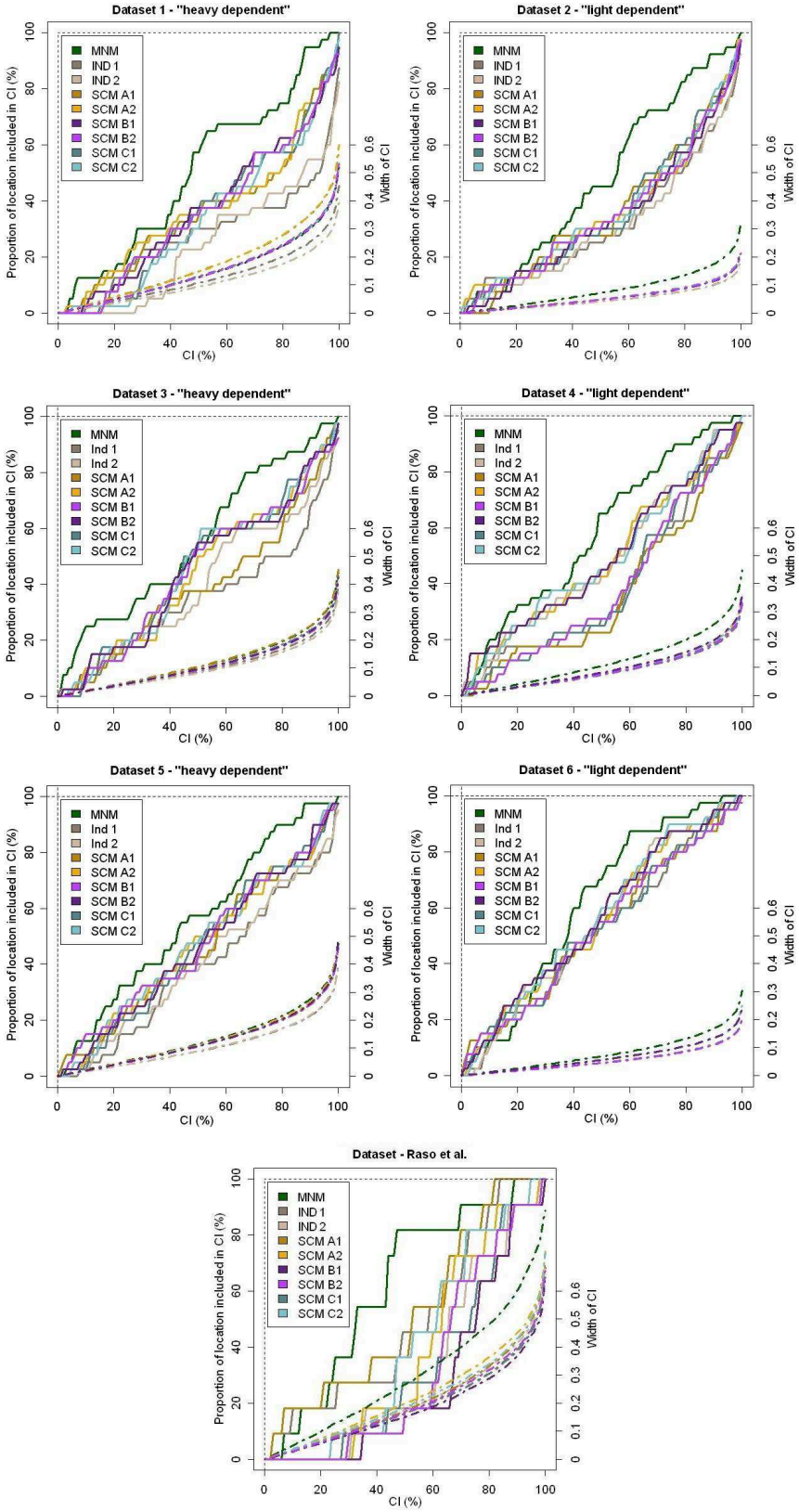

Figure 8.2: Validation results of the CI approach.

correlation. Therefore, shared spatial components for dataset 5 are of less importance than for the other datasets, while shared non-spatial components are important. This seems to be reflected in the estimates for the factor loadings.

Light disease dependency: Table 8.2 and Figure 8.2 (right handed plots) indicate that the MNM is the model with best predictive ability among all validation methods although it does not fit the data best as shown by the DIC measure. The MNM's did not fit the light disease dependency dataset as well as the heavy dependent datasets (compare Table 8.3 and Table 8.1), specially dataset 4 and 6.

The remaining models are less distinguishable than in the case of heavily-dependent data because the differences among all model validation methods are marginal. It follows that the simplest models containing the fewest parameters (IND 1 or 2) are the best models if multinomial modelling is not possible. IND 1 model predicts best for datasets 2 and 6, while IND 2 model is considered the best predictive model for dataset 4. The factor loadings of the corresponding shared component models are nearly always non-significant.

8.3.3 Model prediction comparison

The MNM and the best predicting model out of the remaining ones (SCM B1, IND, SCM C1, IND 2, SCM C2 and IND1 for datasets 1-6 respectively) were used to produce co-infection risk maps via Bayesian kriging (Figure 8.3). For the heavily dependent datasets, the risk maps showed that the pattern of co-infection risk is quite similar. However, the amount of co-infection risk is underestimated by the SCM's if we consider the MNM's as 'gold standard'. Due to the absence of covariates and any spatial effects, the risk surface maps for the IND's show virtually no variation. The corresponding maps of the spatial random effects $\underline{\phi}_3^{(m)}$ of MNM and $\underline{\Phi}$ estimated from SCM's (Figure 8.4) highlight the similarity between the two sets of random effects while modeling co-infection risk.

Table 8.3: Model parameter estimates with 95% CI's based on SCM B1 for dataset 1/3, IND 1 for dataset 2/6/7, IND 2 for dataset 4, SCM C2 for dataset 5 (shared parameters centered between single infection estimates of column 6 and 7).

Data-set	Para-meter	MNM co-infection	MNM mono-infection 1*	MNM mono-infection 2†	Best model single infection 1*	Best model single infection 2†
1	β_0	1.11 (0.89, 1.34)	-0.51 (-0.55, -0.48)	-0.52 (-0.56, -0.49)	0.87 (0.74, 0.98)	0.87 (0.74, 0.98)
	ρ	38.9 (22.7, 73.3)	285.2 (130.3, 443.0)	327.2 (144.7, 444.8)	106.4 (69.8, 156.3)	
	σ^2	0.10 (0.04, 0.21)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)	-	-
	τ^2	0.10 (0.06, 0.15)	0.01 (0.00, 0.02)	0.02 (0.00, 0.03)	0.00 (0.00, 0.01)	0.01 (0.00, 0.01)
	λ	-	-	-	0.39 (0.34, 0.44)	0.37 (0.33, 0.43)
2	β_0	0.02 (-0.03, 0.07)	-0.13 (-0.16, -0.09)	-0.20 (-0.24, -0.17)	0.05 (0.02, 0.08)	-0.01 (-0.04, 0.01)
	ρ	263.6 (114.0, 442.1)	174.4 (61.3, 417.3)	221.3 (85.3, 431.6)	-	-
	σ^2	0.03 (0.01, 0.07)	0.01 (0.00, 0.02)	0.01 (0.00, 0.03)	-	-
	τ^2	0.03 (0.01, 0.07)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)	0.03 (0.03, 0.04)	0.03 (0.02, 0.03)
3	β_0	2.04 (1.86, 2.18)	0.06 (-0.02, 0.12)	0.48 (0.42, 0.54)	1.20 (1.07, 1.33)	1.51 (1.41, 1.60)
	ρ	72.7 (25.5, 165.8)	274.9 (63.0, 442.6)	291.0 (43.2, 441.5)	75.4 (26.7, 163.7)	
	σ^2	0.07 (0.03, 0.14)	0.05 (0.01, 0.13)	0.04 (0.01, 0.12)	-	-
	τ^2	0.04 (0.01, 0.08)	0.07 (0.01, 0.12)	0.08 (0.01, 0.13)	0.04 (0.01, 0.09)	0.05 (0.01, 0.08)
	λ	-	-	-	0.30 (0.20, 0.41)	0.19 (0.10, 0.27)
	δ	-	-	-	0.10 (0.01, 0.30)	0.08 (-0.12, 0.27)
4	β_0	0.00 (-0.13, 0.09)	-0.03 (-0.14, 0.05)	-0.01 (-0.23, 0.10)	0.11 (-0.26, 0.45)	0.13 (-0.10, 0.32)
	ρ	25.9 (21.7, 49.8)	82.4 (22.4, 418.4)	35.8 (22.1, 402.9)	25.4 (21.6, 59.3)	28.9 (21.7, 370.7)
	σ^2	0.11 (0.07, 0.19)	0.09 (0.06, 0.16)	0.10 (0.06, 0.17)	0.11 (0.07, 0.18)	0.09 (0.05, 0.14)
	τ^2	0.08 (0.06, 0.11)	0.09 (0.06, 0.12)	0.10 (0.07, 0.13)	0.07 (0.05, 0.10)	0.07 (0.05, 0.10)
5	β_0	2.02 (1.96, 2.08)	0.03 (-0.09, 0.14)	0.49 (0.41, 0.55)	1.18 (1.10, 1.27)	1.51 (1.45, 1.56)
	ρ	290.2 (96.1, 443.3)	92.7 (26.8, 216.1)	290.6 (54.3, 440.8)	116.7 (29.2, 425.4)	261.7 (41.9, 437.4)
					269.0 (49.8, 443.9)	
	σ^2	0.05 (0.01, 0.12)	0.07 (0.02, 0.13)	0.04 (0.00, 0.13)	0.03 (0.00, 0.06)	0.02 (0.00, 0.06)
	τ^2	0.06 (0.01, 0.12)	0.04 (0.01, 0.09)	0.09 (0.01, 0.14)	0.03 (0.00, 0.08)	0.03 (0.00, 0.06)
	λ	-	-	-	0.17 (0.01, 0.32)	0.08 (-0.13, 0.28)

Continued on next page

Data-set	Para-meter	MNM co-infection	MNM mono-infection 1*	MNM mono-infection 2†	Best model single infection 1*	Best model single infection 2†
	δ	-	-	-	0.24 (0.09, 0.35)	0.20 (0.05, 0.30)
6	β_0	0.14 (0.00, 0.23)	-0.08 (-0.18, 0.00)	0.36 (0.29, 0.48)	-0.16 (-0.19, -0.13)	0.28 (0.25, 0.31)
	ρ	23.5 (21.5, 33.8)	23.3 (21.6, 32.1)	23.1 (22.6, 32.7)	-	-
	σ^2	0.06 (0.04, 0.09)	0.06 (0.04, 0.10)	0.06 (0.04, 0.09)	-	-
	τ^2	0.04 (0.03, 0.05)	0.04 (0.03, 0.05)	0.04 (0.03, 0.05)	0.04 (0.03, 0.05)	0.03 (0.02, 0.04)
Raso	β_0	-0.04 (-1.29, 1.66)	0.14 (-1.12, 3.61)	-0.98 (-1.49, -0.46)	0.41 (-0.35, 1.07)	-0.63 (-0.97, -0.28)
	β_1	-0.48 (-1.85, 1.34)	-0.88 (-2.45, 1.04)	0.46 (-0.09, 0.98)	-1.93 (-2.74, -1.17)	0.34 (-0.09, 0.72)
	β_2	0.61 (-0.18, 1.43)	0.57 (-0.24, 1.43)	0.65 (0.05, 1.29)	0.43 (-0.54, 1.33)	0.55 (0.10, 1.03)
	β_3	-0.44 (-1.41, 0.67)	-0.61 (-1.74, 0.55)	0.28 (-0.49, 1.11)	-0.75 (-1.88, 0.51)	0.24 (-0.37, 0.87)
	β_4	0.22 (-0.89, 1.30)	0.32 (-1.06, 1.80)	0.35 (-0.65, 1.34)	0.36 (-0.96, 1.97)	0.25 (-0.48, 0.98)
	ρ	10.1 (1.01, 248.)	17.8 (2.83, 245.)	1.28 (0.53, 10.0)	-	-
	σ^2	1.57 (0.10, 11.0)	1.89 (0.05, 14.8)	0.26 (0.04, 0.70)	-	-
	τ^2	0.43 (0.06, 1.47)	0.47 (0.07, 1.81)	0.30 (0.05, 0.81)	1.82 (1.14, 2.99)	0.39 (0.25, 0.66)

β_0 constant, β_1 altitude levels $\geq 400m$, β_2 tropical forest, β_3 deforested savannah/crops, β_4 tropical rainforest

*: Corresponds to *S.mansoni* infection in dataset 7.

†: Corresponds to hookworm infection in dataset 7.

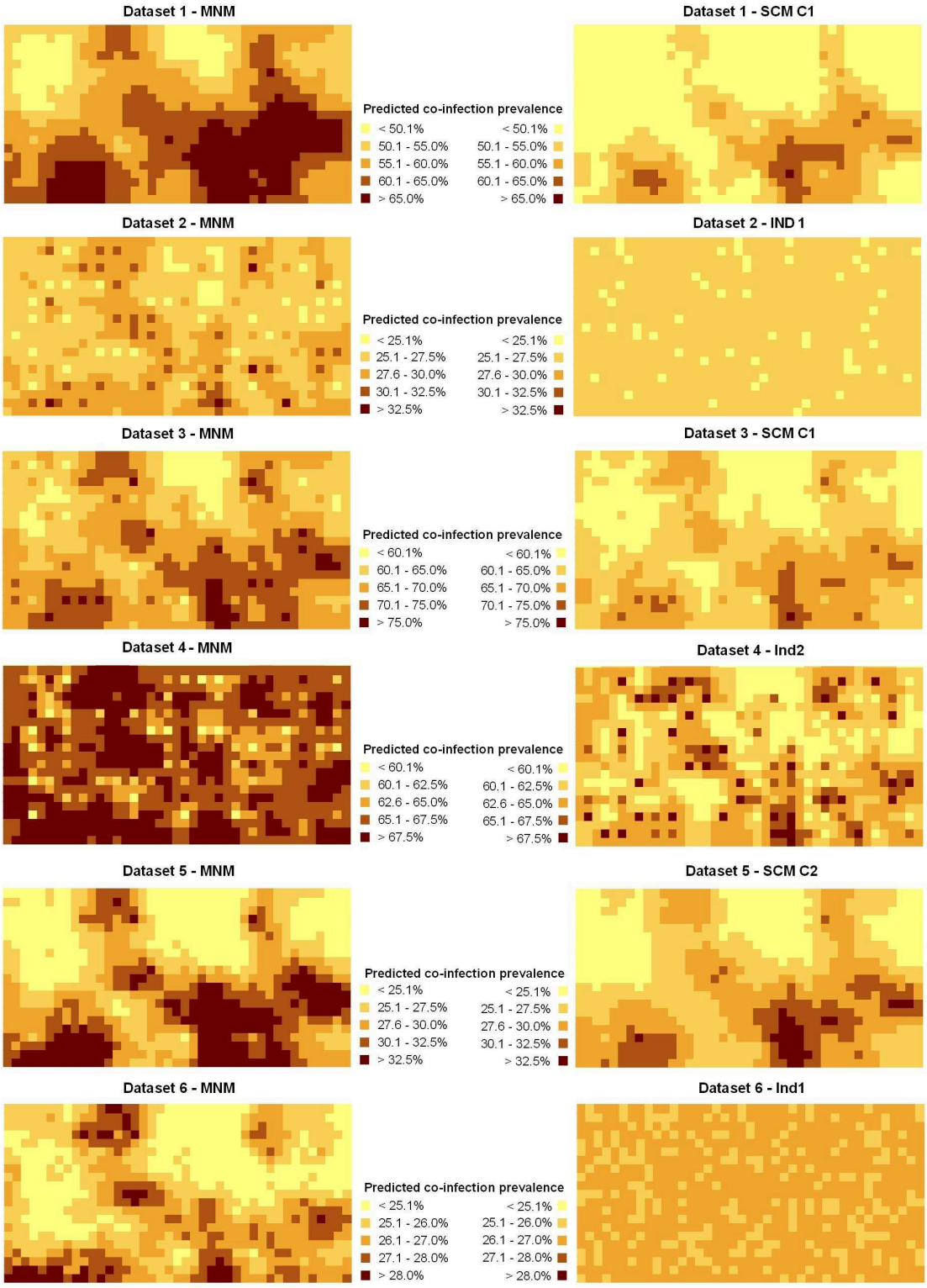

Figure 8.3: Co-infection risk surface for MNM and best fitting model for simulated datasets.

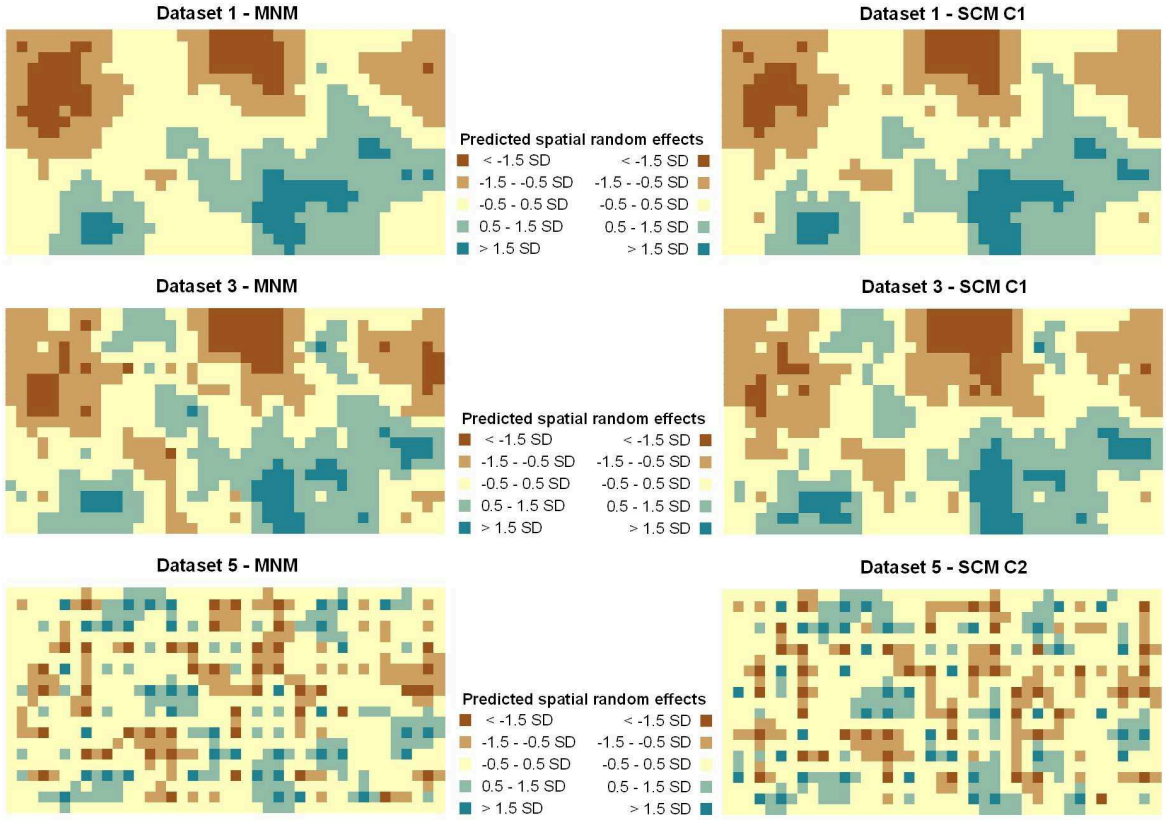

Figure 8.4: Spatial random effect surface for MNM and heavy-dependent simulated datasets.

8.4 Application

Convergence of the applied data was achieved at about 50,000 iterations for the IND's and SCM's and at 80,000 for the MNM. After convergence occurred, every 50th iterations of each chain was stored to collect an uncorrelated sample of 500 iterations from the posterior distributions. Model validation results based on these samples showed that the MNM is the best predicting model based on MAE, KL and CI plot, while it is outperformed by IND 1 for χ^2 (see Table 8.2 and Figure 8.2, lower figure). Irrespective of MNM, IND 1 shows an overall good performance not only in the χ^2 method but also for MAE and CI plot. Therefore, we consider IND 1 as the overall best model to predict co-infection risk from surveys screening for single infections for this dataset. Additionally, this model has the fewest parameters. The resulting co-infection risk plots of the MNM and IND 1 are presented in Figure 8.5 showing similar patters of risk but slightly underestimated by IND 1 compared to MNM.

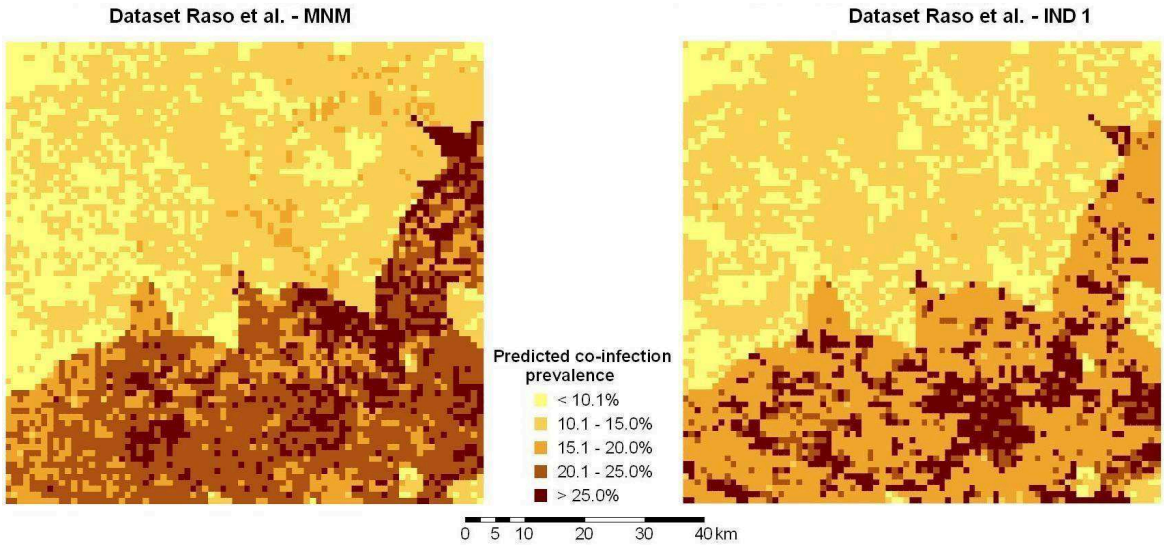

Figure 8.5: Co-infection risk surface for MNM and best fitting model for applied dataset.

Parameter estimates for MNM and IND 1 are given in Table 8.3 (dataset 7). The spatial ranges are estimated to be 0.30 km (0.01 km, 2.97 km) for co-infections, 0.17 km (0.01 km, 1.06 km) for mono-infections with *S.mansoni* and 2.34 km (0.30 km, 5.66 km) for mono-infections with hookworm. These estimates are close to the minimum distance between locations (0.81 km) and the spatial resolution of the prediction map (1 km) suggesting weak spatial correlation. These findings go hand in hand with the previous analysis by Raso et al. (2006b) who used a similar MNM model to estimate co-infection risks in the region of Man. The weak spatial correlation can be explained by the the weak spatial patterns in hookworm (Raso et al., 2006a) and *S.mansoni* (Raso et al., 2005) single infections (spatial range of about 1.8 km and 7.5 km, respectively). In addition, these results explain the good performance of the independent non-spatial model in model validation. IND 1 estimates are usually closer to the corresponding estimates of mono-infection than co-infection. The factor loadings δ_2 and λ_2 of SCM C1 for *S.mansoni* and hookworm single infections are including zero within their 95% CI of which further supports the assumption of independence between the two diseases.

8.5 Conclusion

In this study, we assessed the performance of SCM's to estimate the geographical distribution of co-infection risk from independent surveys screening individuals for single infections. Our results suggest that the MNM's always perform best and can be considered as 'gold

standard'. However, in real situations we rarely have surveys screening simultaneously for multiple infections. Therefore, MNM's are not applicable. Studies often ignore the presence of disease dependence and treat data as truely independent. Our simulation studies suggest that this approach performs well for light-dependent diseases. However, for heavily-dependent ones, SCM's are able to capture disease correlation and give rise to models with better predective ability than independent models. This is important when we are interested in identifying geographical patterns of co-infection risk. In real applications, we are not able to assess whether diseases are heavily- or light-dependent. Therefore, we recommend to fit the SCM as well as IND models on a subset of the data (e.g. 80%) and to assess the predictive ability of the models on the remaining test locations. If there are virtually no differences between the models, the diseases are likely to be independent from each other in the given setting and therefore the independent model is appropriate. Based on our simulations results we also expect that for heavily-dependent data factor loadings tend to be significant while for light-dependent data they are indistinguishable from 0.

The different model validation methods were coherent in the choice of the most appropriate model in case the models were distinguishable. However, the DIC measures frequently disagreed because DIC is a goodness of fit measure and can not be used to compare the predictive ability of the models.

The geographical pattern of the spatial co-infection random effect of the MNM's was similar to those obtained from the shared spatial random effect of SCM's for heavily-dependent diseases. Therefore the latent factor of the shared model can be interpreted as error introduced by co-infection. Concluding, SCM's with shared spatial random effects perform better than independent models in heavily disease-dependent settings. On the other hand, IND's and SCM's are not distinguishable for slightly disease dependent data.

In this study, we assumed that the diseases have all survey locations in common. This assumption might not be true in real applications where locations are not aligned across surveys. In such a case multinomial modelling is not possible but the SCM approach based on single infections can be readily adjusted to this problem in the Bayesian framework. The unknown prevalence has to be predicted and the introduced uncertainty can be incorporated in an error term which will finally result in higher uncertainty of co-infection risk. This work is currently in progress. Furthermore, we restricted our work for this paper to only two diseases due to simplicity. As presented by Held et al. (2005), considering three or even more diseases in shared component models is feasible and straight forward.

Acknowledgements

JU (project no. PPOOB-102883 and PPOOB-119129) and PV (project no. 325200-118379) are grateful to the Swiss National Science Foundation. GR acknowledges financial support from the University of Queensland.

8.6 Appendix

In this section, we provide an example BUGS code for the most complex model used in this manuscript (SCM C2). Implementation of the other models is straight forward by removing the unnecessary parameters.

```
model{
for (i in 1:N){
  # N is the number of survey locations
  # Z1/2 the number of positives for disease 1/2
  # x1/2 the estimated proportion of positives for disease 1/2
  Z1[i] ~ dbin(x1[i],100)
  Z2[i] ~ dbin(x2[i],100)
  logit(x1[i]) <- b[1,1] +v.a[1]*v.s[i] +v.n[1,i] +w.a[1]*w.s[i] +w.n[1,i]
  logit(x2[i]) <- b[1,2] +v.a[2]*v.s[i] +v.n[2,i] +w.a[2]*w.s[i] +w.n[2,i]
}

for (j in 1:2){ b[1,j] ~ dnorm(0,0.01) }

for (i in 1:N){
  # w.n is the non-spatial non-shared random error
mu[i] <- 0.0
w.n[1,i] ~ dnorm(0,tau.wn[1])
w.n[2,i] ~ dnorm(0,tau.wn[2])
  # w.s is the non-spatial shared random error
w.s[i] ~ dnorm(0,1.0)}

# v.n is the spatial non-shared random error
v.n[1,1:N] ~ spatial.exp(mu[], x[], y[], tau.vn[1], rho.vn[1],1)
v.n[2,1:N] ~ spatial.exp(mu[], x[], y[], tau.vn[2], rho.vn[2],1)
# v.s is the spatial shared random error
```

```
v.s[1:N] ~ spatial.exp(mu[], x[], y[], 1.0, rho.vs,1)

# rho is the spatial decay parameter
rho.vs ~ dunif(21.5,450)
rho.vn[1] ~ dunif(21.5,450)
rho.vn[2] ~ dunif(21.5,450)

# tau is the spatial precision and sigma the spatial variance
tau.vn[1] ~ dgamma(2.01,1.01)
tau.vn[2] ~ dgamma(2.01,1.01)
tau.wn[1] ~ dgamma(2.01,1.01)
tau.wn[2] ~ dgamma(2.01,1.01)
sigma.vn[1] <- 1/tau.vn[1]
sigma.vn[2] <- 1/tau.vn[2]
sigma.wn[1] <- 1/tau.wn[1]
sigma.wn[2] <- 1/tau.wn[2]

# w.a is the non-spatial factor loading
w.a[1] ~ dnorm(0,0.01)I(0,)
w.a[2] ~ dnorm(0,0.01)
# v.a is the spatial factor loading
v.a[1] ~ dnorm(0,0.01)I(0,)
v.a[2] ~ dnorm(0,0.01)
}
```

Chapter 9

Discussion

This PhD thesis contributes to the field of schistosomiasis epidemiology with (i) Bayesian isotropic and anisotropic geostatistical models for high spatial resolution schistosomiasis risk mapping and prediction based on age-heterogeneous historical survey data collected over very large number of locations; (ii) statistical methodology for assessing the geographical distribution of co-infection from independent single-disease surveys when diseases are correlated; and (iii) spatially explicit estimation of schistosomiasis risk and number of infected people in 29 countries across West and eastern Africa. Hence, for first time, empirical model-based evidence of schistosomiasis risk and burden in those regions is provided. These estimates are of considerable importance for schistosomiasis control programmes, as they indicate high-risk areas requiring interventions, allow calculations of the number of praziquantel tablets required based on WHO guidelines at the appropriate administrative level, and provide baseline maps to assess effectiveness of interventions on the roadmap towards schistosomiasis elimination.

The methodology and results of our research are described in six manuscripts included as chapters in this thesis. Three manuscripts are published, in PLoS Neglected Tropical Diseases (Schur et al., 2011b), in Parasites & Vectors (Schur et al., 2011d) and in Statistics in Medicine (Schur et al., 2011a), one manuscript is in press, in Acta Tropica (Schur et al., 2011c), and one manuscript has been submitted to Statistics in Medicine. In each chapter, a detailed discussion on the findings is provided. In the following sections, a summary of the main outcomes and important limitations of the respective analyses is presented, which will allow to put forward a set of recommendations for future research and as well as some implications to schistosomiasis control.

9.1 Significance of work and implications for control interventions

After a long period of general neglect, there is growing interest in the control of schistosomiasis and other neglected tropical diseases (Hotez et al., 2007; Fenwick et al., 2009; Utzinger et al., 2009). However, despite successful control efforts in different parts of the world, schistosomiasis remains highly prevalent, particularly in sub-Saharan Africa (Steinmann et al., 2006; Utzinger et al., 2009). Common control strategies are based on large-scale administration of anthelminthic drugs, delivered through the public school system, improved sanitation and hygiene, and vector control campaigns (Davis, 2009). Financial resource allocation and implementation of such control strategies should be driven by evidence-based information on the geographical distribution and disease burden in order to meet the needs

of the local populations and to optimally target control interventions. Currently, decisions are only based on rough estimates due to a lack of accurate and up-to-date disease risk estimates.

High resolution spatially explicit model-based risk maps are essential tools for all phases of control and monitoring activities - starting from the planning, over to the implementation and coordination phase, and even evaluation of interventions - especially in resource-constrained settings. In the planning phase, the maps can guide control interventions in a cost-effective manner by highlighting areas of high risk and areas with imprecise risk estimates that need additional surveys. The maps may also guide the efficient allocation of sparse resources avoiding stock-out problems for equipment and drugs by providing estimates on the number of infected people and treatment needs. Additionally, subsequent mapping efforts could monitor the effect of control activities and assess their impact. Furthermore, maps on the spatial distribution of different diseases in the same region could identify areas of high co-endemicity guiding integrated intervention approaches to improve cost-effectiveness.

In this thesis, the first model-based *S. haematobium* and *S. mansoni* prevalence maps at high spatial resolution are presented for the eastern and western African region. Bayesian geostatistical models were employed, approximating the spatial process from a subset of locations to overcome modelling of geostatistical data collected over large numbers of locations (Banerjee et al., 2008; Gosoniu et al., 2011a; Rumisha et al., 2011). The survey data were obtained from the readily available open-access GNTD database (Hürlimann et al., 2011). Different sets of climatic and other environmental data were implemented in the models to evaluate the effect on schistosomiasis prevalence and to predict the outcome at high spatial resolution in order to detect potential hotspots of schistosomiasis transmission. Small transmission hotspots could arise due to the focality of the schistosomiasis distribution, which is an important epidemiological feature of the disease (Lengeler et al., 2002). In separate analyses, the effect of anisotropy was assessed, shared component geostatistical models were developed, assessing co-endemicity from single disease survey data, and finally age-alignment factors were estimated to obtain age-adjusted disease risk maps. Furthermore, several model validation procedures were used to assess accuracy of the model-based predictions and to compare different models. Gridded population count data for the year 2008 were obtained from LandScan, projected to 2010 by applying averaged national growth rates for the period from 2006 to 2010 and adjusted age groups of interest. Finally, these population count data were combined with the schistosomiasis risk

estimates underlying the prediction maps and to obtain country-specific and population-adjusted prevalence estimates and number of infected individuals for an ensemble of 29 eastern and West African countries.

Prior to our work, existing and widely cited statistics on the number of individuals infected with schistosomiasis as published by Chitsulo et al. (2000), Steinmann et al. (2006) and Utzinger et al. (2009) were mainly based on interpolated disease risk estimates published by Utroska et al. (1989). These estimates are unreliable because they are lacking empirical modelling to account for the disease-environment relation. In addition, environmental transformations, population movement, mass drug administration and other control interventions are likely to have modified the distribution of schistosomiasis (Steinmann et al., 2006; Fenwick, 2006; Fenwick et al., 2009; Utzinger et al., 2009; WHO, 2010; WHO and UNICEF, 2010) outdating estimates based on the ones published by Utroska et al. (1989).

Table 1 compares the burden estimates obtained from our work and the ones from Chitsulo et al. (2000) over all countries in continental Africa (detailed results are provided in Chapters 4 and 6). In total, Chitsulo and colleagues estimated 174 million *Schistosoma spp.* infections in continental Africa, with 129 million alone in West and eastern Africa. Our estimates for West (50.8 million) and eastern (121.8 million) Africa sum up to a total of almost 173 million *Schistosoma* infections, which is 34% higher than the Chitsulo estimates for those regions. However, our eastern Africa estimates are based on age-aligned models and thus correspond to the entire population, while our West Africa estimates refer to individuals $\leq$20 years. Model validation had shown that regional alignment factors resulted in more accurate predictions than country-specific ones. Employing the same alignment factors for West Africa, we obtain approximately 140 million infected individuals of the entire population in the region. This leads to a total of about 262 million infections in both, West and eastern Africa, which would exceed the Chitsulo estimate by more than 100%.

The Chitsulo et al. estimates for the remaining countries in South, Central and North Africa sum up to 45 million additional infections. This estimate does not take into account the large population changes in some highly endemic countries, such as Chad, Democratic Republic of the Congo, Egypt and South Africa (United Nations, 2007). Based on our West and East Africa analysis, we might assume that the number of infected in the remaining African countries could be as high as 90 million (100% more than the Chitsulo estimate). Therefore, the total number of infected with schistosomiasis in continental

Africa might even be higher than 350 million. However, this estimate does not take into account the progress made with national control programmes, for example in Egypt (Bank, 2008). However, the above estimates consider neither the low sensitivity of the diagnostics methods, especially in low infection intensity areas and if repeated samples are not taken (de Vlas and Gryseels, 1992; Marti and Koella, 1993; Engels et al., 1996; Yu et al., 1998; Utzinger et al., 2001), nor the additional small burden due to S. intercalatum, which is however declining (Tchuem Tchuenté et al., 2003).

Table 9.1: Country-specific estimates of the total population and number of infected individuals with schistosomiasis for continental Africa in 1995 and 2010. Disease risk estimates for 2010 are based on the median of model-based predictions adjusted for the total population, while 1995 estimates are survey interpolations presented by Chitsulo et al. (2000).

Country	Total population		*Schistosoma spp.* infected		*S. haematobium*-infected	*S. mansoni*-infected
	1995	2010	1995	2010	2010	2010
Algeria	28.0	33.931	2.10	-	-	-
Angola	10.8	13.226	4.80	-	-	-
Benin	5.5	8.030	1.95	2.124 [a]	1.792 [a]	0.940 [a]
Botswana	1.5	1.889	0.15	-	-	-
Burkina Faso	10.4	16.100	6.24	4.738 [a]	4.282 [a]	1.446 [a]
Burundi	6.3	9.445	0.84	3.806	2.820	1.908
Cameroon	13.3	19.300	3.02	2.668 [a]	2.099 [a]	0.952 [a]
Central African Republic	3.3	4.623	0.33	-	-	-
Chad	6.4	10.568	2.78	-	-	-
Congo	2.6	4.576	0.89	-	-	-
Congo, DRC	49.0	70.691	13.84	-	-	-
Cte d'Ivoire	14.0	19.200	5.6	4.286 [a]	3.229 [a]	2.262 [a]
Djibouti	-	0.512	-	0.147	0.061	0.107
Egypt	57.8	82.376	10.06	-	-	-
Equatorial Guinea	0.4	0.647	0.008	-	-	-
Eritrea	3.6	5.477	0.26	2.329	1.218	1.710
Ethiopia	56.4	89.500	4.00	29.095	16.157	19.746
Gabon	1.1	1.378	0.50	-	-	-
Gambia	1.1	1.770	0.33	0.173 [a]	0.168 [a]	0.005 [a]
Ghana	17.1	22.100	12.40	5.912 [a]	5.077 [a]	2.659 [a]
Guinea	6.6	8.885	1.70	2.259 [a]	1.824 [a]	0.999 [a]
Guinea-Bissau	1.1	1.545	0.33	0.218 [a]	0.203 [a]	0.024 [a]
Kenya	26.7	40.300	6.14	16.693	6.209	13.833
Lesotho	-	2.126	-	-	-	-
Liberia	2.7	2.900	0.648	0.924 [a]	0.658 [a]	0.588 [a]

Continued on next page

Country	Total population		*Schistosoma spp.* infected		*S. haematobium*-infected	*S. mansoni*-infected
	1995	2010	1995	2010	2010	2010
Libya	5.4	6.126	0.27	-	-	-
Malawi	9.8	14.400	4.20	6.883	5.047	3.764
Mali	9.8	13.080	5.88	2.291 [a]	1.997 [a]	0.845 [a]
Mauritania	2.3	3.503	0.63	0.333 [a]	0.299 [a]	0.055 [a]
Morocco	26.6	34.396	0.06	-	-	-
Mozambique	16.2	20.200	11.3	11.224	8.263	6.960
Namibia	1.5	2.078	0.009	-	-	-
Niger	9.0	14.242	2.40	1.397 [a]	1.321 [a]	0.179 [a]
Nigeria	111.3	152.566	25.83	18.754 [a]	15.741 [a]	9.257 [a]
Rwanda	6.4	10.700	0.38	3.93	3.113	1.639
Senegal	8.5	11.200	1.30	1.464 [a]	1.338 [a]	0.183 [a]
Sierra Leone	4.2	6.455	2.50	1.999 [a]	1.792 [a]	0.853 [a]
Somalia	9.5	9.150	1.71	3.890	2.230	2.851
South Africa	41.5	48.766	4.50	-	-	-
Sudan	26.7	42.100	4.85	16.416	9.148	11.976
Swaziland	0.9	1.146	0.23	-	-	-
Tanzania	29.6	42.100	15.24	15.304	9.666	8.119
Togo	4.1	5.548	1.03	1.251 [a]	1.102 [a]	0.419 [a]
Tunisia	9.0	10.398	0.0002	-	-	-
Uganda	19.2	33.600	6.14	8.511	5.450	4.343
Western Sahara	-	0.404	-	-	-	-
Zambia	9.0	10.900	2.39	3.578	2.635	1.706
Zimbabwe	11.0	11.553	4.40	-	-	-
TOTAL	697.2	975.706	174.165	172.597	114.939	100.328

[a] Estimates based on individuals $\leq$20years only.

The global schistosomiasis burden is currently believed to be 4.5 million disability-adjusted life years (DALYs) lost (WHO, 2002), but discussions on the appropriate assignments of disability weights are still ongoing (King et al., 2005; Hotez, 2009). During the 2010 annual meeting of the American Society of Tropical Medicine and Hygiene, Prof. Charles King elaborated in his talk "Revising global burden disease estimates for schistosomiasis" that the term schistosomiasis should not just relate to individuals currently excreting eggs, but also to all individuals who are still suffering from the adverse effects of previous infections, such as stunning or aenemia. Applying this definition, he estimated more than 440 million schistosomiasis cases worldwide and a global burden between 8.9-16.1 million standard DALYs lost annually. Even though this definition and the resulting estimates might be provocative, they highlight that schistosomiasis is causing morbid sequelae that affect quality of life even if the disease is cured. Our new schistosomiasis risk estimates for various countries in Africa can assist in revising and refining the existing

burden estimate on an evidenced-based foundation. Due to the general trend of increasing numbers of infected individuals in the presented countries, the total burden estimates may be even higher.

WHO recommends annual treatment of school-aged children in areas with a schistosomiasis prevalence equal or larger than 50%, biannual treatment in areas with risk 10-49% and treatment at the start and end of schooling in endemic areas with risk <10% (WHO, 2002). Disease risk estimates combined with the above recommendations can be used to estimate the number of treatment needed (Utzinger et al., 2009). Therefore, our work is very important in disease control since it can be used to estimate the number of treatment required at any administrative level varying from country to community. It also helps in optimizing the distribution of drugs which currently are not sufficient to cover the treatment needs of the infected populations.

The launch of the open-access GNTD database is a big step forward in the epidemiological research related to neglected tropical diseases. The database is a rich source of schistosomiasis prevalence data, especially in Africa, and covers several decades of surveys for historical and contemporary mapping purposes (Hürlimann et al., 2011). It makes the data available for research of other groups to further refine the tools we have in estimating disease burden at high spatial scales.

9.2 Limitations

Prevalence data are not reported in literature by standard age groups. In addition, some researchers do not report their data at the geographical unit they were collected, but rather as regional averages, or do not properly state the survey population, diagnostic method or survey date. Efforts were made to contact the authors in order to obtain the original data as accurate as possible, but response rate was generally low. This led to the exclusion of a number of survey locations and a reduction of the final dataset lowering model accuracy. This is of particular importance for regions with already sparse data. Even though we tried to access as much grey literature as possible for the GNTD database, there remained significant areas of sparse data because of a lack of surveys, data inaccessibility or data loss due to inappropriate archiving procedures, civil war or political unrests.

Many locations included in the GNTD database had to be retrospectively geo-referenced due to missing coordinate information (Hürlimann et al., 2011). Frequent methods of geolocating included the use of the GEOnet Names Server (GNS) database, Google Maps,

and the estimation of the village location based on maps of the study region and additional contextual information on the location provided in the corresponding publication. The reliability of the identified coordinates based on these approaches is difficult to assess (Stanislawski et al., 1996; Bonner et al., 2003) and will influence the uncertainty of model parameters, especially the spatial range parameter, and model-based predictions. Furthermore, we assigned the prevalence estimates to specific localities, although surveys cover areas rather than single points on the map. For instance, school prevalence data refer to the area around the school where the children are living and not only the school locality. The size of these areas varies from survey to survey and can hardly be defined. These effects might be addressed by the point-radius method which assigns an area of uncertainty around the geo-located position (Wieczorek et al., 2004). However, we consider the benefits of this method for the present thesis as marginal given the scales of prediction and the large spatial range estimates.

Standard geostatistical models assume that the origin of the survey locations is stochastically independent from the underlying spatial process (Diggle et al., 1998) and all locations in the prediction area are equally likely to be sampled. However, survey locations are typically chosen according to prior expectations on the observed prevalence and are likely to be concentrated in sub-regions above the average prevalence. This issue is referred to as preferential sampling in the statistical literature. Ignoring preferential sampling could lead to overestimation of the model-based predictions due to oversampling of high prevalence values (Diggle et al., 2010). Survey data obtained from the GNTD database contain many surveys with low observed prevalence levels, for example in West Africa 45% and 73% of the survey locations were below 10% for *S. haematobium* for *S. mansoni*, respectively, while in 20% (*S. haematobium*) and 50% (*S. mansoni*) of the locations no infection was found. We believe that many surveys in the GNTD database only reported *Schistosoma* infections as side outcomes, due to the same stool examination methods that allow for the concurrent diagnosis of *S. mansoni* while actually screening for soil-transmitted helminth infections. Therefore, the error due to preferential sampling might be less prominent in the compiled schistosomiasis survey data than for other survey data.

Schistosomiasis is a highly environmentally driven disease due to the intermediate host-parasite relationship (Rollinson et al., 2001; Malone, 2005). Environmental covariates can be used in geostatistical models to describe the transmission processes and to predict schistosomiasis risk at unsampled locations. Some environmental covariates are extracted from satellite data, however, the relation between satellite signals and actual ground conditions

is not constant across large regions and it remains unclear how well satellite information approximates the ground conditions. Research in this area is ongoing to obtain better proxies on ground conditions for epidemiological purposes (Hay et al., 2006; Scharlemann et al., 2008). The relationship between outcome and predictors is often non-linear. A common approach to account for non-linearity, which is often used and we also employed in this thesis, is to categorise covariates. Categorisation might be based for example on prior knowledge of the form of the relationship (which may result in undersampled categories), on quantiles of the data (which might lead to inappropriate cut-offs), or on a combination of these methods. While the choice of such abrupt cut points is unreasonable, the interpretation of the results is straight forward. Non-parametric regressions using spline approaches are an alternative (Gosoniu et al., 2009; Magalhães et al., 2011). For instance, Crainiceanu et al. (2004) proposed a Bayesian approach to penalized splines which was further implemented by Gosoniu et al. (2009) in malaria risk mapping. Climatic environmental data obtained from satellites are available at high temporal resolution. They are summarised over an interval prior to the survey date in order to be linked to the disease data. In malaria risk mapping, lag time analyses have been performed to assess the appropriate period, between certain environmental conditions and infection, during which each environmental covariates with temporal variation should be averaged (Riedel et al., 2010). However in contrast to malaria, schistosomiasis is a chronic disease and the exact date of infection is usually unknown, and hence lag time analysis is not meaningful. Yearly or long-term averages can be used instead to assess general effects on schistosomiasis transmission, but abnormal conditions that might have altered the disease distribution can not be assessed. The choice of the most vital environmental predictors to model schistosomiasis risk is important. So far, analyses were mainly conducted on a set of environmental covariates based on expert opinions and bivariate logistic regression results. Variable choice based on expert opinions might lead to the wrong selection of model parameters by either missing factors of local importance, or including redundant covariates introducing unnecessary uncertainty in the model. Covariate selection via bivariate regressions might also lead to redundant covariates in a multivariate framework and may result in model convergence problems due to correlation between parameters. For the analysis of schistosomiasis in eastern Africa and for the implementation of anisotropy in this thesis, we determined the best set of covariates using Gibbs variable selection (George and McCulloch, 1993). This approach allows to estimate the posterior predictive probability of the models and chooses a parsimonious, however best fitting, set of covariates. In future, large-scale interventions

may confound the environmental effects, as it has been observed in malaria risk mapping (Gosoniu et al., 2010; Riedel et al., 2010).

Schistosomiasis risk varies with age and between gender (Jordan and Webbe, 1982; Hotez et al., 2006a; Davis, 2009). Large-scale schistosomiasis risk modelling based on complied age-heterogeneous survey data should take into account this variation, but stratified results by age and gender are often not published. Inclusion of all data without accounting for age-dependency would result in imprecise schistosomiasis risk estimates, while exclusion of the most heterogeneous surveys in age would result in lower model accuracy due to the reduced number of survey locations, especially in areas with sparse data. For the first time in geostatistical risk mapping of schistosomiasis, we took into account age-heterogeneity between surveys via the estimation of alignment factors relating disease risk between school-aged children, adults and entire communities. This approach improved model-based predictions. However, we assumed constant risk within each of the aforementioned groups. In this respect, we ignored important differences in risk especially among childhood and adolescence. In order to transform the observed prevalence for a given age group into a prevalence of standardised age, mathematical descriptions of schistosomiasis age-prevalence curves could be coupled with geostatistical models to align age-heterogeneous surveys. For instance, Gemperli et al. (2006b) and Gosoniu (2008) obtained age-adjusted malaria risk maps from heterogeneous surveys using mathematical malaria transmission models. Holford and Hardy (1976) developed an immigration-death model to describe the schistosomiasis age-prevalence curves from cross-sectional survey data. Raso et al. (2007a) further extended the aforementioned model to take into account diagnostic sensitivity and formulated it using Bayesian geostatistical models. Furthermore, age-prevalence curves would need to be fitted and implemented for different transmission settings in order to capture the so-called peak shift, that relates to the shifting of transmission peaks to later ages of childhood and adolescence in low transmission areas (Woolhouse, 1998).

Compiled schistosomiasis survey data are not only heterogeneous in age but also in the methods used to diagnose schistosomiasis. Each diagnostic method has a different sensitivity and specificity depending on infection intensity (Bergquist et al., 2009). Multiple samples per individual are analysed to reduce diagnostic error, however the number of sampling efforts is not standardised and depends on available resources. Hence, pooling of prevalence data obtained via different diagnostic methods and sampling efforts is

likely to result in incorrect disease risk estimates. In this thesis, we addressed diagnostic-incomparability by omitting surveys based on non-direct diagnostic techniques, because this techniques have very different sensitivity and specificity as compared to direct techniques. The geostatistical models could be improved by incorporating diagnostic sensitivity and specificity parameters (Wang et al., 2008). However, many surveys have incomplete information on the diagnostic methods, which either results to an exclusion of such surveys, or which requires assumptions on the diagnostic methods and hence introducing further bias.

Another important modelling assumption limiting the power of the models used for large-scale risk mapping was that of stationary spatial processes. Stationary models imply that the spatial correlation is only a function of distance and independent of location and direction (Gosoniu et al., 2009). In malaria, it has been shown by Gosoniu et al. (2009) and Gosoniu and Vounatsou (2011b) that in regions with large differences in environmental conditions, such as ecological zones the relation between environmental and other factors with the disease risk is not constant over the whole area and non-stationary models provide more accurate results than their stationary counterparts. In schistosomiasis, regions with diverse climatic conditions differently influence vector transmission (Stensgaard et al., 2011). For example, ecological zones characterised by dry climatic conditions might be less suitable for the development of the intermediate host snails than moistures zones resulting in smaller estimates of the spatial range parameter; or regions with different main directions of river flow might demand for spatial processes with different angles of anisotropy. In addition, unobserved factors, such as health system performance, vary over the study regions introducing further non-stationary effects.

Schistosomiasis risk is changing with time due to environmental transformations, control interventions, social-economic improvements, population movement and urbanisation, among other reasons. The data included in the GNTD database were collected over several decades and hence, will be influenced by temporal trends. Our models for eastern and West Africa include time as an ordinal covariate allowing for fixed temporal effects in the data, however neglecting potential temporal correlation between survey dates. This assumption leads to an equal contribution of surveys (assuming the same amount of individuals screened) to risk prediction, irrespective of survey date, and treats older surveys as important for contemporary risk mapping as recent surveys. Spatio-temporal models could be implemented instead. However, preliminary residual analyses suggested only weak temporal correlation in the West Africa data.

9.3 Extension of the work

Our work cannot only be improved by overcoming the aforementioned limitations, but also by a number of potential extensions. An immediate extension will focus on the spatial distribution of schistosomiasis in South and Central Africa and the estimation of the number of infected individuals in these countries. Preferably, this analysis should be performed accounting for diagnostic sensitivity and age-heterogeneity. The resulting risk estimates could be further used to calculated annual treatment needs based on either country-specific, district-specific or pixel-level cut-offs according to the WHO schistosomiasis control recommendations (WHO, 2002). In addition, the (population-adjusted) prevalence maps and model-based estimates on the number of infected individuals and treatment needs could be made available via the web with an interface which will allow the users to specify their own geographical area of interest and download the raw prevalence data extracted from the GNTD database together with the area-specific prediction maps and the estimated number of infected individuals and treatment needs. These estimates can support local practitioners to implement control intervention programs and to define the location of future surveys in previous neglected areas.

Model-based predictions could be further validated with the conduction of new surveys in order to probe the predictive ability of the models, especially in areas where uncertainty of the predictions was found to be high and areas lacking contemporary surveys. Geostatistical models should assist in identifying the survey locations based on the estimated spatial process, model uncertainty and population estimates. The additional surveys could then be included in future iterations of the work to improve model-based predictions and to obtain new contemporary maps. Comparisons between the updated schistosomiasis risk maps and the baseline maps would allow evaluation of temporal changes of disease transmission and assessment of the effectiveness of control interventions.

Our spatially explicit large-scale schistosomiasis estimates are based on separate modeling of the two predominant species in Africa, namely *S. haematobium* and *S. mansoni*. The two *Schistosoma* species showed overlapping distribution, known as co-endemicity, leading to simultaneously infected individuals and potentially aggravated morbidity. Despite few studies on endemicity (Raso et al., 2007b) and co-infection (Raso et al., 2006b; Brooker and Clements, 2009) of schistosomiasis with hookworm, co-infection between *Schistosoma* species has not yet been studied and it remains unclear whether simultaneous infections occur randomly or show spatial dependency. In Chapter 8 we have shown the advantages of joint modeling approaches via shared component analyses in studying the geographical

distribution of co-infection under the presence of disease dependency. This approach should be applied in *S. haematobium* and *S. mansoni* risk estimation to assess the level of dependency and to improve model-based estimates on combined schistosomiasis risk omitting the assumption of independence. This analysis will improve our understanding of schistosomiasis transmission and enhance cost-effectiveness of integrated control intervention programmes (Brady et al., 2006).

The spatial distribution of the intermediate host species is, to a large extent, influencing the spatial distribution of schistosomiasis risk. However, our models are only indirectly taking into account snail distribution based on the relation with environmental factors. Recently, Stensgaard et al. (2011) have created continental maps on Biomphalaria presence in Africa that could be linked with *S. mansoni* survey locations. Incorporating the snail distribution as a predictor in our models is likely to improve model predictive ability. Statistical models could be developed to obtain spatially-explicit estimates of the probability of snail presence from data lacking information on snail absence (Elith et al., 2006).

Another extension of the work presented here is the estimation of the future distribution of schistosomiasis risk in Africa using different climate change scenarios to support control and elimination programmes in areas of future disease presence and absence. It has been shown that climate change will affect the intermediate snail species distribution in Africa leading to expanding and contracting areas of *Biomphalaria* (Stensgaard et al., 2011), which is likely to change schistosomiasis distribution. In addition, Yang et al. (2010) have already presented how the spatial *S. japonicum* distribution in P.R. China might change due to different global warming assumptions. Estimating the spatial distribution of disease burden under climate change scenarios will assist in health system preparedness and guide disease elimination programmes.

Chapter 10

Conclusion

Compiled survey data are essential to obtain large-scale estimates on the spatial distribution of diseases despite the above mentioned limitations. The newly established GNTD database initiated by the EU-funded CONTRAST project (`http://www.eu-contrast.eu/`) is currently the only comprehensive collection of historical and contemporary schistosomiasis survey data on global scale that is publicly available. This database will be constantly expanded and updated to serve as indispensable tool for large-scale mapping of neglected tropical diseases. Current efforts to expand the database include the extraction of schistosomiasis and soil-transmitted helminth data in Latin America and China. In addition, a web-based interface is created to enhance data accessibility for external groups via various search functions, enable data contribution from other researchers and improve data entry. However, the work in this thesis has shown that the collection of data has to be accompanied by the development of appropriate statistical models taking into account the data characteristics to obtain accurate risk maps.

Data-driven Bayesian geostatistical models enabled us to obtain empirical, high-resolution infection risk estimates for *S. haematobium* and *S. mansoni* in western and eastern Africa. These are important tools for evidenced-based decision-making on the spatial implementation of future control interventions and to define treatment needs in order to reduce disease burden. The impact of interventions and transmission dynamics can be monitored and evaluated via subsequent updates of the maps. We plan to further improve geostatistical methodology, as outlined in the above sections, to increase accuracy of the model-based predictions and to provide revised maps using improved models and newly collected data. We hope that these efforts will contribute to successful disease control and bring us closer to disease elimination.

Bibliography

Aanensen, D. M., Huntley, D. M., Feil, E. J., al-Own, F., and Spratt, B. G. (2009). EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. *PloS One*, 4(9):e6968.

Abdel-Wahab, M. F., Strickland, G. T., El-Sahly, A., El-Kady, N., Zakaria, S., and Ahmed, L. (1979). Changing pattern of schistosomiasis in egypt 1935–79. *Lancet*, 2(8136):242–244.

Adriko, M., Kazibwe, F., Ogutu, D., Tukahebwa, E. M., Kabatereine, N. B., Kariuki, H. C., and Stothard, J. R. (2011). Urinary schistosomiasis in lango region, uganda 60 years after schwetz: Past and present trends in the face of ongoing control for intestinal schistosomiasis and soil transmitted helminthiasis. *Acta Tropica*.

Allen, A. V. H. and Ridley, D. S. (1970). Further observations on the formol-ether concentration technique for faecal parasites. *Journal of Clinical Pathology*, 23(6):545–546.

Anderson, R. M. and May, R. M. (1985). Helminth infections of humans: mathematical models, population dynamics, and control. *Advances in Parasitology*, 24:1–101.

Arntzen, T., Bakken, S., Caraveo, S., Gutmans, A., Lerdorf, R., and et al. (2001). PHP: a widely used general purpose scripting language. http://www.php.net/.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 1 edition.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 70(4):825–848.

Bank, T. W. (2008). Project performance assessment reports - arab republic of egypt - national schistosomiasis control project. Project Performance Assessment Reports 44466, World Bank.

Bavia, M. E., Hale, L. F., Malone, J. B., Braud, D. H., and Lee, S. M. (1999). Geographic information systems and the environmental risk of schistosomiasis in bahia, brazil. *The American Journal of Tropical Medicine and Hygiene*, 60(4):566–572.

Bavia, M. E., Malone, J. B., Hale, L., Dantas, A., Marroni, L., and Reis, R. (2001). Use of thermal and vegetation index data from earth observing satellites to evaluate the risk of schistosomiasis in bahia, brazil. *Acta Tropica*, 79(1):79–85.

Beck-Wörner, C., Raso, G., Vounatsou, P., N'Goran, E. K., Rigo, G., Parlow, E., and Utzinger, J. (2007). Bayesian spatial risk prediction of *Schistosoma mansoni* infection in western côte d'ivoire using a remotely-sensed digital elevation model. *The American Journal of Tropical Medicine and Hygiene*, 76(5):956–963.

Bergquist, R., Johansen, M. V., and Utzinger, J. (2009). Diagnostic dilemmas in helminthology: what tools to use and when? *Trends in Parasitology*, 25(4):151–156.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., and Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology (Cambridge, Mass.)*, 14(4):408–412.

Booth, M., Vounatsou, P., N'Goran, E. K., Tanner, M., and Utzinger, J. (2003). The influence of sampling effort and the performance of the kato-katz technique in diagnosing *Schistosoma mansoni* and hookworm co-infections in rural côte d'ivoire. *Parasitology*, 127(Pt 6):525–531.

Bradley, D. J. (1972). Regulation of parasite populations. a general theory of the epidemiology and control of parasitic infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 66(5):697–708.

Bradley, D. J., Sturrock, R. F., and Williams, P. N. (1967). The circumstantial epidemiology of s. haematobium in lango district, uganda. *East African Medical Journal*, 44:193–204.

Brady, M. A., Hooper, P. J., and Ottesen, E. A. (2006). Projected benefits from integrating NTD programs in sub-saharan africa. *Trends in Parasitology*, 22(7):285–291.

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

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

Brooker, S. (2010). Estimating the global distribution and disease burden of intestinal

nematode infections: adding up the numbers - a review. *International Journal for Parasitology*, 40(10):1137–1144.

Brooker, S. and Clements, A. C. (2009). Spatial heterogeneity of parasite co-infection: determinants and geostatistical prediction at regional scales. *International Journal for Parasitology*, 39(5):591–597.

Brooker, S., Clements, A. C., Hotez, P. J., Hay, S. I., Tatem, A. J., Bundy, D. A., and Snow, R. W. (2006). The co-distribution of plasmodium falciparum and hookworm among african schoolchildren. *Malaria Journal*, 5(1):99.

Brooker, S., Donnelly, C. A., and Guyatt, H. L. (2000). Estimating the number of helminthic infections in the republic of cameroon from data on infection prevalence in schoolchildren. *Bulletin of the World Health Organization*, 78(12):1456–1465.

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

Brooker, S., Hotez, P. J., and Bundy, D. A. P. (2010). The global atlas of helminth infection: mapping the way forward in neglected tropical disease control. *PLoS Neglected Tropical Diseases*, 4(7):e779.

Brooker, S., Kabatereine, N. B., Gyapong, J. O., Stothard, J. R., and Utzinger, J. (2009a). Rapid mapping of schistosomiasis and other neglected tropical diseases in the context of integrated control programmes in africa. *Parasitology*, 136(13):1707–1718.

Brooker, S., Kabatereine, N. B., Smith, J. L., Mupfasoni, D., Mwanje, M. T., Ndayishimiye, O., Lwambo, N. J., Mbotha, D., Karanja, P., Mwandawiro, C., Muchiri, E., Clements, A. C., Bundy, D. A., and Snow, R. W. (2009b). An updated atlas of human helminth infections: the example of east africa. *International Journal of Health Geographics*, 8:42.

Bundy, D. A. P., Chandiwana, S. K., Homeida, M. M., Yoon, S., and Mott, K. E. (1991). The epidemiological implications of a multiple-infection approach to the control of human helminth infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(2):274–276. PMID: 1887492.

Chitsulo, L., Engels, D., Montresor, A., and Savioli, L. (2000). The global status of schistosomiasis and its control. *Acta Tropica*, 77(1):41–51.

Cioli, D., Botros, S. S., Wheatcroft-Francklow, K., Mbaye, A., Southgate, V., Tchuenté, L. T., Pica-Mattoccia, L., Troiani, A. R., El-Din, S. H. S., Sabra, A. A., Albin, J.,

Engels, D., and Doenhoff, M. J. (2004). Determination of ED50 values for praziquantel in praziquantel-resistant and -susceptible *Schistosoma mansoni* isolates. *International Journal for Parasitology*, 34(8):979–987.

Cioli, D. and Pica-Mattoccia, L. (2003). Praziquantel. *Parasitology Research*, 90 (Supp. 1):S3–9.

Clements, A. C. A., Bosqué-Oliva, E., Sacko, M., Landouré, A., Dembélé, R., Traoré, M., Coulibaly, G., Gabrielli, A. F., Fenwick, A., and Brooker, S. (2009a). A comparative study of the spatial distribution of schistosomiasis in mali in 19841989 and 20042006. *PLoS Neglected Tropical Diseases*, 3(5):e431.

Clements, A. C. A., Deville, M., Ndayishimiye, O., Brooker, S., and Fenwick, A. (2010). Spatial co-distribution of neglected tropical diseases in the east african great lakes region: revisiting the justification for integrated control. *Tropical Medicine & International Health: TM & IH*, 15(2):198–207.

Clements, A. C. A., Firth, S., Dembelé, R., Garba, A., Touré, S., Sacko, M., Landouré, A., Bosqué-Oliva, E., Barnett, A. G., Brooker, S., and Fenwick, A. (2009b). Use of bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in western africa. *Bulletin of the World Health Organization*, 87(12):921–929.

Clements, A. C. A., Garba, A., Sacko, M., Touré, S., Dembélé, R., Landouré, A., Bosqué-Oliva, E., Gabrielli, A. F., and Fenwick, A. (2008). Mapping the probability of schistosomiasis and associated uncertainty, west africa. *Emerging Infectious Diseases*, 14(10):1629–1632.

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

Clements, A. C. A., Moyeed, R., and Brooker, S. (2006b). Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in east africa. *Parasitology*, 133(Pt 6):711–719.

Cohen, J. (2008). Science and society. science goes hollywood: NAS links with entertainment industry. *Science (New York, N.Y.)*, 322(5906):1315.

Conceio, M. J., Argento, C. A., and Corra, A. (2000). Study of *Schistosoma mansoni* isolates from patients with failure of treatment with oxamniquine. *Memrias Do Instituto Oswaldo Cruz*, 95(3):375–380.

Crainiceanu, C., Ruppert, D., and Wand, M. (2004). Bayesian analysis for penalized spline regression using winbugs. Technical report, Berkeley Electronic Press.

Cross, E. R. and Bailey, R. C. (1984). Prediction of areas endemic for schistosomiasis through use of discriminant analysis of environmental data. *Military Medicine*, 149(1):28–30.

Cross, E. R., Sheffield, C., Perrine, R., and Pazzaglia, G. (1984). Predicting areas endemic for schistosomiasis using weather variables and a landsat data base. *Military Medicine*, 149(10):542–544.

Davis, A. (2009). Schistosomiasis. In *Manson's Tropical Diseases. 22nd edition*, Chapter 82, pages 1425–1460. Saunders Elsevier.

de Vlas, S. J. and Gryseels, B. (1992). Underestimation of *Schistosoma mansoni* prevalences. *Parasitology Today (Personal Ed.)*, 8(8):274–277.

Deville, J. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer, New York.

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, 47(3):299–350.

Doenhoff, M. J., Cioli, D., and Utzinger, J. (2008). Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. *Current Opinion in Infectious Diseases*, 21(6):659–667.

Doumenge, J. P., Mott, K. E., Cheung, C., Villenave, D., Chapuis, O., Perrin, M. F., and Reaud-Thomas, G. (1987). Atlas of the global distribution of schistosomiasis / atlas de la répartition mondiale des schistosomiases. Technical report, WHO-CEGET-CNRS, Presses Universitaires de Bordeaux, Bordeaux.

Ecker, M. D. and Gelfand, A. E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, 31(1):67–83.

Ecker, M. D. and Gelfand, A. E. (2003). Spatial modeling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics*, 10(2):165–178.

Elith, J., Graham, C. H., Anderson, R. P., Dudk, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle,

B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., ScachettiPereira, R., Schapire, R. E., Sobern, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, 29(2):129–151.

Emmert, D. B., Stoehr, P. J., Stoesser, G., and Cameron, G. N. (1994). The european bioinformatics institute (EBI) databases. *Nucleic Acids Research*, 22(17):3445–3449.

Engels, D., Sinzinkayo, E., de Vlas, S. J., and Gryseels, B. (1997). Intraspecimen fecal egg count variation in *Schistosoma mansoni* infection. *The American Journal of Tropical Medicine and Hygiene*, 57(5):571–577.

Engels, D., Sinzinkayo, E., and Gryseels, B. (1996). Day-to-day egg count fluctuation in *Schistosoma mansoni* infection and its operational implications. *The American Journal of Tropical Medicine and Hygiene*, 54(4):319–324.

Fenwick, A. (2006). New initiatives against africa's worms. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(3):200–207.

Fenwick, A., Webster, J. P., Bosqué-Oliva, E., Blair, L., Fleming, F. M., Zhang, Y., Garba, A., Stothard, J. R., Gabrielli, A. F., Clements, A. C. A., Kabatereine, N. B., Toure, S., Dembele, R., Nyandindi, U., Mwansa, J., and Koukounari, A. (2009). The schistosomiasis control initiative (sci): rationale, development and implementation from 2002-2008. *Parasitology*, 136(13):1719–1730.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data - or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115–32.

French, M. D., Churcher, T. S., Gambhir, M., Fenwick, A., Webster, J. P., Kabatereine, N. B., and Basáñ ez, M. (2010). Observed reductions in *Schistosoma mansoni* transmission from large-scale administration of praziquantel in uganda: a mathematical modelling study. *PLoS Neglected Tropical Diseases*, 4(11):e897.

Fulford, A. J. C., Butterworth, A. E., Sturrock, R. F., and Ouma, J. H. (1992). On the use of age-intensity data to detect immunity to parasitic infections, with special reference to *Schistosoma mansoni* in kenya. *Parasitology*, 105(02):219–227.

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

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

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

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

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

Giardina, F., Gosoniu, L., Konate, L., and Vounatsou, P. (2011). Estimating the burden of malaria in senegal: Bayesian zero inflated binomial geostatistical modeling of the MIS 2008 data. *PLoS ONE*.

Gosoniu, D., Gosoniu, L., Tille, Y., and Vounatsou, P. (2011a). Subsampling the gaussian process of very large geostatistical data - does the sampling approach matter? *Computational Statistics and Data Analysis*.

Gosoniu, D., Vounatsou, P., Kahn, K., and Tillé, Y. (2011b). Geostatistical modeling of large non-gaussian irregularly distributed data. *Computational Statistics*.

Gosoniu, L. (2008). Development of bayesian geostatistical models with applications in malaria epidemiology. Technical report, University of Basel.

Gosoniu, L., Veta, A. M., and Vounatsou, P. (2010). Bayesian geostatistical modeling of malaria indicator survey data in angola. *PloS One*, 5(3):e9322.

Gosoniu, L. and Vounatsou, P. (2011a). Bayesian geostatistical variable selection: Modeling the liberia malaria indicator survey data. *Malaria Journal*.

Gosoniu, L. and Vounatsou, P. (2011b). Non-stationary partition modeling of geostatistical data for malaria risk mapping. *Journal of Applied Statistics*, 38(1):3.

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

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

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

Gray, D. J., Forsyth, S. J., Li, R. S., McManus, D. P., Li, Y., Chen, H., Zheng, F., and Williams, G. M. (2009). An innovative database for epidemiological field studies of neglected tropical diseases. *PLoS Neglected Tropical Diseases*, 3(5):e413.

Green, E. (2008). District creation and decentralisation in uganda. working paper no. 24 - development as State-Making. Technical report.

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

Gryseels, B., Polman, K., Clerinx, J., and Kestens, L. (2006). Human schistosomiasis. *The Lancet*, 368(9541):1106–1118.

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

Hatz, C. F. (2001). The use of ultrasound in schistosomiasis. *Advances in Parasitology*, 48:225–284.

Hay, S., Tatem, A., Graham, A., Goetz, S., Rogers, D., Simon I. Hay, A. G., and Rogers, D. J. (2006). Global environmental data for mapping infectious disease distribution. In *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, volume Volume 62, pages 37–77. Academic Press.

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R. F., Brooker, S., Smith, D. L., Moyeed, R. A., and Snow, R. W. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Medicine*, 6(3):e1000048.

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

Holford, T. R. and Hardy, R. J. (1976). A stochastic model for the analysis of age-specific prevalence curves in schistosomiasis. *Journal of Chronic Diseases*, 29(7):445–458.

Hotez, P. (2008). Hookworm and poverty. *Annals of the New York Academy of Sciences*, 1136:38–44.

Hotez, P. J. (2009). Mass drug administration and integrated control for the world's high-prevalence neglected tropical diseases. *Clinical Pharmacology and Therapeutics*, 85(6):659–664.

Hotez, P. J., Bundy, D. A., Beegle, K., Brooker, S., Drake, L., de Silva, N., Montresor, A., Engels, D., Jukes, M., Chitsulo, L., Chow, J., Laxminarayan, R., Michaud, C., Bethony, J., Correa-Oliveira, R., Shuhua, X., Fenwick, A., and Savioli, L. (2006a). Helminth infections: soil-transmitted helminth infections and schistosomiasis. In *Disease Control Priorities in Developing Countries. 2nd edition*, Chapter 24. World Bank, Washington (DC).

Hotez, P. J., Molyneux, D. H., Fenwick, A., Kumaresan, J., Sachs, S. E., Sachs, J. D., and Savioli, L. (2007). Control of neglected tropical diseases. *The New England Journal of Medicine*, 357(10):1018–1027.

Hotez, P. J., Molyneux, D. H., Fenwick, A., Ottesen, E., Ehrlich Sachs, S., and Sachs, J. D. (2006b). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS Medicine*, 3(5):e102.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sensing of Environment*, 83(1-2):195–213.

Hürlimann, E., Schur, N., Boutsika, K., Stensgaard, A. S., Laizer, N., Laserna de Himpsl, M., Ziegelbauer, K., Camenzind, L., Simoonga, C., Mushinge, G., Saarnak, C. F. L., Utzinger, J., Kristensen, T. K., and Vounatsou, P. (2011). Toward an open-access, real-time global database for mapping, control, and surveillance of neglected tropical diseases. *PLoS Neglected Tropical Diseases*.

Johnson, D. (2005). Bayesian inference for geostatistical regression models. Technical report, unpublished manuscript, available from http://www. stat. colostate. edu/ nsu/starmap/johnson. spatial. regression. pdf.

Jordan, P. and Webbe, G. (1982). *Schistosomiasis epidemiology, treatment and control*. William Heinemann Medical Books, London.

Katz, N., Chaves, A., and Pellegrino, J. (1972). A simple device for quantitative stool thick-smear technique in *Schistosomiasis mansoni*. *Revista Do Instituto De Medicina Tropical De Sao Paulo*, 14(6):397–400.

Katz, N. and Miura, M. (1954). On the comparison of some stool examination methods. *Jpn J Parasitol*, 3:35.

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

Kazembe, L. N., Muula, A. S., and Simoonga, C. (2009). Joint spatial modelling of common morbidities of childhood fever and diarrhoea in malawi. *Health & Place*, 15(1):165–172.

Kazembe, L. N. and Namangale, J. J. (2007). A bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in malawi. *European Journal of Epidemiology*, 22(8):545–556.

Kim, H., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

King, C. H. (2010). Parasites and poverty: the case of schistosomiasis. *Acta Tropica*, 113(2):95–104.

King, C. H., Dickman, K., and Tisch, D. J. (2005). Reassessment of the cost of chronic helmintic infection: a meta-analysis of disability-related outcomes in endemic schistosomiasis. *Lancet*, 365(9470):1561–1569.

King, C. H., Sturrock, R. F., Kariuki, H. C., and Hamburger, J. (2006). Transmission control for schistosomiasis - why it matters now. *Trends in Parasitology*, 22(12):575–582.

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

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85.

Koroma, J. B., Peterson, J., Gbakima, A. A., Nylander, F. E., Sahr, F., Soares Magalhes, R. J., Zhang, Y., and Hodges, M. H. (2010). Geographical distribution of intestinal schistosomiasis and soil-transmitted helminthiasis and preventive chemotherapy strategies in sierra leone. *PLoS Neglected Tropical Diseases*, 4(11):e891.

Koukounari, A., Gabrielli, A. F., Touré, S., Bosqué-Oliva, E., Zhang, Y., Sellin, B., Donnelly, C. A., Fenwick, A., and Webster, J. P. (2007). *Schistosoma haematobium* infection and morbidity before and after large-scale administration of praziquantel in burkina faso. *Journal of Infectious Diseases*, 196(5):659 –669.

Kristensen, T. K. (2008). African schistosomiasis: refocusing upon the environment. *Newsletter of the Royal Society of Tropical Medicine and Hygiene*, 13:1–8.

Lammie, P. J., Fenwick, A., and Utzinger, J. (2006). A blueprint for success: integration of neglected tropical disease control programmes. *Trends in Parasitology*, 22(7):313–321.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Hammond, M., Hill, C. A., Konopinski, N., Lobo, N. F., MacCallum, R. M., Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D. W., Stinson, E. O., Topalis, P., Birney, E., Gelbart, W. M., Kafatos, F. C., Louis, C., and Collins, F. H. (2009). VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Research*, 37(Database issue):D583–587.

Le Sueur, D., Binka, F., Lengeler, C., De Savigny, D., Snow, B., Teuscher, T., and Toure, Y. (1997). An atlas of malaria in africa. *Africa Health*, 19(2):23–24.

Lee, K., Bacchetti, P., and Sim, I. (2008). Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS Medicine*, 5(9):e191.

Lengeler, C., Mshinda, H., Morona, D., and de Savigny, D. (1993). Urinary schistosomiasis: testing with urine filtration and reagent sticks for haematuria provides a comparable prevalence estimate. *Acta Tropica*, 53(1):39–50.

Lengeler, C., Utzinger, J., and Tanner, M. (2002). Questionnaires for rapid screening of schistosomiasis in sub-Saharan africa. *Bulletin of the World Health Organization*, 80(3):235–242.

Madsen, H. (1985a). *Ecology and control of African freshwater pulmonate snails. Part 2: Basic principles in ecology of freshwater snails.* Danish Bilharziasis Laboratory, Copenhagen.

Madsen, H. (1985b). *Ecology and control of African freshwater pulmonate snails. Part I: Life cycle and methodology.* Danish Bilharziasis Laboratory, Copenhagen.

Magalhães, R. J. S., Clements, A. C. A., Patil, A. P., Gething, P. W., and Brooker, S. (2011). The applications of model-based geostatistics in helminth epidemiology and control. *Advances in Parasitology*, 74:267–296. PMID: 21295680.

Malone, J. B. (2005). Biology-based mapping of vector-borne parasites by geographic information systems and remote sensing. *Parassitologia*, 47(1):27–50.

Malone, J. B., Huh, O. K., Fehler, D. P., Wilson, P. A., Wilensky, D. E., Holmes, R. A., and Elmagdoub, A. I. (1994). Temperature data from satellite imagery and the distribution of schistosomiasis in egypt. *The American Journal of Tropical Medicine and Hygiene*, 50(6):714–722.

Malone, J. B., Yilma, J. M., McCarroll, J. C., Erko, B., Mukaratirwa, S., and Zhou, X. (2001). Satellite climatology and the environmental risk of *Schistosoma mansoni* in ethiopia and east africa. *Acta Tropica*, 79(1):59–72.

Marti, H. and Escher, E. (1990). Saf - an alternative fixation solution for parasitological stool specimens. *Schweizerische Medizinische Wochenschrift*, 120(40):1473–1476.

Marti, H. and Koella, J. C. (1993). Multiple stool examinations for ova and parasites and rate of false-negative results. *Journal of Clinical Microbiology*, 31(11):3044–3045.

Melman, S. D., Steinauer, M. L., Cunningham, C., Kubatko, L. S., Mwangi, I. N., Wynn, N. B., Mutuku, M. W., Karanja, D. M. S., Colley, D. G., Black, C. L., Secor, W. E., Mkoji, G. M., and Loker, E. S. (2009). Reduced susceptibility to praziquantel among naturally occurring kenyan isolates of *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases*, 3(8):e504.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6).

Moffett, A., Strutz, S., Guda, N., González, C., Ferro, M. C., Sánchez-Cordero, V., and Sarkar, S. (2009). A global public database of disease vector and reservoir distributions. *PLoS Neglected Tropical Diseases*, 3(3):e378.

Molyneux, D. H. (2006). Elimination of transmission of lymphatic filariasis in egypt. *Lancet*, 367(9515):966–968.

MySQL (1995). MySQL: the world's most popular open source database. http://www.mysql.com/.

Ndir, O. (2003). Enquete nationale de depistage des bilharzioses chez les enfants d'age scolaire au sénégal. Technical report.

Ouma, J. H. and Waithaka, F. (1978). Prevalence of *Schistosoma mansoni* and *Schistosoma haematobium* in kitui district, kenya. *East African Medical Journal*, 55(2):54–60.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large datasets. *Computational Statistics & Data Analysis*, 51(8):3631–3653.

Pan, J., Nahm, M., Wakim, P., Cushing, C., Poole, L., Tai, B., and Pieper, C. F. (2009). A centralized informatics infrastructure for the national institute on drug abuse clinical trials network. *Clinical Trials (London, England)*, 6(1):67–75.

Pinot de Moira, A., Fulford, A. J. C., Kabatereine, N. B., Kazibwe, F., Ouma, J. H.,

Dunne, D. W., and Booth, M. (2007). Microgeographical and tribal variations in water contact and *Schistosoma mansoni* exposure within a ugandan fishing community. *Tropical Medicine & International Health: TM & IH*, 12(6):724–735.

Polderman, A. M. (1979). Transmission dynamics of endemic schistosomiasis. *Tropical and Geographical Medicine*, 31(4):465–475.

Pullan, R. L., Gething, P. W., Smith, J. L., Mwandawiro, C. S., Sturrock, H. J. W., Gitonga, C. W., Hay, S. I., and Brooker, S. (2011). Spatial modelling of soil-transmitted helminth infections in kenya: a disease control planning tool. *PLoS Neglected Tropical Diseases*, 5(2):e958.

Ramana, J. and Gupta, D. (2009). ProtVirDB: a database of protozoan virulent proteins. *Bioinformatics (Oxford, England)*, 25(12):1568–1569.

Raso, G., Matthys, B., N'Goran, E. K., Tanner, M., Vounatsou, P., and Utzinger, J. (2005). Spatial risk prediction and mapping of *Schistosoma mansoni* infections among schoolchildren living in western côte divoire. *Parasitology*, 131:97–108.

Raso, G., Vounatsou, P., Gosoniu, L., Tanner, M., N'Goran, E. K., and Utzinger, J. (2006a). Risk factors and spatial patterns of hookworm infection among schoolchildren in a rural area of western côte d'ivoire. *International Journal for Parasitology*, 36(2):201–210.

Raso, G., Vounatsou, P., McManus, D. P., N'Goran, E. K., and Utzinger, J. (2007a). A bayesian approach to estimate the age-specific prevalence of *Schistosoma mansoni* and implications for schistosomiasis control. *International Journal for Parasitology*, 37(13):1491–500.

Raso, G., Vounatsou, P., McManus, D. P., and Utzinger, J. (2007b). Bayesian risk maps for *Schistosoma mansoni* and hookworm mono-infections in a setting where both parasites co-exist. *Geospatial Health*, 2(1):85–96.

Raso, G., Vounatsou, P., Singer, B. H., N'Goran, E. K., Tanner, M., and Utzinger, J. (2006b). An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*-hookworm coinfection. *Proceedings of the National Academy of Sciences of the United States of America*, 103(18):6934–6939.

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

Robinson, E., Picon, D., Sturrock, H. J., Sabasio, A., Lado, M., Kolaczinski, J., and Brooker, S. (2009). The performance of haematuria reagent strips for the rapid mapping of urinary schistosomiasis: field experience from southern sudan. *Tropical Medicine & International Health: TM & IH*, 14(12):1484–1487.

Rollinson, D., Stothard, J. R., and Southgate, V. R. (2001). Interactions between intermediate snail hosts of the genus bulinus and schistosomes of the *Schistosoma haematobium* group. *Parasitology*, 123 Suppl:S245–260.

Ross, A. G., Vickers, D., Olds, G. R., Shah, S. M., and McManus, D. P. (2007). Katayama syndrome. *The Lancet Infectious Diseases*, 7(3):218–224.

Rudge, J. W., Stothard, J. R., Basáñez, M., Mgeni, A. F., Khamis, I. S., Khamis, A. N., and Rollinson, D. (2008). Micro-epidemiology of urinary schistosomiasis in zanzibar: Local risk factors associated with distribution of infections among schoolchildren and relevance for control. *Acta Tropica*, 105(1):45–54.

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

Rumisha, S. F., Gosoniu, D., Kasasa, S., Smith, T. A., Abdulla, S., Masanja, H., and Vounatsou, P. (2011). Bayesian modeling of large geostatistical data to estimate seasonal and spatial variation of sporozoite rate. *tatistical Methods & Application*.

Sanderson, E. W., Jaiteh, M., Levy, M. A., Redford, K. H., Wannebo, A. V., and Woolmer, G. (2002). The human footprint and the last of the wild. *BioScience*, 52(10):891–904.

Scharlemann, J. P. W., Benz, D., Hay, S. I., Purse, B. V., Tatem, A. J., Wint, G. R. W., and Rogers, D. J. (2008). Global data for ecology and epidemiology: a novel algorithm for temporal fourier processing MODIS data. *PloS One*, 3(1):e1408.

Schmidt, A. M. and Rodriguez, M. A. (2010). Modelling multivariate counts varying continuously in space. Technical report, Oxford University Press.

Schur, N., Gosoniu, L., Raso, G., Utzinger, J., and Vounatsou, P. (2011a). Modelling the geographical distribution of co-infection risk from single disease survey data. *Statistics in Medicine*, 30(14):1761–1776.

Schur, N., Hürlimann, E., Garba, A., Traoré, M. S., Ndir, O., Ratard, R. C., Tchuem Tchuenté, L., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2011b). Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged 20 years in west africa. *PLoS Neglected Tropical Diseases*, 5(6):e1194.

Schur, N., Hürlimann, E., Stensgaard, A., Chimfwembe, K., Mushinge, G., Simoonga, C., Kabatereine, N. B., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2011c). Spatially explicit *Schistosoma* infection risk in eastern africa based on bayesian geostatistical modelling. *Acta Tropica*.

Schur, N., Utzinger, J., and Vounatsou, P. (2011d). Modelling age-heterogeneous *Schistosoma haematobium* and *S. mansoni* survey data via alignment factors. *Parasites & Vectors*.

Schwetz, J. (1951). Communications on vesical bilharzia in the lango district of uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 44:501–513.

Shears, P. and Lusty, T. (1987). Communicable disease epidemiology following migration: studies from the african famine. *The International Migration Review*, 21(3):783–795.

Sibley, C. H., Barnes, K. I., and Plowe, C. V. (2007). The rationale and plan for creating a world antimalarial resistance network (WARN). *Malaria Journal*, 6:118.

Simoonga, C., Utzinger, J., Brooker, S., Vounatsou, P., Appleton, C., Stensgaard, A., Olsen, A., and Kristensen, T. K. (2009). Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in africa. *Parasitology*, (13):1683–1693.

Smith, J. L., Haddad, D., Polack, S., Harding-Esch, E. M., Hooper, P. J., Mabey, D. C., Solomon, A. W., and Brooker, S. (2011). Mapping the global distribution of trachoma: why an updated atlas is needed. *PLoS Neglected Tropical Diseases*, 5(6):e973.

Smits, H. L. (2009). Prospects for the control of neglected tropical diseases by mass drug administration. *Expert Review of Anti-Infective Therapy*, 7(1):37–56.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):583–639.

Stanislawski, L. V., Dewitt, B. A., and Shrestha, R. L. (1996). Estimating positional accuracy of data layers within a GIS through error propagation. *Photogrammetric engineering and remote sensing*, 62(4):429433.

Steinmann, P., Keiser, J., Bos, R., Tanner, M., and Utzinger, J. (2006). Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet Infectious Diseases*, 6(7):411–425.

Stensgaard, A., Saarnak, C. F. L., Utzinger, J., Vounatsou, P., Simoonga, C., Mushinge,

G., Rahbek, C., Mhlenberg, F., and Kristensen, T. K. (2009). Virtual globes and geospatial health: the potential of new tools in the management and control of vector-borne diseases. *Geospatial Health*, 3(2):127–141.

Stensgaard, A., Utzinger, J., Vounatsou, P., Hürlimann, E., Schur, N., Saarnak, C. F. L., Mushinge, G., Simoonga, C., Kabatereine, N. B., Tchuem Tchuenté, L., Rahbek, C., and Kristensen, T. K. (2011). Large-scale determinants of intestinal schistosomiasis and intermediate host snail distribution across africa: does climate matter? *Acta Tropica*.

Stothard, J. R., Chitsulo, L., Kristensen, T. K., and Utzinger, J. (2009). Control of schistosomiasis in sub-Saharan africa: progress made, new opportunities and remaining challenges. *Parasitology*, 136(13):1665–1675.

Tchuem Tchuenté, L. A., Southgate, V. R., Jourdane, J., Webster, B. L., and Vercruysse, J. (2003). *Schistosoma intercalatum*: an endangered species in cameroon? *Trends in Parasitology*, 19(9):389–393.

Tzala, E. and Best, N. (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, 17(1):97–118.

United Nations (2007). World population prospects: The 2006 revision, highlights. Technical Report ESA/P/WP.202, United Nations, Department of Economics and Social Affairs, Population Division, New York.

Utroska, J. A., Chen, M., Dixon, H., Yoon, S., Helling-Borda, M., Hogerzeil, H. V., and Mott, K. E. (1989). An estimate of the global needs for praziquantel within schistosomiasis control programmes. Technical report, World Health Organization, Geneva.

Utzinger, J., Bergquist, R., Olveda, R., and Zhou, X. (2010). Important helminth infections in southeast asia diversity, potential for control and prospects for elimination. *Advances in Parasitology*, 72:1–30.

Utzinger, J., Booth, M., N'Goran, E. K., Müller, I., Tanner, M., and Lengeler, C. (2001). Relative contribution of day-to-day and intra-specimen variation in faecal egg counts of *Schistosoma mansoni* before and after treatment with praziquantel. *Parasitology*, 122(Pt 5):537–544.

Utzinger, J. and Keiser, J. (2004). Schistosomiasis and soil-transmitted helminthiasis: common drugs for treatment and control. *Expert Opinion on Pharmacotherapy*, 5(2):263–285.

Utzinger, J., N'Goran, E. K., Caffrey, C. R., and Keiser, J. (2011). From innovation to application: Social-ecological context, diagnostics, drugs and integrated control of schistosomiasis. *Acta Tropica*, (in press).

Utzinger, J., Raso, G., Brooker, S., de Savigny, D., Tanner, M., Ornbjerg, N., Singer, B. H., and N'Goran, E. K. (2009). Schistosomiasis and neglected tropical diseases: towards integrated and sustainable control and a word of caution. *Parasitology*, 136(13):1859–1874.

van der Werf, M. J., de Vlas, S. J., Brooker, S., Looman, C. W. N., Nagelkerke, N. J. D., Habbema, J. D. F., and Engels, D. (2003). Quantification of clinical morbidity associated with schistosome infection in sub-saharan africa. *Acta Tropica*, 86(2-3):125–139.

Ver Hoef, J., Cressie, N., Fisher, R., and Case, T. (2001). Uncertainty and spatial linear models for ecological data. In *Spatial uncertainty in ecology: implications for remote sensing and GIS applications*, Chapter 10. Springer-Verlag, New York.

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Surv. Rev.*, XXII(176):88–93.

Vounatsou, P., Raso, G., Tanner, M., N'Goran, E. K., and Utzinger, J. (2009). Bayesian geostatistical modelling for mapping schistosomiasis transmission. *Parasitology*, 136(13):1695–1705.

Wang, X., Zhou, X., Vounatsou, P., Chen, Z., Utzinger, J., Yang, K., Steinmann, P., and Wu, X. (2008). Bayesian spatio-temporal modeling of *Schistosoma japonicum* prevalence data in the absence of a diagnostic 'gold' standard. *PLoS Neglected Tropical Diseases*, 2(6):e250.

Watts, S. J. (1987). Population mobility and disease transmission: the example of guinea worm. *Social Science & Medicine (1982)*, 25(10):1073–1081.

Wenlock, R. W. (1977). The prevalence of hookworm and of *S. haematobium* in rural zambia. *Tropical and Geographical Medicine*, 29(4):415–421.

WHO (2002). Prevention and control of schistosomiasis and Soil-Transmitted helminthiasis: Report of a WHO expert committee. Technical Report WHO Tech Rep Ser 912, WHO, Geneva.

WHO (2006a). Neglected tropical diseases. hidden successes, emerging opportunities. Technical report, Department of Control of Neglected Tropical Diseases, WHO, Geneva, Switzerland.

WHO (2006b). Preventive chemotherapy in human helminthiasis: coordinated use of anthelminthic drugs in control interventions: a manual for health professionals and programme managers. Technical report, World Health Organization, Geneva, Switzerland.

WHO (2008). World malaria report 2008. Technical report, WHO Press.

WHO (2010). Schistosomiasis. *Weekly Epidemiological Record / Health Section of the Secretariat of the League of Nations*, 85(18):158–164.

WHO and UNICEF (2010). Progress on sanitation and drinking-water - 2010 update. Technical report, WHO, UNICEF.

Widenius, M. and Axmark, D. (2002). *Mysql Reference Manual*. O'Reilly & Associates, Inc., 1st edition.

Wieczorek, J., Guo, Q., and Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745767.

Wilkins, H. A., Goll, P., Marshall, T. F., and Moore, P. (1979). The significance of proteinuria and haematuria in *Schistosoma haematobium* infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 73(1):74–80.

Woolhouse, M. E. J. (1998). Patterns in parasite epidemiology: the peak shift. *Parasitology Today*, 14(10):428–434.

Woolhouse, M. E. J., Taylor, P., Matanhire, D., and Chandiwana, S. K. (1991). Acquired immunity and epidemiology of *Schistosoma haematobium*. *Nature*, 351(6329):757–759.

Yang, G., Utzinger, J., Lv, S., Qian, Y., Li, S., Wang, Q., Bergquist, R., Vounatsou, P., Li, W., Yang, K., Zhou, X., and Xiao-Nong Zhou, R. B. (2010). The regional network for asian schistosomiasis and other helminth zoonoses (rnas+): Target diseases in face of climate change. In *Important Helminth Infections in Southeast Asia: Diversity and Potential for Control and Elimination, Part B*, volume Volume 73, pages 101–135. Academic Press.

Yang, G., Vounatsou, P., Zhou, X., Tanner, M., and Utzinger, J. (2005a). A bayesian-based approach for spatio-temporal modeling of county level prevalence of *Schistosoma japonicum* infection in jiangsu province, china. *International Journal for Parasitology*, 35(2):155–162.

Yang, G., Vounatsou, P., Zhou, X., Utzinger, J., and Tanner, M. (2005b). A review of geographic information system and remote sensing with applications to the epidemiology and control of schistosomiasis in china. *Acta Tropica*, 96(2-3):117–129.

Yapi, Y. G., Briët, O. J. T., Diabate, S., Vounatsou, P., Akodo, E., Tanner, M., and Teuscher, T. (2005). Rice irrigation and schistosomiasis in savannah and forest areas of côte d'ivoire. *Acta Tropica*, 93(2):201–211.

Yu, J. M., de Vlas, S. J., Yuan, H. C., and Gryseels, B. (1998). Variations in fecal *Schistosoma japonicum* egg counts. *The American Journal of Tropical Medicine and Hygiene*, 59(3):370–375.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470.

Curriculum vitae

Nadine Schur
Date and place of birth: 29th October 1983 in Bad Muskau, Germany
Nationality: German

Education

2008-2011	PhD studies in Epidemiology at the Swiss Tropical and Public Health Institute, Basel, Switzerland on "Geostatistical modelling of schistosomiasis transmission in Africa" under the joint supervision of PD. Dr. P. Vounatsou and Prof. J. Utzinger
2007-2008	Master of Science (MSc) in Epidemiology at the University of Basel, Basel, Switzerland on "Geographical patterns and predictors of parasitaemia risk and haemoglobin level: Model-based mapping using the Zambia indicator survey data (ZMIS) 2006" under the supervision of PD. Dr. P. Vounatsou
2003-2007	Diploma (Dipl FH) in Biomathematics at the University of Applied Sciences Zittau/Görlitz, Zittau, Germany on "Bayesian geostatistical modelling of the Zambian national malaria indicator survey data"

Professional Activities and Teaching

2010	Lecturer in the Bayesian Disease Mapping course
2003	Statistics Lecturer in the College of Statistics, University of Bucharest
2006–2007	Statistics Lecturer at the European Course in Tropical Epidemiology

Achievements and Conferences

2010	"Runners-Up" in Young Investigator Award session at ASTMH 2010

2010	ASTMH meeting, oral presentation on "Geostatistical model-based estimates of schistosomiasis risk in West Africa"
2010	ISBA meeting 2010, poster presentation on "Modelling the geographical distribution of co-infection risk from single disease survey data"
2009	GnosisGIS symposium 2009, oral presentation on "Bayesian geostatistical modelling of co-infection risk from single disease survey data"

PUBLICATIONS

Schur N, Ndir O, Utzinger J, Vounatsou P (2011). **Bayesian modeling of anisotropic geostatistical data: An application in mapping urinary schistosomiasis in Senegal**. *Statistics in Medicine* (under review).

Hürlimann E, **Schur N**, Boutsika K, Stensgaard AS, Laserna de Himpsl M, Ziegelbauer K, Laizer N, Camenzind L, Di Pasquale A, Ekpo UF, Simoonga C, Mushinge G, Saarnak CFL, Utzinger J, Kristensen TK, Vounatsou P (2011). **Toward an open-access global database for mapping, control, and surveillance of neglected tropical diseases**. *PLoS Neglected Tropical Diseases* (in press).

Stensgaard AS, Utzinger J, Vounatsou P, Hürlimann E, **Schur N**, Saarnak CFL, Mushinge G, Simoonga C, Kabatereine NB, Tchuem Tchuenté LA, Rahbek C, Kristensen TK (2011). **Large-scale determinants of intestinal schistosomiasis and intermediate host snail distribution across Africa: does climate matter?** *Acta Tropica* (under review).

Schur N, Hürlimann E, Stensgaard AS, Chimfwembe K, Mushinge G, Simoonga C, Kabatereine NB, Kristensen TK, Utzinger J, Vounatsou P (2011). **Spatially explicit *Schistosoma infection* risk in eastern Africa based on Bayesian geostatistical modelling**. *Acta Tropica* (in press).

Schur N, Gosoniu L, Utzinger J, Vounatsou P (2011). **Modelling the geographical distribution of co-infection risk from single disease surveys**. *Statistics in Medicine* 30(14): 1761-1776.

Schur N, Hürlimann E, Garba A, Traoré MS, Ndir O, Ratard RC, Tchuem Tchuenté LA, Kristensen TK, Utzinger J, Vounatsou P (2011). **Geostatistical model-based estimates of schistosomiasis risk in West Africa**. *PLoS Neglected Tropical Diseases* 5(6): e1194.

Schur N, Utzinger J, Vounatsou P (2011). **Modelling age-heterogeneous *Schistosoma haematobium* and *S. mansoni* survey data via alignment factors**. *Parasites & Vectors* 4(1): 142.

Riedel N, Vounatsou P, Miller JM, Gosoniu L, Chizema-Kawesha E, Mukonka V, Steketee RW (2010). **Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modelling of the 2006 Zambia national malaria indicator survey (ZMIS)**. *Malaria Journal* 9, 37.