Bayesian geostatistical variable selection and prediction of tropical diseases

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Dimitrios-Alexios Karagiannis-Voules
aus Griechenland

Basel, 2016

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

Basel, den 10. November 2015

Prof. Dr. Jörg Schibler
Dekan

Summary

A global commitment from governmental, non-profit, research and even profit organizations to combat tropical diseases has led to an increase of funding for implementing control interventions. Guidelines for controlling a disease commonly depend on its prevalence or incidence. Information on the disease risk distribution is important for successful control implementation. Spatial statistical modelling provides a framework to predict disease risk at high spatial resolution, assess disease dynamics and evaluate the effects of interventions.

In sub-Saharan Africa, estimates of soil-transmitted helminthiasis risk and of treatment requirements are lacking, mainly due to scarcity of georeferenced data and inaccessibility of the available ones. There is a need to bridge this gap for cost-effective disease control, monitoring and evaluation. Soil-transmitted helminthiasis is a poverty-related disease. Socioeconomic proxies (SES), such as socioeconomic status, access to safe water and sanitation (WASH) facilities, could improve predictive risk modelling. Socioeconomic data are available from household surveys and are georeferenced at village-level. It is unclear whether village-aggregated SES can improve predictions of disease risk. Brazil is one of the most affected countries with leishmaniasis. Despite the implementation of a notifiable system in the country, geostatistical analyses of leishmaniasis incidence are limited to few districts and provinces in the country. A countrywide analysis estimating the geographical and temporal distribution of the disease has not been carried out. There is a lot of progress in malaria control over the last years. Interventions are widely administered and repeated national surveys in Africa are conducted collecting spatial data on the disease risk and on a number of intervention coverage indicators. However, the effects of the ongoing interventions on malaria

risk have not been analysed. Model formulations that can estimate the effects of malaria interventions in space and time have not been established.

This thesis aims to address the above gaps of knowledge by developing data-driven Bayesian geostatistical models. The specific objectives of this research are to: (i) assess the spatiotemporal distribution, identify risk factors and calculate the number of infected Brazilians with leishmaniasis (Chapter 2); (ii) predict the distribution of soil-transmitted helminth (STH) infection risk in sub-Saharan Africa, evaluate temporal trends, and provide spatiotemporally explicit estimates of people infected and of treatment requirements by country (Chapter 3); (iii) predict the distribution of STH risk in Cambodia and evaluate the predictive ability of SES proxies in geostatistical disease modelling using individual and village-specific SES (Chapter 4); (iv) provide geostatistical models with spatially varying covariate effects and assess variable selection formulations to estimate effects of malaria interventions on disease risk (Chapter 5); and (v) develop geostatistical models with spatiotemporally varying covariate effects and evaluate sensitivity of predictive process approximation for variable selection of large data (Chapter 6).

In Chapter 2, we apply Bayesian geostatistical negative binomial models to analyze reported incidence data of cutaneous and visceral leishmaniasis in Brazil covering a 10-year period (2001-2010). Particular emphasis is placed on estimating spatial and temporal patterns. The number of cases are predicted at province and country levels.

In Chapter 3, we analyze soil-transmitted helminth infection risk in sub-Sahara Africa. Data are obtained from a systematic review and analyzed using geostatistical models. Areas where data are lacking but a high infection risk is predicted are highlighted. We calculate anthelmintic treatment needs by country using World Health Organization guidelines.

Chapter 4 presents a geostatistical analysis of soil-transmitted helminth infections in Cambodia. The study pursues an in-depth investigation of the use of socioeconomic predictors in mapping poverty-related diseases. Additional to the country-level analysis with SES aggreagated at village level, separate analyses are carried out using individual-level SES proxies to assess and quantify their associations with

soil-transmitted helminth infections. Analyses using individual and village-specific proxies are compared.

In Chapter 5, we provide geostatistical models with spatially varying coefficients for estimating effects of malaria interventions in space and assess sensitivity of variable selection approaches to model specification. The proposed models were fitted on malaria data from two national surveys in Angola to identify the best proxies of intervention coverage measures on malaria risk and find the provinces in the country that interventions have an important effect on the disease.

In Chapter 6, we develop a computational algorithm that we called iteratively integrated nested Laplace approximations (i-INLA) to perform variable selection of spatiotemporally varying coefficients of non-Gaussian data via a marginal likelihood approximation. We implemented the algorithm on the Angola malaria data to assess effects of interventions in space and time on the dynamics of malaria. We use the predictive process approximation to the spatial components of the models to speed inference. Effects of the predictive process approximation on variable selection are investigated.

This PhD thesis contributes to the fields of Bayesian spatial modelling and spatiotemporal epidemiology of tropical diseases with: (i) methodology for Bayesian variable selection of spatiotemporally varying coefficients allowing flexible inference, especially for computationally intensive geostatistical models of data collected over large number of locations; (ii) sensitivity analysis of Bayesian variable selection formulations of models with spatially varying coefficients; (iii) estimates of incidence rates for cutaneous and visceral leishmaniasis in Brazil depicting the current situation of leishmaniasis in the country; (iv) an open-access georeferenced database cataloguing all available survey data for soil-transmitted helminth infections in sub-Saharan Africa and Cambodia for disease control and research purposes; (v) up-to-date smooth risk maps, and estimates of the number of people infected and of the required treatments of soil-transmitted helminth infections in sub-Saharan Africa and Cambodia; (vi) an evaluation of the predictive ability of cluster-aggregated WASH and other SES-related proxies in disease mapping of poverty-related diseases; and (vii) geostatistical models of malaria risk for estimating effects of malaria intervention coverage measures across space and over time.

Acknowledgements

During this work, there have been many influencers, mentors, inspirers, friends, among others, whose contribution I will try to acknowledge below.

My main supervisor, Dr. Penelope Vounatsou, has been guiding me in all steps of this PhD. Penelope gave me the chance to develop different skills. With her spatial, statistical, epidemiological and scientific expertise she showed me how state-of-the-art research is conducted. I am filled with gratitude for being Penelope's student. I am also grateful to Penelope for building a group with a superb collegial environment and making me a member of it.

I wholeheartedly thank the director of Swiss TPH and my co-supervisor, Prof. Dr. Jürg Utzinger, for his multidirectional support. Jürg's epidemiological and parasitological expertise colored my training. One of his memorable impacts was his motivational (that slippery "l") discussions and emails that kept me going. ...

I am grateful to Dr. Dolores Catelan for accepting the role of external examiner and assessing this thesis.

I would like to acknowledge my educational past because it constitutes an important pillar of my personal development and because I often look back to it and understand more now than then. Specifically, I would like to thank Prof. Dr. Leonhard Held for being my Bayesian mentor. Many thanks to Prof. Dr. Stavros Kourouklis who introduced me to the world of statistics.

Being a member of the Bayesian Disease Modelling Group has been a delightful experience. Federica Giardina, Verena Jürgens and Patricia Biedermann have supported me socially and scientifically. My deepest thanks to Frédérique Chammartin

who was my personal, 24/7, mentor. Due to the continuous growth of the group, it is not possible to mention everyone's contribution and thus succinctly acknowledge: Nadine Schur, Ronaldo Scholte, Arthur Mai, Abbas Adigun, Eric Diboulo, Serena Bianco, the Gintowt couple, Elizaveta Semenova, Christian Hermann, Christos Kokaliaris, Guojing Yang.

The people in front and behind the scenes of Swiss TPH, such as the administration, HR, IT, teaching and other departments, have created an excellent environment for conducting research. The former director of the institute, Prof. Dr. Marcel Tanner, made Swiss TPH feel like home and I am grateful for it. In Swiss TPH, I was honoured to have supportive friends, exchange ideas in the stairs, have inspiring talks or collaborate with: Steffi Knopp, Armelle Forrer, Vreni Jean-Richard, Steffi Mauti, Eveline Hürlimann, Mirko Winkler, Aurelio Di Pasquale, Amanda Ross, Christian Schindler, Konstantina Boutsika, Medea Imboden, Peter Odermatt, Jakob Zinsstag.

During different projects, I had the opportunity to collaborate with many researchers from all over the world. Thank you all for excellent collaborations. I would like to acknowledge the INLA gurus Prof. Dr. Håvard Rue, Dr. Finn Lindgren and Dr. Daniel Simpson for their near-live support with INLA.

I am grateful for the funding I received from the Swiss National Science Foundation and the European Research Council, as well as from the Dissertationsfonds der Universtität Basel/Basler Studienstiftung for printing this thesis. The University of Basel and the Swiss School of Public Health provided a perfect infrastructure to carry out this PhD. The sciCORE stuff, and especially Martin Jacquot, were very helpful with the universities' clusters.

Throughout these years, I received support, for which I am thankful, from my friends: Christos Michos, Christos Chinaris, Gary Cooper, Thanos Marantos, Michalis Chinaris, Kostas Rinis, Eleftheria Kioupritzi, Johannes Dahm, Dimitris Zampetakis, Hesam Montazeri. I apologize in advance for missing some.

Katerina, thank you for standing by me, your incredible patience and caring.

I am grateful to Iro and Mary for their lifetime support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Rationale

Efforts to control or eliminate tropical diseases have been increasing over the last years. The combination of resources from governmental, non-profit, research and even profit organizations illustrate the commitment to combat tropical diseases. After the London Declaration, pharmaceutical companies have committed to donate drugs valued at US$17.8 billion until 2020 (http://unitingtocombatntds.org). In 2012, the United States Agency for International Development recommended the use of more than US$700 million funds to continue and extend existing programmes against neglected tropical diseases (NTDs) and malaria (http://www.gpo.gov). According to the Open Malaria Funding Data Platform (http://www.rollbackmalaria.org/financing/mfdp), the global funds against malaria and infectious diseases have scaled from millions in 2000 to billions in 2013.

Cost-effective implementation of control against a disease is crucial to maximize its return. Information on a disease's distribution can be used as an intervention planning tool. This stands for different stages of a control implementation. Namely, it can be used for evaluating treatment schedules, cost-effectiveness studies, calculation of required treatments, logistics, burden estimation etc. Currently, control interventions of a disease commonly depend on its prevalence or incidence. Therefore disease/infection distribution is important for stakeholders.

In practice, detailed information of a disease's distribution is rarely available for *e.g.* a whole country. Usually, surveys are conducted in villages or schools, or data are reported in some unit. Therefore, information might be missing over large areas of the country while surveyed points might not be representative of the control implementation unit they belong. For this reason, disease mapping has become very popular in the field of tropical diseases. *Disease mapping* has received different interpretations in the literature such as plotting of raw observed survey data, plotting aggregated raw data over an administrative unit and plotting of smooth estimates. The model-based version of the latter is currently considered as the "gold standard". Spatial statistical modelling is a rigorous established methodology that can be used for predicting a disease's risk distribution over a domain, calculating number of infected people, identifying predictors etc.

Hay et al. (2013) scored infectious diseases according to what type of mapping is recommended. Despite the fact that model-based spatial analyses of soil-transmitted helminth (STH) infection risk had been carried out and had demonstrated the usefulness of spatial statistics in the field, Hay et al. (2013) characterized two out of the three main STH infections as "do not map". Of course, the STH community reacted and in a "Simon says" game of replies a score corresponding to "model-based geostatistics" was reassigned. "Model-based geostatistics" was also assigned to malaria (caused by *Plasmodium falciparum* and *P. vivax*) while leishmaniasis was categorized as "niche modelling" (disease occurrence).

This thesis focuses on spatial statistical applications of the following infectious tropical diseases: leishmaniasis, soil-transmitted helminthiasis and malaria. The first two belong to a larger class of neglected tropical diseases while malaria is no longer classified as such.

1.2 Disease characteristics

1.2.1 Leishmaniasis

Leishmaniasis is caused by parasites of the genus *Leishmania* which is transmitted by sandflies. The disease occurs in human in two different clinical forms: (i) cutaneous (CL, referring to the greater group of American tegumentary leishmaniasis), which causes skin or mucosal lesion; and (ii) visceral (VL), which affects organs such as the liver and spleen (Utzinger et al., 2012).

Recently, 98 countries reported endemic transmission, with an estimated 0.7-1.2 and 0.2-0.4 million new cases per year for CL and VL, respectively. Deaths due to VL are estimated between 20,000 and 40,000 (Alvar et al., 2012); a clear drop of the 2002 estimate of 59,000 deaths (WHO, 2002b). The disability-adjusted life years (DALYs) for leishmaniasis in 2010 where estimated to approximately 3 million, showing a decrease of 43.6% since 1990 (Murray et al., 2012). Despite these trends, studies suggest that burden is, in fact, increasing (Desjeux, 2001, 2004). Alvar et al. (2012) warn that the 2010 DALY estimate has not included or supported data collection and field validation.

Globally, there are approximately 300 millions of population at risk of leishmaniasis (WHO, 2002b, 2013c), while 90% of VL cases are reported in six countries: Bangladesh, Brazil, Ethiopia, India, South Sudan and Sudan. The distribution of CL is more widespread. Figure 1.1 depicts the global statuses of endemicity of VL and CL in 2012 according to WHO (http://www.who.int/leishmaniasis/en/).

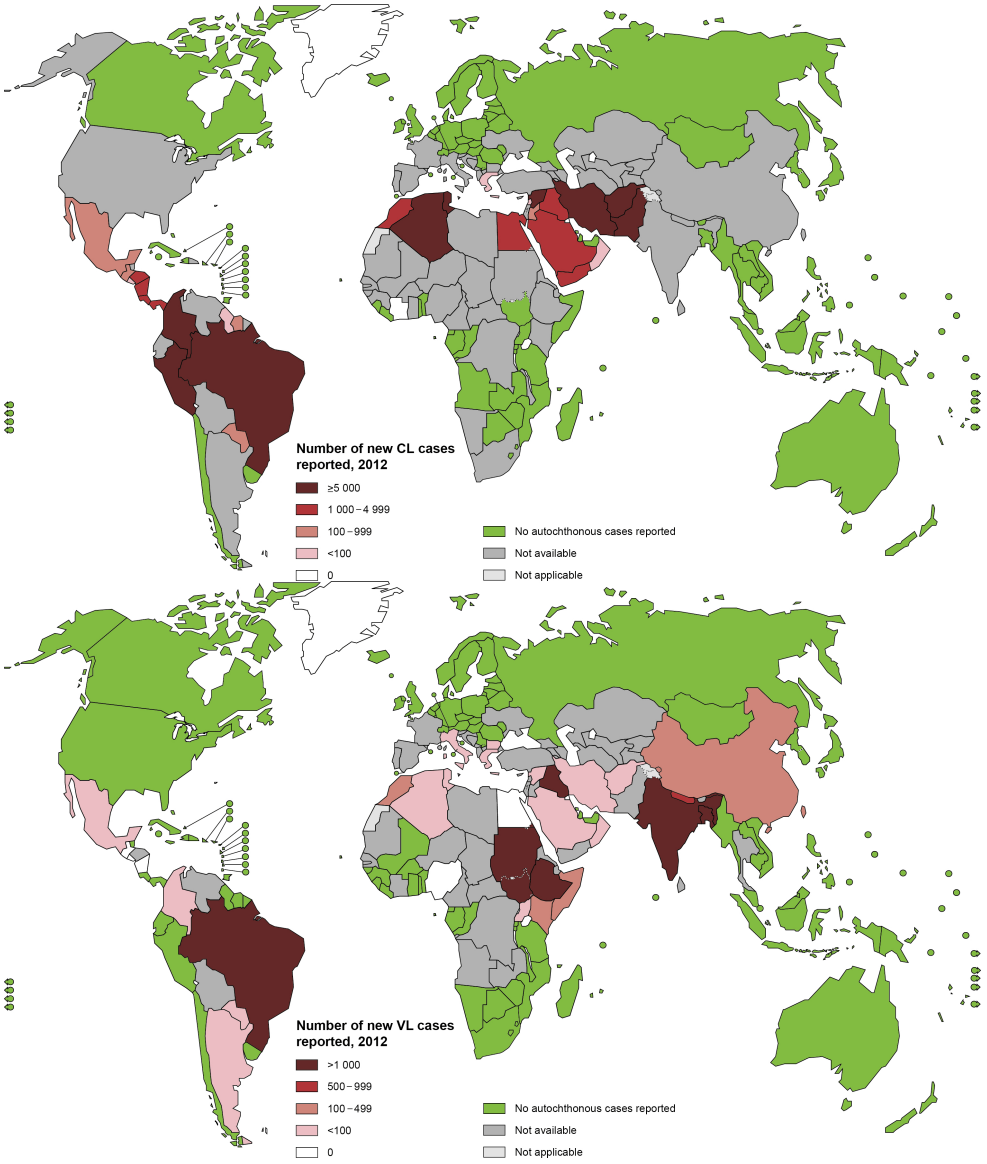


Figure 1.1: Endemicity status of CL (top) and VL (bottom) in 2012.

The control of leishmaniasis is complex due to the heterogeneity of sandfly species,

the different sub-forms of leishmaniasis and the reservoir hosts. The central pillar for control of leishmaniasis is constituted by effective and active case detection, diagnosis and treatment. Additional actions against leishmaniasis are reservoir-related control, vector control (through, for example, indoor spraying or insecticide treated nets) and environmental management (WHO, 2010a).

Desjeux (2004) pointed that the population with the lowest socioeconomic status (SES) is affected the most by leishmaniasis. Numerous studies have assessed the associations between environmental determinants and leishmaniasis. For example, moderate temperature and high vegetation have been associated with high incidence of CL (Valderrama-Ardila et al., 2010). A positive association of precipitation and VL has been reported by Ali-Akbarpour et al. (2012).

1.2.2 Soil-transmitted helminth infections

Soil-transmitted helminth infections refer to a group of parasitic infections caused by intestinal worms. *Ascaris lumbricoides*, *Trichuris trichiura* and hookworm (*Necator americanus* and *Ancylostoma duodenale*) are the three main parasites infecting people. These parasites are transmitted through the fecal-oral route.

Human helminthiases, account for the largest burden of NTDs (Utzinger et al., 2012; Murray et al., 2012; Hotez et al., 2014). Approximately 5 billion people were at risk of soil-transmitted helminthiasis and 1 billion people were estimated to be infected globally with at least one of the three main species in 2010. This estimate resulted in 5 million DALYs (de Silva et al., 2003; Pullan and Brooker, 2012; Murray et al., 2012; Pullan et al., 2014). These numbers decreased since 1990, when 2.5 billion infections were estimated to be infected, yielding 9 million DALYs.

A large proportion of the infections occurs in Asia where more than a quarter of the population is estimated to be infected with at least one intestinal worm. In sub-Saharan Africa, no decreasing trend was observed in the 2010 Global Burden of Disease study (Murray et al., 2012; Pullan et al., 2014) when comparing the situations of 1990 and 2010. The global environmental suitabilities and distributions of *A. lumbricoides*, *T. trichiura* and hookworm infections are illustrated in Figure 1.2 (Pullan and Brooker, 2012; Pullan et al., 2014).

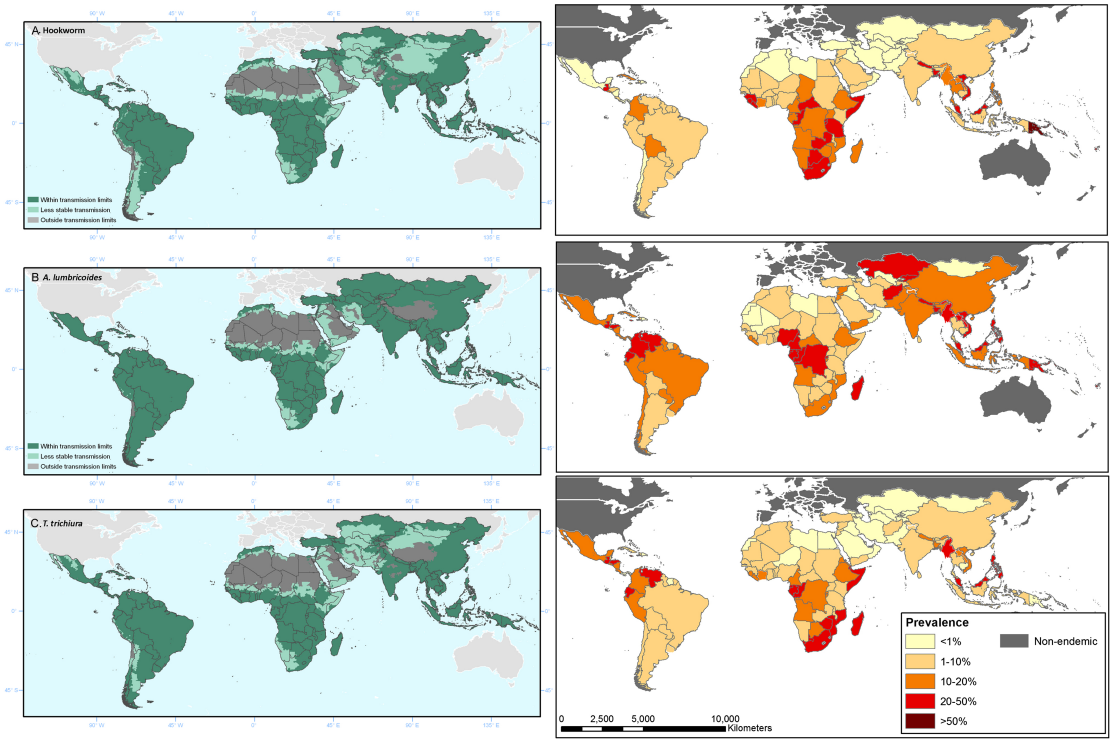


Figure 1.2: Global environmental suitabilities (left) and distributions (right) of hookworm infection (top), *Ascaris lumbricoides* (middle), and *Trichuris trichiura* (bottom).

For the control of soil-transmitted helminthiasis, the World Health Assembly resolution 54.19 in May 2001 (WHO, 2002a; Savioli et al., 2009) has endorsed and urged preventive chemotherapy as well as the access to safe water, sanitation and health education. As a result, annual coverage rates for treatment with albendazole or mebendazole have considerably increased in recent years, although they are still far below the targeted threshold of 75% (WHO, 2010b, 2014). The Global Program to Eliminate Lymphatic Filariasis (GPELF) and the African Programme for Onchocerciasis Control (APOC) have altogether administered billions of tablets of albendazole, mebendazole, and ivermectin treatments which impact on STH prevalence (see, for example, Ottesen et al., 2008). The increase of preventive chemotherapy together with socioeconomic development have lead to a decrease in STH infection prevalence (de Silva et al., 2003; Li et al., 2010; Utzinger et al., 2010). Programme coverages of STH and lymphatic filariasis control programmes as reported from WHO are provided in Section 3.5. Campbell et al. (2014) consider the

water, sanitation and hygiene (WASH) as key factors for successful and sustainable STH control.

As the mode of STH transmission suggests, access to clean water and developed sanitation contributes to interrupting transmission. Therefore socioeconomic status is a determinant of disease risk. Environmental factors are known to influence transmission. Altitude, humidity, temperature and precipitation, among others, have been shown to be associated with STH infection risk (see, for example, Brooker et al., 2006; Chammartin et al., 2013b).

1.2.3 Malaria

Infections by *Plasmodium* parasites cause malaria. Humans are infected through female mosquito (of the genus *Anopheles*) bites. Mosquitoes are the definitive hosts and humans the intermediate ones. The two most common parasites that infect humans are *P. falciparum* and *P. vivax*.

In 2013, malaria incidence and death rates per 100,000 populations were estimated to 2360.42 and 11.78, respectively. The annual change of both rates since 2000 has been estimated to approximately -3% (Murray et al., 2014). The highest incidence rates are estimated for Western sub-Saharan Africa and Oceania. DALYs in 2010 due to malaria have climbed to 82 million from 69 million in 1990, an increase of 19% (Murray et al., 2012). WHO alarms that every minute, a child dies in Africa from malaria (http://www.who.int/mediacentre/factsheets/fs094/en/, accessed July 2015).

Assessing the distribution of malaria has been a topic of research for many years. Lysenko and Semashko (1968) published the first world malaria endemicity map. Malaria was found endemic in many parts of the world such as central and south America, Africa, the Mediterranean region, Asia and Oceania. Since then, the Mapping Malaria Risk in Africa (Le Sueur et al., 1997) and the Malaria Atlas Project (Hay and Snow, 2006) initiatives have re-raised the interest and gave emphasis in mapping malaria (Dalrymple et al., 2015). Two model-based analyses depicted the global distribution of *P. falciparum* malaria in 2007 and 2010 (Hay et al., 2009; Gething et al., 2011). The disease is found mostly in sub-Saharan

Africa and Southeast Asia, Figure 1.3. Gething et al. (2012) evaluated the global distribution of *P. vivax* malaria.

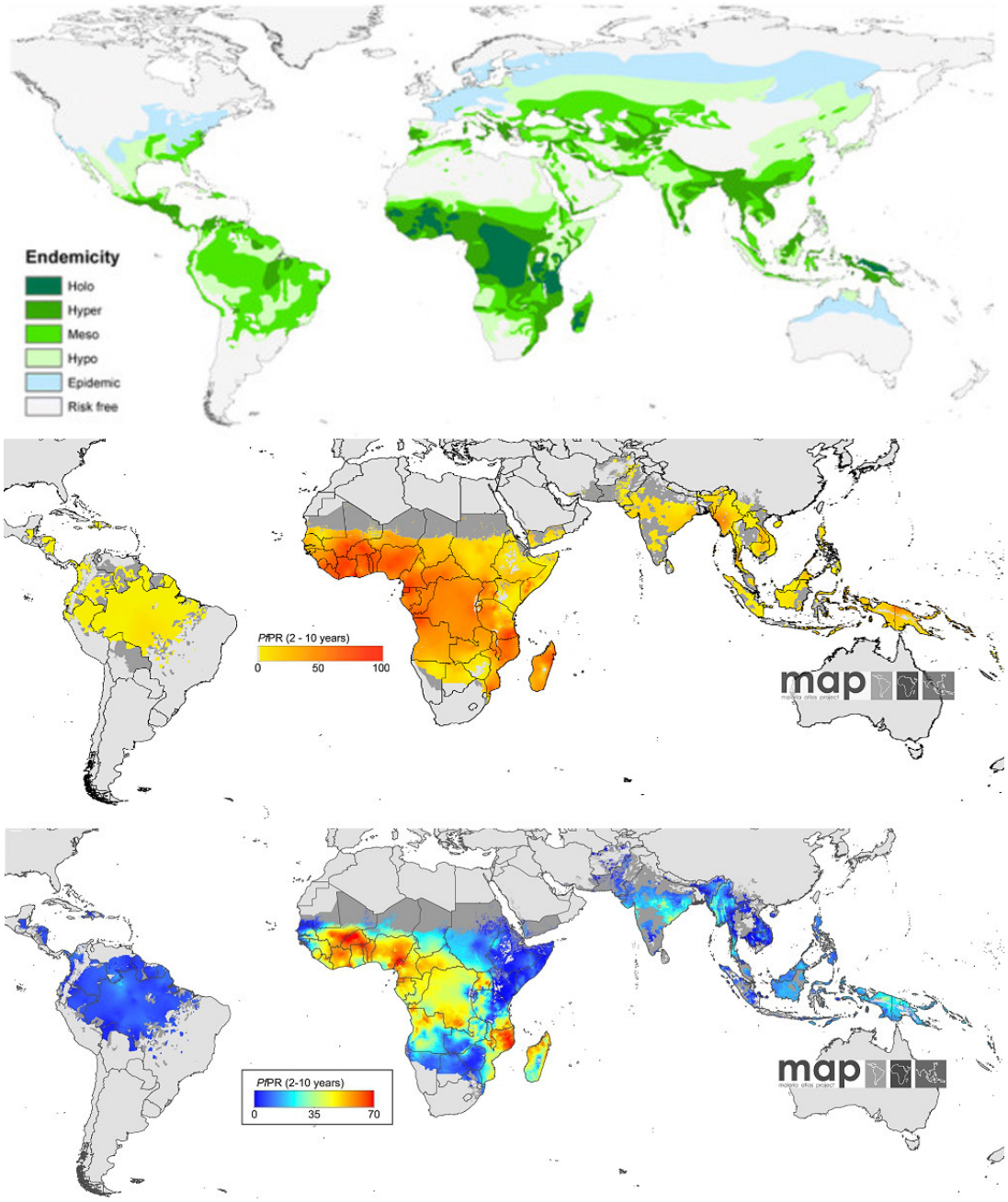


Figure 1.3: Malaria endemicity distribution in 1900's (top, Lysenko and Semashko, 1968, acccessed through Dalrymple et al., 2015), 2007 (middle, Hay et al., 2009) and 2010 (bottom, Gething et al., 2011).

Currently, the predominant means of interrupting and controlling malaria transmission is vector control. This is commonly performed through the distribution of insecticide-treated mosquito nets and indoor insecticide spraying. Although vaccination is not, at the moment, a part of control measures, research towards efficacious vaccines has advanced. The RTS, S/AS01 malaria (*P. falciparum*) vaccine has shown promising results (RTS,S Clinical Trials Partnership, 2015) and WHO states that in 2015 a decision will be taken to include or not this vaccine in existing control tools (http://www.who.int/mediacentre/factsheets/fs094/en/, accessed July 2015). An alternative direction of interventions that would target human rather than parasite factors, and would block the parasite entering the blood cells, has demonstrated significant potential in experiments on humanised mice (Zenonos et al., 2015).

Climatic conditions affect malaria transmission. Importantly, rainfall seasonality influences mosquito population and leads to seasonal transmission, since mosquito bites are increased during and after the rainy season. Gallup and Sachs (2001) characterized malaria as the disease of the global poor in view of being most prevalent in regions with low income countries.

1.3 Proxies of disease risk determinants

1.3.1 Climatic

Leishmaniasis, soil-transmitted helminthiasis and malaria are environmentally driven diseases. Climatic factors influence their transmission and can be associated with disease risk (or incidence) in spatial modelling. These associations are used to predict risk in locations where data are lacking at high spatial and temporal resolution. The use of remote sensing (RS) has lead to an abundance of climate-related data. Proxies of temperature, rainfall, vegetation, land use etc. are some of the factors that are available due to RS at high spatiotemporal resolutions. In addition, model-based surfaces in combination with raw measuremetns may provide supplementary environmental information such as soil characteristics or climatic scenarios.

1.3.2 Socioeconomic

Human factors like occupation, housing and socioeconomic status, play an important role in disease transmission. Therefore, they can be used as risk factors in statistical models. There are numerous wealth indices that may capture the socioeconomic status of an individual. The asset index combines different indicators through a principal component analysis (see, for example, Vyas and Kumaranayake, 2006). Access to improved (WHO and UNICEF, 2006) sanitation and drinking water sources may be considered for explaining STH infection risk. However, WASH-related indicators that could show an impact on STH infections are not yet defined by WHO (Campbell et al., 2014). Measuring SES proxies might be part of the disease-specific survey, which is commonly done for small scales or for well established designs as for malaria, or they might be part of a separate study.

1.3.3 Intervention coverage

The increase of funds, efforts, research and alarms to combat NTDs and malaria have resulted in an increase of interventions worldwide. Temporal and even spatial trends of disease risk might be explained from the implemented interventions. WHO reports country-level control efforts and rarely are data available for provinces or districts. Thus including control measures, for example preventive chemotherapy coverages for STH, in statistical models is not informative. For malaria, though, the Roll Back Malaria Partnership (RBMP) has designed surveys and questionnaires to consistently produce intervention-related data. The RBMP also defined cluster-level malaria intervention indicators that may depict the effect of control strategies (MEASURE Evaluation et al., 2013). These indicators correspond to mosquito nets' coverage of ownership and usage by the total population or specifically by pregnant women and children, as well as indoor residual spraying and case management, among others.

1.4 Geostatistical modelling of disease risk

The geostatistical models used in this thesis belong to the broader class of mixed models. Namely they have fixed and random components. Spatial dependence

is imposed in the random parts. This is commonly achieved by assuming a multivariate Gaussian (prior) distribution of the random, location-level, intercept that incorporates spatial dependence through its covariance matrix. Spatial random slopes may depict the varying, in space, effect of a predictor. Using Gaussian priors for the rest of the fixed or random parts, the model is a latent Gaussian model. The rest of the parameters (such as likelihood or random parameters of the Gaussian priors) will be referred to as hyper-parameters.

Due to the highly parameterized models and the unknown forms of posterior distributions, Bayesian inference in this thesis is conducted with Markov chain Monte Carlo (MCMC) simulations (Gelfand and Smith, 1990) and integrated nested Laplace approximations (INLA, Rue et al., 2009). MCMC is a sampling-based algorithm which samples from the posterior marginal distributions of the parameters. INLA is a deterministic algorithm that makes distributional assumptions (based on Taylor series) for the model's parameters. Specifically, it assumes that the conditional, on hyper-parameters, posterior distribution of the latent Gaussian part of the linear predictor is also Gaussian. Then, the joint posterior of this set of hyper-parameters is evaluated. Another Gaussian assumption for the joint posterior of the hyper-parameters completes the "nestedness" of the computational algorithm.

1.4.1 Large data

Inference requires lots of posterior evaluations to estimate a model's parameters. The geostatistical intercept (and, in addition, any geostatistical slope) has dimension equal to the number of data locations and even a single calculation of the multivariate normal density may be computationally expensive. This is due to the matrix computations that need to take place. This issue is called "the big N problem", where N is the number of locations. There are a number of approximations to achieve faster computations (for a recent review, see Lasinio et al., 2013). Banerjee et al. (2008) propose the predictive process approach that is based on a set of m locations, where $m < N$, on which the random intercept prior distribution is evaluated. Then, an unbiased estimate of the random effect at the N locations is calculated by using properties of the multivariate Gaussian distributions. The

method introduced by Lindgren et al. (2011) replaces the Gaussian random field with a Gaussian Markov random field (Rue and Held, 2005) using the stochastic partial differential equations (SPDE) approach. The Markovian structure of the covariance (precision) matrix accelerates all relevant computations. A method which tappers the covariance matrix by introducing zeroes for nearly independent locations has been proposed by Furrer et al. (2006). Kernel convolutions offer a flexible alternative for faster computations (Higdon et al., 1999). The big N problem has attracted interest for many years with Whittle (1954) using the fast Fourier transform to calculate the density of the multivariate normal distribution.

1.4.2 Misalignment

Data arising from different sources may be characterized by the fact that they are not measured on the same individuals and locations. For example, surveys that collect data on socioeconomic proxies are usually not coupled with parasitological examinations to measure STH infections (and vice versa), although Ziegelbauer et al. (2012); Strunz et al. (2014), among others, have shown such associations. As a result, SES proxies have not been extensively used in spatial analysis of STH infection risk.

Unless there exists an estimate of SES on the STH survey locations, associations cannot be quantified. Under the umbrella of spatial analyses, a natural approach would be to initially conduct a spatial analysis for SES indices and predict at STH survey locations. The problem with this approach is that the distribution of the SES prediction is not taken into account. A joint model of the SES and STH would address this issue. This would solve the location but not the individual misalignment. Namely, a SES prediction simply characterizes the location. Therefore, an association in a joint model could not identify that, for instance, the people with low SES (with an arbitrary unit) are the ones infected with STH in a specific location. A location-level (predicted) SES proxy characterizes the *e.g.* village and not each individual within it. In the case of individual data, such association could be identified. To sum up, to investigate potential associations of a predictor and disease risk in the case of spatial misalignment, a joint analysis should be used (to incorporate prediction distribution) but the results of it must

be carefully interpreted.

1.4.3 Variable selection

Identifying disease risk predictors is an important pillar of disease mapping. Associations of predictors and disease risk can supply additional information to decision makers. Accuracy of risk predictions depends on the usage of appropriate predictors.

Variable selection in disease mapping has been conducted through a variety of approaches. For long time, variable selection in disease mapping has been conducted by not taking into account the spatial intercept in bivariate and multivariate regressions (Clements et al., 2009; Soares Magalhães et al., 2011; Schur et al., 2011a; Raso et al., 2012, , among others). Stepwise selection approaches, ignoring as well the spatial component which is later assumed, have also been followed (see, for example, Clements et al., 2006). However, Chammartin et al. (2013a) used Bayesian geostatistical variable selection and showed that ignoring the geostatistical term might result in selecting a different (presumably wrong) set of predictors.

O'Hara and Sillanpää (2009) reviewed Bayesian variable selection methods. In the common case of indicator variable selection (see section 2.4 in O'Hara and Sillanpää, 2009), the Bayesian hierarchical formulation allows to include a variable selection component in the prior (George and McCulloch, 1996), likelihood, (Kuo and Mallick, 1998) or in both (Dellaportas et al., 2002).

Bayesian variable selection for spatial models has recently attracted some interest. For example, Wagner and Duller (2012) conduct a variable selection of random intercept for which a spatial analogue can be envisaged. For spatially varying coefficients, Reich et al. (2010) performed fixed or random slope selection for multivariate Gaussian response and Boehm Vock et al. (2015) used local variable selection through Gaussian Copula. Bayesian variable selection for large geostatistical data to address the "big N problem" has not yet been explored.

The use of approximate Bayesian inference (INLA) and large data approximations (SPDE in specific) allow fast model evaluations. Evaluating all possible models that are under consideration may be feasible. Therefore, model fitting measures such

as the deviance information criterion (Spiegelhalter et al., 2002) can be used for model selection. Cross-validatory measures constitute an alternative. Validating a model on a set of the data that was not used for fitting is an attractive choice when prediction is the goal. Gneiting and Raftery (2007) suggest the use of proper scoring rules to identify models that are both well-calibrated and sharp. Leave-one-out cross-validated scoring rules are based on each observation's prediction. INLA offers a fast calculation of such measures (Held et al., 2010).

1.5 Goal and objectives

The overarching goal of this PhD thesis is to assess the spatiotemporal dynamics of tropical diseases by applying, implementing and further developing Bayesian geostatistical models.

1.5.1 Specific objectives

The thesis pursues the following interrelated specific objectives:

(i) assess the spatiotemporal distribution, identify risk factors and calculate number of infected Brazilians with leishmaniasis (Chapter 2);

(ii) predict the distribution of STH infection risk in sub-Saharan Africa, evaluate temporal trends, and provide spatiotemporally explicit estimates of people infected and of treatment requirements by country (Chapter 3);

(iii) predict the distribution of STH risk in Cambodia and evaluate the predictive ability of SES proxies in geostatistical disease modelling of STH using individual and location-specific SES (Chapter 4);

(iv) provide geostatistical models with spatially varying covariate effects and assess variable selection formulations to estimate effects of malaria interventions on disease risk (Chapter 5); and

(v) develop geostatistical models with spatiotemporally varying covariate effects and evaluate sensitivity of predictive process approximation for variable selection of large data (Chapter 6).

Chapter 2

Bayesian geostatistical modeling of leishmaniasis incidence in Brazil

Karagiannis-Voules D.A.[1,2], Scholte R.G.C.[1,2,3], Guimarães L.H.[3,4], Utzinger J.[1,2], Vounatsou P.[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Centro de Pesquisas René Rachou, Fiocruz, Belo Horizonte, Brazil
[4] Serviço de Imunologia, Complexo Hospitalar Universitário Prof. Edgard Santos, Universidade Federal da Bahia, Bahia, Brazil

This paper has been published in *PLoS Neglected Tropical Diseases* 2013, 7: e2213.

Abstract

Background: Leishmaniasis is endemic in 98 countries with an estimated 350 million people at risk and approximately 2 million cases annually. Brazil is one of the most severely affected countries.

Methodology: We applied Bayesian geostatistical negative binomial models to analyze reported incidence data of cutaneous and visceral leishmaniasis in Brazil covering a 10-year period (2001-2010). Particular emphasis was placed on spatial and temporal patterns. The models were fitted using integrated nested Laplace approximations to perform fast approximate Bayesian inference. Bayesian variable selection was employed to determine the most important climatic, environmental, and socioeconomic predictors of cutaneous and visceral leishmaniasis.

Principal Findings: For both types of leishmaniasis, precipitation and socioeconomic proxies were identified as important risk factors. The predicted number of cases in 2010 were 30,189 (standard deviation (SD): 7,676) for cutaneous leishmaniasis and 4,889 (SD: 288) for visceral leishmaniasis. Our risk maps predicted the highest numbers of infected people in the states of Minas Gerais and Pará for visceral and cutaneous leishmaniasis, respectively.

Conclusions: Our spatially explicit, high-resolution incidence maps identified priority areas where leishmaniasis control efforts should be targeted with the ultimate goal to reduce disease incidence.

2.1 Introduction

Leishmaniasis is a group of neglected tropical diseases that are caused by parasites of the genus *Leishmania*. The parasites are transmitted by female phlebotomine sandflies and the disease occurs in human in two different clinical forms: (i) cutaneous (CL, referring to the greater group of American tegumentary leishmaniasis), which causes skin or mucosal lesion; and (ii) visceral (VL), which affects organs such as the liver and spleen (Utzinger et al., 2012). The latter, if not diagnosed and treated in the early stages, is usually fatal (Desjeux, 2004; Alves, 2009).

In 2002, the World Health Organization (WHO) estimated that 350 million people were at risk of leishmaniasis, with approximately 2 million (1.5 million CL and 0.5 million VL) cases and 59,000 deaths (WHO, 2002b). Recently, 98 countries reported endemic transmission, with an estimated 0.7-1.2 and 0.2-0.4 million new cases per year for CL and VL, respectively. Deaths due to VL are estimated between 20,000 and 40,000 (Alvar et al., 2012). The burden of leishmaniasis has been increasing worldwide (Desjeux, 2004, 2001). In Brazil, for example, the number of CL cases climbed from 6,335 in 1984 to 30,030 in 1996 (Brandão Filho et al., 1999). From 1990 to 2007 some 560,000 new cases of leishmaniasis were reported, primarily CL (Alves, 2009; Maia-Elkhoury et al., 2008). However, after 2005, the total number of CL cases has dropped and remained stable, just above 20,000.

Strategies for the control of leishmaniasis in Brazil have not changed over the past 60 years, which might explain why incidence did not decrease (Dantas-Torres and Brandão Filho, 2006). According to World Health Assembly (WHA) resolution 60.10, put forward in 2007, a well-defined implementation of a control program for leishmaniasis is still lacking (WHO, 2007). The difficulties in case reporting and detection are the main obstacles for such a program. At the same time, due to heterogeneity between the sandfly species, vector control introduces high costs. Effective control requires reliable maps of the spatial distribution of the disease, as well as the number of affected people, so that treatment and other control interventions can be implemented most cost-effectively.

Bayesian geostatistical models have been applied in the mapping of malaria (Gemperli et al., 2004; Hay et al., 2009; Gosoniu et al., 2012; Raso et al., 2012) and

neglected tropical diseases (Raso et al., 2005; Clements et al., 2009, 2010; Schur et al., 2011a). Geostatistical models relate the disease data with potential predictors and quantify spatial dependence via the covariance matrix of a Gaussian process facilitated by adding random effects at the observed locations. However, covariance matrix computations hamper implementation of the models on data collected over large number of locations ($> 1,000$). Different methodologies have been proposed to address this issue (for a recent review see Lasinio et al. 2013). A predictive process approach, developed by Banerjee et al. (2008), has been successfully applied in infectious disease mapping (see, for example, Schur et al. 2011a). Lindgren et al. (2011) showed that Gaussian Markov random fields (Rue and Held, 2005) can be used in geostatistical settings. Rue et al. (2009) provide fast computational algorithms for latent Gaussian models, based on integrated nested Laplace approximations (INLA).

There are only few studies that assessed the spatiotemporal distribution, including underlying risk factors, of leishmaniasis. Chaves and Pascual (2006) explored the temporal association of CL cases in Costa Rica by taking into account climatic variables. Chaves et al. (2008) used negative binomial models with breakpoints to analyze CL incidence in Costa Rica. Valderrama-Ardila et al. (2010) studied environmental determinants of CL incidence in an area of Colombia, using spatial models. In Colombia, the probability of CL presence based on ecological zones and environmental variables was explored by King et al. (2004). In Argentina, Salomon et al. (2012) modeled CL incidence using maximum entropy modeling. To date, efforts for estimating the associated risk and the predicted spatial distribution of leishmaniasis in Brazil are limited to small geographical areas. For instance, Shimabukuro et al. (2010) analyzed CL transmission in the state of So Paulo by using data on sandfly species presence, while Machado-Coelho et al. (1999) investigated spatiotemporal clustering in south-east Brazil. Jirmanus et al. (2012) examined seasonal variation of CL incidence in Corte de Pedra over a 20-year period and analyzed demographic characteristics of CL patients. Werneck and Maguire (2002) used spatial models, with one socioeconomic and one environmental covariate to explore VL incidence in the city of Teresina. Assunção et al. (2001) predicted VL rates in Belo Horizonte employing spatiotemporal models without

including climatic or socioeconomic covariates. The Ministry of Health (MoH) in Brazil has reported incidence maps for the whole country but without the use of predictors and of Bayesian geostatistical approaches Brasil Ministério da Saúde, Secretaria de Vigilância em Saúde (2007a,b). More recently, Alvar et al. (2012) provided worldwide estimates of leishmaniasis and included incidence maps of Brazil corresponding to raw data aggregated by state.

In this study, we analyzed incidence data of CL and VL obtained by the information system for notifiable diseases (ISND) during 2001 to 2010 from the MoH in Brazil. We employed Bayesian geostastical negative binomial models, fitted via INLA to predict the incidence of the diseases, using climatic, environmental, and socioeconomic covariates. We produced countrywide high resolution maps for leishmaniasis and estimated the number of infected people at the unit of the state. The generated incidence maps and estimates might be useful for decision-makers to prioritize intervention areas, and optimizing resources allocation to render control and elimination efforts most cost-effective.

2.2 Materials and Methods

2.2.1 Ethics Statement

We report a geospatial analysis of CL and VL incidence data in Brazil. The data were readily obtained from existing databases. Hence, there are no specific ethical considerations.

2.2.2 Leishmaniasis Incidence Data

Annual incidence data extracted from ISND, were obtained from 3,895 (for CL) and 2,176 (for VL) municipalities of Brazil. We have considered autochthonous cases. The municipalities chosen for the analysis were the ones with reported cases (including zeros) for at least one year between 2001 and 2010. Figure 2.1 shows the municipalities with incidence data and the 10-year mean incidence rate for both CL and VL.

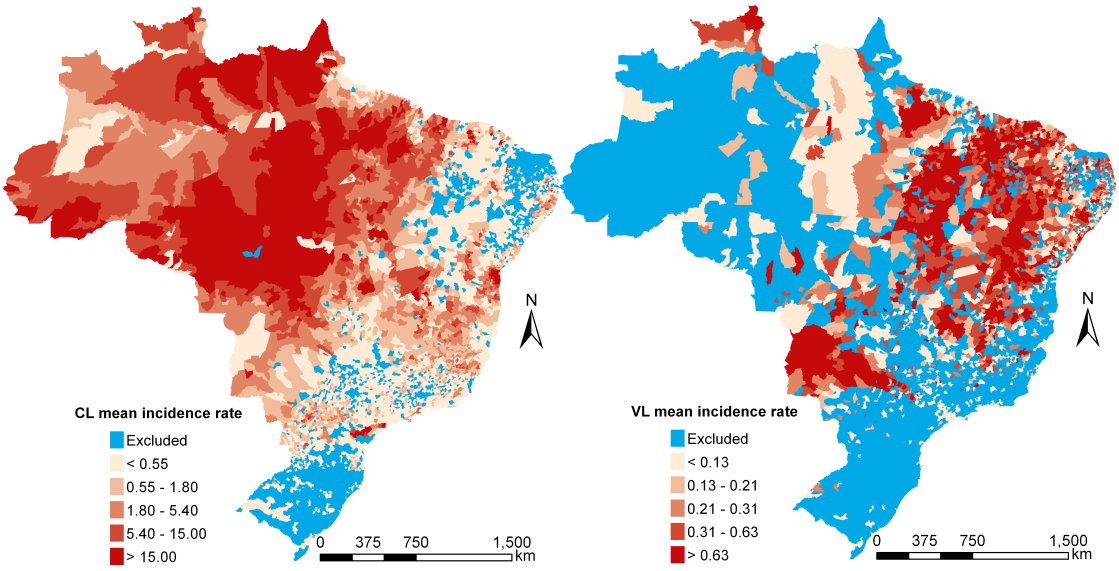


Figure 2.1: Raw incidence rates (per 10,000) averaged over a 10-year period (20012010) for cutaneous leishmaniasis (left) and visceral leishmaniasis (right). Municipalities colored in blue, were excluded from analysis due to missing data.

2.2.3 Climatic and Environmental Data

Climatic data, including altitude, were extracted from Worldclim Global Climate Data (Hijmans et al., 2005). These data consist of 19 bioclimatic variables. Environmental data were obtained from MODIS (Oak Ridge National Laboratory Distributed Active Archive Center, 2011). Land surface temperature (LST) data were used as proxies of day and night temperature. The normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) were considered as proxies for moisture and vegetation. Details of the data sources are summarized in Table 2.1. Municipality level estimates were obtained in ArcMap (Environmental Systems Research Institute, 2010) by aggregating the high resolution data.

Table 2.1: Climatic and environmental predictors used for geostatistical modeling of leishmaniasis in Brazil.

Source	Data type	Data period	Temporal resolution	Spatial resolution
Shuttle Radar Topography Mission data	Digital elevation model	2000	Once	1 km
Moderate Resolution Imaging Spectroradiometer (MODIS)/Terra	Land surface temperature for day and night	2005-2009	8 days	1 km
	Normalized difference vegetation index	2005-2009	16 days	1 km
	Enhanced vegetation index	2005-2009	16 days	1 km
Worldclim global climate	Annual mean temperature	1950-2000	Once	1 km
	Mean temperature diurnal range	1950-2000	Once	1 km
	Isothermality	1950-2000	Once	1 km
	Temperature seasonality	1950-2000	Once	1 km
	Maximum temperature of warmest month	1950-2000	Once	1 km
	Maximum temperature of coldest month	1950-2000	Once	1 km
	Temperature annual range	1950-2000	Once	1 km
	Mean temperature of wettest quarter	1950-2000	Once	1 km
	Mean temperature of driest quarter	1950-2000	Once	1 km
	Mean temperature of warmest quarter	1950-2000	Once	1 km
	Mean temperature of coldest quarter	1950-2000	Once	1 km
	Annual precipitation	1950-2000	Once	1 km
	Precipitation of wettest month	1950-2000	Once	1 km
	Precipitation of driest month	1950-2000	Once	1 km
	Precipitation seasonality	1950-2000	Once	1 km
	Precipitation of wettest quarter	1950-2000	Once	1 km
	Precipitation of driest quarter	1950-2000	Once	1 km
	Precipitation of warmest quarter	1950-2000	Once	1 km
	Precipitation of coldest quarter	1950-2000	Once	1 km

2.2.4 Socioeconomic Data

The socioeconomic indicators used in our study are summarized in Table 2.2. They include: (i) rural population and human development index (HDI) for the year 2000 provided by the Instituto Brasileiro de Geografia e Estatstica (IBGE); (ii) unsatisfied basic needs (UBN) for 2000 provided by the Pan American Health Organization (PAHO/WHO); and (iii) infant mortality rate (IMR) for 2000 and human influence index (HII) for 2005 obtained by the Center for International Earth Science Information Network (CIESIN) (Center for International Earth Science Information Network (CIESIN), Columbia University, 2000; Wildlife Conservation WCS, Center for International Earth Science Information Network (CIESIN), 2005).

Population data for 2010 at municipality level were available from IBGE, while population density at a spatial resolution of 5×5 km was obtained from CIESIN (Center for International Earth Science Information Network (CIESIN), Columbia University, 2005).

Table 2.2: Socioeconomic predictors used for geostatistical modeling of leishmaniasis in Brazil for 2001-2010.

Source	Data type	Data period	Resolution
IBGE (census data)	Population data	2010	Municipality
	Human development index (HDI)	2000	Municipality
	Rural population	2000	Municipality
PAHO (unsatisfied basic needs) (census data)	Bras0_3 (% of pupils enrolled in primary school)	2000	Municipality
	Bras0_4 (% of pupils completing primary school)	2000	Municipality
	Bras0_5 (rate literacy 15 to 24 years)	2000	Municipality
	Bras0_6 (girls and boys primary school)	2000	Municipality
	Bras0_7 (girls and boys high school)	2000	Municipality
	Bras0_8 (girls and boys undergraduate school)	2000	Municipality
	Bras0_9 (relation literacy women and men 15 to 24 years)	2000	Municipality
	Bras_10 (% women with non farming occupation)	2000	Municipality
	Bras0_11 (% people with potable water at home)	2000	Municipality
	Bras0_12 (% people with sanitation at home)	2000	Municipality
	Bras0_13 (% people with energy at home)	2000	Municipality
	Bras0_14 (% people that own their house)	2000	Municipality
	Bras0_15 (index secure tenure house)	2000	Municipality
	Bras0_16 (unemployment rate)	2000	Municipality
	Bras0_17 (% of houses with phone)	2000	Municipality
	Bras0_18 (% of house with computer)	2000	Municipality
	Bras2_11 (% of people overcrowding)	2000	Municipality
	Bras2_15 (% of people subsistence)	2000	Municipality
	Infant mortality rate (IMR)	2000	Municipality
CIESIN	Human influence index (HII)	2005	1 km

2.2.5 Statistical Analysis

The incidence data were modeled via negative binomial regression. Exploratory analysis was carried out in R R (R Core Team, 2014) to assess linearity of the covariates. For continuous covariates, we constructed three new categorical variables with 2, 3, and 4 categories, based on the quantiles of the variables distribution. The Akaikes information criterion (AIC) was used to select between a categorical or a linear form of each variable. To quantify the temporal trend, we included a binary variable, splitting the 10-year period in two phases, 2001-2005 and 2006-2010. Gibbs variable selection (Dellaportas et al., 2002) was performed in WinBUGS (Lunn et al., 2000) with the inclusion of an independent random effect at municipality level and a year specific auto-correlated term. All the covariates were assigned a 0.5 prior probability to be included in the final model. The total number of candidate covariates was 45. The covariates giving rise to the model with the highest posterior probability were subsequently used to fit a Bayesian geostatistical negative binomial model with spatially structured random effects at municipality

level. The spatial correlation was considered to be decreasing with distance between any pair of locations. The temporal random effects were modeled by auto-regressive terms of order 1. More specifically, we assumed that the reported number of CL and VL cases, for location i and year t, follow a negative binomial distribution with mean μ_{it} and dispersion parameter κ. Covariates and random effects were modeled on the log scale of μ_{it} , that is $\log(\mu_{it}) = \log(P_i) + X_{it}^T\beta + w_i + \epsilon_t$, where X_{it} and β are the vectors of covariates and coefficients, respectively, and P_i is the population of the i-th municipality. The spatial random effects $w = (w_1,)$ take into account the spatial dependence of the data by assuming they follow a zero-mean multivariate normal distribution with Matérn covariance function (see, for example, Banerjee et al. 2004a). ϵ_t is the auto-correlated error term with $\epsilon_t \sim \mathcal{N}(\rho\epsilon_{t-1}, \tau_2^2)$ for $t > 1$, and $\epsilon_1 \sim \mathcal{N}(0, \tau_1^2)$ with $\tau_1^2 = \tau_2^2/(1-\rho^2)$, and ρ is the auto-correlation. The large number of municipalities included in our modeling approach challenges geostatistical model fit, and thus resulting in extremely slow Markov chain Monte Carlo (MCMC) runs. To overcome computational burden, we estimated model parameters via INLA, using the homonymous R-package (available at www.r-inla.org). Details on model fit are provided in the Appendix.

Model validation was performed by fitting the model to a randomly selected subset of 80% of the locations and predicting the mean of the remaining 20% (test data). Bayesian credible intervals (BCI) of 95% probability are calculated and the percentage of observations included in these intervals is reported (coverage), as well as the square root of the mean square error (RMSE) of the test data.

A number of municipalities had not reported any cases of leishmaniasis for some years. As it was unclear whether these missing values in our dataset corresponded to true zeros or a lack of reporting cases, a separate analysis was carried out with missing values considered as zeros.

2.3 Results

2.3.1 Descriptive Results

Figure 2.2 shows the annual incidence rates of CL and VL per 10,000 people in Brazil for the period 2001-2010. A decrease of CL rates is observed after 2005,

while VL rates remained stable. The maximum annual number of cases at the unit of the municipality was 1,820 for CL (Manaus) and 262 for VL (Araguaína).

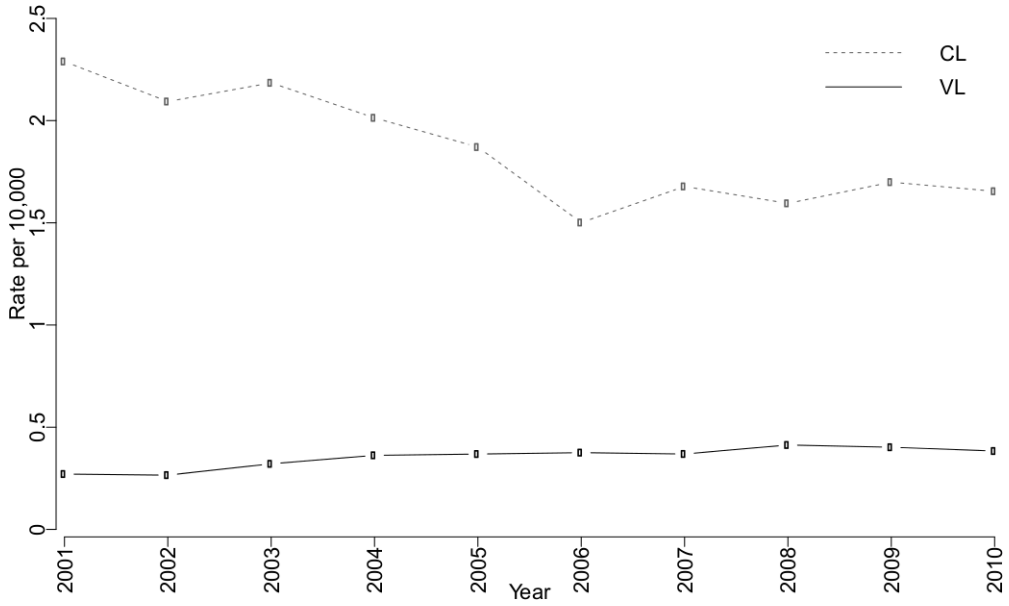


Figure 2.2: Temporal trend of observed countrywide incidence rates per 10,000.

2.3.2 Model estimates

Estimates, BCIs, and confidence intervals (CIs) of the multivariate Bayesian geostatistical and non-spatial models for CL are presented in Table 2.3. After 2005, the incidence of CL dropped by approximately 20%, which is in line with the results shown in Figure 2.2. Higher temperature diurnal range, temperature of wettest quarter, annual precipitation, precipitation seasonality, precipitation of warmest quarter, and EVI are positively associated with CL. On the other hand, higher LST is negatively associated with CL incidence. The following socioeconomic variables were associated with low incidence rates of CL: percentage of people with potable water at home, percentage of people with sanitation, percentage of people that own their house, and HII. A higher incidence rate was observed for men, as revealed by the negative relation between the CL incidence and the percentage of women living in an area.

Table 2.3: Parameter estimates for cutaneous leishmaniasis (CL) in Brazil for 2001-2010.

Variable	Bayesian geostatistical IRR (95% BCI)	Non-spatial IRR (95% CI)
Mean temperature diurnal range (°C)		
< 9.36	1.00	1.00
9.36-10.90	1.46 (1.19, 1.78)	1.00 (0.94, 1.06)
10.90-11.86	1.79 (1.42, 2.27)	1.15 (1.08, 1.22)
> 11.86	2.08 (1.56, 2.75)	1.62 (1.50, 1.75)
Mean temperature of wettest quarter (°C)	1.30 (1.18, 1.44)	1.19 (1.16, 1.22)
Annual precipitation (mm)	1.70 (1.54, 1.88)	1.24 (1.21, 1.27)
Precipitation seasonality	1.71 (1.50, 1.95)	1.13 (1.10, 1.16)
Precipitation of warmest quarter (mm)		
< 207	1.00	1.00
207-369	1.20 (0.99, 1.44)	1.18 (1.12, 1.25)
369-530	1.29 (1.54, 1.88)	1.67 (1.55, 1.81)
> 530	0.88 (0.66, 1.15)	0.74 (0.68, 0.81)
EVI		
< 35.78	1.00	1.00
35.78-39.06	1.31 (1.18, 1.46)	1.70 (1.61, 1.79)
39.06-42.73	1.65 (1.45, 1.89)	1.82 (1.71, 1.93)
> 42.73	2.14 (1.46, 2.54)	2.39 (2.22, 2.57)
Day LST (°C)	0.74 (0.66, 0.83)	0.81 (0.78, 0.83)
% People with potable water at home		
< 40.57	1.00	1.00
40.57-71.72	1.00 (0.90, 1.12)	1.18 (1.12, 1.25)
71.72-95.69	0.78 (0.67, 0.92)	0.56 (0.52, 0.60)
> 95.69	0.68 (0.56, 0.84)	0.39 (0.36, 0.43)
% People with sanitation at home	0.81 (0.76, 0.86)	0.82 (0.79, 0.84)
Proportion of own-rent house	0.92 (0.88, 0.96)	0.90 (0.88, 0.92)
% of women	0.82 (0.77, 0.86)	0.74 (0.72, 0.76)
HII		
< 17.02	1.00	1.00
17.02-20.30	0.86 (0.76, 0.98)	0.79 (0.75, 0.84)
20.30-23.48	0.73 (0.63, 0.85)	0.54 (0.50, 0.57)
> 23.48	0.70 (0.59, 0.83)	0.45 (0.42, 0.48)
Period		
2001-2005	1.00	1.00
2005-2010	0.80 (0.67, 0.95)	0.83 (0.80, 0.86)
	Mean (95% BCI)	
σ^2**(spatial variance)**	1.45 (1.35, 1.56)	
Range (km)	88.3 (82.2, 94.9)	
τ_2^2 **(temporal variance)**	0.02 (0.01, 0.03)	
ρ **(temporal correlation)**	0.74 (0.30, 0.95)	
κ **(dispersion)**	2.23 (2.15, 2.32)	

Parameter estimates of VL are summarized in Table 2.4. The most suitable climatic and environmental factors for VL are: low altitude, low annual precipitation, increased temperature diurnal range, and none extreme precipitation during the warmest quarter. With regard to socioeconomic variables, similar as in CL, effects of the two socioeconomic variables (*i.e.*, percentage of people with sanitation at home and percentage of people that own their house) were associated with lower

incidence of VL. Mean temperature diurnal range was the only climatic variable associated with a lower rate of VL incidence.

Table 2.4: Parameter estimates for visceral leishmaniasis (VL) in Brazil for 2001-2010.

Variable	Bayesian geostatistical IRR (95% BCI)	Non-spatial IRR (95% CI)
Altitude (m)		
< 163	1.00	1.00
163-341	0.93 (0.75, 1.16)	0.76 (0.70, 0.84)
341-560	0.96 (0.74, 1.25)	0.70 (0.63, 0.78)
> 560	0.81 (0.61, 1.09)	0.53 (0.48, 0.60)
Mean temperature diurnal range (°C)		
< 9.00	1.00	1.00
9.00-10.38	1.17 (0.92, 1.48)	1.58 (1.45, 1.73)
10.38-11.80	1.81 (1.33, 2.47)	3.05 (2.79, 3.34)
> 11.80	2.47 (1.74, 3.48)	4.70 (4.26, 5.20)
Annual precipitation (mm)		
< 832	1.00	1.00
832-1212	0.89 (0.73, 1.10)	0.81 (0.74, 0.88)
1212-1512	0.64 (0.48, 0.85)	0.63 (0.57, 0.69)
> 1512	0.59 (0.42, 0.82)	0.59 (0.52, 0.65)
Precipitation of warmest quarter (mm)		
< 130	1.00	1.00
130-205	1.10 (0.89, 1.37)	1.25 (1.15, 1.36)
205-359	0.88 (0.67, 1.15)	1.11 (1.02, 1.21)
> 359	0.54 (0.39, 0.76)	0.68 (0.60, 0.76)
Precipitation of coldest quarter (mm)	1.12 (0.97, 1.29)	1.26 (1.20, 1.31)
%People with sanitation at house		
< 2	1.00	1.00
2-25	0.91 (0.82, 1.02)	0.89 (0.83, 0.96)
> 25	0.62 (0.54, 0.73)	0.60 (0.55, 0.66)
Proportion of own-rent house		
< 81.51	1.00	1.00
81.51-87.23	0.88 (0.77, 1.00)	1.04 (0.96, 1.13)
87.23-90.76	0.88 (0.77, 1.01)	0.99 (0.91, 1.07)
> 90.76	0.71 (0.61, 0.83)	0.86 (0.79, 0.94)
Period		
2001-2005	1.00	1.00
2006-2010	1.16 (0.94, 1.35)	1.24 (1.18, 1.31)
	Mean (95% BCI)	
σ^2 **(spatial variance)**	1.09 (0.97, 1.23)	
Range (km)	109.1 (96.3, 124.6)	
τ_2^2 **(temporal variance)**	0.01 (0.00, 0.03)	
ρ **(temporal correlation)**	0.35 (-0.25, 0.86)	
κ **(dispersion)**	1.74 (1.62, 1.88)	

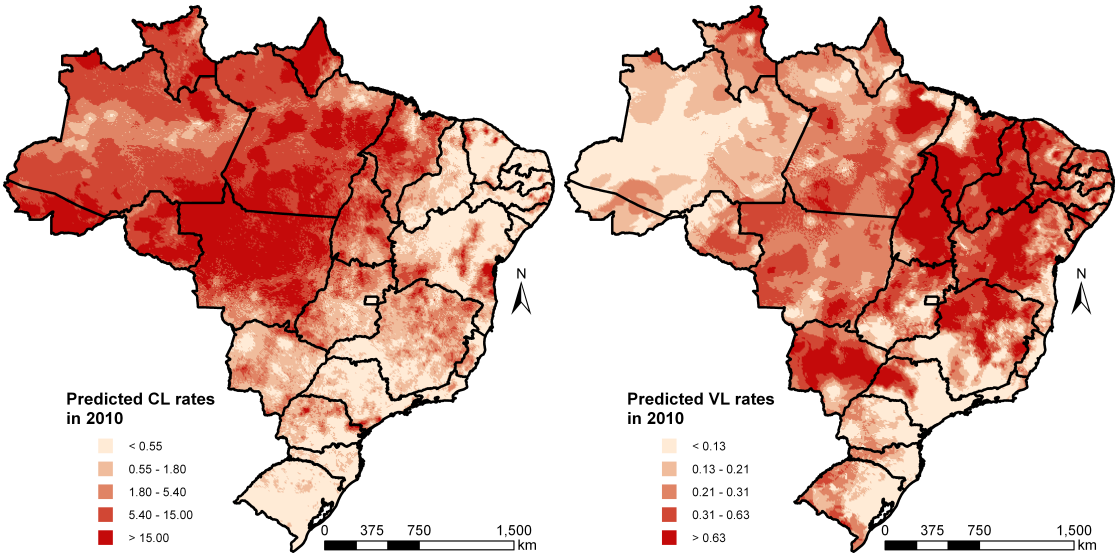
For both diseases the spatial variance was higher than the temporal one. Estimates of the range parameter indicate that spatial correlation becomes negligible for distances above 88.3 and 109.1 km for CL and VL, respectively.

2.3.3 Model Validation

The model of CL had a RMSE of 14.2 when predicted over the 20% randomly selected locations. One third of the cases (34%) were included in 95% BCIs of the posterior predictive distribution. The respective estimates for VL were 4.11 and 23%.

2.3.4 Incidence Maps

Model-based predictions were obtained over a grid of 136,841 pixels at 8×8 km spatial resolution. The rates (per 10,000 people) of the predictions for CL and VL in 2010 are depicted in Figures 2.3a and 2.3b, respectively.



(a) Cutaneous leishmaniasis. **(b)** Visceral leishmaniasis.

Figure 2.3: Geostatistical model-based predicted incidence rates per 10,000 in Brazil in 2010.

The decreasing trend of CL cases is apparent by comparing the maps for the year of 2010 (Figure 2.3a) with that of 2001 (Figure 2.4). For instance, in 2010 lower rates were observed in west and north-west Brazil in the states of Amazonas and Roraima.

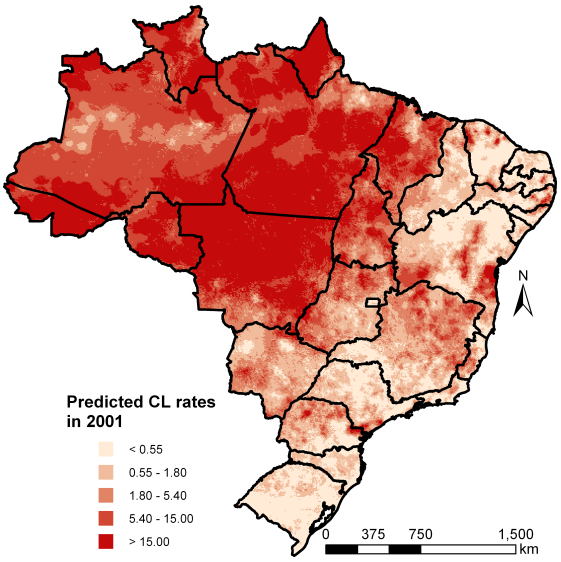


Figure 2.4: Geostatistical model-based predicted incidence rates per 10,000 for cutaneous leishmaniasis in Brazil in 2001.

Incidence maps under the assumption that missing cases were zeros are provided in the Appendix.

2.3.5 Country and State Estimates

The incidence rate map was overlaid with the population map of Brazil to estimate the number of cases per pixel. By aggregating the number of pixels per state, we estimated the number of infected people for both diseases (Table 5). The total number of cases predicted for 2010 was 30,189 (standard deviation (SD): 7,676) for CL and 4,889 (SD: 288) for VL. The highest prediction for CL occurred in the state of Par (4,332), while for VL in Minas Gerais (693). The corresponding country and state estimates under the assumption that missing cases were zeros are reported in the Appendix.

Table 2.5: Country and state predicted cases of cutaneous leishmaniasis (CL) and visceral leishmaniasis (VL) in Brazil in 2010.

State	CL cases (SD)	VL cases (SD)
Acre	1,511.0 (647.3)	7.8 (3.0)
Alagoas	151.7 (30.4)	115.4 (20.4)
Amapá	466.5 (52.1)	8.3 (4.9)
Amazonas	1,829.1 (858.0)	26.6 (6.2)
Bahia	3,402.3 (905.0)	467.1 (50.7)
Ceará	1,637.2 (345.0)	599.4 (100.6)
Distrito Federal	67.3 (32.4)	14.1 (5.6)
Espírito Santo	248.6 (75.7)	31.6 (8.0)
Goiás	634.6 (169.0)	89.0 (11.5)
Maranhão	3,417.3 (855.3)	500.0 (59.8)
Mato Grosso	3,383.2 (1461.1)	68.3 (9.9)
Mato Grosso do Sul	258.0 (488.8)	204.1 (65.0)
Minas Gerais	1,947.6 (110.4)	692.7 (67.7)
Pará	4,331.6 (1129.0)	406.6 (52.1)
Paríba	190.1 (195.7)	79.1 (10.7)
Paraná	1,082.6 (412.6)	82.3 (17.7)
Pernambuco	895.0 (40.2)	184.2 (25.4)
Piauí	199.4 (55.7)	276.0 (40.0)
Rio de Janeiro	281.7 (748.3)	48.0 (16.5)
Rio Grande do Norte	77.3 (17.5)	108.0 (15.2)
Rio Grande do Sul	182.4 (58.8)	109.0 (24.1)
Rondônia	1,896.8 (724.2)	32.1 (9.6)
Roraima	173.8 (171.5)	7.9 (2.6)
Santa Catarina	194.3 (78.8)	61.0 (18.7)
São Paulo	1,006.8 (90.6)	343.4 (28.3)
Sergipe	70.2 (15.7)	68.1 (11.4)
Tocantins	652.9 (229.7)	258.5 (40.1)
Total	30,189.1 (7675.8)	4,888.7 (288.3)

2.4 Discussion

We provide countrywide, model-based incidence maps for both cutaneous and visceral leishmaniasis in Brazil, at a high spatial resolution (8×8 km). Furthermore, we explored the underlying spatial processes, identified risk factors, and displayed high incidence areas. Taken together, our investigations provide a deeper understanding of the determinants of the two diseases. We employed Bayesian geostatistical models fitted on readily available incidence data from the MoH in Brazil, and used Bayesian variable selection to identify environmental and socioeconomic predictors. Although analyses for mapping leishmaniasis incidence data at state level were previously conducted, they rarely used rigorous statistical modeling approaches to take into account spatiotemporal correlations. However, ignoring correlation, risk factor analyses and predictions may be incorrect.

Our results indicate that humid warm climates with high vegetation indexes are associated with high incidence of CL. In contrast, high temperatures are associated with lower incidence of CL. A study in sub-Andean zone in Colombia (Valderrama-Ardila et al., 2010) also reported a negative association between incidence of CL and temperatures exceeding a minimum cut-off of 20.6°C. The association between vegetation and CL incidence found in our study, corroborates previous observations (Valderrama-Ardila et al., 2010) and may point to the role of deforestation driving CL outbreaks due to vector proliferations (Pupo Nogueira Neto et al., 1998). Our analysis suggests a higher incidence rate for males, which has also been reported by the MoH in Brazil (Brasil Ministério da Saúde, Secretaria de Vigilância em Saúde, 2007b). These observations might be explained by gender-specific occupational exposure within endemic areas (Klaus et al., 1999). The climatic conditions suitable for VL transmission are different to those of CL. A spatial analysis, done for the Islamic Republic of Iran, including environmental covariates, revealed that precipitation was positively associated with CL incidence (Ali-Akbarpour et al., 2012). On the other hand, the incidence of VL was not associated with the presence of vegetation and the role of annual precipitation is negative, which might reflect extreme conditions. An inverse relation of VL incidence and the mean of 3-year precipitation has been reported in a previous study in north-east Brazil (Thompson et al., 2002). VL shows higher incidence rates in lowlands as revealed by the negative altitude effect, which is in accordance with previous observations (Elnaiem et al., 2003).

There was an association between socioeconomic factors with the diseases incidence, confirming earlier reports that the population with the lowest socioeconomic status is affected the most (Desjeux, 2004). Indeed, the higher the proportion of people with access to clean water and improved sanitation, the lower the infection rate. In fact, control programs which focus on improving sanitation were associated with lower incidence rates. The intimate connection between poor living conditions and leishmaniasis has been discussed before (Werneck and Maguire, 2002).

Our analysis underscores the importance of rigorous geostatistical modeling in identifying factors related to transmission. Results from non-spatial analogue models may identify different predictors or even estimate a different direction of

the effects. The strong spatial correlations estimated by our models may suggest that we missed out important spatially structured predictors. For instance, vector and reservoir presence would drive such models. In addition, the analysis was based on incidence data aggregated over municipalities. Since the observed data are already available at municipality level, it is unlikely that predictions at the same level would be more informative. The strength of the predictive models is their ability to generate estimates in areas where no data are available. Data at higher spatial resolution may be able to obtain more precise estimates.

Incidence data were missing for some municipalities and some years in the 10-year observation period. These missing values could indicate true zero cases; however zeros have been recorded in the dataset in addition to the missing data. In our analysis we treated non-reported cases as missing. This may partially explain the overestimation of the total number of cases. To address this issue, we carried out a separate analysis, assuming that non-reported cases are zeros. The point estimates of predicted cases per state and the smooth maps are given in the Appendix. This analysis provided estimates of the total numbers of cases in the country which were closer to the reported ones in ISND. Maia-Elkhoury et al. (2007) estimated 42% and 45% (depending on source comparison) of under-reporting for VL in ISND using a capture-recapture method. Alvar et al. (2012) pointed that these percentages correspond to 1.3-1.7-fold degrees of under-reporting. Our total VL predicted cases fall within this interval. We are not aware of similar estimation of CL under-reporting for ISND. By assuming a similar amount of under-reporting for CL (due to the same source), the total number of predicted cases of our analysis lies within the above interval. Overestimation of the predicted cases may also arise because the incidence is very low and models cannot predict exact zeros. An estimate slightly higher than zero at pixel level will overestimate the total number of cases. The more pixels aggregated, the larger the overestimation. Hence, the model will overestimate, for example, treatment needs. Rounding to zero pixel-level cases predicted less than 0.1, the total number of model-based estimates of VL cases at country level drops to 3,320 from 4,889 and for CL to 28,164 from 30,189. However, this cut-off is arbitrary. For decision making, thresholds of predicted cases could be applied. These could be defined by some optimality criteria, which balance

cost of not providing timely treatment on one hand and cost of administering drugs which were not required on the other hand.

Our study has several limitations that are offered for consideration. Brazil is the fifth largest country of the world and can be divided into different ecological zones. We assumed a single relation between risk factors and the incidence of leishmaniasis, which might not be able to capture properly the geographical distribution. Non-stationary models allowing for different spatial dependencies and covariate distribution in a specific area (Banerjee et al., 2004b; Gelfand et al., 2003) may improve predictive ability. We did not include a space-time interaction, but instead assumed a constant spatial process over time. To perform such an analysis, data are needed for specific time periods and for each municipality. In our study, this would require either dropping a large number of municipalities from the study or incorporating the estimation of their values in the modeling process. The latter might result in identifiability problems of the parameters, and hence, we only considered additive effects. We assumed constant effects of the predictors over time and therefore could not explain the temporal trends of CL from the trends of the predictors considered in the study. The coverages of the test data for both diseases might seem low, but do not account for the zero cases. The 2.5% quantile cannot be zero, and thus all the zero incidence cases will be missed. To illustrate this, we rounded the lower quantile (which of course increases the credibility level) and recalculated the coverages resulting in 66% for CL and 71% for VL. In addition, 50% and 38% municipalities had 0 reported cases for CL and VL, respectively. Giardina et al. (2012) showed that zero-inflated (ZI) models gave better predictions than standard geostatistical models for predicting malaria risk using sparse malaria survey data. ZI models with an invariable probability of ZI were also fitted, but according to the deviance information criterion (DIC) they showed similar fits to the data and the probability of ZI was very low (of the magnitude 10^{-6}). Cross-validatory measures (*i.e.*, coverage and RMSE) did not improve when ZI models were fitted. Non-linearity was addressed by categorizing the predictors. Alternative approaches (*i.e.*, polynomial terms or splines) may provide more flexible ways to model the relation between disease and predictors, and potentially give more accurate estimates. We have chosen categorical covariates

because they offer easier epidemiological interpretation.

In conclusion, we present the first high-resolution model-based estimates of CL and VL in Brazil. We used INLA, a novel inferential approach in the field of neglected tropical diseases. Our incidence maps, together with the predicted number of CL and VL cases, constitute useful tools for decision making and prioritization of disease control intervention. Recent developments in Bayesian geostatistical computation (*e.g.*, INLA) already enable analyses of surveillance data in almost real time. Updates of these maps could be automatized, and hence performed shortly after data collection and reporting. We anticipate that in near future surveillance programs will integrate these methods in their systems. The possibility to aggregate over any desired level, such as the catchment area of health facilities, would further help planning drug delivery and other control measures. In particular, these maps could identify communities where enhanced prevention measures are warranted. Environmental predictors are important for identifying high incidence areas, while improving socioeconomic status might constitute the single most important factor to enhance control programs. The current methodology should be further developed to address the aforementioned limitations and provide more accurate spatial and temporal predictions of leishmaniasis incidence.

Acknowledgements

We acknowledge the help of Finn Lindgren, Daniel Simpson, and Håvard Rue for their inputs on INLA methodology and code and for providing access to a remote server. We thank the Ministry of Health in Brazil for providing the annual incidence data of leishmaniasis. We thank the PAHO Neglected Infectious Diseases Programme for their contribution to the data collection.

2.5 Appendix

In this section we present the model formulation, a brief description of the INLA approximation to estimate the marginal posterior distributions of the model parameters, and provide implementation details for the analysis of leishmaniasis data. Extensive theoretical explanations about INLA in a spatiotemporal setting have

been presented elsewhere (Cameletti et al., 2013).

Model formulation and INLA

Let Y_{it} be the number of cases for municipality i at year t. We assume that the Y_{it}'s are generated by a negative binomial distribution, *i.e.* $Y_{it} \sim \mathcal{NB}(\mu_{it}, k)$ with mean μ_{it} and dispersion parameter k. The linear predictor $\eta_{it} = \log(\mu_{it}) = \log(P_i) + X_{it}^T\beta + w_i + e_t$ includes an offset term for the population P_i, the vector X_{it}^T of covariates and their respective coefficients β, spatially and temporally structured random effects w_i and e_t, respectively. We consider that the vector of w_i arises from a multivariate normal distribution $w \sim \mathcal{MVN}(0, \Sigma)$ with Matérn covariance function between locations i, j that is, $\Sigma_{ij} = \frac{\sigma^2(\kappa d_{ij})^\nu K_\nu(\kappa d_{ij})}{\Gamma(\nu)2^{\nu-1}}$, where σ^2 is the spatial process variance, d_{ij} is the distance between the centroids of i and j, κ is a scaling parameter, ν is a smoothing parameter fixed to 1 in our application and K_ν is the modified Bessel function of second kind and order ν. The Matérn specification of the covariance matrix implies that the spatial range r, that is the distance at which spatial correlation becomes negligible (*i.e.* smaller than 10%) is $r = \frac{\sqrt{8}}{\kappa}$. We adopted a stationary autoregressive AR(1) process for e_t such that, $e_t \sim \mathcal{N}(\rho e_{t-1}, \tau_2^2)$ for $t > 1$ and $e_t \sim \mathcal{N}(0, \tau_1^2)$, where $\tau_1^2 = \tau_2^2/(1-\rho^2)$ and ρ the auto-correlation parameter, constraint in the interval $(-1, 1)$. We complete Bayesian model formulation by specifying prior distributions for the remaining five hyperparameters. In particular, we choose log-gamma priors for, $\tau_2^{-2}, \sigma^{-2}, r$ and k parametrized in the log scale, that is, $\log\left(\tau_2^{-2}\right), \log\left(\sigma_2^{-2}\right) \sim logGa(1, 0.0005), \log(k) \sim logGa(1, 1), \log(r) \sim logGa(1, 0.01)$. A normal prior distribution is used for ρ re-parametrized in order to be defined in $\mathbb{R}$, that is $\log\left(\frac{1+\rho}{1-\rho}\right) \sim \mathcal{N}(0, 6.66)$. Normal priors $\mathcal{N}(0, 0.001)$ were also assigned for the regression coefficients and a vague normal one for the intercept.

Bayesian inference using SPDE/INLA

Bayesian inference estimates the marginal posterior distributions $p(\phi_j|y) = \int p(\phi_j|\theta, y)p(\theta|y)d\theta$ of the elements of the parameter vector $\phi = (\beta, w, e)^T$, where θ is the vector of hyperparameters and y are the data. Geostatistical models often rely on Markov

chain Monte Carlo (MCMC) simulation to estimate $p(\phi_j|y)$. However, computations involving the spatial covariance matrix are not feasible for large number of locations. Lindgren et al. (2011) proposed the stochastic partial differential approach which represents the above Gaussian spatial process by a Gaussian Markov random field. Hense Σ is approximated by the covariance matrix Q^{-1} of the GMRF, which provides directly the inverse of Q, overcoming a computationally intensive matrix operation. The spatial process representation is based on a partition of the study region into a set of non-intersecting triangles. Subsequently, INLA can be used for fast Bayesian inference. INLA approximates the above integral by $\hat{p}(\phi_j|y) = \sum_k \hat{p}(\phi_j|\theta_k, y)\hat{p}(\theta_k|y)\omega_k$. $\hat{p}(\theta_k|y)$ is calculated from the Laplace approximation of $p(\theta|y)$, that is $p(\theta|y) \propto \frac{p(\phi,\theta|y)}{p_G(\phi|\theta,y)}|_{\phi=\phi_M}$, where $p_G(\phi|\theta, y)$ is the Gaussian approximation of $p(\phi|\theta, y)$ and ϕ_M is the mode of $p(\phi|\theta, y)$. The prediction of the spatial random effect on a grid of locations is performed by projecting the triangular random effects on the grid and calculating a weighted sum of the values at the vertices. The weights are the barycentric coordinates of each grid point. Estimates of the total number of cases across states or the whole country can be obtained by summing pixel-level predictions. The INLA package does not provide directly variation measures for joint distributions and therefore, it cannot estimate the variance of the above quantities. However, it can estimate the variance of linear combinations of η_{it} for a given time point t (eg 2010). Using the Taylor expanstion, the variance of the total predicted cases is given by: $Var\left(\sum_i \exp\left(\eta_{it}\right)\right) \approx Var\left(\sum_i \exp\left(E\left(\eta_{it}\right)\right)\eta_{it}\right)$ where the weights $\exp\left(E\left(\eta_{it}\right)\right)$ of the linear combination are the point predictions at the pixel i. INLA can estimate the right part of the above equation in a second model fit which includes the prediction grid with missing values in the response. Additional linear combinations were defined to calculate the variance of the cases per state in a similar manner.

INLA **implementation**

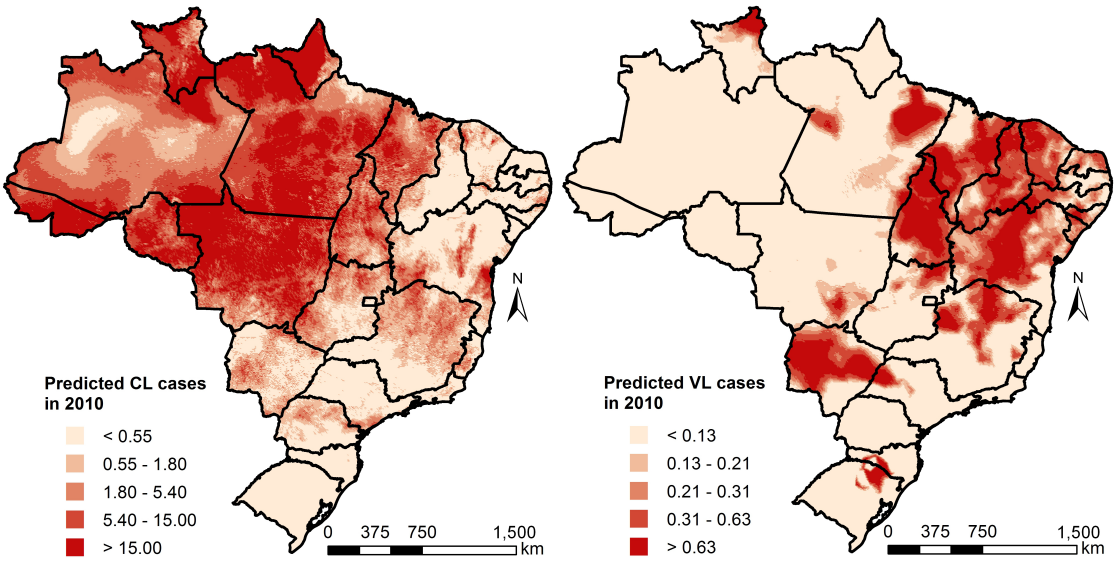
The data file contained standardized continuous predictors and the dummy (0/1) variables of the categorical ones. We assigned a missing value to the response of a randomly selected set of 20% of the data. The response was predicted for these points and used to calculate cross-validatory measures. The R package

maps was used to define the boundaries of our region that was triangulated. The inla.mesh.create.helper() and inla.spde2.matern() functions, of the INLA package, were applied to construct the domain (mesh) and define the covariance function of the spatial process. The inla() was called to perform approximate Bayesian inference and obtain summaries for the coefficients and the hyperparameters. The grid of prediction was constructed with the inla.mesh.projector(). inla.mesh.project() projected the mean of the latent spatial effect on the grid. Using ArcMap (Environmental Systems Research Institute, 2010), covariate values and the population data were extracted at the grid points which are later read in R. The mean of the linear predictor was calculated and summarized over the states to approximated the predicted cases. Finally, a second inla() call enabled the estimation of the variance of the cases aggregated over the whole country and states.

Predicted cases by state and incidence maps under the assumption that missing values are zeros

Table 2.6: Geostatistical model-based predicted incidence rates per 10,000 for cutaneous leishmaniasis in Brazil in 2010.

State	CL cases	VL cases
Acre	1791.0	0.1
Alagoas	85.0	96.5
Amapá	462.6	0.5
Amazonas	2810.5	2.2
Bahia	2237.6	410.0
Ceará	1152.7	505.1
Distrito Federal	81.4	12.6
Espírito Santo	232.0	3.6
Goiás	478.9	35.1
Maranhão	3016.7	401.6
Mato Grosso	3440.4	27.5
Mato Grosso do Sul	192.7	203.6
Minas Gerais	1619.6	417.9
Pará	4227.8	484.3
Paraíba	63.9	40.0
Paraná	645.6	2.9
Pernambuco	605.2	143.9
Piauí	198.3	263.4
Rio de Janeiro	298.2	4.5
Rio Grande do Norte	11.7	87.1
Rio Grande do Sul	10.8	146.2
Rondônia	1743.8	0.9
Roraima	249.9	7.6
Santa Catarina	61.2	48.9
São Paulo	796.2	155.7
Sergipe	27.9	58.9
Tocantins	743.5	256.6
Total	27285.0	3817.4



(a) Cutaneous leishmaniasis. **(b)** Visceral leishmaniasis.

Figure 2.5: Geostatistical model-based predicted incidence rates per 10,000 in Brazil in 2010.

Chapter 3

Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis

Karagiannis-Voules D.A.[1,2], Biedermann P.[1,2], Ekpo U.F., Garba A., Langer E.[1,2], Mathieu E., Midzi N., Mwinzi P., Poldermann A.M., Raso G.[1,2], Sacko M., Talla I., Tchuem-Tchuenté L.A., Touré S., Winkler M.S.[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

[3] Department of Biological Sciences, Federal University of Agriculture, Abeokuta, Nigeria

[4] Réseau International Schistosomiases, Environnement, Aménagements et Lutte, Niamey, Niger

[5] Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA

[6] Department of Medical Microbiology, College of Health Sciences, University of Zimbabwe, Harare, Zimbabwe
[7] Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya
[8] Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
[9] Institut National de Recherche en Santé Publique, Bamako, Mali
[10] Direction de la Lutte contre la Maladie, Ministère de la Santé, Dakar, Senegal
[11] Laboratory of Parasitology and Ecology, University of Yaoundé I, Yaoundé, Cameroon
[12] Center for Schistosomiasis and Parasitology, Yaound, Cameroon
[13] Programme National de Lutte contre la Schistosomiase, Ministère de la Santé, Ouagadougou, Burkina Faso

This paper has been published in *The Lancet Infectious Diseases* 2015, 15: 74-84

Abstract

Background: Interest is growing in predictive risk mapping for neglected tropical diseases (NTDs), particularly to scale up preventive chemotherapy, surveillance, and elimination efforts. Soil-transmitted helminths (hookworm, *Ascaris lumbricoides*, and *Trichuris trichiura*) are the most widespread NTDs, but broad geographical analyses are scarce. We aimed to predict the spatial and temporal distribution of soil-transmitted helminth infections, including the number of infected people and treatment needs, across sub-Saharan Africa.

Methods: We systematically searched PubMed, Web of Knowledge, and African Journal Online from inception to Dec 31, 2013, without language restrictions, to identify georeferenced surveys. We extracted data from household surveys on sources of drinking water, sanitation, and womens level of education. Bayesian geostatistical models were used to align the data in space and estimate risk of with hookworm, *A lumbricoides*, and *T trichiura* over a grid of roughly 1 million pixels at a spatial resolution of 5×5 km. We calculated anthelmintic treatment needs on the basis of WHO guidelines (treatment of all school-aged children once per year where prevalence in this population is 2050% or twice per year if prevalence is greater than 50%).

Findings: We identified 459 relevant survey reports that referenced 6040 unique locations. We estimate that the prevalence of hookworm, *A lumbricoides*, and *T trichiura* among school-aged children from 2000 onwards was 16.5%, 6.6%, and 4.4%. These estimates are between 52% and 74% lower than those in surveys done before 2000, and have become similar to values for the entire communities. We estimated that 126 million doses of anthelmintic treatments are required per year.

Interpretation: Patterns of soil-transmitted helminth infection in sub-Saharan Africa have changed and the prevalence of infection has declined substantially in this millennium, probably due to socioeconomic development and large-scale deworming programmes. The global control strategy should be reassessed, with emphasis given also to adults to progress towards local elimination.

Funding: Swiss National Science Foundation and European Research Council.

3.1 Introduction

Over the past 10 years, interest has grown in better understanding the extent of neglected tropical diseases (NTDs) (Molyneux, 2004; Hotez et al., 2006; Utzinger et al., 2012; Hotez, 2013). Spatially explicit information on the distribution of NTDs is crucial to improve control and elimination efforts (Brooker et al., 2009b; Chammartin et al., 2013a). Advances have been made with spatial modelling, including risk profiling of leishmaniasis at the district level Shimabukuro et al. (2010), predictive risk mapping of loiasis at the national level (Diggle et al., 2007), cross-national models of schistosomiasis (Clements et al., 2009; Schur et al., 2011b; Ekpo et al., 2013), a subcontinental map of soil-transmitted helminth infection (Chammartin et al., 2013b), and a continental future projection of lymphatic filariasis (Slater and Michael, 2013). Additionally, the modelling results have enabled estimation of the number of infected people at different geographical scales, which facilitates calculation of treatment and other intervention needs and their costs (Schur et al., 2012).

For human helminthiases, which account for the largest burden of NTDs (Utzinger et al., 2012; Murray et al., 2012; Hotez et al., 2014), WHO recommends periodic administration of anthelmintic drugs on the basis of prevalence of infection at a given location to control morbidity (WHO, 2006). Predictions of infection risk in areas where prevalence data are lacking can be supplied by spatial statistical models. Studies have provided model-based risk maps and estimates over large scales in South America (Chammartin et al., 2013b) and China (Lai et al., 2013). The authors constructed gridded estimates of population-adjusted prevalence and identified high-risk areas that should be prioritised for control interventions. The work also highlighted the need for doing surveys in areas where data are unavailable or extremely scarce.

Sub-Saharan Africa is among the regions with the highest prevalence of soil-transmitted helminth infections, but progress to reduce the burden has been slower there than in any other region of the world (de Silva et al., 2003; Bethony et al.,

2006). Country-wide analyses of soil-transmitted helminthiasis risk have been done (Pullan et al., 2011). However, a cross-national geostatistical analysis to estimate spatiotemporal patterns and provide country-specific infection estimates at high spatial resolution across sub-Saharan Africa has not yet been done (Brooker et al., 2009a). A Bayesian geostatistical analysis of the number of people infected with soil-transmitted helminths in sub-Saharan Africa was done as a part of the Global Burden of Disease 2010 study (Murray et al., 2012) using data from the Global Atlas of Helminth Infection (Brooker et al., 2010; Pullan et al., 2014). Risk maps have been provided by the Global Atlas of Helminth Infection. The Global Neglected Tropical Diseases (GNTD) database compiles open-access geographically referenced prevalence data for soil-transmitted helminth infections and other NTDs that can be used by researchers and control managers to obtain spatially and temporally explicit estimates of at-risk areas (Hürlimann et al., 2011; Saarnak et al., 2013).

We did a systematic review and extracted data from geographically referenced surveys that reported prevalence of hookworm, *Ascaris lumbricoides*, and *Trichuris trichiura* infections in sub-Saharan Africa. We did a meta-analysis of the data with Bayesian geostatistical models and provided high-resolution risk maps. We also assessed the potentials of education attainment, water, and sanitation-related indicators to increase the predictive ability of the models. Additionally, we estimated the annual treatment needs across the region according to WHO guidelines for preventive chemotherapy (WHO, 2006).

3.2 Methods

Systematic review

We did a systematic review in accordance with the PRISMA guidelines (Moher et al., 2009). We systematically searched PubMed, Web of Knowledge, and African Journal Online from inception to Dec 31, 2013, with no restrictions applied for date of survey or language of publication. We included 43 sub-Saharan African countries (Appendix). We used the following search terms: angola* (OR benin*, OR botswana, OR burkina faso, OR upper volta, OR burundi, OR cte divoire, OR cote divoire, OR ivory coast, OR cameroon, OR camerun, OR kamerun, OR central

african republic, OR chad, OR congo, OR zaire, OR djibouti, OR equatorial guinea, OR eritrea, OR ethiopia, OR gabon, OR gambia, OR guinea, OR guinea-bissau, OR kenya, OR lesotho, OR liberia, OR malawi, OR mali, OR mauritania, OR mozambique, OR namibia, OR niger*, OR rwanda, OR senegal, OR sierra leone, OR somalia, OR south africa, OR south sudan, OR sudan, OR swaziland, OR tanzania, OR togo, OR uganda, OR zambia, OR zimbabwe, OR rhodesia) AND helminth* (OR ascari*, OR trichur*, OR hookworm*, OR necator, OR ankylostom*, OR ancylostom*, OR strongy*, OR hymenolepis, OR toxocara, OR enterobius*, OR geohelminth*, OR nematode*). We also searched the grey literature, including personal collections and reports from control programmes and ministries of health.

Data extraction

We adapted the protocol by Chammartin et al. (2013b) to extract the data. We initially reviewed titles and abstracts, if available, and excluded studies of animals, plants, and genetics, case reports, in-vitro studies, and those that did not mention surveys of soil-transmitted helminthiasis. Quality assessment of retrieved items was based on 30% of articles selected at random. Full-text articles were excluded if they did not report prevalence data, were based on a specific group of patients (eg, hospital patients, those infected with HIV, neonates, etc), were case-control studies, clinical trials, or pharmacological studies (except control groups without anthelmintic intervention), were done in displaced populations (eg, travellers, military), and if the population had undergone deworming in the past 12 months. Extracted data were systematically entered into the GNTD database and geographically referenced with information provided in the reports and various online map and travel guide resources (eg, Wikimapia, Google Maps, iGuide Interactive Travel Guide). We assigned centroids for administrative units on the basis of administrative boundaries in the Database of Global Administrative Areas (version 2).

Relevant prevalence data were extracted and entered in the GNTD database with information on the source (authors, journal, publication date), survey (date, type of survey), location (coordinates, name, administrative unit), and parasitology (species, number of people positive or examined, prevalence, age, diagnostic tool).

If information was missing and papers had been published in the past 20 years, we contacted the authors. Quality of prevalence data extraction was assessed (Chammartin et al., 2013b) and all coordinates were double-checked in Google Maps. We included surveys in the meta-analysis if the sample size was greater than ten individuals. If the date of the survey was missing, we used date of publication instead. Data were screened by location to check for duplicates, in which case the survey with the greater amount of information was used for analysis.

Environmental, socioeconomic, and population data

We consulted WorldClimGlobal Climate Data to obtain data on the proxies of temperature, precipitation, and altitude. Soil moisture and acidity values were downloaded from the Nelson Institute Center for Sustainability and the Global Environment.

Household data were compiled from readily available demographic and health surveys, multiple indicator cluster surveys, world health surveys, and living standards measurement study on sources of drinking and non-drinking water, sanitation facilities, and educational level of women. We used the classification of the Joint Monitoring Programme for Water Supply and Sanitation of WHO and UNICEF (WHO and UNICEF, 2006) to identify households with access to improved drinking-water sources and sanitation. By aggregating household indicators at village level, we constructed proxies of socioeconomic status: the percentage of households with access to improved drinking-water sources, the percentage of households with access to improved sanitation, and the percentage of women who had attended at least primary school.

Locations were classified as rural or urban, according to data downloaded from the Center for International Earth Science Information Network (Center for International Earth Science Information Network (CIESIN), Columbia University, 2005). We obtained population densities in 2000 and 2010 from Worldpop and country-specific percentages of the population younger than 20 years from the United States Census Bureau International Database http://www.census.gov/population/international/data/idb/. For Sudan and South Sudan, percentages of the population aged younger than 20 years in 2008 were used instead of values from 2000

because this was the year with the earliest available data. Links to the databases and resources used in this study are provided in the Appendix.

Statistical analysis

In the Bayesian binomial geostatistical analysis, the number of infected people among those surveyed was used as the outcome and environmental and socioeconomic proxies were used as predictors. Additionally, we applied Bayesian binomial geostatistical models to obtain percentages for the socioeconomic proxies (improved drinking-water source, improved sanitation, and womens educational attainment). We used integrated nested Laplace approximations (INLA) (Rue et al., 2009) and the stochastic partial differential equations approach (Lindgren et al., 2011) to do fast approximate Bayesian inference. Analyses were done in R (version 3.1.1) and the INLA package. Details for implementing geostatistical models with INLA are provided elsewhere (Lindgren et al., 2011; Cameletti et al., 2013; Karagiannis-Voules et al., 2013).

Socioeconomic indicators were not available at epidemiological survey locations. To align the data, we used Bayesian geostatistical models and obtained high resolution estimates for the indicators, with the urban classification as a predictor. Climatic predictors were highly correlated. To avoid collinearity, we specified groups of highly correlated covariates (Pearsons correlation coefficient greater than 0.9). Within each group we selected the variable and its functional form that best predicted the data according to the (leave-one-out) cross-validated logarithmic score (Gneiting and Raftery, 2007; Held et al., 2010) calculated from a bivariate Bayesian geostatistical logistic regression model. The functional forms assessed were linear and categorical (three or four categories, dependent on the quantiles of each variables distribution). Non-linear effects were modelled by spline approximations with random walk processes of order one and two (Rue and Held, 2005). If a random walk was selected and the effect resembled a known functional form, it was substituted by the specific function to ease computations. The variable and form with the lowest mean logarithmic scores were selected from each group. To identify the set of most important environmental and socioeconomic covariates, we fitted geostatistical models with all possible combinations of covariates and selected those

with the best mean logarithmic scores.

All models included survey period as a binary covariate (before 2000 or from 2000 onwards), and interactions were assessed with survey type (school-based, defined as surveys done in schools or those focusing on populations younger than 20 years, or community-based). To incorporate the uncertainty of the socioeconomic indicators predictive value, we fitted a joint model of the indicator and prevalence if the best model included any socioeconomic predictor. The joint model uses the local mean adjustment of the socioeconomic indicator at the epidemiological survey locations, estimated from the posterior predictive distribution of its spatial process. Furthermore, we fitted models with period-specific socioeconomic proxies obtained from geostatistical models, each with a single (continent-specific) temporal trend and a common spatial process for both periods. To take into account that the distance between two locations on the Earth is not a straight line, we used a distance measure that is defined on the spheres surface (Lindgren et al., 2011).

The models were used to predict the risk of species-specific soil-transmitted helminth infection on a 5×5 km grid of 960,132 pixels. By overlaying the predicted risk surfaces with the population density grids and the census-based population percentages, we calculated population-adjusted prevalence by country and subregion (southern, western, eastern, and middle, as defined by the United Nations Statistics Division, see http://unstats.un.org/unsd/methods/m49/m49regin.htm, and modified to include Sudan in the eastern subregion). A set of 300 random samples, simulated from the joint posterior predictive distribution, was used to estimate infection risk and number of people infected and for uncertainty calculations.

We based our calculation of anthelmintic treatment needs on the WHO guidelines (WHO, 2006), which suggest treating all school-aged children once per year in communities where the infection prevalence in the school-aged population is 2050% or twice per year if prevalence is greater than 50%. For each predicted pixel-level prevalence that was higher than 20% or 50%, the treatment needs were equal to one or two times the school-aged population within that pixel, respectively (Schur et al., 2012). We used the WHO definition of school-aged population (age 514 years) and values were obtained from the United States Census Bureau International Database.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data and had final responsibility for the decision to submit for publication.

3.3 Results

Of 6221 identified data sources, we extracted information from 459 (Figure 3.1; Appendix). 51% of surveys were done before 2000. The total number of unique survey locations was 6040, of which 785 (13%) corresponded to urban settlements.

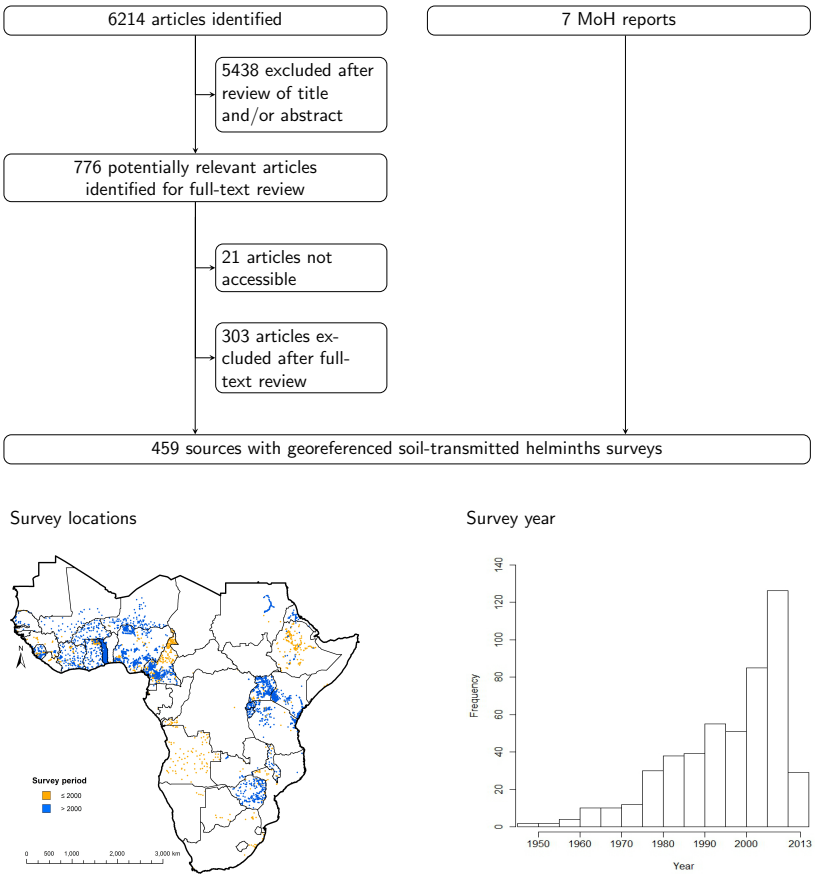


Figure 3.1: Literature search and selection, survey locations, and survey years.

Raw observed data for prevalence of species-specific soil-transmitted helminth infections are shown in Figure 3.2. Most data were derived from national surveys done in Cameroon, Kenya, Nigeria, and Togo. No data were available for Democratic Republic of the Congo (DR Congo), Djibouti, Lesotho, South Sudan, or Swaziland.

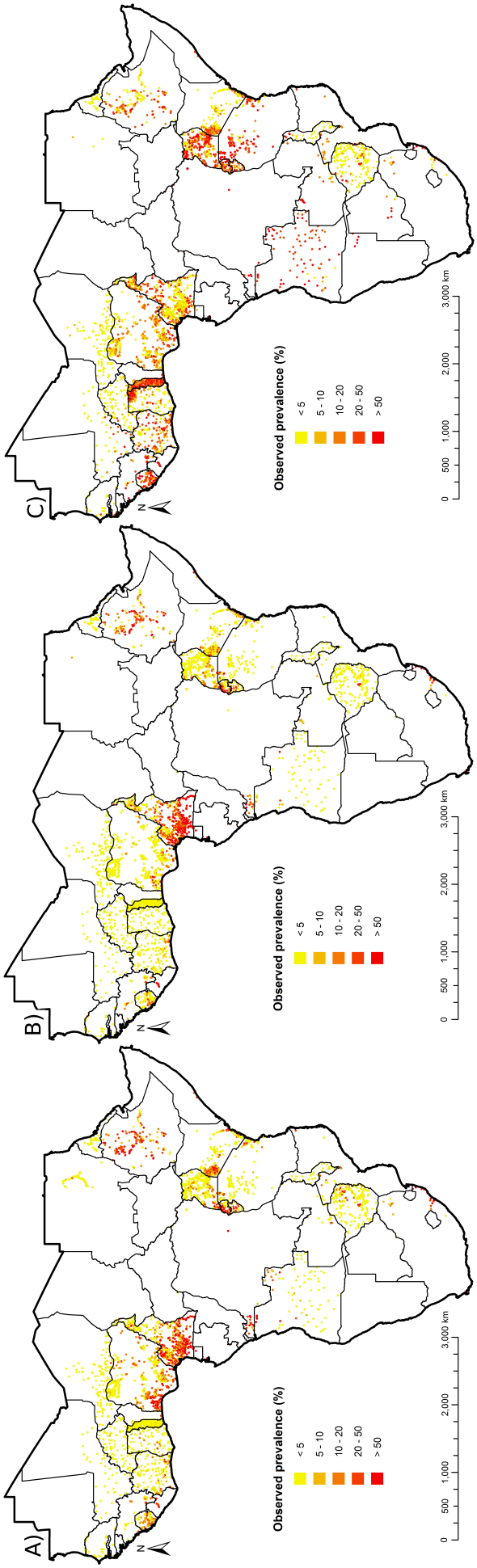


Figure 3.2: Raw observed prevalence of soil-transmitted helminth infections in sub-Saharan Africa.

Raw socioeconomic and population data were available for 32,618 locations (Appendix). The medians of the percentage of households with access to improved drinking-water sources and use of improved sanitation were 80% and 12%, respectively (Appendix). All three socioeconomic variables were positively associated with the urban classification (data not shown). Five countries had no geographically referenced data for the socioeconomic proxies. Of the remaining 38, only 12 had data from both survey periods (Appendix). Hence, our geo statistical models could only yield period-specific estimates for socioeconomic proxies and allowed assessment of only continent-level rather than country level temporal trends. Exploratory analysis, however, suggested that some countries had positive changes and others had negative changes (Appendix).

After taking into account highly correlated predictors and identifying their best functional form from bivariate Bayesian geostatistical models, we fitted all possible combinations of around 12 predictors per helminth species, which gave rise to 4096 models. The best predictive model and the estimated parameters of the predictors for each species are shown in Table 3.1. The socioeconomic proxy improved drinking-water source was included in the best model for hookworm. Negative trends were found for surveys done from 2000 onwards for all three soil-transmitted helminth species. Surveys of school-aged children revealed higher prevalence for *A. lumbricoides* and *T. trichiura* than community-based surveys done before 2000, whereas the survey type had no effect on the prevalence of hookworm infection. Hookworm was negatively associated with urban settlements and locations with high percentages of improved drinking-water sources. Non-linearity of the soil moisture resembled a parabolic function, indicating that extreme dry or wet soils are associated with the absence of hookworm infection. Low variations in temperature and precipitation in the warmest quarter were associated with increased prevalence of *A. lumbricoides*. The factors associated with increased risk of *T. trichiura* infection were high precipitation in the warmest quarter, high temperature in the coldest quarter, and low variation in precipitation.

Table 3.1: Posterior estimates of the final geostatistical models for risk of species-specific soil-transmitted helminth infection in sub-Saharan Africa.

Hookworm*	**Median estimate (95% CI)**
Urban-rural classification	
Rural	0
Urban	-0.43 (-0.59, -0.27)
Survey period (year)	
Before 2000	0
From 2000 onwards	-1.44 (-1.52, -1.35)
Survey type	
Community-based	0
School-based	-0.04 (-0.08, 0.00)
Survey period × survey type	0.02 (0.10, 0.06)
Mean adjustment of improved drinking-water sources	0.07 (0.11, 0.02)
Spatial variance†	5.06 (4.74, 5.45)
Spatial range (km)†	29.2 (27.6, 31.0)
Ascaris lumbricoides	
Isothermality (%)	
< 66.9	0
66.9-74.5	1.28 (0.99, 1.58)
> 74.5	1.57 (1.24, 1.90)
Precipitation of warmest quarter (mm)	
< 173	0
173-277	1.34 (1.06, 1.63)
277-1151	1.94 (1.62, 2.27)
> 1151	2.07 (1.75, 2.40)
Survey period (year)	
Before 2000	0
From 2000 onwards	-1.41 (-1.54, -1.28)
Survey type	
Community-based	0
School-based	0.53 (0.47, 0.59)
Survey period × survey type	0.63 (0.76 , 0.50)
Spatial variance†	6.39 (5.90, 6.93)
Spatial range (km)†	40.2 (36.8, 42.8)
Trichuris trichiura	
Mean temperature of coldest quarter (°C)	
< 21.4	0
21.4-23.6	0.12 (-0.10, 0.33)
23.6-25.1	0.08 (-0.19, 0.35)
> 25.1	0.22 (-0.15, 0.58)
Precipitation of warmest quarter (mm)	
< 175	0
175-279	1.29 (0.84, 1.74)
279-1051	1.23 (0.77, 1.69)
> 1051	1.48 (1.02, 1.94)
Soil acidity (pH)	
< 5.6	0
5.6-6.1	-1.06 (-1.42, -0.71)
> 6.1	0.18 (-0.22, 0.58)
Precipitation seasonality	-0.04 (-0.04, -0.03)
Survey period (year)	
Before 2000	0
From 2000 onwards	-1.57 (-1.72, -1.42)
Survey type	
Community-based	
School-based	0.75 (0.68, 0.83)
Survey period × survey type	-0.88 (-1.03, -0.73)
Spatial variance†	7.59 (6.90, 8.34)
Spatial range (km)†	46.8 (43.3, 51.0)

*see Appendix for effect of soil moisture. †parameter of the spatial process.

Period-specific socioeconomic predictors did not change model-based estimates of covariate effects and predictions of infection risk compared with constant socioeconomic predictors. Hence, we report predictions of infection risk based on models with common socioeconomic predictors for the two time periods , see Figure 3.3. Hookworm risk was high in western and middle Africa, with the highest-risk areas being in Sierra Leone, Togo, and around Lake Victoria. *A. lumbricoides* and *T. trichiura* followed similar spatial patterns with high prevalence seen in Cameroon, Ethiopia, and at the borders of Burundi, DR Congo, and Rwanda. A high-resolution map of the percentage of households with access to improved drinking-water sources is provided in the Appendix.

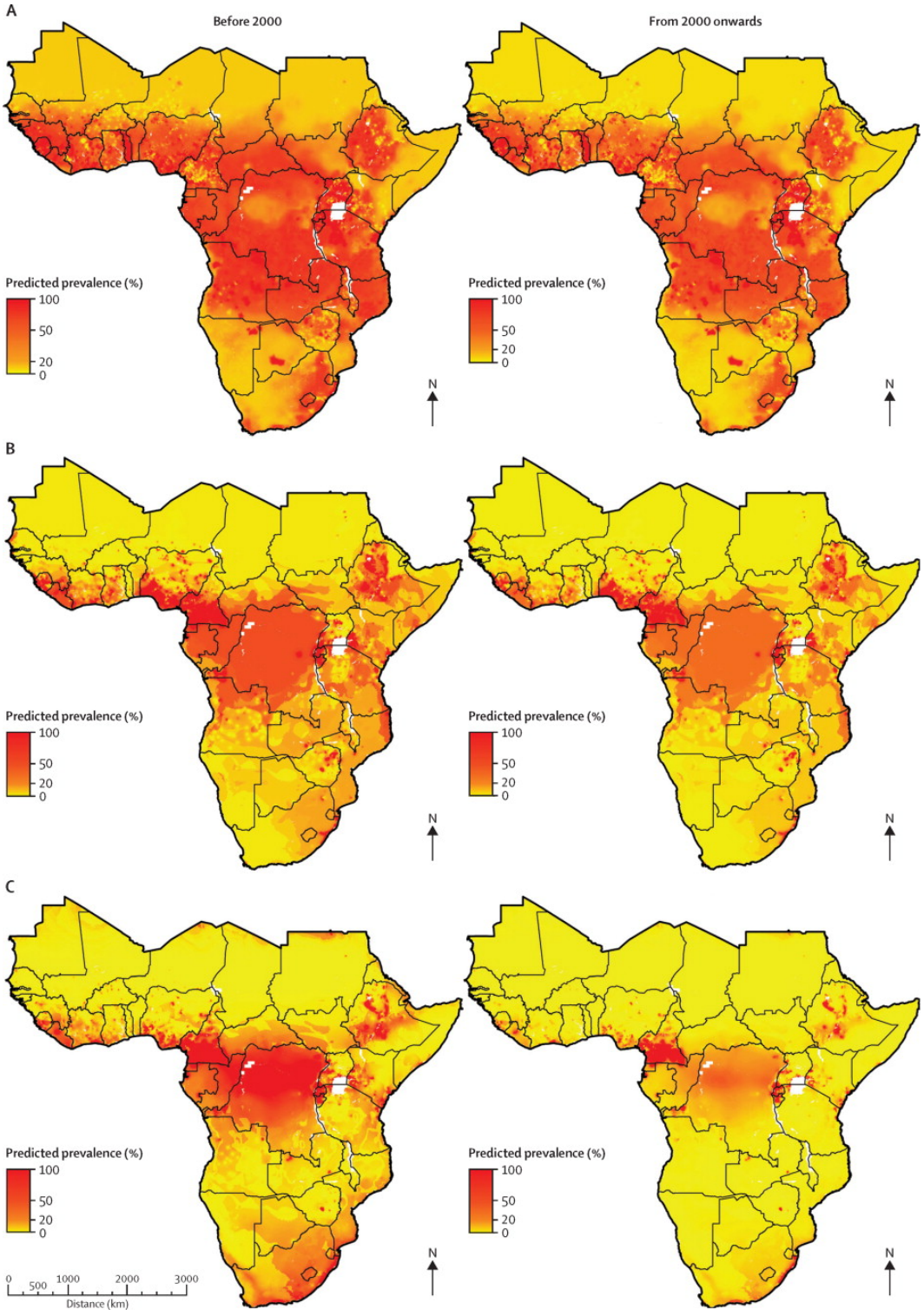


Figure 3.3: Median predicted risk estimates for soil-transmitted helminth infections in sub-Saharan Africa before 2000 and from 2000 onwards. (A) Hookworm. (B) *Ascaris lumbricoides*. (C) *Trichuris trichiura*.

Population-adjusted prevalence of infection for each of the three soil-transmitted helminth species was stratified by country and subregion (Tables 3.2 and 3.3). The highest prevalence of soil-transmitted helminth infection was predicted in western Africa, followed by eastern, southern, and middle Africa. Sierra Leone and Togo were predicted to have the highest hookworm prevalence. Gabon and Rwanda had the highest risks of infection with *A. lumbricoides* and *T. trichiura* . Overall, we estimated that among the roughly 800 million people in the 43 countries of sub-Saharan Africa 130 million, 53 million, and 37 million people were infected with hookworm, *A. lumbricoides*, and *T. trichiura* , respectively. Overall annual treatment needs were estimated to be nearly 126 million doses, and the total number of school-aged children needing treatment was estimated to be 91 million. More than 15 million treatments would be needed for DR Congo, Ethiopia, and Nigeria, which corresponds to roughly 12%, 12%, and 20% of the total treatment needs, respectively.

Table 3.2: Population-adjusted prevalence of species-specific soil-transmitted helminth infections by survey period and subregion.

Before 2000				
	Population	Hookworm	*A. lumbricoides*	*T. trichiura*
Population aged <20 years				
Eastern	121,340,394	34.3 (32.9-35.7)	19.7 (18.9-20.7)	18.1 (17.1-19.2)
Middle	64,945,309	34.5 (33.0-36.0)	18.5 (17.5-19.6)	16.4 (15.3-17.4)
Southern	38,987,696	34.2 (32.7-35.6)	18.9 (17.9-20.2)	17.5 (16.4-18.6)
Western	119,268,659	34.1 (32.7-35.5)	19.7 (18.9-20.7)	17.6 (16.7-18.6)
Population aged >20 years				
Eastern	98,948,578	34.6 (32.9-36.0)	16.0 (15.2-16.8)	12.9 (12.2-13.8)
Middle	52,856,357	34.7 (33.1-36.3)	15.1 (14.1-16.2)	11.4 (10.5-12.4)
Southern	31,878,914	34.3 (32.5-36.0)	15.4 (14.5-16.4)	12.3 (11.3-13.4)
Western	97,017,768	34.4 (32.8-35.9)	15.8 (15.0-16.5)	12.3 (11.5-13.1)
From 2000 onwards				
	Population	Hookworm	*A. lumbricoides*	*T. trichiura*
Population aged <20 years				
Eastern	150,202,927	16.5 (15.6-17.6)	6.7 (6.3-7.3)	4.7 (4.3-5.2)
Middle	81,824,674	16.7 (15.6-18.0)	6.3 (5.7-7.0)	4.1 (3.7-4.8)
Southern	48,847,355	16.6 (15.3-17.8)	6.4 (5.9-7.0)	4.2 (3.8-4.7)
Western	147,346,980	16.4 (15.5-17.5)	6.6 (6.1-7.1)	4.4 (4.0-4.9)
Population aged >20 years				
Eastern	127,312,077	16.7 (15.8-17.8)	7.2 (6.6-7.8)	5.3 (4.8-5.9)
Middle	69,305,534	17.0 (16.0-18.3)	6.6 (6.0-7.4)	4.6 (4.1-5.3)
Southern	41,601,143	16.8 (15.7-17.9)	6.8 (6.3-7.5)	4.8 (4.2-5.4)
Western	124,494,350	16.7 (15.7-17.9)	7.1 (6.5-7.7)	4.9 (4.5-5.5)

Table 3.3: Population-adjusted prevalence of soil-transmitted helminth infections from 2000 onwards and annual anthelmintic treatment needs.

Country	Population aged <20 years (1000s)	Hookworm	*A. lumbricoides*	*T. trichiura*	All soil-transmitted helminths	Treatment needs for school-aged children (1000s)
Angola	7,872	15.9 (13.5-20.5)	3.4 (2.4-5.1)	1.8 (1.1-2.9)	20.1 (17.4-25.1)	1,835 (1,477-2,388)
Benin	3,953	18.3 (12.3-26.8)	4.3 (1.7-9.7)	1.4 (0.4-4.7)	23.5 (17.0-32.4)	1,086 (670-1,602)
Botswana	929	4.4 (3.4-6.2)	1.4 (0.8-3.3)	2.0 (1.0-4.8)	7.8 (6.0-11.0)	49 (27-95)
Burkina Faso	9,122	9.9 (7.0-15.7)	0.4 (0.3-1.0)	0.4 (0.2-1.1)	10.7 (7.9-16.6)	902 (540-1,711)
Burundi	4,826	30.4 (23.1-38.9)	16.2 (13.9-18.8)	13.5 (11.1-17.1)	49.0 (43.1-55.3)	3,079 (2,684-3,438)
Cameroon	9,199	9.9 (8.3-11.3)	10.4 (9.5-11.6)	12.5 (11.7-13.4)	28.0 (26.3-29.5)	3,340 (3,094-3,550)
Central African Republic	2,037	15.5 (12.1-19.9)	4.0 (2.7-5.8)	4.0 (2.6-6.4)	22.0 (18.7-26.4)	512 (388-660)
Chad	6,405	7.4 (5.6-10.2)	1.2 (0.7-2.6)	0.3 (0.2-0.8)	9.0 (7.0-11.9)	415 (214-689)
Congo	1,869	12.9 (7.6-34.5)	4.8 (2.1-24.5)	5.0 (2.2-15.4)	24.8 (14.7-46.1)	461 (203-942)
Cte d'Ivoire	9,537	23.7 (19.8-27.6)	4.4 (3.5-5.4)	5.6 (4.6-7.0)	30.0 (26.9-33.2)	3,617 (3,197-4,162)
Djibouti	135	3.4 (1.0-10.6)	0.5 (0.1-2.2)	2.6 (0.6-12.4)	6.9 (2.8-17.4)	5 (0-26)
DR Congo	37,088	17.9 (15.5-21.3)	9.2 (7.4-11.3)	10.5 (8.8-13.2)	33.0 (30.0-36.5)	15,551 (13,668-17,628)
Equatorial Guinea	273	11.2 (6.0-21.1)	14.6 (7.5-28.0)	17.6 (9.5-32.1)	37.0 (26.5-51.2)	125 (82-188)
Eritrea	2,693	3.3 (1.8-5.9)	0.7 (0.3-1.6)	0.5 (0.2-1.2)	4.5 (2.9-7.7)	58 (15-159)
Ethiopia	44,433	17.7 (15.1-20.0)	8.5 (7.1-10.2)	6.1 (4.7-7.9)	28.8 (25.7-31.2)	15,592 (13,586-17,237)
Gabon	592	26.0 (12.9-40.6)	14.4 (6.0-31.8)	26.0 (12.5-36.4)	47.8 (36.2-57.9)	347 (262-420)
Gambia	984	20.3 (7.3-44.6)	1.6 (0.5-13.8)	0.1 (0.0-1.7)	22.7 (9.5-46.0)	276 (46-614)
Ghana	11,641	14.4 (11.7-19.3)	4.3 (2.8-8.7)	1.5 (0.8-3.0)	19.8 (16.1-24.8)	2,468 (1,775-3,431)
Guinea	4,502	21.9 (17.3-26.9)	3.8 (2.6-6.0)	1.7 (1.0-3.9)	26.2 (21.9-31.4)	1,452 (1,174-1,789)
Guinea-Bissau	466	14.0 (7.3-27.2)	0.8 (0.2-3.0)	0.1 (0.0-0.7)	15.3 (8.3-28.4)	65 (16-170)
Kenya	20,188	16.9 (14.3-20.0)	11.5 (9.9-13.4)	5.0 (4.0-6.8)	29.2 (26.1-32.4)	7,376 (6,502-8,330)
Lesotho	1,011	21.3 (10.8-35.6)	2.4 (0.6-9.9)	5.7 (1.3-18.4)	29.3 (17.9-43.1)	360 (190-551)
Liberia	1,410	21.9 (16.4-28.8)	8.6 (5.3-13.6)	6.6 (3.5-11.8)	33.4 (27.4-40.7)	636 (504-790)
Malawi	8,475	14.9 (10.9-20.0)	2.1 (1.2-3.9)	0.6 (0.2-1.7)	17.3 (13.1-22.4)	1,578 (1,075-2,184)
Mali	8,986	10.4 (8.4-12.8)	0.6 (0.4-0.9)	0.3 (0.2-0.6)	11.2 (9.1-13.4)	1,012 (712-1,307)
Mauritania	1,938	2.8 (1.5-12.2)	0.5 (0.3-1.2)	0.2 (0.1-0.7)	3.7 (2.3-12.9)	20 (7-291)
Mozambique	12,417	15.6 (12.5-19.5)	3.8 (2.6-6.3)	3.4 (2.2-5.2)	21.9 (18.5-25.8)	3,162 (2,503-3,935)
Namibia	973	7.2 (4.0-13.7)	1.1 (0.5-2.8)	0.9 (0.4-3.0)	9.4 (5.6-15.6)	82 (31-175)
Niger	8,878	4.0 (2.7-5.7)	0.5 (0.3-0.9)	0.1 (0.1-0.3)	4.6 (3.3-6.4)	237 (114-459)
Nigeria	81,508	16.9 (14.7-19.2)	9.8 (8.1-11.3)	3.0 (2.2-4.6)	26.7 (24.1-28.9)	26,290 (23,258-28,907)
Rwanda	5,477	31.9 (24.2-39.6)	31.7 (28.1-37.4)	23.3 (20.9-27.2)	62.6 (56.3-68.5)	4,243 (3,826-4,598)
Senegal	6,632	8.2 (5.1-15.2)	2.7 (1.6-6.3)	0.4 (0.2-1.2)	11.6 (8.0-17.9)	664 (326-1,410)
Sierra Leone	2,461	34.3 (29.1-40.6)	6.3 (4.8-8.1)	2.5 (1.8-4.0)	39.9 (34.9-45.7)	1,288 (1,091-1,484)
Somalia	4,091	5.5 (4.2-8.0)	2.4 (1.6-4.5)	4.2 (2.8-6.5)	11.7 (9.4-15.0)	496 (394-708)
South Africa	18,219	11.6 (8.8-16.0)	5.7 (4.2-8.1)	8.1 (6.0-13.8)	22.9 (19.5-29.0)	4,653 (3,744-6,378)
South Sudan	5,617	8.6 (6.8-10.8)	2.2 (1.6-3.2)	1.9 (1.3-3.1)	12.3 (10.3-15.0)	549 (391-785)
Sudan	19,292	2.4 (1.8-3.2)	0.6 (0.4-1.2)	0.2 (0.1-0.6)	3.3 (2.6-4.4)	110 (51-296)
Swaziland	631	18.4 (7.7-38.8)	2.3 (0.4-12.3)	2.9 (0.4-14.5)	24.7 (11.9-45.1)	185 (67-358)
Tanzania	23,633	24.3 (21.6-27.7)	3.8 (2.9-4.8)	2.0 (1.4-3.9)	28.6 (25.6-32.9)	8,296 (7,300-9,688)
Togo	2,838	34.4 (31.3-38.4)	0.7 (0.4-1.5)	0.4 (0.2-0.9)	35.3 (32.0-39.1)	1,280 (1,131-1,495)
Uganda	20,639	31.0 (27.5-33.8)	4.5 (3.5-5.8)	3.7 (2.9-5.1)	36.6 (33.4-39.3)	9,457 (8,556-10,306)
Zambia	7,478	18.2 (13.9-27.5)	2.6 (1.7-6.7)	1.0 (0.6-2.0)	21.7 (16.7-30.2)	1,839 (1,238-2,763)
Zimbabwe	6,875	10.4 (7.7-14.7)	2.7 (2.0-4.1)	0.7 (0.4-1.4)	13.4 (10.8-17.8)	933 (664-1,395)
Total	428,222	16.5 (15.6-17.6)	6.6 (6.1-7.0)	4.4 (4.1-4.9)	24.7 (23.7-25.7)	126,466 (120,241-132,307)

3.4 Discussion

We did a systematic review and a geostatistical meta-analysis of surveys of soil-transmitted helminth infections in sub-Saharan Africa. Analyses based on geostatistical models are the most rigorous approaches for risk profiling of NTDs at different geographical scales. Our models included temporal terms, socioeconomic proxies, and environmental predictors. By exhaustive fitting of all possible Bayesian geostatistical models, we identified one for each soil-transmitted helminth species that was used to predict infection risk at high spatial resolution. A decrease in prevalence from 2000 onwards is predicted for sub-Saharan Africa, which matches the findings from other regions, such as Cambodia (Karagiannis-Voules et al., 2015b), China (Lai et al., 2013), and South America (Chammartin et al., 2013b).

The use of environmental and socioeconomic factors allowed us to predict the infection risk in areas where no surveys have been done. We predicted moderate prevalence in regions where data are scarce. For instance, the risk of infection is high for all three soil-transmitted helminth species in DR Congo and extends into Gabon for hookworm and *A. lumbricoides*. Surveys are warranted in areas with sparse data to update predictions and models. Model-based estimates should in turn be iteratively updated (Kabore et al., 2013) to support monitoring and surveillance efforts.

Our analyses revealed several insights that are noteworthy. First, the predicted prevalence for all three soil-transmitted helminth species in southern Africa was much lower than previously reported (Pullan et al., 2014). The previous estimates were based on only 45 survey locations across southern Africa, including Zimbabwe. Our analysis is based on more than 200 unique survey locations in this subregion, most of which had survey data obtained after 2008. Hence, the risk of soil-transmitted helminth infection in southern Africa might previously have been overestimated. Second, in eastern Africa (excluding Madagascar), our estimated prevalence was lower than that predicted before but the higher value might have been driven by high prevalence of soil-transmitted helminth infection in Madagascar (Kightlinger et al., 1995). Third, in western Africa, we retrieved roughly double the amount of point prevalence measures and predicted lower prevalence for *A. lumbricoide*s than did Pullan and colleagues Pullan et al. (2014). Fourth, in

middle Africa (excluding Cameroon), only few geographically referenced surveys were available for inclusion in our analyses and, therefore, our estimates are prone to notable uncertainty.

The risk of infection with all three soil-transmitted helminth species declined from 2000 onwards, probably due to socioeconomic development (Sundaram et al., 2011) and intensified control measures (Organization et al., 2013). We excluded surveys done within 12 months after deworming to avoid potential bias from the direct effect of anthelmintic treatment. Within 1 year of treatment, prevalence of *A. lumbricoides* and *T. trichiura* approach pretreatment levels, whereas the prevalence of hookworm infection remains reduced by about half (Jia et al., 2012). Hence, a negative temporal trend is expected to relate to deworming, other interventions, and socioeconomic development. In sub-Saharan Africa, a slight increase in the prevalence of soil-transmitted helminth infections was reported after a comparison of data from 1994 and 2003, but prevalence decreased in all other regions worldwide in the same period.19 A later analysis by Pullan et al. (2014), however, showed no temporal trend for soil-transmitted helminth infection.

Poverty and socioeconomic status can be measured through many indices. We used readily available household data and standard classifications to construct proxies for drinking water and sanitation, but we did not detect strong associations. Other possible proxies, such as use of treated water or access to sanitation (Strunz et al., 2014), might improve prediction, as associations with soil-transmitted helminth infections have been noted in assessments of individual-level data. The aggregation of socioeconomic factors at village level and their spatial misalignment with the data for soil-transmitted helminth infections resulted in substantial variation within and between villages, which renders the identification of any effects difficult (Karagiannis-Voules et al., 2015b). This heterogeneity might also explain why period-specific socioeconomic predictors did not improve prediction of risk of infection.

Since 2000, ministries of health, WHO, and other national, international, and non-governmental organisations have stepped up control against soil-transmitted helminthiasis and other NTDs, emphasising preventive chemotherapy. According to data reported by WHO, in 201012, the total of children younger than 15 years

in sub-Saharan Africa who were treated once with albendazole or mebendazole was more than 70 million in each year (see www.who.int/neglected_diseases/preventive_chemotherapy/sth/en/). In 2012, the subcontinental preventive chemotherapy coverage reached 26%. Initiation of major control interventions, however, has differed between countries, and not all have reported administered treatments to WHO. Therefore, we cannot take into account treatment coverage at the national or subnational level. Apart control programmes for soil-transmitted helminthiasis, the Global Program to Eliminate Lymphatic Filariasis (GPELF) and the African Programme for Onchocerciasis Control (APOC) have administered hundreds of millions of tablets of albendazole, mebendazole, and ivermectin treatments in the past decade (Appendix) (WHO, 2012b, 2013a). Although albendazole and mebendazole are not as efficacious against *T. trichiura*, as against hookworm and *A. lumbricoides*, combination of either of these drugs with ivermectin results in reasonable efficacy against *T. trichiura* (Moncayo et al., 2008; Wen et al., 2008; Massa et al., 2009; Knopp et al., 2010).4952 Since its establishment, APOC has administered more than 80 million doses of ivermectin (WHO, 2012b) and GPELF has widely administered combination therapy with albendazole and ivermectin among other treatments (WHO, 2013b). WHO estimates that, in 2011, the number of Africans covered by preventive chemotherapy for at least one disease was higher than 200 million (WHO, 2013a). Since not all treatments are reported to WHO, the true number of people receiving anthelmintic treatment might be substantially greater (Organization et al., 2013; Gallo et al., 2013).

Our lower prevalence, compared with values reported previously, and the achieved coverage estimates (for the African region) of the WHO progress report in 2012 (WHO, 2012c), suggest that the 2020 target set by WHO of preventive chemotherapy reaching at least 75% coverage in all countries is on track. Additionally, our estimation of 91 million school-aged children needing treatment is less than that reported by WHO for the African region in 2011. We calculated the estimated treatment needs at pixel level under two assumptions: infections with the three different species are independent, and the prevalence for the age groups younger than 20 years and 514 years are the same. Nevertheless, these estimates provide important baseline information for decision-makers for initiating and designing

control interventions.

Risk predictions for *A. lumbricoides* and *T. trichiura* among the school-aged population were substantially higher before 2000 than from 2000 onwards. The interaction between trend and survey type might be indicative of the school-based deworming efforts that were intensified since World Health Assembly resolution 54.19 was put forward in May, 2001 (WHO, 2002a). The difference in the prevalence of soil-transmitted helminth infections, according to survey type, has become negligible from 2000 onwards. This finding suggests that prevalence in the school-aged population dropped to a level that matches the entire community. Thus, the emphasis on school-based deworming is worthy of reassessment. More aggressive targeting of the other populations defined by WHO as eligible for intervention (preschool-aged children, women of childbearing age, and adults, particularly those with high occupational exposure)(WHO, 2006) might be necessary. Similar suggestions have been made by Anderson and colleagues (Anderson et al., 2013), who suggested that school-based deworming could have time-limited benefits for the greater community. The proposal of new treatment guidelines will need additional studies to assess how changes in the prevalence an intensity of infection depend on different treatment schedules for different subgroups of the population at specific prevalence levels (Keiser and Utzinger, 2008).

Data compilation and meta-analyses are prone to bias. We adhered to a predefined data extraction protocol to limit potential sources of bias in our analyses. Several methodological improvements have been discussed elsewhere (Chammartin et al., 2013b,a) and relate to incorporating diagnostic sensitivity and the relation between age and prevalence into geostatistical modelling.

We assumed that the relation between prevalence and the predictors and the considered interactions were constant across sub-Saharan Africa. The study area, however, is large and the relation between the predictors and infection risk might vary in space because, for instance, unmeasured factors, such as intervention levels or health-system performance, vary in space. We did not model varying covariate effects across space (Gelfand et al., 2003). We did, however, fit models incorporating smooth changes of spatial process parameters in space (ie, non-stationary models), and these did not improve predictive performance.

We are aware of large-scale surveys done in African countries that could not be included in this analysis because the data were not readily available in geographically referenced forms. In Mozambique and Burkina Faso, for example, two surveys done after 2005 included 1275 and 130 schools, respectively (Augusto et al., 2009; Coulibaly et al., 2011). In Cte dIvoire, a health and demographic surveillance system has been established in the Taabo area in the south-central part of the country, and has been used to construct a household-level database in 2008 (Fürst et al., 2012). Another two surveys in Cte dIvoire included more than 80 schools (Ouattara et al., 2008, 2010). The data from these surveys will be important to incorporate in future model-based predictions. Furthermore, some surveys from the peer reviewed literature were excluded from our analysis due to incomplete information. The need to report complete survey information should be emphasised to assist spatial analysis of aggregated survey data (Brooker et al., 2009a; Saarnak et al., 2013).

High-resolution spatially explicit risk predictions and maps can assist control programmes to select treatment strategies based on endemicity levels and to design future surveys. Estimated numbers of infected people can help international funding agencies to allocate resources to the countries. Information on the number of required treatments can be useful to drug producers and drug donors. Our findings contribute to the international efforts to reach the WHO-defined milestone of mapping soil-transmitted helminth infections to identify areas requiring preventive chemotherapy, and to monitor programmes aimed at achieving the 2020 target of control and elimination of NTDs (WHO, 2012c). Together with the analysis by Chammartin and colleagues in South America (Chammartin et al., 2013b), and Lai et al. (2013) in China, a global stepping stone towards model-based soil-transmitted helminth infection risk estimates has now been built.

Contributors

D-AK-V processed and analysed the data, interpreted the results, and wrote the manuscript. D-AK-V, PB, and EL contributed to the systematic review and data extraction. PV extracted the data on water supply, sanitation, and education and processed these data with D-AK-V. D-AK-V, JU, and PV developed the protocol

and search strategy for the systematic review. UFE, AG, EM, NM, PM, AMP, GR, MS, IT, LATT, ST, and MSW provided substantial data. PV assisted in the meta-analysis. JU and PV conceptualised the project and revised the manuscript. All authors approved the fi nal version of the manuscript before submission.

Declaration of interests

We declare no competing interests.

Acknowledgments

This study was funded by the Swiss National Science Foundation (PDFMP3-137156) and the European Research Council (323180). PB was partly funded by UBS Optimus Foundation (project number 5879). UFE received funding from the European Foundations Initiatives for African Research into Neglected Tropical Diseases. We thank our collaborators in Africa who contributed geographically referenced soil-transmitted helminthiasis survey data for the Global Neglected Tropical Diseases database. We also thank the Demographic and Health Surveys, Multiple Indicator Cluster Surveys, World Health Surveys, and Living Standards Measurement Study projects for making the water and sanitation data freely accessible.

3.5 Appendix

Sources and links

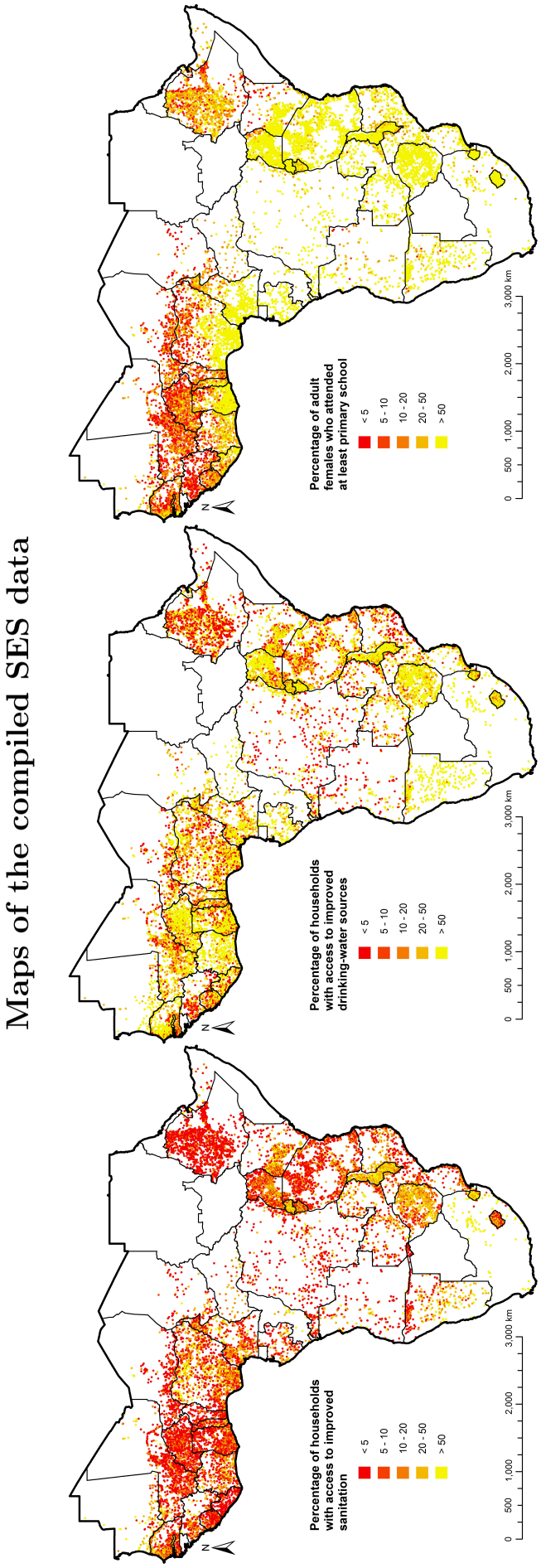
Source	Link
Global Neglected Tropical Diseases database	www.gntd.org
WorldClim	www.worldclim.org
Demographic and Health Surveys	www.measuredhs.com
Multiple Indicator Cluster Surveys	www.childinfo.org/mics.html
World Health Surveys	www.who.int/healthinfo/survey/en/index.html
Living Standards Measurement Study	http://econ.worldbank.org
Worldpop	www.worldpop.org.uk
United States Census Bureau International Database	http://www.census.gov/population/international/data/idb/informationGateway.php
R	http://cran.r-project.org
Integrated nested Laplace approximations	www.r-inla.org
United Nations Statistics Division	http://unstats.un.org/unsd/methods/m49/m49regin.htm
Preventive chemotherapy databank of WHO	www.who.int/neglected_diseases/preventive_chemotherapy/sth/en/index.html

Sources and years of the SES per country

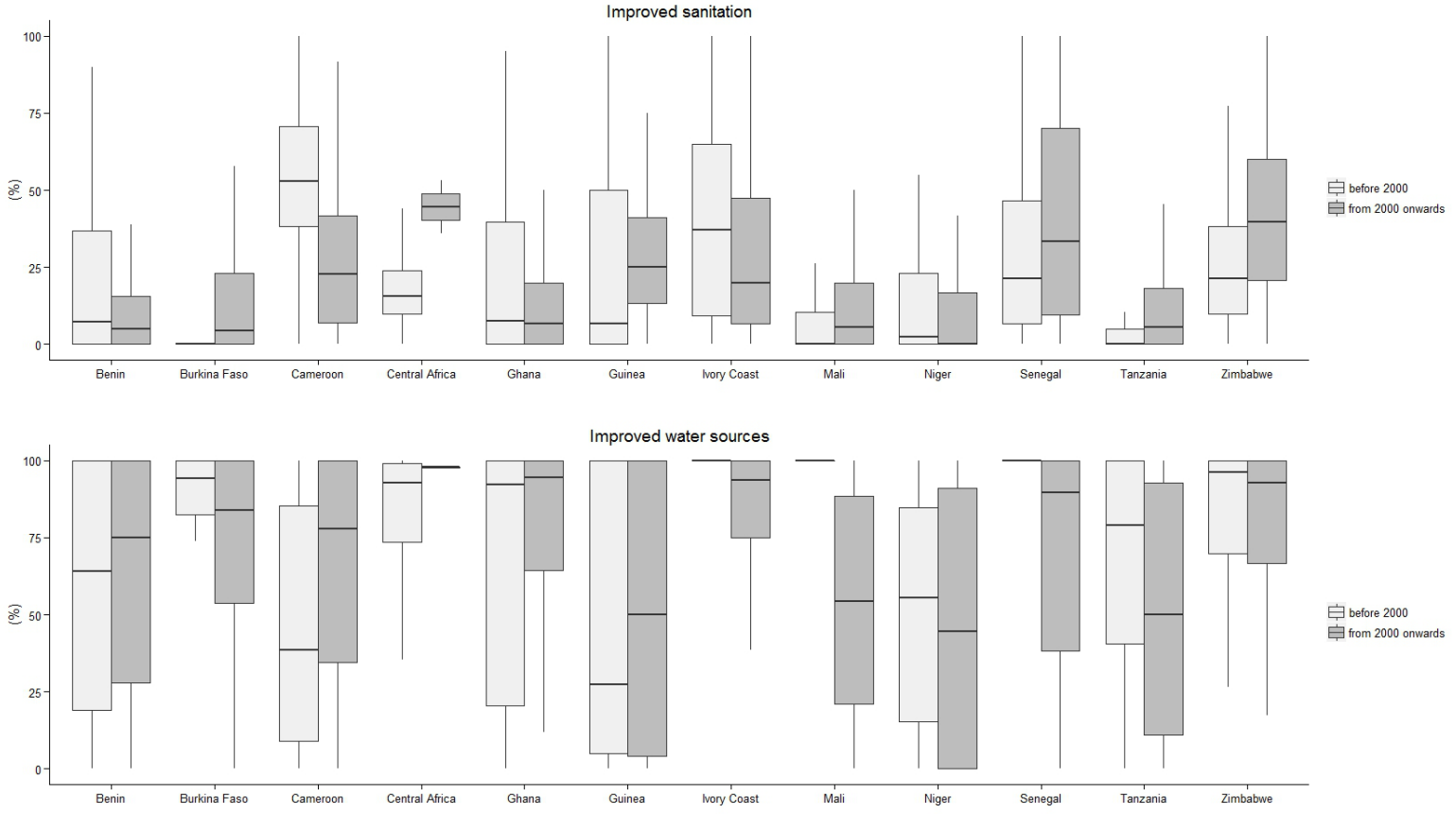
Country	Sources	Survey years
Angola	DHS,MICS	2000,2006,2011
Benin	DHS	1996,2001
Botswana		
Burkina Faso	DHS,WHS	1993,1998,2003,2010
Burundi	DHS,MICS	2000,2005,2010,2012
Cameroon	DHS,MICS	1991,2004,2006,2011
Central African Republic	DHS,MICS	1994,2000,2006
Chad	WHS	2003
Congo DRC	DHS,LSMS,WHS	2000,2003,2007,2010
Congo (The Republic of)		
Cte D'Ivoire	DHS,MICS,WHS	1994,1998,2000,2003,2006,2011
Djibouti	MICS	2006
Equatorial Guinea		
Eritrea		
Ethiopia	DHS,LSMS,WHS	2000,2003,2005,2011
Gabon	DHS	2012
Gambia	MICS	2000,2005
Ghana	DHS,WHS	1993,1998,2003,2008
Guinea-Bissau	MICS	2000,2006
Guinea	DHS	1999,2005
Kenya	DHS,MICS,WHS	2000,2003,2008,2010
Lesotho	DHS	2004,2009
Liberia	DHS	2007,2009,2011
Malawi	DHS,LSMS,WHS	2000,2003,2004,2010,2012
Mali	DHS,WHS	1995,2001,2003,2006
Mauritania	MICS,WHS	2003,2007
Mozambique	DHS	2008,2009,2011
Namibia	DHS,WHS	2000,2003,2006
Niger	DHS,LSMS,MICS	1992,1998,2000,2011
Nigeria	DHS,MICS	2003,2007,2008,2010,2011
Rwanda	DHS	2005,2007,2010
Senegal	DHS,MICS,WHS	1992,1997,2000,2003,2005,2008,2010
Sierra Leone	DHS,MICS	2005,2007,2010
Somalia	MICS	2006,2008,2010
South Africa	WHS	2003
Sudan		
Swaziland	DHS,MICS,WHS	2003,2006,2010
Tanzania	DHS	1999,2003,2007,2010,2011
Togo	DHS	1998
Uganda	DHS	2000,2006,2009,2010,2011
Zambia	DHS,MICS,WHS	2000,2003,2007
Zimbabwe	DHS,MICS,WHS	1999,2003,2005,2009,2010

Survey locations stratified by geographical unit, country and soil-transmitted helminth species

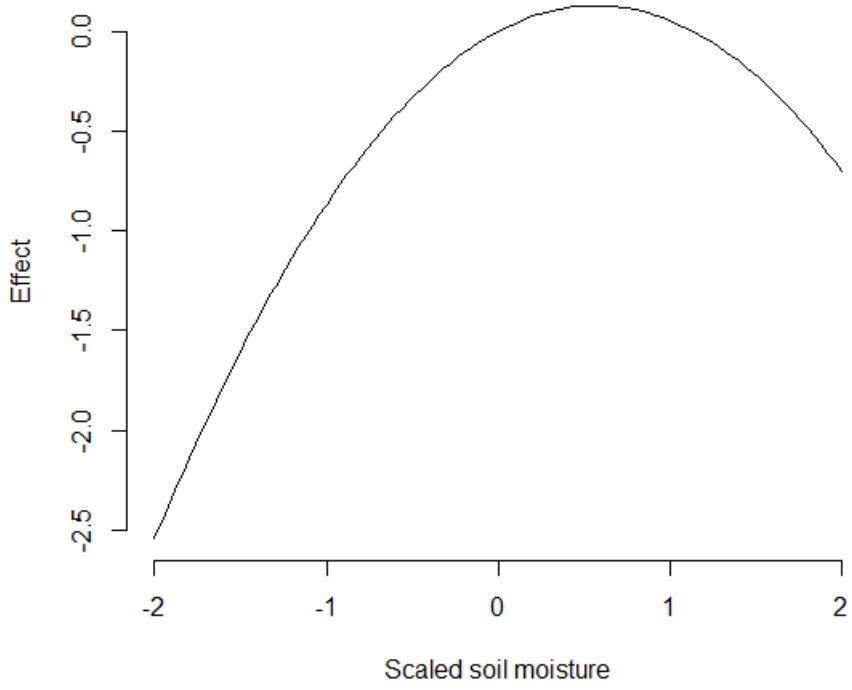
Country	No. of relevant reports	*Ascaris lumbricoides*			*Trichuris trichiura*			Hookworm		
		Unique	Municipality	Point	Unique	Municipality	Point	Unique	Municipality	Point
Angola	4	82	5	77	70	5	65	87	5	82
Benin	4	7	1	6	7	1	6	7	1	6
Botswana	2	1	0	1	1	0	1	7	0	7
Burkina Faso	4	92	0	92	92	0	92	94	0	94
Burundi	1	22	0	22	22	0	22	22	0	22
Cameroon	23	776	0	776	775	0	775	777	0	777
Central African Republic	1	1	0	1	1	0	1	1	0	1
Chad	1	1	0	1	0	0	0	1	0	1
Congo DRC	10	30	0	30	28	0	28	18	0	18
Congo (The Republic of)	0	-	-	-	-	-	-	-	-	-
Côte d'Ivoire	32	252	7	245	247	7	240	255	9	246
Djibouti	0	-	-	-	-	-	-	-	-	-
Equatorial Guinea	1	1	1	0	1	1	0	1	1	0
Eritrea	1	40	0	40	40	0	40	40	0	40
Ethiopia	63	164	6	158	154	2	152	183	6	177
Gabon	3	7	0	7	7	0	7	7	0	7
Gambia	3	1	0	1	0	0	0	3	0	3
Ghana	12	83	0	83	84	0	84	330	0	330
Guinea-Bissau	1	1	0	1	1	0	1	1	0	1
Guinea	5	46	0	46	46	0	46	46	0	46
Kenya	33	738	5	733	735	4	731	736	5	731
Lesotho	0	-	-	-	-	-	-	-	-	-
Liberia	3	12	0	12	10	0	10	12	0	12
Malawi	6	31	0	31	29	0	29	37	1	36
Mali	3	38	38	0	38	38	0	40	38	2
Mauritania	2	9	0	9	9	0	9	9	0	9
Mozambique	4	17	0	17	8	0	8	16	0	16
Namibia	1	0	0	0	0	0	0	5	0	5
Niger	4	133	0	133	132	0	132	134	0	134
Nigeria	121	765	26	739	712	17	695	745	17	728
Rwanda	1	30	30	0	30	30	0	30	30	0
Senegal	11	37	0	37	48	0	48	42	0	42
Sierra Leone	11	86	12	74	85	12	73	85	12	73
Somalia	4	4	0	4	4	0	4	4	0	4
South Africa	7	48	0	48	30	0	30	47	0	47
South Sudan	0	-	-	-	-	-	-	-	-	-
Sudan	6	89	0	89	1	0	1	3	0	3
Swaziland	0	-	-	-	-	-	-	-	-	-
Tanzania	33	156	7	149	145	5	140	160	9	151
Togo	3	1,091	0	1,091	1,091	0	1,091	1,092	0	1,092
Uganda	19	489	2	487	490	2	488	515	2	513
Zambia	9	25	2	23	28	1	27	34	3	31
Zimbabwe	7	216	0	216	196	0	196	208	0	208
Total	459	5,621	142	5,479	5,397	125	5,272	5,834	139	5,695


Maps of the compiled SES data
Percentage of households with access to improved sanitation
Percentage of households with access to improved drinking-water sources
Percentage of adult females who attended at least primary school
< 5
5 - 10
10 - 20
20 - 50
> 50
0
500
1,000
2,000
3,000 km
N

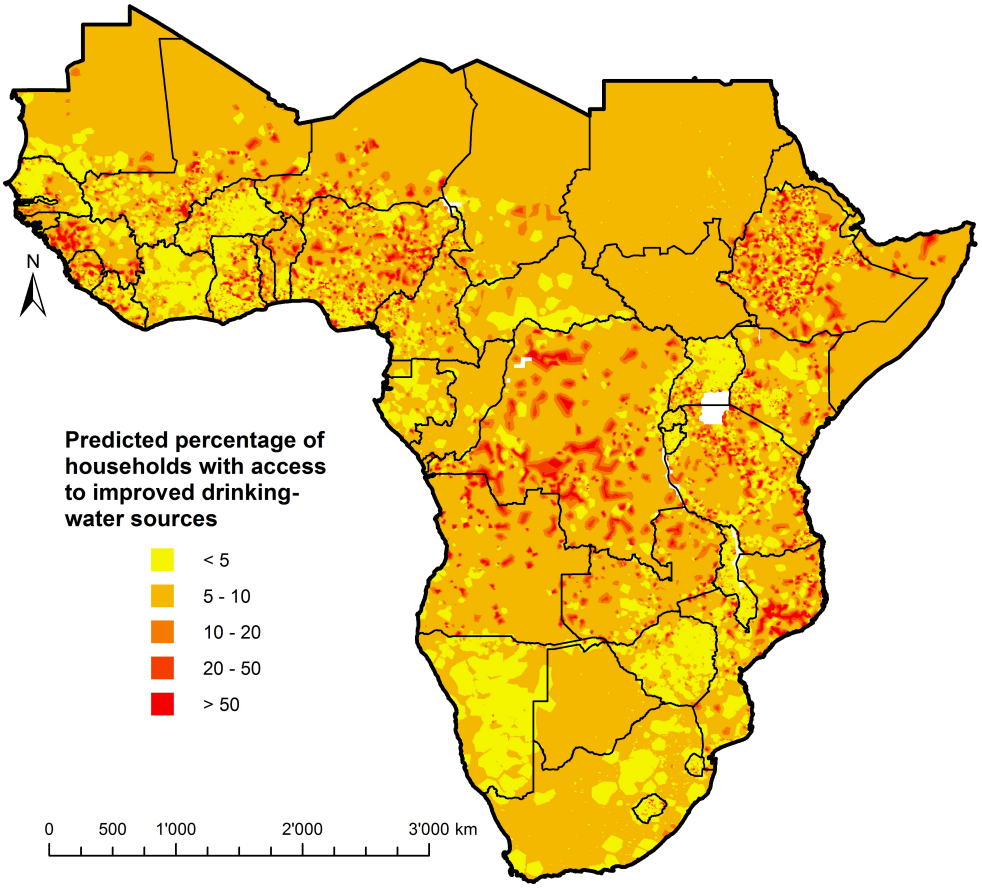
Differences of improved sanitation and drinking-water source in the two periods



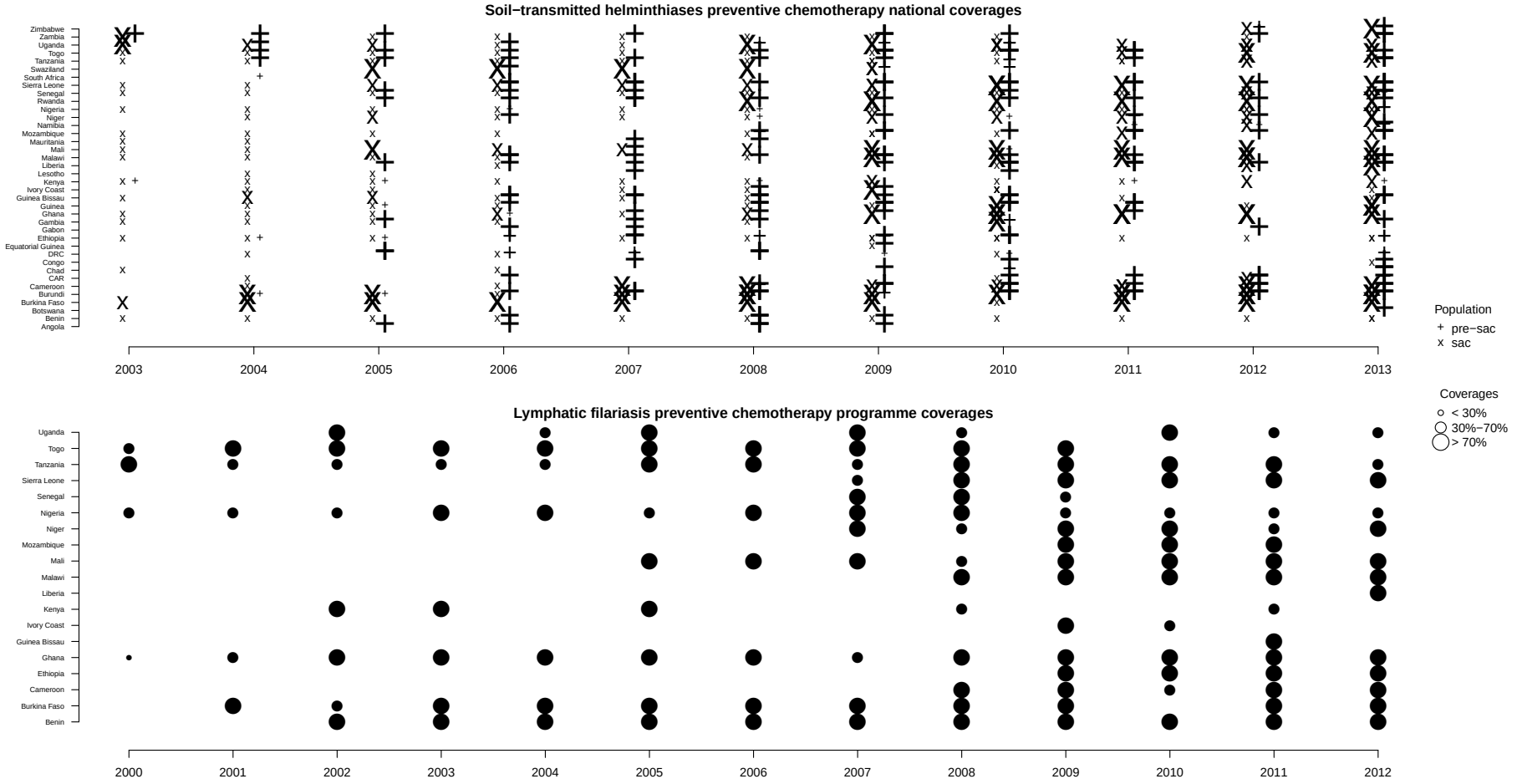
Non-linear effect of moisture on hookworm



Predicted proportion of households with access to improved drinking-water sources



Treatment coverages by country and year



Note that this figure is based on the WHO July 2015 data and includes coverages of 2013 as well. The figure in the published manuscript is based on less data, *i.e.* data available at the time of writing.

Chapter 4

Geostatistical modelling of soil-transmitted helminth infection in Cambodia: do socioeconomic factors improve predictions?

Karagiannis-Voules D.A.[1,2], Odermatt P.[1,2], Biedermann P.[1,2], Khieu V.[1,2,3], Schär F.[1,2], Muth S.[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] National Centre for Parasitology, Entomology and Malaria Control, Ministry of Health, Phnom Penh, Cambodia

This paper has been published in *Acta Tropica* 2015, 141: 204-12.

Abstract

Soil-transmitted helminth infections are intimately connected with poverty. Yet, there is a paucity of using socioeconomic proxies in spatially explicit risk profiling. We compiled household-level socio-economic data pertaining to sanitation, drinking-water, education and nutrition from readily available Demographic and Health Surveys, Multiple Indicator Cluster Surveys and World Health Surveys for Cambodia and aggregated the data at village level. We conducted a systematic review to identify parasitological surveys and made every effort possible to extract, georeference and upload the data in the open source Global Neglected Tropical Diseases database. Bayesian geostatistical models were employed to spatially align the village-aggregated socioeconomic predictors with the soil-transmitted helminth infection data. The risk of soil-transmitted helminth infection was predicted at a grid of 1×1 km covering Cambodia. Additionally, two separate individual-level spatial analyses were carried out, for Takeo and Preah Vihear provinces, to assess and quantify the association between soil-transmitted helminth infection and socioeconomic indicators at an individual level. Overall, we obtained socioeconomic proxies from 1624 locations across the country. Surveys focussing on soil-transmitted helminth infections were extracted from 16 sources reporting data from 238 unique locations. We found that the risk of soil-transmitted helminth infection from 2000 onwards was considerably lower than in surveys conducted earlier. Population-adjusted prevalences for school-aged children from 2000 onwards were 28.7% for hookworm, 1.5% for *Ascaris lumbricoides* and 0.9% for *Trichuris trichiura*. Surprisingly, at the country-wide analyses, we did not find any significant association between soil-transmitted helminth infection and village-aggregated socioeconomic proxies. Based also on the individual-level analyses we conclude that socioeconomic proxies might not be good predictors at an aggregated large-scale analysis due to their large between- and within-village heterogeneity. Specific information of both the infection risk and potential predictors might be needed to obtain any existing association. The presented soil-transmitted helminth infection risk estimates for Cambodia can be used for guiding and evaluating control and elimination efforts.

4.1 Introduction

There is growing interest in spatial modelling of soil-transmitted helminth infections in the current era of major control and elimination efforts against neglected tropical diseases (WHO, 2012a). Indeed, model-based estimates provide a deeper insight of disease distribution in space and time. Hence, such information is essential for disease control managers to provide assistance where and when to focus interventions, including the number of treatments needed (Schur et al., 2011a), and for monitoring purposes (Montresor et al., 2013). Several recent studies provided model-based estimates of soil-transmitted helminth infection at national or sub-continental level; for example, in the Peoples Republic of China (Lai et al., 2013), in South America (Chammartin et al., 2013b) and in sub-Saharan Africa (Karagiannis-Voules et al., 2015a).

Interestingly, these studies consistently showed a decreasing temporal trend of soil-transmitted helminth infection prevalence when comparing data from before 2000 with data from 2000 onwards. It is conceivable that socioeconomic development, coupled with intensified control efforts emphasising preventive chemotherapy (WHO, 2012c; Gallo et al., 2013) are the root causes of these declining trends of soil-transmitted helminthiasis (de Silva et al., 2003; Li et al., 2010; Utzinger et al., 2010). Preventive chemotherapy has been endorsed by World Health Assembly (WHA) resolution 54.19 in May 2001 (WHO, 2002a; Savioli et al., 2009) and annual coverage rates for treatment of school-aged children with albendazole or mebendazole have considerably increased in recent years, although they are still far below the targeted threshold of 75% (WHO, 2010b, 2014). Importantly, WHA resolution 54.19 also urged the promotion of access to safe water, sanitation and health education to combat soil-transmitted helminthiasis and other neglected tropical diseases (WHO, 2002a). Despite the fact that helminth infections are significantly associated with water, sanitation and hygiene(WASH) indicators (Ziegelbauer et al., 2012; Strunz et al., 2014), they have not been used in spatial modelling of soil-transmitted helminthiasis. The prior lack of detailed georeferenced data pertaining to WASH and socioeconomic proxies precluded such analyses. However, the number of projects collecting information on WASH, peoples education attainment and nutrition, and the considerable increase in open-access data repository now permit fundamentally

different ways of scientific inquiry. Indeed, in some countries, the aforementioned proxies are readily available at higher spatial resolution than that of soil-transmitted helminth prevalence, and can thus be explored as potential predictors of infection risk.

In 2003, UNICEF released a report entitled Mapping human helminth infections in Southeast Asia (UNICEF, 2003) that made use of non-spatial models to predict - at high spatial resolution the infection prevalence of soil-transmitted helminths. A relation of environmental predictors and prevalence in Vietnam was applied to the broader Southeast Asia due to lack of data in the rest of the countries. Within the Global Burden of Disease 2010 study (Murray et al., 2012), soil-transmitted helminth prevalence estimates in Southeast Asia were the highest within Asia (Pullan et al., 2014). Globally, the prevalence of *Ascaris lumbricoides* was the second highest among sub-regions after central sub-Saharan Africa. Several countries in Southeast Asia have reached high percentages of preventive chemotherapy coverage (Jex et al., 2011; WHO, 2014). Despite this progress, the burden of soil-transmitted helminthiasis remains unacceptably high (Jex et al., 2011). Particularly in Cambodia, according to the preventive chemotherapy databank of the World Health Organization (WHO; www.who.int/neglecteddiseases/preventivechemotherapy/sth/en/),treatment coverage reached 100% and 80% already in 2006 for preschool-aged children and school-aged children, respectively. Apart from 2009, coverage stayed at such high levels until 2012 and conceivably for 2013 (data will soon become available). However, parasitological results from surveys are still reporting high prevalences (see, for example, Chhakda et al. 2006; Khieu et al. 2013).

The objectives of the current study were: (i) to construct socioeconomic proxies related to water, sanitation, education and nutrition in Cambodia; (ii) to assess their predictive ability in geostatistical risk modelling of soil-transmitted helminth infection; and (iii) to obtain high-resolution estimates of soil-transmitted helminth infection risk in the country, adjusted for environmental and socioeconomic predictors. To create socioeconomic proxies, we compiled household survey data and aggregated them at village level. The predictive ability of these proxies was assessed by carrying out two types of analyses: (i) at country level using village-aggregated

proxies and (ii) at province level using individual-level and village-aggregated proxies for two provinces. We conducted a systematic review of soil-transmitted helminth infections in Cambodia and employed Bayesian geostatistical models adjusted for environmental variables for spatial and temporal risk profiling.

4.2 Methods

4.2.1 Environmental data

Temperature, precipitation and altitude data were obtained from WorldClim (www.worldclim.org). Soil moisture and acidity data were downloaded from the International Soil Reference and Information Centre (www.isric.org).

4.2.2 Socioeconomic and population data

Household data were compiled from readily available Demographic and Health Surveys (DHS; www.measuredhs.com), Multiple Indicator Cluster Surveys(MICS; www.childinfo.org/mics.html) and World Health Surveys(WHS; www.who.int/healthinfo/survey/en/index.html). These international programmes collect data for Southeast Asia. We aggregated the household data at village level and constructed the following proxies of socioeconomic status: (i) percentage of households with improved sanitation; (ii) percentage of households with access to improved drinking-water sources; (iii) infant mortality rate (*i.e.* the probability to die between birth and the first birthday); (iv) percentage of primary school-aged population that is attending primary school (referred to as net attendance rate, NAR); (v) percentage of adult females who attended at least primary school (referred to as attainment); (vi) percentage of adult females who are able to read a whole sentence (referred to as literacy); and (vii) asset index (*i.e.* proportion of people in the poorest category as defined by DHS). The classification of improved sanitation and drinking-water source was performed using the criteria of the Joint Monitoring Programme for Water Supply and Sanitation of WHO and UNICEF (WHO and UNICEF, 2006).

Nutritional indicators were built using the WHO child growth standards (WHO Multicentre Growth Reference Study Group, 2006) and correspond to normalised scores (z-scores) of physical variables adjusted for age; namely, (i) z-score of weight

adjusted for age; (ii) z-score of height adjusted for age; and (iii)z-score of body mass index (BMI) adjusted for age. We used the igrowup R-package provided by WHO (available at: www.who.int/childgrowth/software/en/).

Locations were classified as either rural or urban according to information provided by the Socioeconomic Data and Applications Center of the Center for International Earth Science Information Network (www.sedac.ciesin.columbia.edu). From this source we also obtained a high resolution human influence index grid. Population density for the year 2010 was obtained from Worldpop(www.worldpop.org.uk). Country-specific percentages of people under 20 years of age for the year 2010 were collected from the United States Census Bureau International Database (www.census.gov/population/international/data/idb/informationGateway.php).

4.2.3 Literature review and data extraction

We searched PubMed (www.ncbi.nlm.nih.gov/pubmed) and Web of Knowledge (www.webofknowledge.com) for parasitological surveys, conducted in Cambodia, reporting prevalence of soil-transmitted helminth infections (*A. lumbricoides*, *Trichuris trichiura* and hookworm). The search contained all peer-reviewed literature from inception to 14 February 2014. Additionally, we searched the grey literature, such as reports from Ministries of Health (MoH), helminthiases control programmes and authors personal collections.

We used and further adapted protocols put forth by Chammartin et al. (2013b) and Karagiannis-Voules et al. (2015a) to extract and georeference data from the literature review. Details of this procedure are given in the Appendix. We entered the data in the Global Neglected Tropical Diseases database (see www.gntd.org Hürlimann et al., 2011; Saarnak et al., 2013).

4.2.4 Statistical analysis

The soil-transmitted helminth data and the compiled socioeconomic proxies were analysed using an approach presented elsewhere (Karagiannis-Voules et al., 2015a). In brief, the extracted soil-transmitted helminth prevalence data were analysed through Bayesian binomial geostatistical models, employing environmental and

socioeconomic data as predictors. All computations were performed through integrated nested Laplace approximations (INLA; Rue et al., 2009) and the stochastic partial differential equations approach (Lindgren et al., 2011). We used the software R (R Core Team, 2014) and the INLA package (available at: www.r-inla.org). For more information on implementing geostatistical models with INLA, the reader is referred to Lindgren et al. (2011); Cameletti et al. (2013); Karagiannis-Voules et al. (2013).

To select the set of covariates that best predicts soil-transmitted helminth infection prevalence data, we used a cross-validated logarithmic score proposed by Gneiting and Raftery (2007) and Held et al. (2010) and performed the following steps: (i) selected variable and its functional form (linear, categorical with three or four categories and spline approximations through random walk processes see Rue and Held 2005); from sets of highly correlated covariates (*i.e.* Pearsons correlation coefficient >0.9) to avoid collinearity; and (ii) fitted geostatistical models with all possible combinations of covariates. In both steps, models with the lowest mean logarithmic score were selected.

Country-wide spatial analyses using village-aggregated socioeconomic proxies

In order to spatially align the socioeconomic proxies with the soil-transmitted helminth infection prevalence data, we used Bayesian binomial and Gaussian (for the nutritional z-scores) geostatistical models and predicted these proxies at the disease locations, using the urban classification as a predictor. We then performed the aforementioned model selection. To incorporate the prediction uncertainty of the socioeconomic indicators, we fitted a joint model of the indicators and the prevalence, if the best model included such predictors.

All models included survey period as a binary covariate (cut-off, year 2000), survey type and their interaction. The cut-off year 2000 is chosen due to WHA resolution 54.19, put forward in May 2001, that urged member states to step up preventive chemotherapy. The survey type was considered as a covariate with two levels, corresponding to school-based (defined as surveys conducted in schools or surveys focusing on population aged below 20 years)and community-based surveys.

The best models were used to predict the risk of soil-transmitted helminth infection at a grid of 1 × 1 km that included 183,543 pixels for Cambodia. By overlaying the predicted risk surfaces with the population density grids, and the census-based population percentages, we were able to calculate population-adjusted soil-transmitted helminth infection prevalence at the unit of the province.

Provincial spatial analyses

To further investigate the relationship between potential socioeconomic predictors and soil-transmitted helminth infection risk, we performed the aforementioned Bayesian geostatistical risk factor analysis, using data from two surveys conducted in Takeo and Preah Vihear provinces. Details of the surveys have been presented elsewhere (Khieu et al., 2014b,a). Importantly, while implementing these parasitological surveys, a questionnaire was administered concurrently and participants were asked for several socioeconomic indicators. We included the following indicators as potential predictors of hookworm infection: (i) presence of a latrine at home (binary variable; yes or no); (ii) usual place of defecation (categorical variable; household compound, forest, rice field, toilet); (iii) water source (binary variable; improved or unimproved); (iv) educational attainment (categorical variable; no school, primary school, secondary school, high school, university); and (v) asset index (derived from a principal component analysis, as detailed by Vyas and Kumaranayake, 2006). With regard to improved water sources, the following features were included: dam, lake, pond, private pond, private well, village pond and village well with pump. On the other hand, canal, lake, rain water, river and village well (without pump) were classified as unimproved. Thus, for subsequent analyses both infection and socioeconomic status are known at individual-level. The models included a binary variable for sex and a spline approximation, through a second order random walk process (Rue and Held, 2005), for age.

Geostatistical analyses carried out over large areas often include aggregated socioeconomic data (*e.g.* at village level). These data are either available at the location where infection or disease is observed or obtained from model-based predictions when infection or disease and predictors are not spatially aligned. Data aggregation

ignores within-village variation due to loss of individual information and misalignment might lead to inaccurate predictions due to potentially large between-village variability. The surveys in Takeo and Preah Vihear provinces offer the possibility to resemble such conditions of large-scale studies and compare analyses based on individual-level and aggregated socioeconomic data. Hence, we conducted such a comparison, using Bayesian geo-statistical models by (i) aggregating disease and socioeconomic data at village level and performing the model selection approach described earlier, and (ii) assuming that socioeconomic variables are not available in 10 randomly selected villages and modelled them jointly with hookworm infection prevalence. The later step was repeated 100 times to assess sampling variability and explore the change of effects for different sets of villages at which the socioeconomic data were assumed missing. For both steps, percentages of the individual socioeconomic proxies listed above, were used.For example, percentage of people with latrine at home, percentage of people who usually defecate in toilet, percentage of those with access to improved water sources at the lowest asset index category, etc.

4.3 Results

4.3.1 Exploratory analysis

4.3.2 Socioeconomic proxies

DHS, MICS and WHS collect data all over Southeast Asia (see Appendix). In total, we compiled sanitation, drinking-water, education and nutrition data from 5,687 locations across Southeast Asia. Survey data sources, years, total number of locations and data summaries, stratified by country, are provided in Table 4.1. More than a fourth of the total locations are concentrated in Cambodia.The lowest mean percentages of people with access to improved sanitation and drinking-water sources are observed in Cambodia. Overall, roughly a fourth and half of the population in Cambodia had access to improved sanitation and water, respectively. Nutritional data were only available for Cambodia and Timor Leste. Cambodia and Lao Peoples Democratic Republic (Lao PDR) have the lowest percentages of households with access to improved sanitation and improved drinking-water sources.

People in these two countries have better access to water than sanitation. The opposite is observed in Indonesia, Myanmar and Thailand, where the proportion of people with sanitation is higher than access to drinking-water.

Table 4.1: Survey period, sources, locations and summary measures of socioeconomic proxies for nine countries of Southeast Asia.

Country	Survey year(s)	Source	Total points	Mean % of improved sanitation	Mean % of improved drinking water sources	Mean % females educational attainment	Mean % of literate females	Mean net attendance rate
Cambodia	2000, 2005, 2010	DHS	1,624	26.9	51.8	76.7	49.1	76.5
Indonesia	2002	DHS	1,305	81.3	70.9	92.1	79.8	-
Lao PDR	2003	WHS	200	43.4	56.2	56.9	-	-
Malaysia	2003	WHS	372	98.5	98.4	71.7	-	-
Myanmar	2003	WHS	110	89.4	68.7	67.8	-	-
Philippines	2003	DHS, WHS	1,448	77.3	91.1	97.2	91.7	91.3
Thailand	2005	MICS	76	96.3	59.8	96.6	90.1	99.8
Timor Leste	2009	DHS	454	44.3	61.9	69.8	57.6	73.5
Vietnam	2003	WHS	98	75.6	89.7	89.1	-	-

Figure 4.1 depicts the raw observed socioeconomic data in Cambodia. Percentage of improved sanitation is low in numerous villages in the country. Access to improved drinking-water sources is particularly high in the south-eastern part of the country, close to the capital Phnom Penh. Apart from north-eastern Cambodia, education levels are high. Mean nutritional z-scores in Cambodia show large small-scale spatial heterogeneity with almost no visible patterns. Geostatistical models indicate that the following indicators were associated (results not shown) with the urban classification: percentage of households with access to improved sanitation, percentage of households with access to improved drinking-water sources, asset index, infant mortality rate and all of the education proxies investigated.

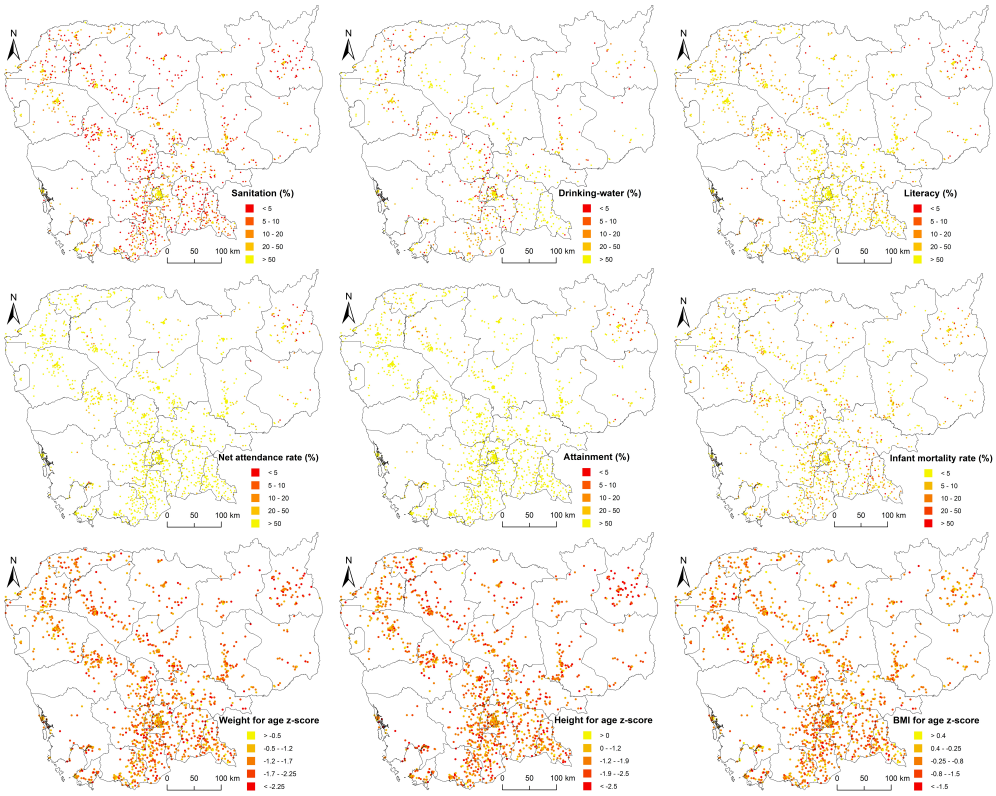


Figure 4.1: Observed socioeconomic proxies across Cambodia.

Boxplots demonstrating the distribution of the socioeconomic proxies according to the two urbanity classes in Cambodia are given in Figure 4.2. Z-scores of nutritional indicators across Cambodia are centred below 0.

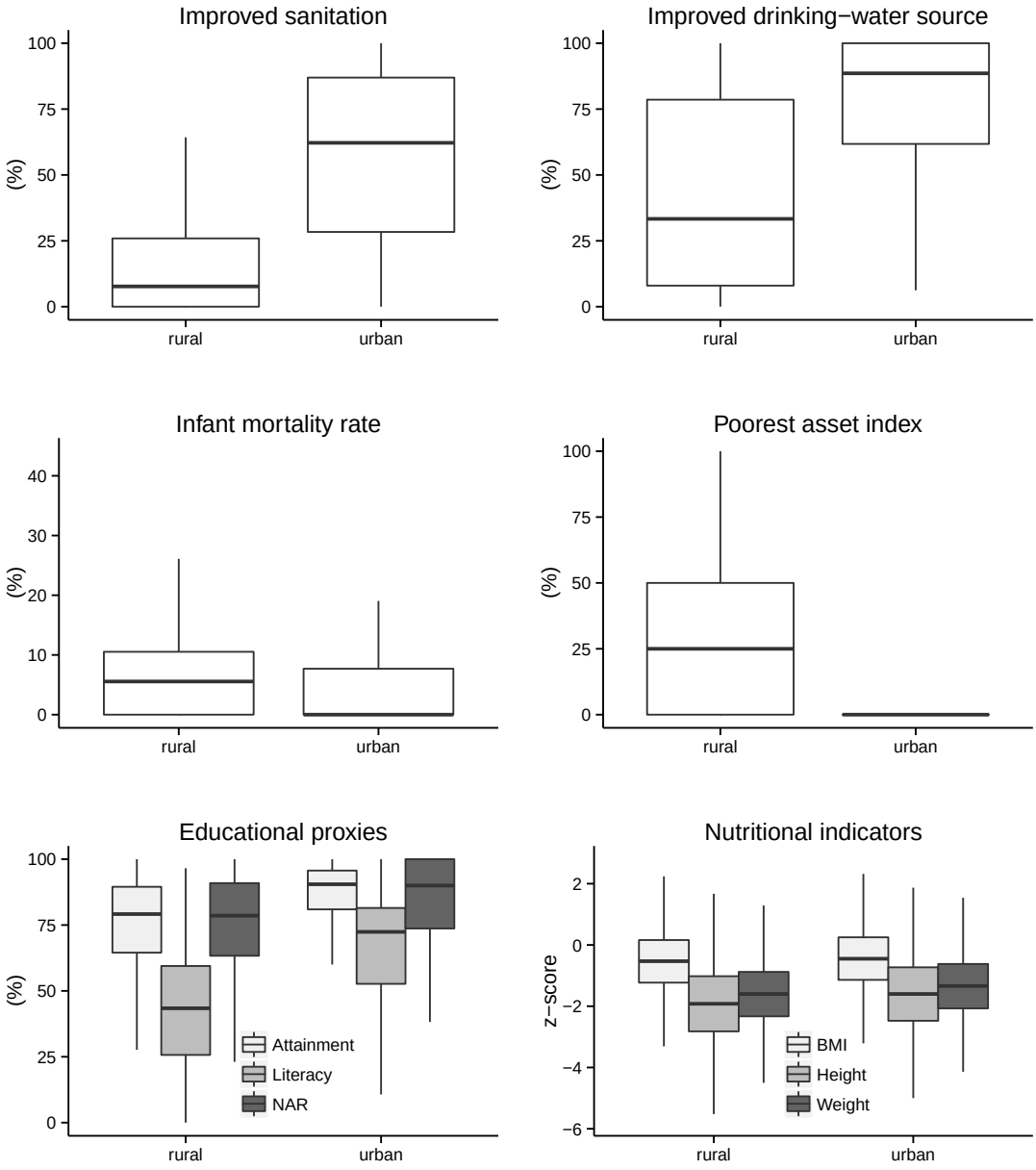


Figure 4.2: Comparison of sanitation, water, education and nutrition in rural and urban settings in Cambodia.

Soil-transmitted helminth infection prevalence data in Cambodia

Overall, we identified 78 sources with potentially relevant soil-transmitted helminth infection prevalence data in Cambodia. Sixteen of these sources contained relevant

survey data, resulting in 238 unique locations. Figure 4.3 shows the observed prevalence data for the three common species of soil-transmitted helminths. In brief, low prevalences of *A. lumbricoides* were observed in the Preah Vihear and Takeo provinces. The prevalence of hookworm infection is high throughout the country with the exception of low-prevalence locations concentrated in the south-centre of the country.

4.3.3 Geostatistical model-based results

Country-wide analyses using village-aggregated socioeconomic proxies

Taken together, we fitted more than 700,000 models for each of the three soil-transmitted helminth species and the estimates of the final models are given in Table 4.2. *A. lumbricoides* was the only species for which a socioeconomic proxy emerged as a potential predictor to explain its spatial distribution, namely females educational attainment. We estimated sharp declines in the prevalence of *A. lumbricoides* and *T. trichiura* from 2000 onwards. A smaller temporal trend is estimated for the risk of hookworm infection. School-aged children are estimated with a smaller prevalence than community-based estimates before 2000 and at similar levels from 2000 onwards for *A. lumbricoides* and *T. trichiura*. For hookworm, no differences in the two population type estimates were found. Urban settlements are associated with lower risk of hookworm infection.

Table 4.2: Posterior estimates (median; 95% credible interval) of the final geostatistical models for species-specific soil-transmitted helminth infections in Cambodia.

Ascaris lumbricoides	Estimate (95% CI)
Annual mean temperature (°C)	
< 27.2	0
27.2-27.4	-0.94 (-1.96, 0.05)
27.4-27.6	-0.03 (-1.11, 1.04)
> 27.6	0.66 (-0.48, 1.79)
Altitude (m)	-0.20 (-0.69, 0.27)
Mean adjustment for females education	-0.37 (-0.79, 0.04)
Survey period (year)	
Before 2000	0
From 2000 onwards	-5.34 (-6.15, -4.57)
Survey type	
Community-based	0
School-based	-0.51 (-0.72, -0.30)
Survey period × survey type	1.38 (0.74, 2.06)
Spatial variance†	1.73 (1.25, 2.41)
Spatial range (km)†	9.16 (6.15, 13.80)
Trichuris trichiura	
Altitude (m)	
< 11	0
11-21	-0.87 (-1.43, -0.31)
21-64	-1.02 (-1.66, -0.39)
> 64	-1.20 (-1.98, -0.44)
Survey period (year)	
Before 2000	0
From 2000 onwards	-3.19 (-3.92, -2.49)
Survey type	
Community-based	0
School-based	-0.95 (-1.27, -0.64)
Survey period × survey type	1.59 (0.34, 2.26)
Spatial variance†	1.67 (1.19, 2.38)
Spatial range (km)†	5.97 (3.77, 9.36)
Hookworm	
Urban-rural classification	
Rural	0
Urban	-0.59 (-1.08, -0.09)
Annual mean temperature (°C)	
< 27.3	0
27.3 - 27.6	-0.29 (-0.62, 0.03)
> 27.6	-1.07 (-1.44, -0.71)
Survey period (year)	
Before 2000	
From 2000 onwards	-0.24 (-0.55, 0.08)
Survey type	
Community-based	0
School-based	-0.19 (-0.42, 0.03)
Survey period × survey type	0.08 (-0.17, 0.33)
Spatial variance†	1.20 (0.90, 1.63)
Spatial range (km)†	3.77 (2.71, 5.16)

† parameter of the spatial process.

The predicted prevalence of soil-transmitted helminth infections from 2000 onwards for the population segment aged below 20 years is shown in Figure 4.3. A high maximum prevalence is predicted for hookworm (84%), while the maximum predicted prevalence for *T. trichiura* and *A. lumbricoides* is considerably lower (24% and 23%, respectively). For the south-central part of Cambodia, the predicted prevalence of *A. lumbricoides* and *T. trichiura* is relatively high (> 15%), while the predicted prevalence of hookworm in the same part is quite low (< 5%).

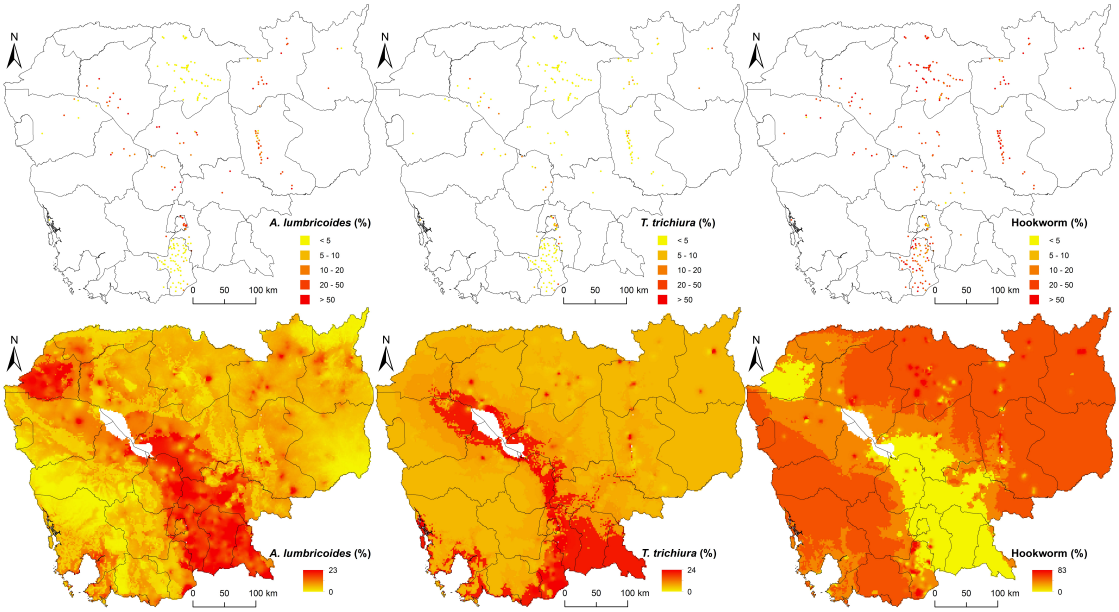


Figure 4.3: Observed soil-transmitted helminth prevalences and model-based predictions for school-aged children in Cambodia for 2000 onwards.

We converted spatial risk profiles into number of people infected, by multiplying with the population grid and calculated population-adjusted prevalence estimates at the unit of the province. The data are summarised in Table 4.3. For the whole country, we obtain prevalences estimates for hookworm, *A. lumbricoides* and *T. trichiura* of 28.7%, 1.5% and 0.9%, respectively. The estimated overall prevalence of any soil-transmitted helminth infection is 30.5%. The highest prevalence for any soil-transmitted helminth infection is predicted for Mondulkiri province, while the lowest for Phnom Penh.

Table 4.3: Population-adjusted prevalence (%) of soil-transmitted helminth infection for school-aged children from 2000 onwards in Cambodian provinces.

Province	Population <20 years	*Ascaris lumbricoides*	*Trichuris trichiura*	Hookworm	All soil-transmitted helminths[†]
Banteay Meanchey	314,767	1.90 (0.90, 4.64)	0.62 (0.39, 1.21)	24.48 (19.90, 29.84)	26.61 (22.23, 32.60)
Battambang	476,250	0.86 (0.46, 1.87)	0.67 (0.36, 1.22)	38.91 (33.25, 44.39)	39.96 (34.54, 45.29)
Kampong Cham	781,090	2.22 (1.26, 4.40)	0.66 (0.39, 1.14)	27.27 (23.07, 32.54)	29.72 (25.66, 34.72)
Kampong Chhnang	217,745	1.47 (0.67, 3.42)	0.93 (0.48, 1.65)	24.09 (19.19, 29.59)	26.05 (21.32, 31.69)
Kampong Speu	326,923	0.88 (0.44, 1.82)	0.54 (0.31, 0.99)	36.53 (30.53, 41.40)	37.44 (31.40, 42.40)
Kampong Thom	292,604	1.51 (0.86, 2.66)	0.67 (0.40, 1.11)	28.04 (23.69, 32.30)	29.53 (25.48, 33.88)
Kandal	589,426	2.04 (1.06, 4.14)	1.32 (0.82, 2.19)	22.23 (17.42, 27.53)	24.84 (20.02, 30.42)
Kep	16,650	0.91 (0.15, 4.55)	1.01 (0.28, 3.19)	42.89 (27.65, 57.39)	44.09 (28.92, 59.20)
Koh Kong	66,462	1.08 (0.39, 2.89)	1.08 (0.64, 2.24)	38.41 (32.32, 44.39)	39.79 (33.72, 45.60)
Kratie	148,332	0.88 (0.50, 1.65)	0.52 (0.31, 0.84)	36.97 (32.42, 42.09)	37.88 (33.56, 43.19)
Mondulkiri	28,467	0.50 (0.12, 25.28)	0.48 (0.26, 0.92)	43.34 (37.33, 49.97)	44.51 (38.14, 56.03)
Oddar Meanchey	79,993	0.78 (0.31, 1.95)	0.52 (0.28, 1.03)	38.06 (31.79, 43.97)	38.97 (33.26, 44.93)
Pailin	33,076	0.36 (0.08, 3.38)	0.44 (0.18, 1.34)	42.86 (32.09, 54.63)	43.46 (32.63, 55.17)
Phnom Penh	619,860	0.87 (0.35, 2.07)	1.05 (0.52, 2.17)	11.42 (7.04, 18.50)	13.33 (8.82, 20.54)
Preah Sihanouk	91,294	1.06 (0.36, 3.32)	0.87 (0.41, 1.96)	30.42 (22.05, 41.25)	31.93 (24.02, 42.58)
Prey Veng	441,357	2.25 (0.96, 4.86)	1.25 (0.74, 2.22)	22.82 (17.30, 28.07)	25.62 (20.61, 30.91)
Pursat	184,810	0.98 (0.48, 2.06)	0.69 (0.38, 1.25)	36.77 (31.39, 42.23)	37.93 (32.69, 43.54)
Ratanakiri	69,959	0.60 (0.16, 5.31)	0.47 (0.22, 1.03)	42.45 (34.52, 50.03)	43.30 (35.65, 51.95)
Siem Reap	423,821	1.04 (0.58, 1.84)	0.71 (0.46, 1.13)	41.62 (36.86, 46.93)	42.70 (37.97, 48.04)
Stung Treng	51,761	0.88 (0.51, 1.64)	0.50 (0.28, 0.90)	39.51 (34.13, 45.75)	40.52 (35.27, 46.66)
Svay Rieng	225,379	1.98 (0.80, 5.07)	1.45 (0.78, 2.66)	23.50 (17.22, 29.87)	26.35 (19.98, 32.79)
Takeo	395,543	1.14 (0.72, 1.84)	1.16 (0.75, 1.93)	33.66 (30.05, 37.62)	35.30 (31.96, 39.17)
Total	5,875,567	1.45 (0.73, 3.24)	0.88 (0.51, 1.59)	28.67 (23.65, 34.26)	30.51 (25.62, 36.18)

[†] Overall prevalence was calculated under the assumption that the three species are independent.

Provincial analyses

The additional individual-level spatial risk factor analyses of hookworm risk resulted in final models with coefficients presented in Table 4.4. Both models included age as smooth random walk process of order 2, which is depicted in the Appendix. The effect of age on hookworm risk shows a bimodal shape with a first peak at age 1525 years and a second peak at age 70 years and above. Males were at a higher risk of hookworm infection than females in both provinces. Hookworm in Preah Vihear is associated with the usual defecation place and, with a reference category of behind the house, all types have a negative coefficient. The smallest coefficient is estimated for people that usually use a toilet for defecation. In both provinces, higher asset index was associated with lower hookworm infection risk. In Takeo province, two additional socioeconomic factors were identified to be linked to hookworm infection. While the existence of latrine at home was negatively associated with hookworm prevalence, the use of unimproved water sources was associated with a higher risk of hookworm infection.

The best model using village aggregated socioeconomic and hookworm data from Takeo province did not identify any socio-economic predictors, while the model using data from Preah Vihear included two socioeconomic proxies. In the model for Preah Vihear the percentage of people who have a latrine at home was negatively associated with hookworm infection, while the percentage of people at the poorest asset index category had a positive effect (results are included in the Appendix). To resemble the misalignment of the socioeconomic and disease data of large-scale surveys, we jointly modelled the socioeconomic proxies identified for Preah Vihear with hookworm infection and assumed that 10 villages selected at random did not have socioeconomic data. We repeated the random selection 100 times. The effects and their 95% credible intervals of the two socioeconomic proxies are given in the Appendix. The effects of both variables are not consistent and 0 is included in many credible intervals.

Table 4.4: Posterior estimates (median; 95% credible interval) of the final geostatistical individual-level models for hookworm infection in Takeo and Preah Vihear provinces, Cambodia, for 2011 and 2010, respectively.

Takeo province	Estimate (95% CI)
Altitude (m)	
< 8	0
8-10	-0.87 (-1.43, -0.31)
10-15	-1.02 (-1.66, -0.39)
> 15	-1.20 (-1.98, -0.44)
Human influence index	
< 25.2	0
25.2-26.55	-0.97 (-1.49, -0.46)
> 26.55	-0.38 (-0.95, 0.19)
Latrine at home	
No	0
Yes	-0.37 (-0.60, -0.15)
Main water source	
Improved	0
Unimproved	0.31 (0.02, 0.60)
Socioeconomic status (based on asset index)	
Poor	0
Less poor	-0.24 (-0.48, 0.00)
Least poor	-0.28 (-0.54, -0.03)
Sex	
Female	0
Male	0.42 (0.24, 0.61)
Spatial variance†	0.74 (0.39, 1.44)
Spatial range (km)†	12.70 (6.40, 23.39)
Preah Vihear province	Estimate (95% CI)
Annual precipitation (mm)	
< 1,600	0
1,600-1,650	-0.87 (-1.54, -0.20)
1,650-1,700	-0.61 (-1.27, 0.04)
> 1,700	-0.86 (-1.54, -0.18)
Usual place of defecation	
Household compound	0
Forest	-0.35 (-0.58, -0.11)
Rice field	-0.31 (-0.64, 0.02)
Toilet	-0.75 (-1.12, -0.38)
Socioeconomic status (based on asset index)	
Poor	0
Less poor	-0.01 (-0.23, 0.21)
Least poor	-0.16 (-0.39, 0.06)
Sex	
Female	0
Male	0.41 (0.24, 0.58)
Spatial variance†	0.99 (0.49, 2.08)
Spatial range (km)†	2.52 (1.26, 4.97)

† parameter of the spatial process

4.4 Discussion

We compiled a large ensemble of socioeconomic data for Southeast Asia using a host of readily available databases. In parallel, we conducted a systematic review to identify surveys of soil-transmitted helminth infections in Cambodia. The data were georeferenced and subjected to geostatistical analyses at the unit of the village to explore whether specific socioeconomic proxies can improve upon spatial risk profiling, while adjusting for environmental covariates. Moreover, we focussed on two provinces and did individual-level analysis to determine the association of household socioeconomic indicators and soil-transmitted helminth infection. We estimated population-adjusted prevalences for Cambodia and assessed the usage of socioeconomic proxies at different scales for geostatistical analyses. The risk of all three soil-transmitted helminth infections in Cambodia has considerably declined from 2000 onwards, probably due to a combination of overall socioeconomic development and escalating anthelminthic treatment coverage rates. Indeed, treatment coverage reached 100% already in 2006 and the administration of mebendazole had initially been promoted through the schistosomiasis control programme (Sinuon et al., 2007). Over the past several years, treatment coverage of school-aged children with either albendazole of mebendazole reached levels of 75% (WHO, 2014). Furthermore, programme coverage (which targets whole communities) against lymphatic filariasis in Cambodia has remained above 70% from 2005 to 2009 (http://www.who.int/neglecteddiseases/preventivechemotherapy/lf/en/). Interestingly, before 2000, school-aged children had lower prevalence rates compared to entire communities for both *A. lumbricoides* and *T. trichiura*. Accounting for an interaction of survey type and study period shows that, after 2000, community level prevalences dropped and reached similar prevalences as observed in school-aged children. In other regions of the world, the opposite observations have been made; initially, prevalences of soil-transmitted helminths were higher among school-aged children, but as control efforts emphasising preventive chemotherapy in the school-aged population went to scale, prevalence rates in the school-aged children and entire communities approached each other. This issue has been well documented for sub-Saharan Africa, when comparing data among school-aged children and adults for the time before 2000 and from 2000 onwards (Karagiannis-Voules et al.,

2015a).

Clear effects of socioeconomic proxies studied here on the risk of soil-transmitted helminth infections were not identified in the country-wide analyses, including village-aggregated and misaligned socioeconomic predictors, despite the common belief that soil-transmitted helminth infections are intimately connected with poverty (Hotez, 2008). For instance, Ziegelbauer et al. (2012) meta-analysed individual-based studies and showed that people who have access to and use latrines are at a significantly lower odds of a soil-transmitted helminth infection than their counterparts who have no latrine. In a more recent meta-analysis, Strunz et al. (2014) showed that piped water access was negatively associated with *A. lumbricoides* and *T. trichiura* infections. In a study of hookworm intensity, education of primary caregivers of children had a negative effect (Pullan et al., 2010). In a geostatistical analysis of hookworm prevalence in West Africa, which predicted socioeconomic proxies that were used as predictors of infection risk, several associations were identified (Magalhães et al., 2011). However, this analysis did not take into account the prediction error in a joint modelling approach.

Socioeconomic proxies might not be good predictors at an aggregated large-scale analysis due to considerable between- and within-village heterogeneity. First, the spatial misalignment of socioeconomic and soil-transmitted helminth infection prevalence data gives rise to several issues. As shown in Figure 4.1, almost all socioeconomic proxies showed a high degree of small-scale heterogeneity, and hence, visible patterns were negligible. Large between-locality variability is responsible for high prediction uncertainty of socioeconomic data, which might confound its effect on soil-transmitted helminths. Second, aggregation of individual data for specific localities resulted in substantial loss of variability and information. For instance, individual nutritional z-scores ranged from -6 to $+6$, while location means were not reaching these extremes (Figures 4.1 and 4.2). It follows that the use of such predictors for large-scale mapping of soil-transmitted helminthiasis and perhaps other neglected tropical diseases might not capture the expected associations between risk of infection and socioeconomic status.

It is conceivable that, unless individual data on both infection and socioeconomic predictors are available, as in the studies mentioned above, finding clear associations

might be setting-specific.This claim is further supported by the two individual-level analyses performed for the provinces of Takeo and Preah Vihear. The best models for hookworm infection included predictors such as the usual defecation place, the existence of a latrine at home, the main water source and asset index, which were available at individual-level. On the other hand, the analyses of the same surveys with aggregated data or misaligned predictors (*i.e.* data artificially resembling the conditions which characterise large-scale studies)indicated no socioeconomic effects for Takeo. However, under-standing the reasons that village-specific socioeconomic proxies may not be good predictors of the infection risk is complicated by intensified control which can blur the exposure-disease relations.

In this context, however, it is interesting to note that hook-worm species and their transmission dynamics to humans might vary geographically. In Southeast Asia, the transmission of zoonotic *Ancylostoma ceylanicum* from dogs to humans has been demonstrated (Traub et al., 2008). In a recent study in Preah Vihear province, zoonotic *A. ceylanicum* was diagnosed in about half of the hookworm-infected individuals (Inpankaew et al., 2014). Therefore, even if improved WASH facilities are used by the communities, contamination of the environment with hookworm is assured by dogs.

In recent years, several surveys have been carried out in Cambodia, but the data were not available for the current analysis. For instance, the World Vision Cambodia surveyed 1880 children in three provinces (George et al., 2012). Furthermore, the schistosomiasis control programme also reported soil-transmitted helminth infection prevalences in 2007 (Sinuon et al., 2007). Efforts should be made to obtain and georeference these additional data, so that future model-based analyses of soil-transmitted helminth infection risk in Cambodia are further enhanced. In addition, the used survey data are extracted from different sources. Thus, there are potential biases related to spatial coverage of survey locations and between-surveys heterogeneities in the age groups sampled and diagnostic tools used that have been discussed elsewhere (Chammartin et al., 2013b; Karagiannis-Voules et al., 2015a). We followed an extraction protocol (see Appendix) to limit such sources.

In conclusion, our analyses contribute to a deeper under-standing of socioeconomic predictors for large-scale model-based geostatistical analysis of soil-transmitted

helminthiasis. Individual information of both infection risk and potential predictors are still needed to identify significant and biologically meaningful associations between parasite infection and socioeconomic variables. Clearly, the presented risk maps for the three common soil-transmitted helminth infections for Cambodia can be utilised for prioritising control efforts, spatially designing new surveys and serve as a benchmark for long-term surveillance. Along with previous attempts to map and predict the spatial and temporal distribution of soil-transmitted helminth infections elsewhere in Asia, Africa and Latin America (Chammartin et al., 2013b; Lai et al., 2013; Pullan et al., 2014; Karagiannis-Voules et al., 2015a), the work presented here contributes to a new global trend of elaborate spatial analyses of soil-transmitted helminth infections.

Author contributions

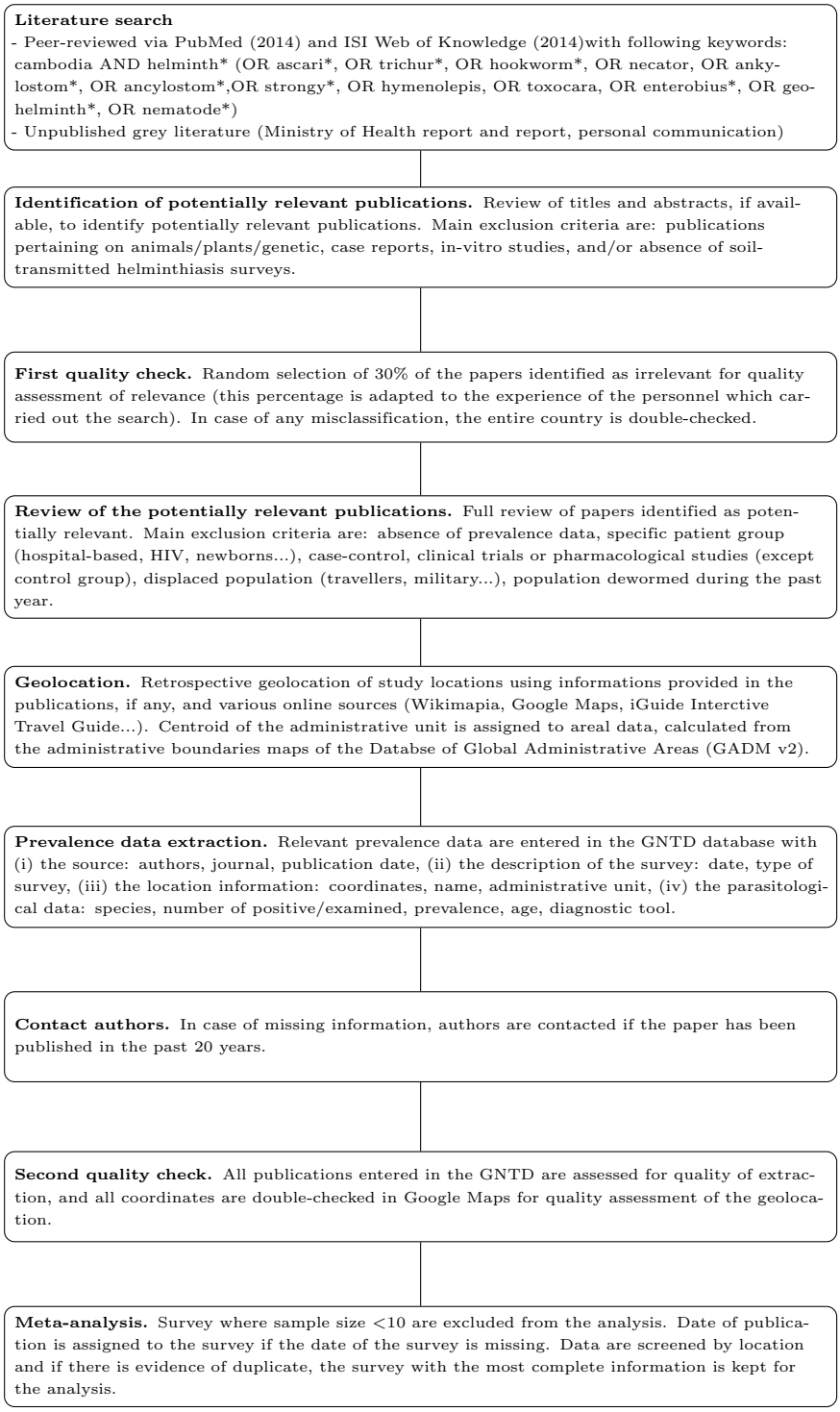
DAKV processed and analysed the data, interpreted the results and wrote the first draft of the manuscript. PB performed the systematic literature search. DAKV and PB extracted the data. PV extracted the water, sanitation and education data. DAKV and PV processed the water, sanitation and education data. DAKV, JU and PV developed the protocol and search strategy for the systematic review. VK, FS and PO provided substantial amount of data. VK and MS assisted in geolocation of surveys. PV and JU conceptualised the project and revised the manuscript. All authors approved the final version of the manuscript prior to submission.

Acknowledgments

We thank the Demographic and Health Surveys, Multiple Indicator Cluster Surveys, World Health Surveys for making the water, sanitation, education and nutrition data available. This investigation received financial support from the Swiss National Science Foundation (http://www.snf.ch, project no. PDFMP3-137156) and the European Research Council (http://erc.europa.eu, grant no.323180).

4.5 Appendix

Extraction protocol



Effect of age in individual level analyses

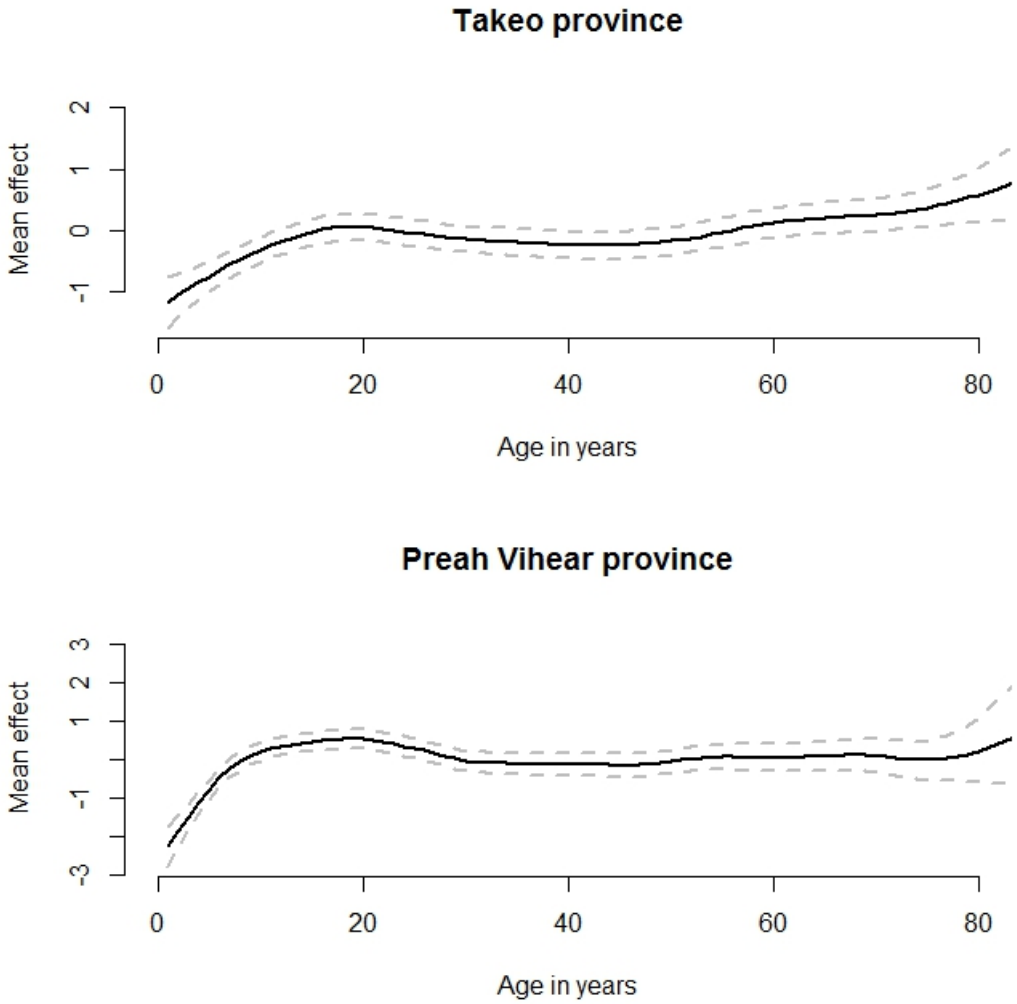
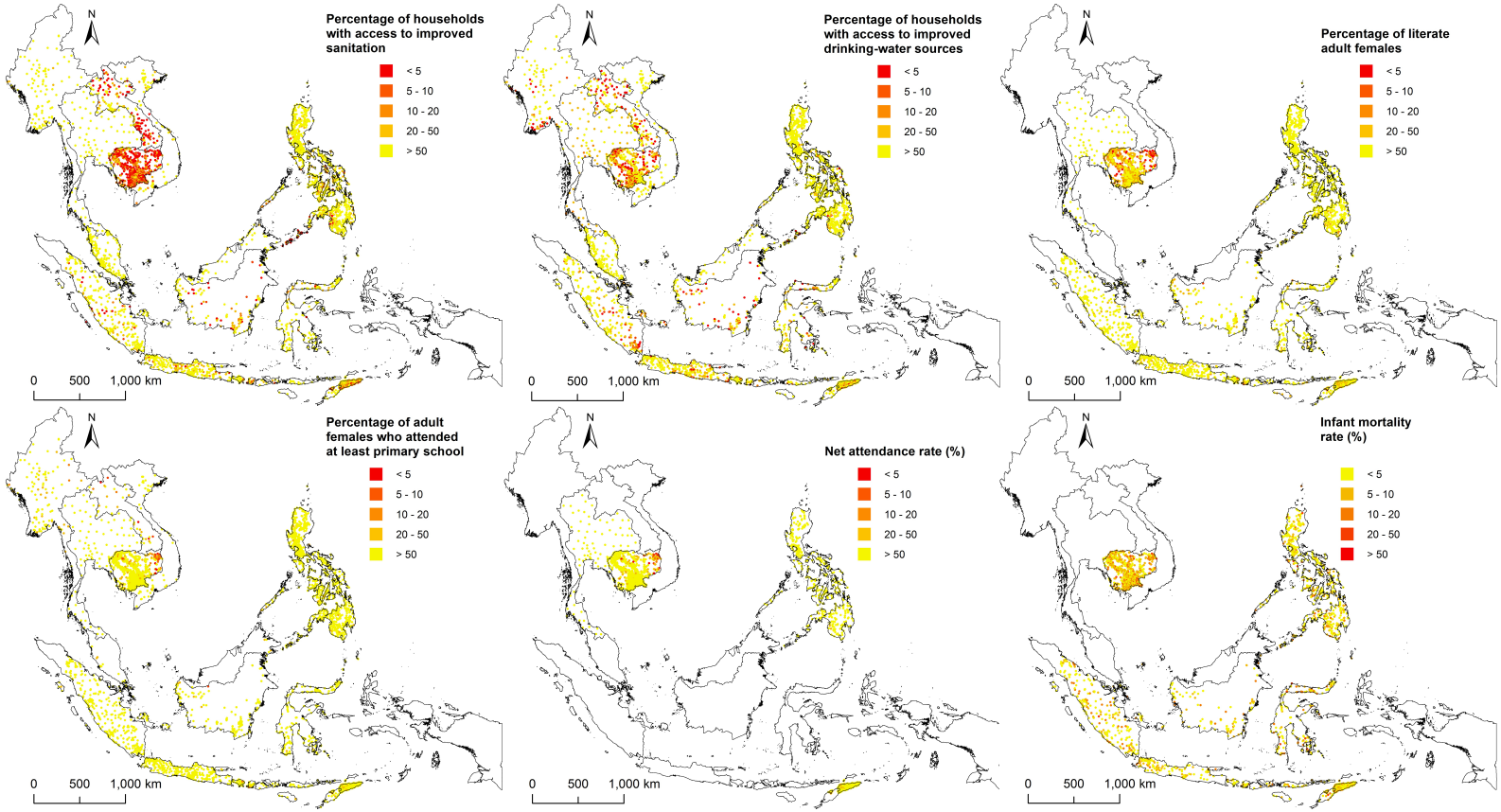


Figure 4.4: The smooth effect of age on hookworm risk in the two individual-level analyses in Takeo and Preah Vihear provinces for 2011 and 2010, respectively

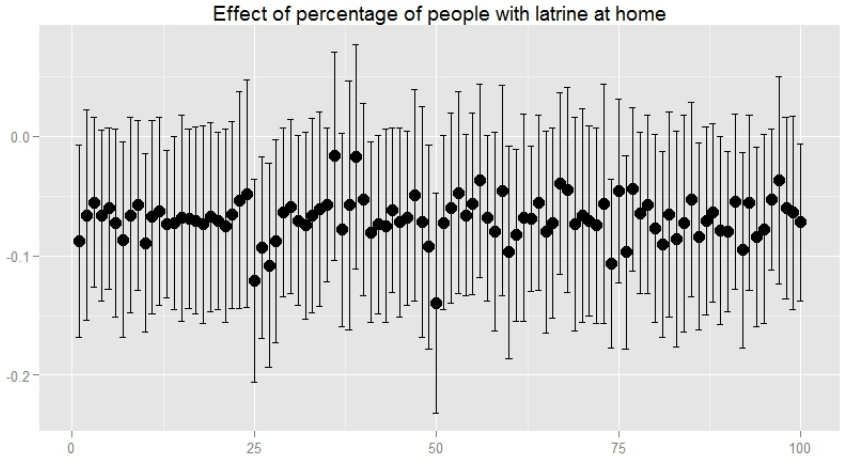
Socioeconomic proxies in Southeast Asia

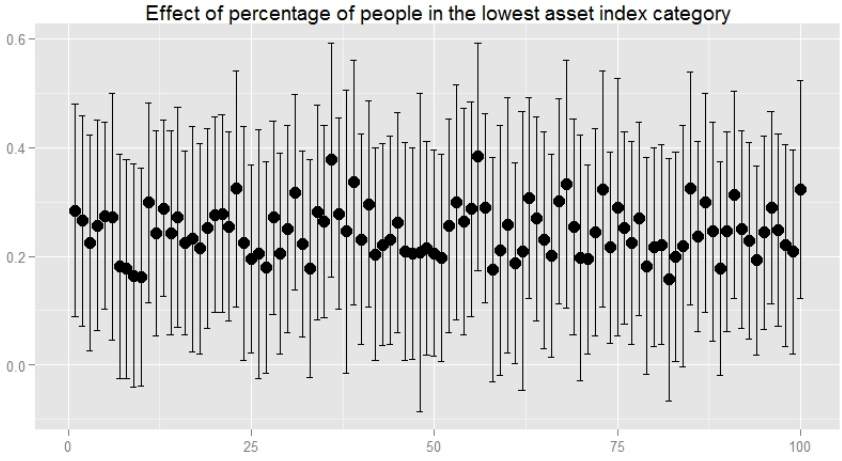


Model results from aggregating the individual-level data

Takeo province	Estimate (95% CI)
Altitude (m)	
< 8	0
8-10	-0.67 (-1.32, -0.02)
10-15	1.03 (0.34, 1.71)
> 15	1.04 (0.34, 1.74)
Human influence index	
< 25.2	0
25.2-26.55	-0.92 (-1.43, -0.41)
> 26.55	-0.32 (-0.88, 0.25)
Spatial variance	0.75 (0.40, 1.41)
Spatial range (km)	11.67 (6.05, 20.92)
Preah Vihear province	
Annual precipitation (mm)	
< 1600	0
1,600-1,650	-0.52 (-1.00, -0.02)
1,650-1,700	-0.43 (-0.95, 0.08)
$> 1,700$	-0.29 (-0.82, 0.24)
Precipitation of warmest quarter (mm)	
< 259	0
259-264	-0.04 (-0.60, 0.53)
264-281	-0.79 (-1.29, -0.30)
> 281	-1.07 (-1.57, -0.57)
Percentage of people in the lowest asset index category	1.25 (0.38, 2.12)
Percentage of people with latrine at home	-0.89 (-1.78, -0.02)
Spatial variance	0.42 (0.19, 0.96)
Spatial range (km)	1.85 (0.83, 3.98)

Socioeconomic effects on hookworm using joint models for Preah Vihear province





Chapter 5

Bayesian variable selection for spatially varying coefficients: an application in malaria intervention effects in Angola

Karagiannis-Voules D.A.[1,2], Vounatsou P.[1,2]

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

This manuscript will be submitted to *Statistics in Medicine*.

Abstract

A number of country-wide surveys are conducted in Africa collecting georeferenced data on disease risk and interventions, such as mosquito bednet coverage, insecticide residual spraying and access to malaria treatment. Control programs are interested in assessing effects of interventions not only at country level but also at sub-national level where the interventions are delivered. There are different measures quantifying the coverage of bednet ownership and usage. Studies have shown that different indicators are able to capture the effect of interventions on malaria risk in different surveys. Statistical models are able to estimate covariate effects varying in space and identify the most important intervention indicators via variable selection approaches. We provide geostatistical model formulations for variable selection of spatially varying covariate effects and assess sensitivity of inference to model specification. We analyze data from the two most recent national malaria surveys in Angola, carried out in 2006 and 2011. We implement the proposed models to identify the most important intervention indicators and to assess the effects of their changes between the two time points on the dynamics of parasitaemia risk at national and sub-national level. The models adjust for climatic confounders. Results showed that an increase in nets-to-people ratio had an important contribution in the parasitemia risk reduction within the five year period in Huambo province. Models with spatially varying effects and variable selection schemes can be included in routinely fitted geostatistical models of malaria survey data to identify areas with successful implementation of control programs.

5.1 Introduction

Geostatistical modelling of malaria risk or incidence is an important tool for burden estimation and disease surveillance. It has been used to model disease data covering different spatial scales. Gosoniu et al. (2010) predicted malaria risk and the number of infected children, aged less than 5 years, in Angola and Samadoulougou et al. (2014) conducted a similar analysis in Burkina Faso. Giardina et al. (2012) used zero-inflated geostatistical models to predict malaria risk in Senegal. A sub-continental study depicted the spatial distribution of malaria in Africa (Noor et al., 2014). Gething et al. (2011, 2012) produced endemicity world malaria maps for *Plasmodium falciparum* and *P. vivax*, respectively. Recently, Bhatt et al. (2015) showed a 50% decrease of prevalence in sub-Saharan Africa since 2000 using geostatistical models.

Malaria is environmentally driven disease and environmental proxies are used as predictors. Incorporating effects of control interventions in such model-based analyses, while adjusting for environmental and socioeconomic predictors, may provide additional information on control progress and improve risk modelling. National malaria surveys are conducted to obtain information on both the disease and control interventions. They have been developed by the Roll Back Malaria Partnership (RBMP) with a consistent design to assist monitoring and evaluation and to assess progress towards the targets of the Global Malaria Action Plan.

Based on these surveys, different indicators of control interventions have been defined related to insecticide-treated mosquito nets (ITN) ownership and usage as well as indoor insecticide residual spraying and access to malaria treatment (MEASURE Evaluation et al., 2013). Studies have shown that different indicators may have an important effect on malaria in different surveys. Variable selection methods can be used to choose the most important indicators.

Furthermore, the effect of such an intervention proxy may not be constant but vary in space. Giardina et al. (2014), showed that malaria intervention measures might not be associated with the disease prevalence if they were considered constant across the study area. Their contribution to the prevalence model was apparent if the effects were allowed to vary according to province or district. The authors

used a conditional autoregressive (CAR) process to model the change of the effect in space and highlight the regions where interventions were less (or more) effective than the countrywide average. However, the above study selected the best indicators of intervention effects conditional on climatic predictors ignoring the spatially varying effects in the variable selection. A more rigorous approach would incorporate variable selection within a geostatistical model that includes locally varying coefficients and adjust for all other predictors.

Bayesian variable selection approaches (see, for example, O'Hara and Sillanpää, 2009) can be applied to fixed and random effects in a model. For a multivariate response with geostatistical random slopes (Gelfand et al., 2003), Reich et al. (2010) formulated a Bayesian variable selection approach which introduces indicators for each predictor that allow covariates to enter in the model with either a fixed or a varying effect (or excluded) and are constant across space. Boehm Vock et al. (2015) assumed that indicators vary independently in space and incorporated a spatial dependence of effects through a Gaussian copula. A spatial dependence on the inclusion indicators was recommended by Lum (2012).

The objective of the current study is to assess the sensitivity of the different Bayesian variable selection approaches in capturing the association of malaria risk and predictors, and to estimate the effects of malaria interventions in space.

5.2 Methods

5.2.1 Data

We obtained parasitological data in Angola for 2006 and 2011 from the Demographic and Health Surveys (DHS) Program http://www.dhsprogram.com/. The data were collected from children aged below 5 years. The complete dataset contained 5096 children living in 337 clusters, of which 113 and 224 were surveyed in 2006 and 2011, respectively. Parasitemia in 2006 was measured through a rapid diagnostic test while in 2011 with microscopy. Additional information on household socioeconomic proxies and malaria interventions was collected. In this study, we use the proportion of mothers without any education, rural or urban classification, intervention proxies related to ITN ownership and usage, access to malaria treatment (case management)

and a socioeconomic proxy defined by the proportion of househlds within the two lowest quintiles of the asset index.

We constructed malaria intervention proxies using indicators suggested by RBMP (see page 7 in MEASURE Evaluation et al., 2013). Based on the available data, we consider indicators that correspond to two proxies of ITN usage: (i) percentage of people who slept under an ITN the night before the survey (USE1), and (ii) percentage of children, aged below 5 years, who slept under an ITN the night before the survey (USE2); three proxies of ITN ownership: (i) percentage of households in a cluster with at least one ITN (OWN1), (ii) percentage of households in a cluster with at least one ITN for every two people (OWN2) and (iii) mean nets to people ratio (OWN3); and one proxy of case management: percentage of children, aged below 5 years, that received Artemisinin-based combination therapy (or other appropriate treatment) among those with fever in the last 2 weeks who received any antimalarial drugs (CASE1).

The observed data on malaria prevalence as well as the difference of 2006 and 2011 intervention coverage indicators are depicted in Figure 6.1. We did not include in the analysis Cabinda province because it is spatially separated from the rest of the country.

To adjust for environmental predictors we used climatic proxies of temperature, rainfall, altitude, distance to water and vegetation averaged over the year prior to each survey. Specifically, land surface temperature at day (lstd), as well as the normalized difference vegetation index (ndvi) were obtained from MODIS (Oak Ridge National Laboratory Distributed Active Archive Center, 2011). Using the same source and the land cover classification of water, we defined the distance to the closest water body. Altitude was downloaded from http://srtm.csi.cgiar.org. Rainfall was downloaded from the Famine Early Warning System Network of the United States Agency for International Development http://earlywarning.usgs.gov/fews/index.php. For details of the spatial and temporal resolutions of the predictors see, for example, Diboulo et al. (2015).

The covariates used in the models correspond to the difference between 2006 and 2011 at a given location. For the urban classification, altitude and distance to water

we used the 2011 value instead of the difference due to their negligible changes over time.

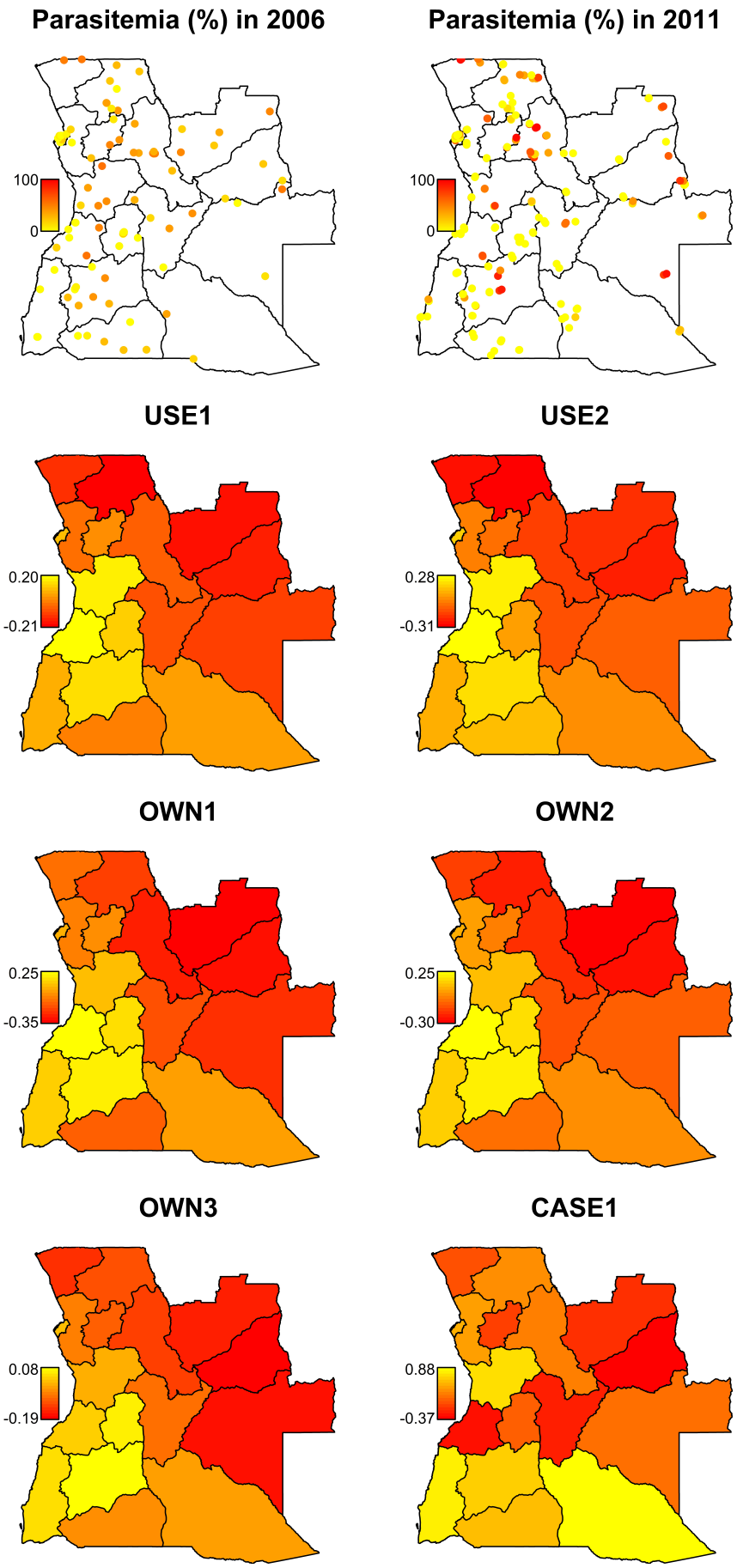


Figure 5.1: Raw data of malaria parasite prevalence in 2006 and 2011 surveys and of the difference of intervention coverage indicators between the two surveys.

5.2.2 Spatially varying regression model

Let Y_i be the response variable observed at location s_i belonging in province m $(m = 1 \dots M)$, $X_{N\times P}$ and $D_{N\times K}$ are design matrices. The linear predictor, which includes fixed and random effects, is formulated as:

$$\eta_i = g\left(E\left(Y_i\right)\right) = b_0 + \mathbf{X}_i^T\mathbf{b} + w_{i0} + \sum_{\substack{k=1\\ i\in m}}^{K} D_{ik}w_{m_ik} \tag{5.1}$$

where $\mathbf{b}$ is the vector of fixed effects, b_0 and $\mathbf{w}_0$ are the fixed and random intercepts. The varying effects are $\mathbf{w}_k$, $k = 1, ..., K$. In our application, Y_i is the number of infected children with malaria and is binomially distributed, while g is the *logit* transformation. We follow a Bayesian hierarchical formulation and define prior distributions for the model parameters.

Each $\mathbf{w}_k$ represents locally varying coefficients for covariate k and has dimension to the number of provinces M. Commonly, it is assumed as a realization of a Gaussian process, *i.e.* $\mathbf{w}_k|\beta_k, \mathbf{\Sigma}_k \sim \mathcal{N}_d\left(\boldsymbol{\beta}^{(k)}, \mathbf{\Sigma}_k\right)$ with $\boldsymbol{\beta}^{(k)}$ being a vector of β_k which is the fixed slope. Through the covariance matrix $\mathbf{\Sigma}_k$, different structures can be accommodated. For instance, if $\mathbf{\Sigma}_k = \mathbf{0}$ then there is no deviation from the mean β_k and only a fixed slope exists. Alternatively, if $\mathbf{\Sigma}_k$ is a diagonal matrix then $\mathbf{w}_k$ are exchangeable and deviate from β_k independently. In case $\mathbf{w}_k$ is defined by location s_i, a geostatistical random slope could be modeled. Due to the fact that malaria control is implemented over administrative units, an effect that varies by province m could depict provinces where interventions have an effect on malaria or not. The province-specific effects can be modelled by a conditional autoregressive Gaussian process, *i.e.* $\mathbf{\Sigma}_k^{-1} = \sigma_k^{-2}\left(\mathbf{R} - \rho_k\mathbf{\Omega}\right)$, with $\mathbf{R}$ being a diagonal matrix with entries the sum of the neighbours of each province and $\mathbf{\Omega}$ a proximity matrix (Banerjee et al., 2014).

Through $\mathbf{w}_0$ we model a geostatistical intercept assuming a Gaussian prior $\mathbf{w}_0|\mathbf{\Sigma}_0 \sim \mathcal{N}_N\left(\mathbf{0}, \mathbf{\Sigma}_0\right)$. We use the exponential correlation function $\Sigma_0\{i,j\} = \sigma_0^2 \exp\left(-\rho_0 d_{ij}\right)$ where σ_0^2 and ρ_0 are the variance and spatial decay parameters of the Gaussian process, and d_{ij} is the Euclidean distance between locations s_i and s_j.

We assign a normal prior distribution to each b_p, that is $b_p|\tau_p^2 \sim \mathcal{N}\left(0, \tau_p^2\right)$ (similarly

for β_k). For variances (σ_0^2, σ_k^2) we use the inverse gamma distribution with shape and scale parameters equal to 2 and 1, respectively. A uniform prior in $(0.7, 100)$ is assigned to ρ_0 and in $(-1, 1)$ to ρ_k.

We model the malaria prevalence of 2011 and adjust for the levels of 2006 by including an offset term of the prevalence on the *logit* scale. The two surveys were carried out on different set of locations. We spatially align the data by predicting the prevalence of 2006 at the 2011 locations using a geostatistical model that included climatic predictors. These predictors were selected via a stochastic search variable selection (SSVS, George and McCulloch, 1996).

5.2.3 Bayesian variable selection

To extend the above model and incorporate variable selection in the fixed effects, we introduce an indicator γ_p for each X_p. We perform a SSVS and define a spike and slab prior $b_p|\delta_p, \tau_p^2 \sim \mathcal{N}\left(0, \delta_p\tau_p^2 + (1-\delta_p)\, u_0\tau_p^2\right)$ (see, for example, George and McCulloch, 1996). u_0 is a shrinkage factor which we fix to 0.001. δ_p is assigned a Bernoulli prior with 0.5 probability of success.

Similarly, for each random slope $\mathbf{w}_k$, we introduce a vector of indicators $\boldsymbol{\gamma}_k$ and redefine the prior of $\mathbf{w}_k$ conditional on $\boldsymbol{\gamma}_k$. The vector $\boldsymbol{\gamma}_k$ may illustrate whether predictor k has a fixed or random slope. The indicators $\boldsymbol{\gamma}_k$ may be constant across all provinces or may be province-specific.

To achieve a specification that assumes common indicators across provinces, we define $\boldsymbol{\gamma}_k = \left\{\gamma_k^{(1)}, \gamma_k^{(2)}\right\}$. $\gamma_k^{(1)}$ is introduced for the fixed effect β_k and $\gamma_k^{(2)}$ for the random slope $\mathbf{w}_k$. A spike and slab prior is used for β_k, that is $\beta_k|\gamma_k^{(1)}, \tau_k^2 \sim \mathcal{N}\left(0, \gamma_k^{(1)}\tau_k^2 + \left(1-\gamma_k^{(1)}\right) u_0\tau_k^2\right)$. The $\mathbf{w}_k$ can be represented through its variance and σ_k^2 can be either treated in the linear predictor as a covariate with a spike and slab prior (Wagner and Duller, 2012), or can be replaced by $\sigma_k^2 = \gamma_k^{(2)}\tilde{\sigma}_k^2 + \left(1-\gamma_k^{(2)}\right) u_0\tilde{\sigma}_k^2$ with the inverse gamma prior placed on $\tilde{\sigma}_k^2$. We assign a multinomial prior distribution to $\left\{\gamma_k^{(1)}, \gamma_k^{(2)}\right\}$ with possible events: $\left\{\left(\gamma_k^{(1)}=0, \gamma_k^{(2)}=0\right), \left(\gamma_k^{(1)}=1, \gamma_k^{(2)}=0\right), \left(\gamma_k^{(1)}=1, \gamma_k^{(2)}=1\right)\right\}$. The probabilities of the multinomial distribution are chosen to allow 50% exclusion. The remaining 50% is equally splitted in the last two events. The event $\left(\gamma_k^{(1)}=0, \gamma_k^{(2)}=1\right)$ is assigned 0 prior

probability due to the fact that $\mathbf{w}_k$ is the deviation from β_k and should enter in the model in case the fixed slope is not 0. We will refer to this formulation as "Model1". Reich et al. (2010) used a discrete spike and based inferences on the marginal posterior of $\left(\gamma_k^{(1)}, \gamma_k^{(2)}\right)$ that has a closed form under a Gaussian likelihood.

To relax the assumption of global inclusion or exclusion of a predictor and allow locally varying indicators, we introduce $\boldsymbol{\gamma}_k = \{\gamma_{1k}, \ldots, \gamma_{Mk}\}$. Then, $\mathbf{w}_k$ is not defined jointly but each of its elements (w_{mk}) is defined conditionally on γ_{mk}, that is $w_{mk}|\gamma_{mk}, \beta_k, \tau_k^2 \sim \mathcal{N}(\gamma_{mk}\beta_k, \gamma_{mk}\sigma_k^2 + (1-\gamma_{mk})u_0\sigma_k^2)$. We use two specifications for $\boldsymbol{\gamma}_k$. Firstly (Model2), we impose a spatial structure in $\boldsymbol{\gamma}_k$ by assigning a Bernoulli prior to each $\gamma_{mk} \sim \mathcal{B}(p_{mk})$ and adopt the probit link for p_{mk}, that is $\Phi^{-1}(p_{mk}) = \epsilon_{mk}$ and $\boldsymbol{\epsilon}_k$ is a CAR-structured random intercept, *i.e.* $\boldsymbol{\epsilon}_k|\boldsymbol{\Sigma}_k \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}_k)$. For a non-Gaussian likelihood, Lum (2012) used data augmentation to exploit conjugacy for this formulation. Secondly (Model3), we define independent inclusion indicators $\gamma_{mk} \sim \mathcal{B}(p_k)$ and impose a CAR dependence on ws by using a Gaussian copula. Specifically, the marginal (over γ_{mk}) prior distribution f_w of w_{mk} is $w_{mk}|p_{mk}, \beta_k, \tau_k^2 \sim p_k\mathcal{N}(\beta_k, \tau_k^2) + (1-p_k)\mathcal{N}(0, u_0\tau_k^2)$. A spatial structure is imposed in the ws by introducing latent variables θ_{mk} such that each $\boldsymbol{\theta}_k$ follows the structure of interest; in our case a CAR. w_{mk} are retrieved by back transforming the θ_{mk} using the cumulative distribution function of the standard normal distribution and the inverse of the cumulative distribution function of f_w. A geostatistical marginal (over γ_{mk}) approach based on a Gaussian copula for variable selection has been proposed by Boehm Vock et al. (2015).

A summary of the formulations of the described methodologies is provided in Table 5.1. We implement the above models in JAGS (Plummer, 2003). The samplers run for 1 million iterations with a single chain and the estimates were obtained from the last 30,000 samples. The predictors with a posterior mean inclusion probability $E(\gamma_p)$ greater than 0.5 were selected for fitting the final model. To compare the models retrieved by the variable selection we use the deviance information criterion (Spiegelhalter et al., 2002; Plummer, 2008).

Table 5.1: Summary of the formulations used in the Bayesian variable selection.

	Fixed slope	Random slope	Inclusion indicators	Interpretation
Model1	$\beta_k \vert \gamma_{1k}, \tau_k^2 \sim \mathcal{N}\left(0, \gamma_{1k}\tau_k^2 + (1-\gamma_{1k})\, u_0 \tau_k^2\right)$	$\mathbf{w}_k \vert \gamma_{2k}, \sigma_k^2, \rho_k \sim \mathcal{N}_d\left(\boldsymbol{\beta}_k, \sigma_k^{-2}\left(\mathbf{R} - \rho_k \boldsymbol{\Omega}\right)\right)$, with $\sigma_k^2 = \gamma_{2k}\tilde{\sigma}_k^2 + (1-\gamma_{2k})\, u_0 \tilde{\sigma}_k^2$	multinomial prior on $(\gamma_{1k}, \gamma_{2k})$ with events $\{(\gamma_{1k}=0, \gamma_{2k}=0), (\gamma_{1k}=1, \gamma_{2k}=0), (\gamma_{1k}=1, \gamma_{2k}=1)\}$ assigned 0.5, 0.25, and 0.25 probabilities, respectively	Global inclusion or exclusion of the random slope.
Model2	$\beta_k \sim \mathcal{N}(0, 100)$	$w_{mk} \vert \gamma_{mk}, \beta_k, \sigma_k^2 \sim \mathcal{N}(\gamma_{mk}\beta_k, \gamma_{mk}\sigma_k^2 + (1-\gamma_{mk}) u_0 \sigma_k^2)$	$\gamma_{mk} \vert p_{mk} \sim \mathcal{B}(p_{mk})$ with $\Phi^{-1}(p_{mk}) = \epsilon_{mk}$ and $\boldsymbol{\epsilon}_k \vert \boldsymbol{\Sigma}_k \sim \mathcal{N}_d\left(\mathbf{0}, \boldsymbol{\Sigma}_k\right)$	Spatially correlated inclusion probabilities. Smoothness in the effects enters implicitly in the mean of ws.
Model3	$\beta_k \sim \mathcal{N}(0, 100)$	$w_{mk} \vert \gamma_{mk}, \beta_k, \sigma_k^2 \sim \mathcal{N}(\gamma_{mk}\beta_k, \gamma_{mk}\sigma_k^2 + (1-\gamma_{mk}) u_0 \sigma_k^2)$ and $\theta_{mk} = \Phi^{-1}\left(F\left(w_{mk}\right)\right)$ with $\boldsymbol{\theta}_k \vert \rho_k \sim \mathcal{N}_d\left(\mathbf{0}, \left(\mathbf{R} - \rho_k \boldsymbol{\Omega}\right)\right)$	$\gamma_{mk} \vert p_k \sim \mathcal{B}(p_k)$ and $p_k \sim \mathcal{U}(0,1)$	Independent local inclusion of the random slope, with the spatial structure imposed through θs.

5.3 Results

Table 5.2 contains the posterior mean inclusion probabilities, *i.e.* $E(\gamma_p|\mathbf{y})$ for the predictors with fixed effects based on the three models. A mean inclusion probability above 0.5 was estimated for ndvi by all approaches. Model2 has additionally identified the lstd and distance to nearest water body.

Table 5.2: Fixed effects posterior mean inclusion probabilities

Fixed effects	Model1	Model2	Model3
urban	0.39	0.36	0.38
lstd	0.40	0.69	0.37
ndvi	0.74	0.54	1
rainfall	0.35	0.24	0.16
altitude	0.35	0.32	0.17
Distwater	0.34	1	0.13
asset	0.20	0.14	0.13
perc no educ	0.27	0.17	0.15

Random slope posterior mean inclusion probabilities are provided in Table 5.3. From the six varying effects of the net indicators, Model1 suggests that none should be included either with a fixed or random effect. For some of the provinces and net indicators, the posterior mean inclusion probabilities of Model2 are higher than 0.5 despite of an effect close to 0. The p_k of Model3 suggests that all net indicators, apart from OWN3, are excluded. For OWN3, a negative effect (median=-1.91 and CI:-3.66, -0.81) is estimated for Huambo province indicating that the decrease of prevalence in the province is attributed to the increase of intervention coverage during 2006 and 2011.

Table 5.3: Posterior inclusion probabilities for the 6 ITN indicators, stratified by variable selection.

Indicator	USE1($k=1$)			USE2($k=2$)			OWN1($k=3$)		
	Model1	Model2	Model3	Model1	Model2	Model3	Model1	Model2	Model3
$\gamma_k^{(1)}$	0.19		0.02 (0, 0.13)	0.19		0.01 (0, 0.17)	0.21		0.02 (0, 0.19)
$\gamma_k^{(2)}$	0		$E(\gamma_{jk})$	0		$E(\gamma_{jk})$	0		$E(\gamma_{jk})$
Bengo (γ_{1k})		0.4	0		0.49	0.01		0.45	0.01
Benguela (γ_{2k})		0.35	0		0.39	0.01		0.46	0.02
Bié (γ_{3k})		0.42	0		0.51	0.01		0.46	0.02
Cuando Cubango (γ_{4k})		0.39	0		0.5	0		0.52	0.03
Cuanza Norte (γ_{5k})		0.41	0		0.5	0.02		0.46	0.01
Cuanza Sul (γ_{6k})		0.39	0		0.47	0.01		0.48	0.02
Cunene (γ_{7k})		0.43	0		0.5	0.01		0.48	0.02
Huambo (γ_{8k})		0.43	0.14		0.48	0.07		0.61	0.09
Huíla (γ_{9k})		0.38	0		0.52	0.01		0.31	0.01
Luanda (γ_{10k})		0.39	0		0.47	0.01		0.45	0.01
Lunda Norte (γ_{11k})		0.36	0.01		0.46	0.01		0.39	0.01
Lunda Sul (γ_{12k})		0.41	0		0.55	0.01		0.49	0.01
Malanje (γ_{13k})		0.38	0		0.55	0.01		0.42	0.01
Moxico (γ_{14k})		0.37	0.01		0.5	0.02		0.49	0.01
Namibe (γ_{15k})		0.4	0.01		0.51	0.01		0.45	0.02
Uíge (γ_{16k})		0.35	0		0.49	0		0.47	0.01
Zaire (γ_{17k})		0.45	0		0.5	0.01		0.49	0.01

Variable	OWN2($k=4$)			OWN3($k=5$)			CASE1($k=6$)		
	Model1	Model2	Model3	Model1	Model2	Model3	Model1	Model2	Model3
$\gamma_k^{(1)}$	0.17		0.02 (0, 0.24)	0.34		0.06 (0, 0.21)	0.33		0.01 (0, 0.18)
$\gamma_k^{(2)}$	0		$E(\gamma_{jk})$	0		$E(\gamma_{jk})$	0		$E(\gamma_{jk})$
Bengo (γ_{1k})		0.43	0.02		0.55	0.15		0.48	0
Benguela (γ_{2k})		0.36	0.01		0.52	0		0.47	0.01
Bié (γ_{3k})		0.42	0.02		0.5	0.01		0.52	0.02
Cuando Cubango (γ_{4k})		0.51	0.01		0.49	0		0.43	0.01
Cuanza Norte (γ_{5k})		0.52	0.04		0.51	0.08		0.68	0.01
Cuanza Sul (γ_{6k})		0.51	0.02		0.56	0		0.46	0.01
Cunene (γ_{7k})		0.46	0.03		0.5	0.17		0.65	0.01
Huambo (γ_{8k})		0.53	0.04		0.67	0.79		0.48	0.01
Huíla (γ_{9k})		0.3	0.01		0.42	0		0.43	0.01
Luanda (γ_{10k})		0.5	0.03		0.37	0		0.44	0.01
Lunda Norte (γ_{11k})		0.45	0.03		0.43	0		0.52	0.01
Lunda Sul (γ_{12k})		0.49	0.01		0.5	0		0.42	0.01
Malanje (γ_{13k})		0.47	0.01		0.37	0		0.49	0.01
Moxico (γ_{14k})		0.46	0.02		0.39	0		0.5	0.01
Namibe (γ_{15k})		0.43	0.03		0.44	0.01		0.37	0.01
Uíge (γ_{16k})		0.42	0.01		0.47	0		0.51	0.02
Zaire (γ_{17k})		0.5	0.02		0.57	0.01		0.55	0.02

In all models and covariates, ρ_k is estimated close to 0. The geostatistical intercept's spatial range is estimated to 131km (CI: 62, 394), 53km (CI: 41, 78) and 153km (CI: 103, 256) by Model1, Model2 and Model3, respectively. The corresponding variance estimates are 3.8 (CI: 2.09, 6.21), 2.36 (CI: 1.84, 3.24) and 2.34 (CI: 1.69, 3.42). The spatial parameters are summarized in Table 5.4.

Table 5.4: Estimates of the spatial parameters.

	σ^2		
	Model1	Model2	Model3
w_0	3.8 (2.09, 6.21)	2.36 (1.84, 3.24)	2.34 (1.69, 3.42)
USE1	0.59 (0.17, 3.99)	0.54 (0.17, 2.93)	0.59 (0.18, 4.18)
USE2	0.57 (0.17, 3.74)	0.54 (0.18, 2.75)	0.6 (0.18, 4.24)
OWN1	0.63 (0.19, 3.2)	0.56 (0.17, 3.38)	0.6 (0.18, 3.91)
OWN2	0.58 (0.17, 8.51)	0.55 (0.18, 3.09)	0.57 (0.17, 3.98)
OWN3	0.74 (0.18, 10.31)	0.54 (0.18, 2.94)	0.57 (0.17, 3.67)
CASE1	0.68 (0.19, 3.73)	0.5 (0.17, 2.37)	0.56 (0.17, 3.7)
	ρ		
	Model1	Model2	Model3
USE1	-0.07 (-0.96, 0.91)	0.21 (-0.93, 1)	-0.02 (-0.91, 0.83)
USE2	-0.01 (-0.95, 0.94)	-0.04 (-0.95, 0.93)	-0.02 (-0.9, 0.83)
OWN1	-0.01 (-0.95, 0.94)	0 (-0.95, 0.95)	-0.02 (-0.9, 0.83)
OWN2	0 (-0.95, 0.95)	0.06 (-0.95, 0.98)	-0.01 (-0.91, 0.83)
OWN3	0 (-0.95, 0.94)	0 (-0.95, 0.95)	0.06 (-0.89, 0.86)
CASE1	0.03 (-0.95, 0.96)	0.04 (-0.95, 0.98)	-0.01 (-0.91, 0.83)

The model identified by the formulation of Model3 had the best fit according to the DIC that was estimated to be 320. The median effect of ndvi was 0.36 (CI: 0.11, 0.58), suggesting a positive association with malaria risk. Models from the formulations of Model1 and Model2 had a DIC equal to 429 and 451, respectively.

5.4 Discussion

To our knowledge, this is the first effort to assess the sensitivity of variable selection methods in identifying spatially varying effects. We assessed the effects of changes in malaria intervention coverage on the difference of parasitemia risk between 2011 and 2006. Model3 had the best fit and pointed one province for which an important negative effect of OWN3 (mean nets-to-people ratio) was estimated. As discussed in Giardina et al. (2014), only a fixed effect for an indicator might not be able to find an association while a varying effect can depict places in which a control

intervention had an important impact. The parametrization of the CAR with the additional parameter ρ_k showed that an exchangeable structure might be more appropriate than a spatial one.

In our application, Model2 estimated high inclusion probabilities despite of an unimportant effect (see Appendix). The p_k of Model3 has to be carefully interpreted. In case $p_k \simeq 0$, a global exclusion can be inferred but a slight deviation from 0 might show that some provinces have an important effect as demonstrated by the posterior means of γ_{mk}. Model1 has a stricter inclusion criterion; that is, a varying effect has to be included for all provinces and, presumably for this reason, it did not identify any varying effect.

The shrinkage factor u_0 influences inference. A value of 0 would lead to a Dirac prior for which a marginalization would be required. We chose a value of 0.001 based on what we would consider to be an important effect. An inverse gamma prior of the variances results to t distributed slab marginal distributions. Wagner and Duller (2012) used exponential and and degenerate priors for these variances that lead to Laplace and Gaussian distributed slabs, respectively. A spatial analogue of their approach could be envisaged and could show a sensitivity on this prior. Effects of indicators were assumed to vary by province. District-level CAR effects could not be modeled due to the fact that approximately half of the districts are lacking intervention data. Going to a finer level, a point-level geostatistical effect is worth exploring for which perhaps the Reich et al. (2010) could perform well due to its lighter parametrization. The use of point-level effects may be supported by studies of ITNs' effectiveness, suggesting that there is a community-wide benefit from ITNs usage (Killeen et al., 2007). We incorporated the 2006 survey by using the median predicted *logit* prevalence as an offset. The median is a point estimate and therefore the whole predictive distribution is not taken into account. In addition, modelling the difference of the intervention indicators' did not allow an investigation of the difference in the effects of the two time points that could be possible with *e.g.* a spatiotemporally varying extension of the methods used.

Identifying malaria control indicators that contribute to malaria risk is important for decision makers. It is of great significance to note not only that an indicator might not have a global fixed effect but also that it might not be spatially structured.

Avoiding the use of a variable selection that accommodates locally varying effects may have an implication in the risk factor analysis and on the importance of the estimated effects.

As funds to combat tropical diseases increase, see for example the Open Malaria Funding Data Platform (http://www.rollbackmalaria.org/financing/mfdp), interventions for other diseases such as soil-transmitted helminthiasis and schistosomiasis will be widely administered. Therefore, data availability will increase for more diseases and the proposed model formulations can become part of monitoring and evaluation to point out areas of need of an additional action. Moreover, due to the fact that there are common control interventions between tropical diseases (such as ITNs for mosquito-related or sanitation improvement for helminths) a joint, unified evaluation could indicate cross-effective interventions.

Finally, we compared three stochastic search variable selection methods for identifying net-related interventions in Angola with potentially spatially varying effects. We showed that identifying predictors is sensitive to the variable selection formulation. In our application, the model with locally varying inclusion indicators and spatially structured effects imposed through a copula performed best. Including such variable selection schemes in routine intervention adminstration could identify where areas with successful interventions.

5.5 Appendix

Table 5.5: Estimates (median and 95% CI) of the varying effects w_{kj} obtained from Model2.

Province	USE1 ($k = 1$)	USE2 ($k = 2$)	OWN1 ($k = 3$)	OWN2 ($k = 4$)	OWN3 ($k = 5$)	CASE1 ($k = 6$)
Bengo	0 (-2.67, 2.39)	0 (-2.71, 2.29)	0 (-2.59, 2.41)	0.01 (-1.29, 2.95)	-0.01 (-3.68, 1.95)	-0.02 (-2.37, 0.7)
Benguela	0 (-2.5, 2.17)	0 (-2.29, 1.47)	0 (-2.76, 3.11)	0 (-2.31, 1.92)	-0.01 (-3.49, 1.91)	-0.02 (-2.58, 0.44)
Bié	0 (-2.92, 2.14)	-0.01 (-2.73, 2.15)	0 (-2.81, 2.95)	0 (-2.68, 2.72)	-0.01 (-3.51, 1.9)	-0.03 (-2.3, 0.52)
Cuando Cubango	0 (-2.84, 2.1)	0 (-2.61, 2.92)	-0.01 (-3.14, 2.88)	0 (-3.11, 3.13)	-0.01 (-2.82, 1.85)	-0.02 (-2.11, 0.44)
Cuanza Norte	0 (-2.31, 2.66)	0 (-2.38, 2.73)	0.01 (-1.88, 3.3)	0.01 (-2.76, 3.66)	-0.01 (-3.64, 1.95)	-0.81 (-2.72, 0.12)
Cuanza Sul	0 (-1.52, 2.51)	-0.01 (-2.78, 1.45)	-0.01 (-2.64, 2.23)	0 (-2.96, 3.12)	-0.02 (-4.04, 1.84)	-0.02 (-2.48, 0.58)
Cunene	0 (-3.27, 2.1)	-0.02 (-3.21, 1.18)	0.01 (-2.23, 2.94)	0.01 (-1.86, 3.24)	-0.01 (-3.71, 1.61)	-0.5 (-3.35, 0.69)
Huambo	-0.01 (-3.53, 1.43)	-0.01 (-2.9, 1.93)	-0.04 (-4.1, 1.72)	-0.02 (-7.98, 1.11)	-0.89 (-4.78, 0.7)	-0.02 (-2.87, 1.13)
Huíla	0 (-2.04, 2.42)	0 (-2.29, 3.21)	0 (-2.05, 1.35)	0 (-0.8, 2.62)	0 (-2.39, 2.09)	-0.01 (-2.51, 1.01)
Luanda	0.01 (-2.09, 2.55)	0.01 (-1.77, 2.89)	0 (-2.02, 3.25)	0.01 (-2.39, 2.96)	0 (-2.26, 1.61)	-0.01 (-2.52, 1.05)
Lunda Norte	0 (-2.2, 2.01)	0 (-2.49, 2.37)	0 (-1.84, 2.42)	0 (-2.47, 2.57)	-0.01 (-2.84, 1.35)	-0.02 (-2.75, 0.93)
Lunda Sul	0 (-2.07, 2.72)	0 (-2.61, 3.57)	0 (-2.59, 2.69)	0.02 (-1.14, 3.81)	-0.01 (-2.6, 1.6)	-0.01 (-2.19, 1)
Malanje	0 (-2.82, 2.34)	0 (-2.7, 2.72)	0 (-2.11, 2.44)	0.01 (-1.44, 2.85)	0 (-1.84, 2.02)	-0.02 (-2.94, 1.01)
Moxico	0 (-2.62, 2.09)	-0.01 (-2.97, 2.75)	-0.01 (-2.61, 3.06)	0 (-2.76, 2.61)	-0.01 (-2.61, 1.04)	-0.02 (-2.89, 0.92)
Namibe	0 (-2.55, 2.42)	0 (-2.47, 3.08)	0 (-2.74, 2.25)	0 (-2.55, 2.44)	-0.01 (-2.76, 1.69)	-0.01 (-2.07, 0.69)
Uíge	0 (-1.44, 2.05)	0 (-2.2, 2.08)	0.01 (-1.72, 3.61)	0.01 (-2.34, 2.93)	0 (-3.32, 2.21)	-0.02 (-2.95, 1)
Zaire	-0.01 (-2.52, 2.02)	-0.01 (-2.4, 1.48)	0 (-2.76, 2.55)	0 (-2.73, 3.31)	-0.03 (-3.63, 1.52)	-0.03 (-3.13, 0.77)

Chapter 6

Marginal likelihood-based Bayesian variable selection of spatiotemporally varying coefficients with predictive processes

Karagiannis-Voules D.A.[1,2], Vounatsou P.[1,2]

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

This manuscript has been presented as a poster in the autumn meeting on latent Gaussian models (17-18 September 2015, Trondheim, Norway) and received the award of "Special Mention".

Abstract

National malaria control programs aim to reduce malaria burden by implementing effective, spatially targeted interventions. Data on the disease risk and on the proportion of the population that is covered by the various interventions are collected systematically by national malaria surveys. We employ Bayesian variable selection within geostatistical models with spatiotemporally varying coefficients to analyse the spatiotemporal effects of malaria interventions using data from two national surveys in Angola. We fit the models derived from all possible combinations of the predictors that include intervention coverage measures and climatic factors. Indicators are introduced to define inclusion/exclusion of each variable and each model's prior probability. For potentially spatiotemporally varying effects, we adopt a multinomial prior allowing either exclusion or inclusion of a non-varying or inclusion of a spatiotemporally varying effect. We develop an iteratively integrated nested Laplace approximation (i-INLA) to the marginal likelihood of each model. We use predictive process approximations to address intensive computations that arise in modelling large geostatistical data by estimating the spatial Gaussian processes involved in a model from a set of locations (knots) with lower size than that of the observed data. We assess the sensitivity of variable selection to the knots' size by comparing with a model fitted on the full set of locations. Knot selection led to different models, however all models identified the same predictor with spatiotemporally varying effects. Our algorithm offers an approximation to the marginal likelihood and can be combined with stochastic search over the model space as well as Bayesian model averaging.

6.1 Introduction

Efforts of international organizations to control malaria have lead to large-scale implementation of interventions. Repeated surveys are conducted for monitoring and evaluation purposes. Interventions are expected to have an effect in disease risk and should be taken into account into geostatistical predictive risk modelling. Their effects are also likely to vary within a country and over time. There are several ways to measure interventions (MEASURE Evaluation et al., 2013) and their effects on malaria burden depend on the intervention measure (Giardina et al., 2014).

Giardina et al. (2014) used geostatistical models with spatially varying regression coefficients to estimate effects of malaria interventions at sub-national levels. They assumed a spatial structure in the effects modelled by conditional autoregressive (CAR) processes. However, it is well-known that mosquito nets have not only an individual protective effect but also a community-wide benefit (see Howard et al., 2000; Hawley et al., 2003; Killeen et al., 2007, among others). Therefore, the assumption of a constant intervention effect within an area, that a CAR structure implies, may not be justifiable. The inclusion of geostatistical random slopes in modelling malaria risk could address this issue. Furthermore, incorporating Bayesian variable selection methods within geostatistical models (Chapter 5) would allow identification of possibly varying effects of important intervention measures. Extending the above models with spatially varying effects to spatiotemporal analogues can help explain heterogeneity in space and time.

Despite the vast development of Bayesian variable selection approaches (for a review, see O'Hara and Sillanpää, 2009), geostatistical disease mapping is rather lacking rigorous inference pertaining to variable selection. It is common in many applications to select the variables that enter in the geostatistical model by omitting the spatial intercept (see, for example, Clements et al., 2006, 2009; Soares Magalhães et al., 2011; Schur et al., 2011a; Raso et al., 2012, among others). Chammartin et al. (2013a) applied Bayesian geostatistical variable selection and showed that ignoring the geostatistical term might result in selecting a different set of predictors.

Wagner and Duller (2012) conducted variable selection of an unstructured random

intercept. For spatially varying coefficients, Reich et al. (2010) performed fixed and random slope selection for a multivariate Gaussian response. Boehm Vock et al. (2015) used local variable selection through a Gaussian copula. Reich's approach was based on a spike and slab prior with a discrete spike. In such formulations, integration of a predictor's fixed and random slope is needed to calculate a conditional (on all other parameters), marginal (over the fixed and random slope) likelihood.

In the case of a Gaussian likelihood, this integral has a closed form (Reich et al., 2010) as does the integral over all spatial intercepts and slopes which is the conditional, on hyper-parameters (such as variances, ranges or likelihood parameters), marginal likelihood. For non-Gaussian likelihoods, that are common in modelling survey or count disease data, there is no closed form of this integral. For generalized linear models with hyper-g priors, Sabanés Bové and Held (2011) used a Laplace approximation (Tierney and Kadane, 1986) to this conditional (in this case, on g) marginal likelihood. The authors integrated over g numerically to approximate the marginal likelihood and to perform variable selection. The resulting algorithm is an integrated Laplace approximation and has been used also in generalized additive linear models (Sabanés Bové et al., 2011).

Models with spatiotemporal effects include large number of hyper-parameters, rendering a numerical integration computationally expensive. We use an iterated Laplace approximation for the integration over the hyper-parameters. This method is based on a Gaussian mixture approximation of the hyper-parameters' marginal posterior distribution. The marginal likelihood is, thus, calculated via an iteratively integrated nested Laplace approximation (i-INLA). A single mixture component would lead to the special case of an integrated nested Laplace approximation (INLA, Rue et al., 2009).

Our approach, requires estimating the mode(s) of the hyper-parameters' marginal posterior and calculating the Hessian. Matrix calculations involved in geostatistical modelling slow down computation. We overcome large matrix computations using predictive processes for all spatiotemporally varying coefficients. The predictive process approximation (Banerjee et al., 2008) is based on a selection of locations, knots, on which the random effects are placed. The effects on the full dataset are

then estimated using properties of the multivariate normal distribution.

We implement the proposed methodology to estimated effects of insecticede treated net (ITN) indicators in Angola and investigate the influence of predictive processes in Bayesian variable selection for spatiotemporally varying coefficients.

6.2 Data

Data concerning malaria infection, age, urban-rural classification and ITN proxies, for Angola in 2006 and 2011 were obtained from the Demographic and Health Surveys (DHS) Programm (http://www.dhsprogram.com). The complete dataset contained 5160 children, aged below 5 years, living in 342 clusters, 115 of these surveyed in 2006 and 227 in 2011. Locations in the two surveys are unique, resulting to space-time misalignment. Forty one percent of the clusters were classified as urban. Age was categorized by year of life. From the first to the fifth year, there were 14.7%, 23.5%, 21.7%, 21.1%, 19% of children surveyed in both time points. Based on explanatory analysis, age category and urban class were included in all models to reduce model space.

We followed RBMP's (MEASURE Evaluation et al., 2013) guidelines and computed ITN-related intervention measures. Complete data for both periods and all locations existed for the following five indicators: (i) percentage of people who slept under an ITN the night before the survey; (ii) percentage of children, aged below 5 years, who slept under an ITN the night before the survey; (iii) percentage of households in a cluster with at least one ITN, (iv) percentage of households in a cluster with at least one ITN for every two people; and (v) mean nets to people ratio.

The observed data on malaria prevalence as well as the 5 ITN measures of 2006 and 2011 are depicted in Figure 6.1. The total malaria prevalence in 2006 was 23.3% and dropped to 9.1% in 2011. The mean prevalence of both years was 13%. In urban areas the prevalence was lower (3.5%) compared to rural settlements (19.5%).

To adjust for environmental predictors we used proxies of temperature, rainfall, altitude, distance to water and vegetation averaged over the year previous to each survey. Specifically, land surface temperature at night (lstn), as well as the

normalized difference vegetation index (ndvi) were obtained from MODIS (Oak Ridge National Laboratory Distributed Active Archive Center, 2011). Using the same source and the land cover classification of water, we defined the distance to the closest water body. Altitude was obtained from http://srtm.csi.cgiar.org using the R package raster (Hijmans, 2014). Rainfall was downloaded from the Famine Early Warning System Network of the United States Agency for International Development http://earlywarning.usgs.gov/fews/index.php.

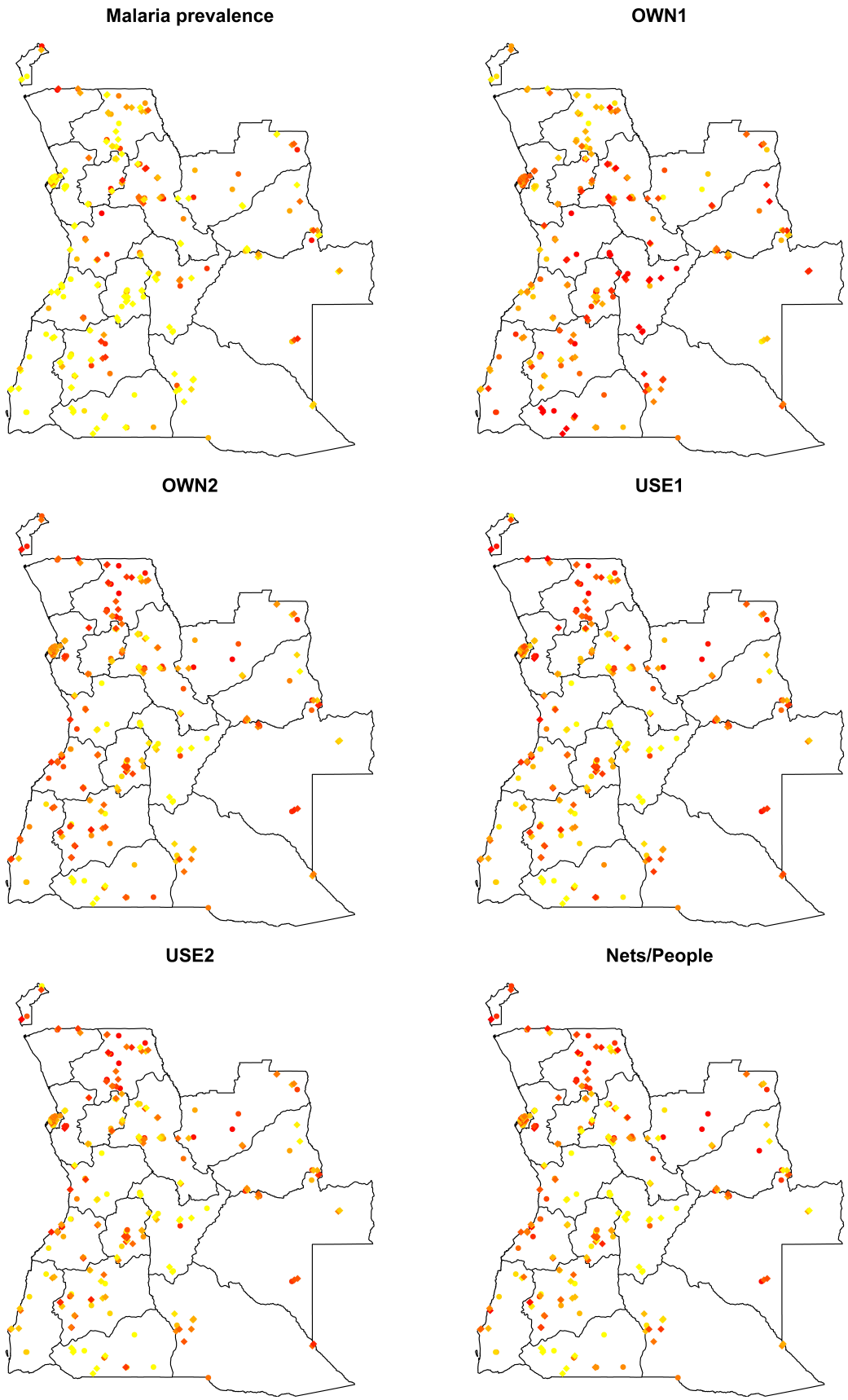


Figure 6.1: Observed data of parasitemia and ITN coverage measures in Angola obtained from the 2006 and 2011 surveys.

6.3 Model specification

6.3.1 Spatiotemporally varying coefficients with predictive processes

Let Y_{it} be the binomially distributed response denoting the number of infected children observed at location s_i $(i = 1, ..., n)$ and time t $(t = 1, ..., T)$, $\mathbf{X}$ and $\mathbf{D}$ be the $n \times p$ and $n \times K$ design matrices of predictors with fixed and random slopes, respectively. The linear predictor with spatiotemporally varying coefficients is formulated as:

$$\eta_{it} = g\left(E\left(Y_{it}\right)\right) = \beta_0 + \mathbf{X}_{it}^T\boldsymbol{\beta} + w_{it0} + \mathbf{D}_{it}^T\tilde{\boldsymbol{\beta}} + \sum_{k=1}^{K} D_{itk}w_{itk} \tag{6.1}$$

We consider the random intercept $\mathbf{w}_0$ to be a realization of a spatiotemporal Gaussian process, that is $\mathbf{w}_0|\boldsymbol{\Sigma}_0 \sim \mathcal{N}_{Tn}\left(0, \boldsymbol{\Sigma}_0\right)$. The spatiotemporal dependence is taken into account through the covariance matrix $\boldsymbol{\Sigma}_0 = \mathbf{Q}_0 \otimes \mathbf{C}_0$. We assume that $\mathbf{C}_0$ is a $n \times n$ spatially structured matrix with $\langle \mathbf{C}_0 \rangle_{ij} = \sigma_0^2 \exp\left(-\phi_0 d_{ij}\right)$ and that $\mathbf{Q}_0$ is a temporal $T \times T$ correlation matrix with $\langle Q_0 \rangle_{tt'} = \rho_0^{|t-t'|}$. The spatial variance and decay parameters are σ_0^2 and ϕ_0, ρ_0 is an autocorrelation parameter and d_{ij} is the Euclidean distance between locations s_i and s_j. Similar specification is followed for the random slopes $\mathbf{w}_k$.

We assign to fixed effects $\beta_0, \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}$ vague normal priors. The Bayesian hierarchical formulation is completed by defining priors for the hyper-parameters $\sigma_0^2, \phi_0, \rho_0, \sigma_k^2, \phi_k, \rho_k$. Commonly, inverse gamma priors are assigned to variances and uniform priors to decay and autocorrelation parameters. Here, we choose to re-parameterize the hyper-parameters in order to define variables in the real line with normal priors. We will use $\boldsymbol{\theta}$ for the vector of the transformed hyper-parameters that could also include any likelihood parameters.

The linear predictor can be formulated in a vector format as $\boldsymbol{\eta} = \tilde{\mathbf{X}}\boldsymbol{\alpha}$ with $\tilde{\mathbf{X}}$ being a grand design matrix and $\boldsymbol{\alpha} = \left(\beta_0, \boldsymbol{\beta}^T, \mathbf{w}_0^T, \tilde{\boldsymbol{\beta}}^T, \mathbf{w}_k^T\right)^T$ the vector of coefficients with a multivariate normal zero-centered prior distribution and a covariance matrix that is block-diagonal with blocks corresponding to the covariance matrices of each

of its elements.

Due to the fact that $\boldsymbol{\Sigma}_0$ and each $\boldsymbol{\Sigma}_k$ would require the inversion of $n \times n$ matrix, we use predictive processes as proposed by Banerjee et al. (2008). The methodology is based on a selection of m knots $\mathbf{s}^*$, that may or may not be part of the n locations, on which the geostatistical random effect is placed. Then, an unbiased estimate of the random effect can be calculated through $\hat{\mathbf{w}}_0$ and $\hat{\mathbf{w}}_k$ where, for example, $\hat{\mathbf{w}}_0 = \left\{\mathbf{Q}_0 \otimes \left(\mathbf{C}_0(\mathbf{s}, \mathbf{s}^*)\mathbf{C}_0^{-1}(\mathbf{s}^*, \mathbf{s}^*)\right)\right\} \mathbf{w}_0^*$. This approximation reduces the dimension of $\boldsymbol{\alpha}$ which is now defined as $\boldsymbol{\alpha} = \left(\beta_0, \boldsymbol{\beta}^T, \mathbf{w}_0^{*T}, \tilde{\boldsymbol{\beta}}^T, \mathbf{w}_k^{*T}\right)^T$. An advantage of the predictive process is that it does not require ad-hoc treatment for spatial and temporal misalignment that are present in our study. The above specification of the predictive process is valid in our case due to the fact that we choose same $\mathbf{s}^*$ for both time points. More generally, one could use space-time knots whose locations differ in space and time.

6.3.2 Variable selection

We perform variable selection to identify the predictors in $\mathbf{X}$ and $\mathbf{D}$ that contribute in explaining malaria risk. Therefore, we conduct model selection by assigning an inclusion prior to each predictor. This is achieved by introducing binary indicators, that essentially index a model's specification and indicate inclusion or exclusion of each predictor.

The random slope $\mathbf{w}_k$ depicts deviations from the fixed one $\tilde{\beta}_k$. To incorporate such interpretation in the model priors, we introduce binary indicators (γ_{1k} and γ_{2k}) for each potentially varying effect. A multinomial prior is assigned to the pair of indicators with possible events $\{(\gamma_{1k} = 0, \gamma_{2k} = 0), (\gamma_{1k} = 1, \gamma_{2k} = 0), (\gamma_{1k} = 1, \gamma_{2k} = 1)\}$ and prior probability of 0.5, 0.25 and 0.25, respectively (Reich et al., 2010). The first event denotes exclusion of the predictor, the second inclusion of a fixed slope and the third inclusion of random slope. For the predictors in $\mathbf{X}$ a binary indicator with probability of inclusion equal to 0.5 is introduced.

If $\boldsymbol{\gamma}$ is the vector of all indicators, then a model's prior probability $p(M)$ is calculated by multiplying the prior of each element of $\boldsymbol{\gamma}$. The model's (unnormalized) posterior probability is then given by $p(M|\mathbf{y}) = f(\mathbf{y}|M)\, p(M)$. A normalization over

the models under consideration can is straightforward. $f(\mathbf{y}|M)$ is the marginal likelihood of model M and is needed to identify the model with the highest posterior probability.

6.3.3 Marginal likelihood approximation

Under a binomial likelihood, a closed form of the marginal likelihood does not exist. To calculate it, first, we use a Gaussian approximation ($f_G(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}, M)$) of the full conditional posterior distribution of the coefficients in model M, $f(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}, M) = \frac{f(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\theta},M)f(\boldsymbol{\alpha}|\boldsymbol{\theta},M)}{f(\mathbf{y}|\boldsymbol{\theta},M)}$ by optimizing the numerator with respect to $\boldsymbol{\alpha}$ and calculating the Hessian at the mode. Since $f(\boldsymbol{\alpha}|\boldsymbol{\theta}, M)$ is *a-priori* Gaussian, the Hessian at the mode of $\boldsymbol{\alpha}$'s full conditional can be calculated fast (Rue et al., 2009).

Given $f_G(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}, M)$, a natural approach to calculate the marginal likelihood would be to use the Laplace approximation, $f_{La}(\mathbf{y}|\boldsymbol{\theta}, M)$, of $f(\mathbf{y}|\boldsymbol{\theta}, M)$ and combine it with another Gaussian approximation, $f_G(\boldsymbol{\theta}|\mathbf{y}, M)$, for the marginal posterior of the hyper-parameters $\tilde{f}(\boldsymbol{\theta}|\mathbf{y}, M) = \frac{f_{La}(\mathbf{y}|\boldsymbol{\theta},M)f(\boldsymbol{\theta}|M)}{f(\mathbf{y}|M)}$. This would be an integrated nested Laplace approximation (Rue et al., 2009) of the marginal likelihood.

Here, to improve the approximation of $f(\mathbf{y}|M)$, we use an iterated Laplace approximation (Bornkamp, 2011). Namely, the marginal posterior of $\boldsymbol{\theta}$ is approximated by a mixture of multivariate Gaussian distributions, with a number of components that is calculated iteratively. Initially, a single component approximation is considered and its difference from $\tilde{f}(\boldsymbol{\theta}|\mathbf{y}, M)$ is calculated on a randomly selected grid of $\boldsymbol{\theta}$. At the largest difference, a mode search and Gaussian fitting takes place. A second Gaussian component is added to $\boldsymbol{\theta}|\mathbf{y}, M$ at this mode and a new marginal likelihood is estimated. If the new marginal likelihood differs by more than 1% from the initial one, then the second Gaussian component is kept and another grid of $\boldsymbol{\theta}$ is selected to iterate the algorithm until convergence or until a prespecified number of Gaussian components is reached (for more details see Section 2 in Bornkamp, 2011). Therefore, the overall approach is an iteratively integrated nested Laplace approximation of the marginal likelihood.

6.3.4 Implementation

We consider a normal prior for the coefficients $\boldsymbol{\alpha}$, that is

$$\boldsymbol{\alpha}|\boldsymbol{\theta}, M \sim \mathcal{N}\left(\mathbf{0}, (100 \oplus 100\mathbf{I}_p \oplus \boldsymbol{\Sigma}_0 \oplus 100\mathbf{I}_K \oplus \boldsymbol{\Sigma}_k)\right)$$

We assume for $\log(\sigma^2)$, a log inverse gamma such that its exponent is inverse gamma distributed with shape parameter equal to 2 and rate equal to 1. For the spatial decay parameters, we consider $\log\left(\frac{\phi-0.5}{50-\phi}\right) \sim \mathcal{N}(0, 1.4)$ and for autocorrelation parameters $\log\left(\frac{\rho+1}{1-\rho}\right) \sim \mathcal{N}(0, 1.4)$. In this study, we fit the models with all possible combination of predictors but allow maximum 1 random slope due the high correlation of predictors in $\mathbf{D}$.

Inference was conducted in R (R Core Team, 2014). The Gaussian approximation of $\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}, M$, was implemented in the INLA package (available at www.r-inla.org). For the Gaussian mixture approximation of $\boldsymbol{\theta}|\mathbf{y}, M$, we used the iterLap (Bornkamp, 2011) package and modified it to avoid unnecessary calculations as well as to include optimization using quadratic approximations through the minqa package (Bates et al., 2014). We set the number of the $\boldsymbol{\theta}|\mathbf{y}, M$ Gaussian mixture components to be maximum 3 and the $\boldsymbol{\theta}$ grid to 250 points. The presented results for $\boldsymbol{\alpha}$ are based on an empirical Bayes integration over the hyper-parameters' mode(s) and for $\boldsymbol{\theta}$ on 1000 random samples from the mixture Gaussian transformed to the natural scale. To select the knots, we used the space-filling design (Johnson et al., 1990), implemented in the fields package (Nychka et al., 2015). We used 50 and 100 knots for each time point (*i.e.* 100 and 200 space-time knots but we will refer to them using the spatial dimension) as well as the full set of locations (342 unique spatial locations). We predict the mean of spatiotemporally varying coefficients for the two time points in a 5×5 km grid covering the country. The prediction of the mean is based on the formulation of the predictive process.

6.4 Application

We have applied the above model to analyze the malaria survey data. We fitted 3 models with 50, 100 knots and the full dataset. The two different knot sizes

and the full set of locations led to different models with maximum *a posteriori* (MAP) probability. The three MAP models included USE1 with a spatiotemporally varying effect. Figure 6.2 depicts the log unnormalized posterior model probabilities stratified by ITN indicator, knots size and random slope type (fixed or random). In general, larger posterior probabilities were observed with increasing knots size. For the models using the full locations, the higher probabilities are observed for USE1 and USE2. No such pattern was apparent for any knots' size. The normalized posterior probabilities of the three MAP models were 0.89, 0.6 and 0.52 for knots 50, 100 and the full set of locations, respectively. The model selected from the full set of locations was ranked 4th best in the variable selection based on 50 knots and 434th when 100 knots were considered and had normalized posterior probabilities of 0.0024 and $2 \cdot 10^{-7}$, respectively.

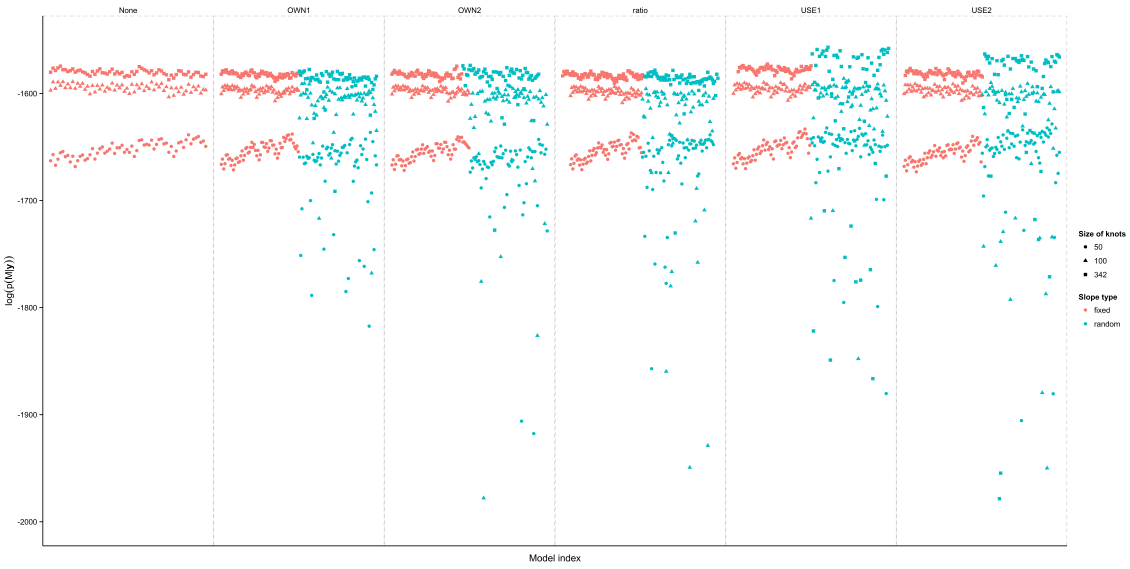


Figure 6.2: Un-normalized posterior probability in the log scale (*i.e.* $\log\left(p\left(M|\mathbf{y}\right)\right)$) stratified by ITN indicator, their effect type and size of knots for all possible models. For illustration purposes, few models with $\log\left(p\left(M|\mathbf{y}\right)\right) < -2000$ are not depicted.

Using the knots size of 50, the MAP model included ndvi, rainfall, altitude and distance to nearest water body. Rainfall and ndvi were positively associated with malaria risk, while the rest of the predictors were negatively associated. Compared to the above predictors, the MAP model of the 100 knots did not include nvdi

and altitude. The MAP model using all locations included rainfall, altitude and USE1. The effects of the three MAP models are provided in Table 6.1. Similar direction of the effects were observed for the common predictors in the three MAP models. Specifically, the effects of age suggested that older children are at higher risk. Urban classification was negatively associated with malaria risk.

Table 6.1: Posterior estimates (median and 95% credible interval) of the fixed effects in the three MAP models.

Variable	50 knots	100 knots	all locations
Age			
$<$ 12 months	0	0	0
12-23	0.14 (-0.20, 0.48)	0.16 (-0.18, 0.51)	0.15 (-0.20, -0.50)
24-35	0.56 (0.22, 0.90)	0.56 (0.21, 0.91)	0.55 (0.19, 0.91)
36-47	0.77 (0.43, 1.10)	0.84 (0.50, 1.18)	0.82 (0.48, 1.17)
48-59	0.93 (0.60, 1.27)	0.96 (0.61, 1.30)	0.95 (0.60, 1.30)
Urban settlement	-1.24 (-1.86, -0.62)	-1.29 (-1.92, -0.67)	-1.07 (-1.89, -0.26)
ndvi	0.39 (0.06, 0.71)		
rainfall	1.18 (0.67, 1.68)	1.05 (0.46, 1.64)	1.22 (0.58, 1.86)
altitude	-1.17 (-1.72, -0.61)		-1.03 (-1.72, -0.34)
distance to water	-0.90 (-1.26, -0.54)	-0.60 (-1.10, -0.09)	
use1	-0.27 (-0.52, -0.03)	-0.26 (-0.49, -0.02)	-0.40 (-0.75, -0.06)

Estimates of the hyper-parameters are provided in Table 6.2. The spatial range of the random intercept is estimate to 247, 293 and 248 km in the models with 50, 100 knots and the full set of locations, respectively. Its variance is smaller in the full set of locations. The autocorrelation parameters shows a small dependence in the two time points for both the intercept and the effect of USE1. The range of USE1 differs between the 3 models. The smallest range is estimated in the full model and it is approximately 7.5 km.

Table 6.2: Posterior estimates (median and 95% credible interval) of the hyper-parameters in the three MAP models.

Variable	50 knots	100 knots	all locations
σ_0^2	4.23 (2.79, 6.17)	4.03 (2.97, 5.31)	2.79 (2.10, 3.72)
$3/\phi_0$ (km)	247.58 (171.45, 341.05)	293.61 (196.66, 407.82)	248.13 (158.81, 353.63)
ρ_0	0.13 (0.06, 0.18)	0.39 (0.26, 0.51)	0.25 (0.16, 0.33)
σ_1^2	0.79 (0.70, 0.89)	0.73 (0.61, 0.84)	1.41 (1.33, 1.49)
$3/\phi_1$ (km)	126.40 (77.56, 196.20)	8.60 (7.44, 11.21)	7.46 (6.71, 18.65)
ρ_1	0.60 (0.11, 0.87)	0.14 (0.07, 0.20)	0.28 (0.13, 0.39)

Prediction of the effect of USE1 (fixed and random effect) on a grid of 5×5 km is provided in Figure 6.3. Small spatial range of the full model resulted in small scale heterogeneities. At distances larger than the range of each model, the mean fixed effect ($\tilde{\beta}_k$) is predicted.

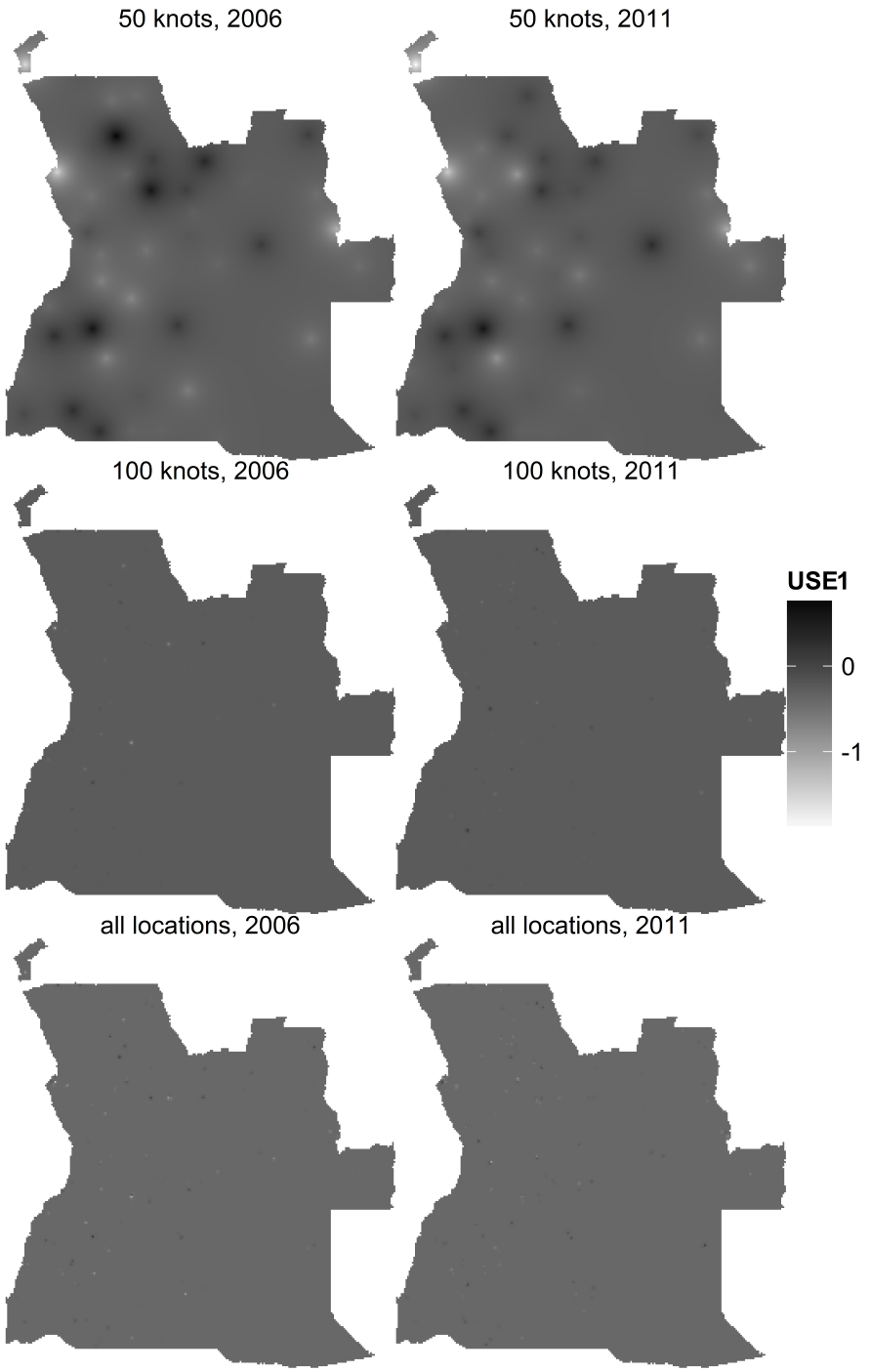


Figure 6.3: Mean predicted effect of USE1 for the two time periods using the three MAP models.

6.5 Discussion

We used a marginal likelihood calculation through an iteratively integrated nested Laplace approximation, to identify potentially spatiotemporally varying effects of net indicators in Angola using predictive processes.

The results suggested that predictive processes may lead to different model identification. In fact, different knot sizes lead to different models. Apart from the size of the knots, the knot selection design could influence inference. We used the space-filling design but, to our experience, it is rather unlikely that other designs would yield similar results as the full model. We also assumed the same knots for both time points. An option, that could address both issues, would be to assume that the locations of the knots in space and time are random *e.g.* a log-Gaussian (predictive) process (Guhaniyogi et al., 2011).

In our application, the total number of models was small enough to allow an exhaustive model space exploration. Alternatively, a stochastic model space search that is tuning-free has been recommended by Sabanés Bové et al. (2011) for generalized additive models using g-priors and could be also used in our setting. A relevant issue, is that we forced to include maximum 1 varying slope. This was done due to the fact that ITN indicators were highly correlated. To smooth away, but yet allow the existence of correlated predictors if data support so, a g-prior could be used for $\tilde{\boldsymbol{\beta}}_K$. This, would take into account the point-level correlation but not the between point correlation. An extension could be to define a multivariate analogue of a g-prior, and/or use not independent, cross-covarying, effects (Gneiting et al., 2012). For instance, including a cross-covarying effects would allow information of each of the indicators' effects to be shared between locations. Although this might seem computationally more demanding due to the increased number of hyper-parameters, a simple parameterization with e.g. assuming that the random slopes are independent realizations of the same Gaussian process (*i.e.* sharing the same hyper-parameters) among varying effects would still allow flexible modelling.

A logical extension of this work would be to assume that $\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}$ is also a mixture of multivariate normal distributions. Coupled with the adopted approach of the hyper-parameters, it would define an integrated nested iterated Laplace approximation.

However, as the dimension of $\boldsymbol{\alpha}$ is much larger than the one of $\boldsymbol{\theta}$, searching for a point in $\boldsymbol{\alpha}$'s domain to re-optimize would be extremely time consuming. Perhaps using the Gaussian approximation of $\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}$ but alternative methodologies of the marginal distributions of the hyper-parameters could reveal a limitation on the Gaussian mixture approximation. For instance, one could use variational Bayes or Hamiltonian Monte Carlo for exploring the $\boldsymbol{\theta}|\mathbf{y}$ and calculating the marginal likelihood.

Identifying intervention coverage indicators that contribute to malaria risk is important for monitoring and evaluation. Interventions might not be uniformly effective across an areal unit such as province or district. Furthermore, in the case of malaria, as discussed, for example, by Killeen et al. (2007), ITNs can have an effect to surrounding settlements and to the broader community. This is emphasized by the small spatial range estimated for the varying effect of USE1 from the full model. A CAR structure of the effects would not be able to capture the community-wide effect of the implemented interventions. As data accumulation and ITN distribution increases, the models could be extended to incorporate covariates for the varying effects that could explain the small-scale heterogeneities of the effects in space and time.

Finally, under an i-INLA approach for a marginal likelihood-based Bayesian variable selection, we showed that the predictive process approximation might lead to a different model than using the full dataset. In our case, all knot sizes retrieved the same varying ITN indicator but this might not always be true. In addition, for the cases where the full model is not possible to be estimated, the retrieved marginal likelihoods can be used for Bayesian model averaging in order to incorporate the model uncertainty in an alternative way. In future studies, we will investigate alternative integrations and "big n" approximations to address the aforementioned issues.

Chapter 7

Discussion

This PhD thesis contributes to the fields of Bayesian spatial modelling and spatiotemporal epidemiology of tropical diseases with: (i) methodology for Bayesian variable selection of spatiotemporally varying coefficients allowing flexible inference, especially for computationally intensive geostatistical models of data collected over large number of locations; (ii) sensitivity analysis of Bayesian variable selection formulations of models with spatially varying coefficients; (iii) estimates of incidence rates for cutaneous and visceral leishmaniasis in Brazil depicting the current situation of leishmaniasis in the country; (iv) an open-access georeferenced database cataloguing all available survey data for soil-transmitted helminth infections in sub-Saharan Africa and Cambodia for disease control and research purposes; (v) up-to-date smooth risk maps, and estimates of number of people infected and number of required treatments of soil-transmitted helminth infections in sub-Sahara Africa and Cambodia; (vi) an evaluation of the predictive ability of cluster-aggregated WASH and other SES-related proxies in disease mapping of poverty-related diseases; and (vii) geostatistical models of malaria risk for estimating effects of malaria intervention coverage measures across space and over time.

Chapters 2-6 correspond to manuscripts which include detailed conclusions. The purpose of this chapter is to summarize the principal findings, bring forward the main highlights, discuss limitations and propose extensions of the work.

7.1 Significance of the work

7.1.1 Statistical methods: variable selection of spatiotemporally varying coefficients

In Chapters 5 and 6, we develop models for Bayesian variable selection of spatiotemporally varying coefficients and apply them in the field of malaria epidemiology. We assess the sensitivity of inference to different variable selection formulations (Chapter 5) and to predictive process approximations for large data (Chapter 6).

Bayesian variable selection of models with random effects has recently received some interest. In the spatial statistics field, random effects model spatially varying coefficients. There are a number of different formulations to conduct stochastic

search variable selection of spatially varying coefficients using a spike and slab prior by introducing indicators for each variable that define probabilities of inclusion into the model. Reich et al. (2010) use a multinomial indicator to allow inclusion of a fixed or random slope assuming that the effect enters in the model for all or none of the locations. The authors consider a discrete spike and a Gaussian likelihood. To relax this assumption, a local indicator can be introduced. The spatial dependence can, then, be incorporated either in the effect or in the indicator. The first is proposed by Boehm Vock et al. (2015) and the latter by Lum (2012).

In Chapter 5, we assess sensitivity of inference to different formulations of stochastic search variable selection with spatially varying effects. We overcome computational problems that arise under non-Gaussian likelihoods, as well as data augmentation to achieve conjugacy (Lum, 2012), by proposing a conditional specification of the spatial random slopes on the local indicators. Our approach enables straightforward implementation in standard Bayesian software such as JAGS (Plummer, 2003).

A discrete spike for non-Gaussian likelihoods would require a numerical integration over a model's parameters. One could integrate over a single covariate's effects and conduct a stochastic search variable selection conditioning on the rest of the parameters (as in Reich et al. 2010 for Gaussian likelihood). We follow another approach by integrating over all parameters of a model and calculate its marginal likelihood. Then, by evaluating all possible models we perform model selection. In Chapter 6, we take advantage of the latent Gaussian model class, that our models belong to, and integrate over the latent Gaussian field using a Laplace approximation. This is computationally inexpensive and has been proposed by Rue et al. (2009). The hyper-parameters (any parameters of the Gaussian field or likelihood parameters) could also be integrated with a Laplace approximation leading to an INLA (Rue et al., 2009) of the marginal likelihood. We improve the latter approximation by using iterated Laplace approximations (Bornkamp, 2011) and name our algorithm an iteratively integrated nested Laplace approximation (i-INLA). To reduce the inferential computational cost, we use the predictive process approximation (Banerjee et al., 2008) for the random slopes and show the effect of approximation (*i.e.* dimension of knots) on model selection. Our proposed algorithm, can be implemented in existing R packages and can be combined with

Bayesian model averaging.

7.1.2 Epidemiological methods: planning and evaluation of interventions

We introduce innovative statistical methodologies in tropical disease epidemiology to address epidemiological questions in the control of NTD and malaria. In particular, in Chapter 2 we conduct the first, to our knowledge, geostatistical analysis using INLA and SPDE in the NTDs field. We implement variable selection using the INLA package by fitting the models with all possible combinations of STH predictors in Chapters 3 and 4. We apply novel approaches of Bayesian variable selection for spatiotemporally varying covariate effects in malaria Epidemiology in Chapters 5 and 6.

Throughout this thesis, we analyze disease data that are observed in a large number of locations (*i.e.* up to 6 thousand). The computational cost of parameter estimation of a single spatial model, as well as prediction at high resolution, depends on this number. Approximate Bayesian inference based on INLA and SPDE enables us to reduce model fit and prediction to few hours from weeks that an MCMC-based calculation would require. Taking advantage of the computational gain we are able to perform geostatistical variable selection based on cross-validatory criteria and select the model with the best predictive ability among all possible models than can be formulated by our predictors.

The work presented in Chapter 4 contributes to understanding the use of socioeconomic proxies in geostatistical modelling of poverty-related diseases. Soil-transmitted helminthiasis, among other tropical diseases, is associated with low socioeconomic status (see, for example, Ziegelbauer et al., 2012). Socio-economic data are available from household surveys at locations which are not aligned with the disease data. As introduced in Section 1.3, this misalignment can be addressed with joint models of location-specific SES proxies and disease risk. However, aggregated SES proxies at locations do not reflect individual exposures. In addition, individually measured SES and disease infection is rarely available for a large number of locations to allow spatial analyses. In Chapter 4, we evaluate the predictive ability of location-specific and individual SES on STH infection risk on a

dataset from Cambodia with individual data. We show that using joint modelling of misaligned SES and disease survey locations may not capture the effect of *e.g.* asset index or sanitation on STH risk. Geostatistical models based on the individual data were able to estimate the expected effect of these proxies. Furthermore, the location-aggregated SES measures led to inconclusive evidence of SES effect.

Interventions against malaria have been intensified in the recent years. Mass mosquito nets' administration, indoor insecticide residual spraying *etc.* are widely implemented. Such interventions are the main cause of the approximately 50% decrease of *P. falciparum* infection prevalence in Africa from 2000 to 2015 (Bhatt et al., 2015). The effectiveness of interventions, though, is less likely to be the the same across space (Giardina et al., 2014). Furthermore malaria bednet coverage is measured by different proxies (MEASURE Evaluation et al., 2013) related to bednet ownership or use. The work of (Giardina et al., 2014) showed that intervention effects differ according to the proxy used in the statistical analysis. In Chapter 5 we propose model formulations for selecting province-specific effects of interventions and assess the effects of changes of intervention coverage measures on the changes of parasitemia risk within a 6-year period in Angola. However, the effects of interventions are likely to vary within province. In fact, studies have shown a community effect of malaria interventions (Killeen et al., 2007). However, modelling the effects of interventions at community level is computationally demanding because spatially varying covariate effects should be estimated using Gaussian processes over a very large number of locations. We address this issue with predictive process approximations of the point-level spatiotemporally varying effects in Chapter 6.

7.1.3 Compilation of helminthiasis survey data

In the field of neglected tropical diseases, georeferenced data availability has been scarce. On the one hand, national surveys are not conducted systematically as, for example, in malaria. On the other hand, accessibility of existing survey data is difficult. Many small-scale survey data are reported in the literature, however they are not readily available for disease mapping purposes.

The Global Neglected Tropical Diseases (GNTD) database (Hürlimann et al., 2011)

compiles and provides freely and publicly accessible georeferenced data of infection prevalence reported in the literature, national control programs, researchers and organizations. We contributed to the GNTD database by conducting two large systematic reviews pertaining to parasitological surveys reporting soil-transmitted helminth infection prevalence in sub-Saharan Africa and Cambodia (Chapters 3 and 4).

These data constitute a comprehensive collection of historical and contemporary information of STH prevalence. The results of the systematic reviews include detailed information such as diagnostic tests, age, sex, survey type *etc.* More importantly, the surveys are georeferenced and can be used in spatial analyses. In particular, we identified 537 sources with relevant data pertaining to STH infections out of more than 6,000 screened references. In total, approximately 6 thousand unique locations were georeferenced.

7.1.4 Tropical epidemiology: disease control and intervention planning

The results of this thesis can assist control programmes to select treatment strategies, funding agencies to allocate resources and drug donors to plan cost-effectively. Risk estimates provide baseline information for monitoring and evaluation. Estimates of the effects of malaria interventions in space identify areas of successful disease control.

The study presented in Chapter 2 highlights the situation of leishmaniasis in Brazil. Incidence rates have decreased since 2000 but remained stable from 2005 onwards. They are, though, at similar levels with late 1980's (Brandão Filho et al., 1999). The information system for notifiable diseases in Brazil has been shown to be prone to under-reporting of visceral leishmaniasis cases (Maia-Elkhoury et al., 2007). Our model-based spatially explicit predictions of disease estimated higher number of disease cases than the ones officially reported. This finding underscores the problem of under-reporting which is crucial for control planning.

For helminthiasis control, we provide up-to-date population-adjusted STH risk estimates in Cambodia (Chapter 4) as well as treatment needs based on WHO

guidelines (WHO, 2006) in sub-Saharan Africa (Chapter 3). In sub-Saharan Africa, we estimate annual requirements of 126 million tablets of albendazole or mebendazole. The predicted number of children requiring treatments is approximately 91 million. Across the sub-continent, we highlight areas where data are scarce and model-based estimates suggest high STH prevalences. For instance, in the Republic of the Congo and the Democratic Republic of the Congo we estimate overall STH prevalences above 20%, which is the WHO threshold for annual treatment. This shows the need to conduct more surveys and alarms about the public health concern of high STH risk in the two countries. The two studies (Chapters 3 and 4) suggest a clear decrease of STH risk from 2000 onwards. The temporal trend can be explained by socioeconomic development as well as intensified administration of preventive chemotherapy since the World Health Assembly resolution 54.19 of 2001 (WHO, 2002a). These results complement the globally decreasing trend of STH infections found also in South America (Chammartin et al., 2013b) and China (Lai et al., 2013). However, a geostatistical analysis of STH risk in sub-Saharan Africa, incorporated in the 2010 Global Burden of Disease Study (Murray et al., 2014), suggested that no temporal trend existed between 1990 and 2010 (Pullan et al., 2014). This is a surprising finding and contradicts the expected effect of the administered hundreds of millions of treatments targeting STH as well as billions of administered tablets targeting onchocerciasis and lymphatic filariasis that have an effect on soil-transmitted helminthiasis (Ottesen et al., 2008). Furthermore, in sub-Saharan Africa we estimated that *A. lumbricoides* and *T. trichiura* prevalences between school-aged children and the broader community are at similar levels from 2000 onwards. This suggests that current control strategies should extend the focus of treatment from children to adults in order to progress towards local elimination. Therefore, Chapter 3 contributes with more accurate, spatially explicit, age and population-specific STH risk estimates for the sub-continent. It also provides pixel-based treatment requirements by country to assist control planning. These results are needed to assess prevalence changes in future surveys and relate these changes with implemented control.

Chapter 5 provides information on the effects of malaria interventions on the disease dynamics in Angola at sub-national level. Estimates of province-specific

effects, adjusted for climatic factors, depict those regions where an intervention had an effect on parasitemia risk reduction and those that control appears to be less effective.

7.2 Limitations

Historical helminthiasis data compilation and meta-analyses are prone to bias due to the different levels of heterogeneity. Specifically, parasitological surveys target different age groups, are based on diagnostic tests with different diagnostic accuracy and data are reported at different spatial units such as school, village, district *etc.* Some of these limitations can be addressed by statistical modelling.

Age heterogeneity among survey locations can be taken into account by incorporating mathematical models within statistical ones to align to a common age group. The main difficulty with this approach is that the age-prevalence dependence changes in space. Estimation is feasible if there are survey data available across the study area and age groups, which is rarely the case. Spatial models can straightforwardly adjust for diagnostic error by including the sensitivity and specificity of the test as parameters in the model. Nikolay et al. (2014) conducted a meta-analysis of diagnostic accuracy of different diagnostic tools used for STH infection. This information can be used to define prior distributions in Bayesian geostatistical models. However, diagnostic test accuracy depends on intensity of infections and data reported in the literature may provide incomplete information about diagnostic tests making difficult diagnostic error adjustment. The issue of modelling data obtained at different spatial survey units is commonly referred in statistical literature as "change of support". Taking into account the different spatial units in geostatistical models may improve parameter estimation.

Administering preventive chemotherapy and improving sanitation as well as access to safe water are the main control measures against soil-transmitted helminth infections. Places were such control is already implemented are expected to have lower infection risk. This information could be introduced via covariates into spatial modelling. However, disease control data availability is scarce. Currently, WHO reports preventive chemotherapy coverage by country. This does not allow

to investigate within country variability and to associate STH risk with the stage of a control program in, *e.g.*, a district. Efforts to collect georeferenced treatment data have been initiated (http://www.deworminginventory.org).

In Chapters 3 and 4, we estimate the overall STH risk under the assumption of independence among species-specific infections. Consequently, we do not incorporate species co-infection and STH risk might be slightly overestimated. A correction factor has been proposed for overall STH prevalence but it was based on a non-rigorous calculation (de Silva and Hall, 2010).

7.3 Extension

This research offers a pillar which future work can be based on. Our STH risk modelling in sub-Saharan Africa and Cambodia combined with modelling efforts in South America (Chammartin et al., 2013b), China (Lai et al., 2013) and ongoing work in South and Southeast Asia can provide global model-based estimates of the disease risk. The inclusion in DALYs estimation would constitute a significant step towards updating disease burden.

In the lack of co-infection data across Africa, a helminthiasis co-distribution study can be conducted using our STH estimates and the schistosomiasis risk predictions of Lai et al. (2015). The resulted maps could guide the integration of STH and schistosomiasis control programs.

The modelling and software developed in this research can be implemented in an online application to provide the necessary tools to control managers. This application combined with the GNTD open-access database and WHO's preventive chemotherapy databank, would result in updating disease estimates and evaluating control implementation in almost real time.

There are many approaches of dealing with correlated predictors in statistical modelling. The use of g-priors (see, for example, Sabanés Bové et al., 2011) would smooth away but yet allow the inclusion of correlated predictors in spatial models. A data-driven modification of g-priors could also smooth together correlated predictors (Krishna et al., 2009). On a similar note, varying coefficients of predictors could be cross-correlated. The use of Matérn cross-covariance functions (Gneiting et al.,

2012) could allow information of effects to be drawn among predictors and locations. A g-prior for random effects has not been yet implemented. Its combination with Matérn cross-covariance functions could be envisaged.

Bibliography

Ali-Akbarpour, M., Mohammadbeigi, A., Tabatabaee, S. H. R., and Hatam, G. (2012). Spatial analysis of eco-environmental risk factors of cutaneous leishmaniasis in Southern Iran. *Journal of Cutaneous and Aesthetic Surgery*, 5: 30–35.

Alvar, J., Vélez, D. I., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., den Boer, M., and the WHO Leishmaniasis Control Team (2012). Leishmaniasis worldwide and global estimates of its incidence. *PLoS One*, 7: e35671.

Alves, W. A. (2009). Leishmaniose visceral americana: situação atual no Brasil leishmaniasis: current situation in Brazil. *Boletim Epidemiológico Paulista*, 6: 25–29.

Anderson, R. M., Truscott, J. E., Pullan, R. L., Brooker, S. J., and Hollingsworth, T. D. (2013). How effective is school-based deworming for the community-wide control of soil-transmitted helminths? *PLoS Neglected Tropical Diseases*, 7: e2027.

Assunção, R. M., Reis, I. A., and Oliveira, C. D. (2001). Diffusion and prediction of leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine*, 20: 2319–2335.

Augusto, G., Nalá, R., Casmo, V., Sabonete, A., Mapaco, L., and Monteiro, J. (2009). Geographic distribution and prevalence of schistosomiasis and soil-transmitted helminths among schoolchildren in mozambique. *American Journal of Tropical Medicine and Hygiene*, 81: 799–803.

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

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida, second 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: 825–848.

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

Bates, D., Mullen, K. M., Nash, J. C., and Varadhan, R. (2014). minqa: Derivative-free optimization algorithms by quadratic approximation. http://CRAN.R-project.org/package=minqa, R package version 1.2.4.

Bethony, J., Brooker, S., Albonico, M., Geiger, S. M., Loukas, A., Diemert, D., and Hotez, P. J. (2006). Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *The Lancet*, 367: 1521–1532.

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C., Henry, A., Eckhoff, P., et al. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526: 207–211.

Boehm Vock, L. F., Reich, B. J., Fuentes, M., and Dominici, F. (2015). Spatial variable selection methods for investigating acute health effects of fine particulate matter components. *Biometrics*, 71: 167–177.

Bornkamp, B. (2011). Approximating probability densities by iterated Laplace approximations. *Journal of Computational and Graphical Statistics*, 20: 656669.

Brandão Filho, S. P., Campbell-Lendrum, D., Brito, M. E., Shaw, J. J., and Davies, C. R. (1999). Epidemiological surveys confirm an increasing burden of cutaneous leishmaniasis in north-east Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93: 488–494.

Brasil Ministério da Saúde, Secretaria de Vigilância em Saúde (2007a). Guida de vigilância epidemiológica. http://portal.saude.gov.br/portal/arquivos/pdf/gve_7ed_web_atual.pdf, accessed: 18 March 2013.

Brasil Ministério da Saúde, Secretaria de Vigilância em Saúde (2007b). Manual de vigilância da leishmaniose tegumentar Americana. http://portal.saude.gov.br/portal/arquivos/pdf/manual_lta_2ed.pdf, accessed: 18 March 2013.

Brooker, S., Clements, A. C., and Bundy, D. A. (2006). Global epidemiology, ecology and control of soil-transmitted helminth infections. *Advances in Parasitology*, 62: 221–261.

Brooker, S., Hotez, P. J., and Bundy, D. A. (2009a). An updated atlas of human helminth infections: the example of East Africa. *International Journal of Health Geographics*, 8: 42.

Brooker, S., Hotez, P. J., and Bundy, D. A. (2010). The Global Atlas of Helminth Infection: mapping the way forward in neglected tropical disease control. *PLoS Neglected Tropical Diseases*, 4: e779.

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

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Advances in Statistical Analysis*, 97: 109–131.

Campbell, S. J., Savage, G. B., Gray, D. J., Atkinson, J., Soares Magalhães, R. J., Nery, S. V., McCarthy, J. S., Velleman, Y., Wicken, J. H., Traub, R. J., et al. (2014). Water, sanitation, and hygiene (WASH): a critical component for sustainable soil-transmitted helminth and schistosomiasis control. *PLoS Neglected Tropical Diseases*, 8: e2651.

Center for International Earth Science Information Network (CIESIN), Columbia University (2000). Global subnational infant mortality rates.

Center for International Earth Science Information Network (CIESIN), Columbia University (2005). Gridded population of the world: future estimates (GPWFE).

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

Chammartin, F., Scholte, R. G., Guimarães, L. H., Tanner, M., Utzinger, J., and Vounatsou, P. (2013b). Soil-transmitted helminth infection in South America:

a systematic review and geostatistical meta-analysis. *The Lancet Infectious Diseases*, 13: 507–518.

Chaves, L. F., Cohen, J. M., Pascual, M., and Wilson, M. L. (2008). Social exclusion modifies climate and deforestation impacts on a vector-borne disease. *PLoS Neglected Tropical Diseases*, 2: e176.

Chaves, L. F. and Pascual, M. (2006). Climate cycles and forecasts of cutaneous leishmaniasis, a nonstationary vector-borne disease. *PLoS Medicine*, 3: e295.

Chhakda, T., Muth, S., Socheat, D., and Odermatt, P. (2006). Intestinal parasites in school-aged children in villages bordering Tonle Sap Lake, Cambodia. *Southeast Asian Journal of Tropical Medicine and Public Health*, 37: 859.

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

Clements, A. C. A., Deville, M.-A., 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 and International Health*, 15: 198–207.

Clements, A. C. A., Firth, S., Dembélé, R., Garba, A., Touré, S., Sacko, M., Landouré, A., B.-O. E., Barnett, A. G., Brooker, S., and Fenwick, A. (2009). Use of Bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in western Africa. *Bulletin of the World Health Organization*, 87: 921–929.

Coulibaly, S., Maiga, M., Sawadogo, M., Magnussen, P., Guiguemde, R., and Some, I. (2011). Epidemiology of intestinal helminths, schistosomiasis and ectoparasites in schoolchildren in burkina faso. *Tropical Medicine and International Health*, 16 (suppl I): 98–99.

Dalrymple, U., Mappin, B., and Gething, P. W. (2015). Malaria mapping: understanding the global endemicity of falciparum and vivax malaria. *BMC Medicine*, 13: 140.

Dantas-Torres, F. and Brandão Filho, S. P. (2006). Visceral leishmaniasis in Brazil: revisiting paradigms of epidemiology and control. *Revista do Instituto de Medicina Tropical de São Paulo*, 48: 151–156.

de Silva, N. R., Brooker, S., Hotez, P. J., Montresor, A., Engels, D., and Savioli, L. (2003). Soil-transmitted helminth infections: updating the global picture. *Trends in Parasitology*, 19: 547–551.

de Silva, N. R. and Hall, A. (2010). Using the prevalence of individual species of intestinal nematode worms to estimate the combined prevalence of any species. *PLoS Neglected Tropical Diseases*, 4: e655.

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

Desjeux, P. (2001). The increase in risk factors for leishmaniasis worldwide. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 95: 239–243.

Desjeux, P. (2004). Leishmaniasis: current situation and new perspectives. *Comparative Immunology, Microbiology and Infectious Diseases*, 27: 305–318.

Diboulo, E., Sié, A., Diadier, D. A., Karagiannis-Voules, D. A., Yé, Y., and Vounatsou, P. (2015). Bayesian variable selection in modelling geographical heterogeneity in malaria transmission from sparse data: an application to Nouna Health and Demographic Surveillance System (HDSS) data, Burkina Faso. *Parasites & Vectors*, 8: 118.

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

Ekpo, U. F., Hürlimann, E., Schur, N., Oluwole, A. S., Abe, E. M., Mafe, M. A., Nebe, O. J., Isiyaku, S., Olamiju, F., Kadiri, M., et al. (2013). Mapping and prediction of schistosomiasis in Nigeria using compiled survey data and Bayesian geospatial modelling. *Geospatial Health*, 7: 355–366.

Elnaiem, D.-E. A., Schorscher, J., Bendall, A., Obsomer, V., Osman, M. E., Mekkawi, A. M., Connor, S. J., Ashford, R. W., and Thomson, M. C. (2003). Risk mapping of visceral leishmaniasis: the role of local variation in rainfall and altitude on the presence and incidence of kala-azar in eastern Sudan. *American Journal of Tropical Medicine and Hygiene*, 68: 10–17.

Environmental Systems Research Institute (2010). ArcGIS Desktop: Release 10.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15.

Fürst, T., Silué, K. D., Ouattara, M., N'Goran, D. N., Adiossan, L. G., N'Guessan, Y., Zouzou, F., Koné, S., N'Goran, E. K., and Utzinger, J. (2012). Schistosomiasis, soil-transmitted helminthiasis, and sociodemographic factors influence quality of life of adults in Côte d'Ivoire. *PLoS Neglected Tropical Diseases*, 6: e1855.

Gallo, K., Mikhailov, A., Hailemeskal, M. B., Koporc, K., Mbabazi, P. S., and Addiss, D. (2013). Contributions of non-governmental organizations to WHO targets for control of soil-transmitted helminthiases. *American Journal of Tropical Medicine and Hygiene*, 89: 1186–1189.

Gallup, J. L. and Sachs, J. D. (2001). The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, 64: 85–96.

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

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

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

George, E. I. and McCulloch, R. E. (1996). Stochastic search variable selection. In *Markov chain Monte Carlo in practice*, pages 203–214. Springer.

George, J., Yiannakis, M., Main, B., Devenish, R., Anderson, C., An, U. S., Williams, S. M., and Gibson, R. S. (2012). Genetic hemoglobin disorders, infection, and deficiencies of iron and vitamin A determine anemia in young Cambodian children. *Journal of Nutrition*, 142: 781–787.

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

Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I., Johnston, G. L., Tatem, A. J., and Hay, S. I. (2011). A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal*, 10: 1475–2875.

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

Giardina, F., Kasasa, S., Sié, A., Utzinger, J., Tanner, M., and Vounatsou, P. (2014). Effects of vector-control interventions on changes in risk of malaria parasitaemia in sub-Saharan Africa: a spatial and temporal analysis. *The Lancet Global Health*, 2: e601–e615.

Gneiting, T., Kleiber, W., and Schlather, M. (2012). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105: 1167–1177.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102: 359–378.

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

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

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Gelfand, A. E. (2011). Adaptive Gaussian predictive process models for large spatial datasets. *Environmetrics*, 22: 997–1007.

Hawley, W. A., Phillips-Howard, P. A., ter Kuile, F. O., Terlouw, D. J., Vulule, J. M., Ombok, M., Nahlen, B. L., Gimnig, J. E., Kariuki, S. K., Kolczak, M. S., et al. (2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *The American Journal of Tropical Medicine and Hygiene*, 68: 121–127.

Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., Brownstein, J. S., Collier, N., Myers, M. F., George, D. B., and W, G. P. (2013).

Global mapping of infectious disease. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368: 20120250.

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: e1000048.

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

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In *Statistical modelling and regression structures*, pages 91–110. Springer.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics*, 6: 761–768.

Hijmans, R. J. (2014). raster: Geographic data analysis and modeling. http://CRAN.R-project.org/package=raster, R package version 2.3-12.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25: 1965–1978.

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

Hotez, P. J. (2013). NTDs V.2.0:blue marble health—neglected tropical disease control and elimination in a shifting health policy landscape. *PLoS Neglected Tropical Diseases*, 7: e2570.

Hotez, P. J., Alvarado, M., Basáñez, M.-G., Bolliger, I., Bourne, R., Boussinesq, M., Brooker, S. J., Brown, A. S., Buckle, G., Budke, C. M., et al. (2014). The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. *PLoS Neglected Tropical Diseases*, 8: e2865.

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

Howard, S., Omumbo, J., Nevill, C., Some, E., Donnelly, C., and Snow, R. (2000). Evidence for a mass community effect of insecticide-treated bednets on the incidence of malaria on the Kenyan coast. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94: 357–360.

Hürlimann, E., Schur, N., Boutsika, K., Stensgaard, A.-S., de Himpsl, M. L., Ziegelbauer, K., Laizer, N., Camenzind, L., Di Pasquale, A., Ekpo, U. F., et al. (2011). Toward an open-access global database for mapping, control, and surveillance of neglected tropical diseases. *PLoS Neglected Tropical Diseases*, 5: e1404.

Inpankaew, T., Schär, F., Dalsgaard, A., Khieu, V., Chimnoi, W., Chhoun, C., Sok, D., Marti, H., Muth, S., Odermatt, P., et al. (2014). High prevalence of *Ancylostoma ceylanicum* hookworm infections in humans, Cambodia, 2012. *Emerging Infectious Diseases*, 20: 976.

Jex, A. R., Lim, Y. A., Bethony, J., Hotez, P. J., Young, N. D., and Gasser, R. B. (2011). Soil-transmitted helminths of humans in Southeast Asia—towards integrated control. *Advances in Parasitology*, 74: 231.

Jia, T.-W., Melville, S., Utzinger, J., King, C. H., and Zhou, X.-N. (2012). Soil-transmitted helminth reinfection after drug treatment: a systematic review and meta-analysis. *PLoS Neglected Tropical Diseases*, 6: e1621.

Jirmanus, L., Glesby, M. J., Guimarães, L. H., Lago, E., Rosa, M. E., Machado, P. R., and Carvalho, E. M. (2012). Epidemiological and clinical changes in American tegumentary leishmaniasis in an area of *Leishmania* (*Viannia*) *braziliensis* transmission over a 20-year period. *American Journal of Tropical Medicine and Hygiene*, 86: 426–433.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26: 131–148.

Kabore, A., Biritwum, N.-K., Downs, P. W., Magalhaes, R. J. S., Zhang, Y., and Ottesen, E. A. (2013). Predictive vs. empiric assessment of schistosomiasis: implications for treatment projections in Ghana. *PLoS Neglected Tropical Diseases*, 7: e2051.

Karagiannis-Voules, D.-A., Biedermann, P., Ekpo, U. F., Garba, A., Langer, E., Mathieu, E., Midzi, N., Mwinzi, P., Polderman, A. M., Raso, G., Sacko, M., Talla,

I., Tchuem Tchuenté, L.-A., S., Winkler, M. S., Utzinger, J., and Vounatsou, P. (2015a). Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis. *The Lancet Infectious Diseases*, 15: 74–84.

Karagiannis-Voules, D.-A., Odermatt, P., Biedermann, P., Khieu, V., Schär, F., Muth, S., Utzinger, J., and Vounatsou, P. (2015b). Geostatistical modelling of soil-transmitted helminth infection in Cambodia: Do socioeconomic factors improve predictions? *Acta Tropica*, 141: 204–212.

Karagiannis-Voules, D.-A., Scholte, R. G., Guimarães, L. H., Utzinger, J., and Vounatsou, P. (2013). Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLoS Neglected Tropical Diseases*, 7: e2213.

Keiser, J. and Utzinger, J. (2008). Efficacy of current drugs against soil-transmitted helminth infections: systematic review and meta-analysis. *Journal of the American Medical Association*, 299: 1937–1948.

Khieu, V., Schär, F., Forrer, A., Hattendorf, J., Marti, H., Duong, S., Vounatsou, P., Muth, S., and Odermatt, P. (2014a). High prevalence and spatial distribution of *Strongyloides stercoralis* in rural Cambodia. *PLoS Neglected Tropical Diseases*, 8: e2854.

Khieu, V., Schär, F., Marti, H., Bless, P. J., Char, M. C., Muth, S., and Odermatt, P. (2014b). Prevalence and risk factors of *Strongyloides stercoralis* in Takeo province, Cambodia. *Parasites & Vectors*, 7: 221.

Khieu, V., Schär, F., Marti, H., Sayasone, S., Duong, S., Muth, S., and Odermatt, P. (2013). Diagnosis, treatment and risk factors of *Strongyloides stercoralis* in schoolchildren in Cambodia. *PLoS Neglected Tropical Diseases*, 7: e2035.

Kightlinger, L. K., Seed, J. R., and Kightlinger, M. B. (1995). The epidemiology of *Ascaris lumbricoides*, *Trichuris trichiura*, and hookworm in children in the Ranomafana rainforest, Madagascar. *Journal of Parasitology*, pages 159–169.

Killeen, G. F., Smith, T. A., Ferguson, H. M., Mshinda, H., Abdulla, S., Lengeler, C., and Kachur, S. P. (2007). Preventing childhood malaria in Africa by protecting adults from mosquitoes with insecticide-treated nets. *PLoS Medicine*, 4: e229.

King, R. J., Campbell-Lendrum, D., and Davies, C. R. (2004). Predicting geographic

variation in cutaneous leishmaniasis, Colombia. *Emerging Infectious Diseases*, 10: 598–607.

Klaus, S. N., Frankenburg, S., and Ingber, A. (1999). Epidemiology of cutaneous leishmaniasis. *Clinics in Dermatology*, 17: 257–260.

Knopp, S., Mohammed, K. A., Speich, B., Hattendorf, J., Khamis, I. S., Khamis, A. N., Stothard, J. R., Rollinson, D., Marti, H., and Utzinger, J. (2010). Albendazole and mebendazole administered alone or in combination with ivermectin against *Trichuris trichiura*: a randomized controlled trial. *Clinical Infectious Diseases*, 51: 1420–1428.

Krishna, A., Bondell, H. D., and Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference*, 139: 2665–2674.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60: 65–81.

Lai, Y.-S., Biedermann, P., Ekpo, U. F., Garba, A., Mathieu, E., Midzi, N., Mwinzi, P., N'Goran, E. K., Raso, G., Assaré, R. K., et al. (2015). Spatial distribution of schistosomiasis and treatment needs in sub-Saharan Africa: a systematic review and geostatistical analysis. *The Lancet Infectious Diseases*.

Lai, Y.-S., Zhou, X.-N., Utzinger, J., and Vounatsou, P. (2013). Bayesian geostatistical modelling of soil-transmitted helminth survey data in the People's Republic of China. *Parasites & Vectors*, 6: 359.

Lasinio, G. J., Mastrantonio, G., and Pollice, A. (2013). Discussing the "big n problem". *Statistical Methods & Applications*, 22: 97–112.

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: 23.

Li, T., He, S., Zhao, H., Zhao, G., and Zhu, X.-Q. (2010). Major trends in human parasitic diseases in China. *Trends in Parasitology*, 26: 264–270.

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

Lum, K. (2012). Bayesian variable selection for spatially dependent generalized linear models. *arXiv preprint arXiv:1209.0661*.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10: 325–337.

Lysenko, A. and Semashko, I. (1968). Geography of malaria. A medico-geographic profile of an ancient disease. *Itogi Nauki: Medicinskaja Geografija*, pages 25–146.

Machado-Coelho, G. L., Assunção, R., Mayrink, W., and Caiaffa, W. T. (1999). American cutaneous leishmaniasis in Southeast Brazil: space-time clustering. *International Journal of Epidemiology*, 28: 982–989.

Magalhães, R. J. S., Barnett, A. G., and Clements, A. C. (2011). Geographical analysis of the role of water supply and sanitation in the risk of helminth infections of children in West Africa. *Proceedings of the National Academy of Sciences*, 108: 20084–20089.

Maia-Elkhoury, A. N. S., Alves, W. A., Sousa-Gomes, M. L. d., Sena, J. M. d., and Luna, E. A. (2008). Visceral leishmaniasis in Brazil: trends and challenges. *Cadernos de Saúde Pública*, 24: 2941–2947.

Maia-Elkhoury, A. N. S., Carmo, E. H., Sousa-Gomes, M. L., and Mota, E. (2007). Analysis of visceral leishmaniasis reports by the capture-recapture method. *Revista de Saúde Pública*, 41: 931–937.

Massa, K., Magnussen, P., Sheshe, A., Ntakamulenga, R., Ndawi, B., and Olsen, A. (2009). The combined effect of the lymphatic filariasis elimination programme and the schistosomiasis and soil-transmitted helminthiasis control programme on soil-transmitted helminthiasis in schoolchildren in Tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103: 25–30.

MEASURE Evaluation, MEASURE DHS, President's Malaria Initiative, Roll Back Malaria Partnership, UNICEF, and WHO (2013). Household survey indicators for malaria control.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D., and Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *British Medical Journal*, 339: b2535.

Molyneux, D. H. (2004). "Neglected" diseases but unrecognised successes—challenges and opportunities for infectious disease control. *The Lancet*, 364: 380–383.

Moncayo, A. L., Vaca, M., Amorim, L., Rodriguez, A., Erazo, S., Oviedo, G., Quinzo, I., Padilla, M., Chico, M., Lovato, R., et al. (2008). Impact of long-term treatment with ivermectin on the prevalence and intensity of soil-transmitted helminth infections. *PLoS Neglected Tropical Diseases*, 2: e293.

Montresor, A., Gabrielli, A. F., Yajima, A., Lethanh, N., Biggs, B., Casey, G., Tihn, T., Engels, D., and Savioli, L. (2013). Markov model to forecast the change in prevalence of soil-transmitted helminths during a control programme: a case study in Vietnam. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 107: 313–318.

Murray, C. J., Ortblad, K. F., Guinovart, C., Lim, S. S., Wolock, T. M., Roberts, D. A., Dansereau, E. A., Graetz, N., Barber, R. M., Brown, J. C., et al. (2014). Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 384: 1005–1070.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380: 2197–2223.

Nikolay, B., Brooker, S. J., and Pullan, R. L. (2014). Sensitivity of diagnostic tests for human soil-transmitted helminth infections: a meta-analysis in the absence of a true gold standard. *International Journal for Parasitology*, 44: 765–774.

Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., Fall, I. S., and Snow, R. W. (2014). The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity. *The Lancet*, 383: 1739–1747.

Nychka, D., Furrer, R., and Sain, S. (2015). fields: Tools for spatial data. http://CRAN.R-project.org/package=fields, R package version 8.2-1.

Oak Ridge National Laboratory Distributed Active Archive Center (2011). MODIS subsetted land products, Collection 5.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4: 85–117.

Organization, W. H. et al. (2013). Soil-transmitted helminthiases: number of children treated in 2011. *Weekly Epidemiological Record*, 88: 145–152.

Ottesen, E. A., Hooper, P. J., Bradley, M., and Biswas, G. (2008). The global programme to eliminate lymphatic filariasis: health impact after 8 years. *PLoS Neglected Tropical Diseases*, 2: e317.

Ouattara, M., N'Guéssan, N. A., Yapi, A., and N'Goran, E. K. (2010). Prevalence and spatial distribution of *Entamoeba histolytica/dispar* and *Giardia lamblia* among schoolchildren in Agboville area (Côte d'Ivoire). *PLoS Neglected Tropical Diseases*, 4: e574.

Ouattara, M., Silué, K. D., NGuéssan, A. N., Yapi, A., Barbara, M., Raso, G., Utzinger, J., and NGoran, É. (2008). Prévalences et polyparasitisme des protozoaires intestinaux et répartition spatiale d*Entamoeba histolytica/Entamoeba dispar* et *Giardia intestinalis* chez des élèves en zone rurale de la région de Man en Côte d'Ivoire. *Cahiers d'Études et de Recherches Francophones/Santé*, 18: 215–222.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, 9: 523–539.

Pullan, R. L. and Brooker, S. J. (2012). The global limits and population at risk of soil-transmitted helminth infections in 2010. *Parasites & Vectors*, 5: 81.

Pullan, R. L., Gething, P. W., Smith, J. L., Mwandawiro, C. S., Sturrock, H. J., 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: e958.

Pullan, R. L., Kabatereine, N. B., Quinnell, R. J., and Brooker, S. (2010). Spatial and genetic epidemiology of hookworm in a rural community in Uganda. *PLoS Neglected Tropical Diseases*, 4: e713.

Pullan, R. L., Smith, J. L., Jasrasaria, R., and Brooker, S. J. (2014). Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasites & Vectors*, 7: 37.

Pupo Nogueira Neto, J., Basso, G., Cipoli, A. P., and El Kadre, L. (1998). American cutaneous leishmaniasis in the state of São Paulo, Brazil–epidemiology in transformation. *Annals of Agricultural and Environmental Medicine*, 5: 1–5.

R Core Team (2014). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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 d'Ivoire. *Parasitology*, 131: 97–108.

Raso, G., Schur, N., Utzinger, J., Koudou, B. G., Tchicaya, E. S., Rohner, F., N'Goran, E. K., Silu, K. D., Matthys, B., Assi, S., Tanner, M., and Vounatsou, P. (2012). Mapping malaria risk among children in Côte d'Ivoire using Bayesian geo-statistical models. *Malaria Journal*, 11: 160.

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

RTS,S Clinical Trials Partnership (2015). Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in africa: final results of a phase 3, individually randomised, controlled trial. *The Lancet*, 386: 31–45.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC Press, London.

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

Saarnak, C. F., Utzinger, J., and Kristensen, T. K. (2013). Collection, verification,

sharing and dissemination of data: the contrast experience. *Acta Tropica*, 128: 407–411.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6: 387–410.

Sabanés Bové, D., Held, L., and Kauermann, G. (2011). Mixtures of g-priors for generalised additive model selection with penalised splines. *arXiv preprint arXiv:1108.3520*.

Salomon, O. D., Quintana, M. G., Mastrangelo, A. V., and Fernandez, M. S. (2012). Leishmaniasis and climate change–case study: Argentina. *Journal of Tropical Medicine*, 2012: 601242.

Samadoulougou, S., Maheu-Giroux, M., Kirakoya-Samadoulougou, F., Keukeleire, M. D., Castro, M. C., and Robert, A. (2014). Multilevel and geo-statistical modeling of malaria risk in children of Burkina Faso. *Parasites & Vectors*, 7: 350.

Savioli, L., Gabrielli, A., Montresor, A., Chitsulo, L., and Engels, D. (2009). Schistosomiasis control in Africa: 8 years after World Health Assembly Resolution 54·19. *Parasitology*, 136: 1677–1681.

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

Schur, N., Hürlimann, E., Stensgaard, A.-S., Chimfwembe, K., Mushinge, G., Simoonga, C., Kabatereine, N. B., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2011b). Spatially explicit *Schistosoma* infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Tropica*, 128: 365–377.

Schur, N., Vounatsou, P., and Utzinger, J. (2012). Determining treatment needs at different spatial scales using geostatistical model-based risk estimates of schistosomiasis. *PLoS Neglected Tropical Diseases*, 6: e1773.

Shimabukuro, P. H. F., da Silva, T. R. R., Fonseca, F. O. R., Baton, L. A., and Galati, E. A. B. (2010). Geographical distribution of American cutaneous

leishmaniasis and its phlebotomine vectors (diptera: *Psychodidae*) in the state of São Paulo, Brazil. *Parasites & Vectors*, 3: 121.

Sinuon, M., Tsuyuoka, R., Socheat, D., Odermatt, P., Ohmae, H., Matsuda, H., Montresor, A., and Palmer, K. (2007). Control of *Schistosoma mekongi* in Cambodia: results of eight years of control activities in the two endemic provinces. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101: 34–39.

Slater, H. and Michael, E. (2013). Mapping, Bayesian geostatistical analysis and spatial prediction of lymphatic filariasis prevalence in Africa. *PLoS One*, 8: e71574.

Soares Magalhães, R. J., Biritwum, N.-K., Gyapong, J. O., Brooker, S., Zhang, Y., Blair, L., Fenwick, A., and Clements, A. (2011). Mapping helminth co-infection and co-intensity: geostatistical prediction in Ghana. *PLoS Neglected Tropical Diseases*, 5: e1200.

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

Strunz, E. C., Addiss, D. G., Stocks, M. E., Ogden, S., Utzinger, J., and Freeman, M. C. (2014). Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. *PLoS Medicine*, 11: e1001620.

Sundaram, J. K., Schwank, O., and Von Arnim, R. (2011). Globalization and development in sub-Saharan Africa. Technical report.

Thompson, R. A., Wellington de Oliveira Lima, J., Maguire, J. H., Braud, D. H., and Scholl, D. T. (2002). Climatic and demographic determinants of American visceral leishmaniasis in northeastern Brazil using remote sensing technology for environmental categorization of rain and region influences on leishmaniasis. *American Journal of Tropical Medicine and Hygiene*, 67: 648–655.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81: 82–86.

Traub, R. J., Inpankaew, T., Sutthikornchai, C., Sukthana, Y., and Thompson, R. A. (2008). PCR-based coprodiagnostic tools reveal dogs as reservoirs of zoonotic

ancylostomiasis caused by *Ancylostoma ceylanicum* in temple communities in Bangkok. *Veterinary Parasitology*, 155: 67–73.

UNICEF (2003). Mapping human helminth infections in Southeast Asia. United Nations Childrens Fund, Bangkok, Thailand.

Utzinger, J., Becker, S. L., Knopp, S., Blum, J., Neumayr, A. L., Keiser, J., and Hatz, C. F. (2012). Neglected tropical diseases: diagnosis, clinical management, treatment and control. *Swiss Medical Weekly*, 142: w13727.

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

Valderrama-Ardila, C., Alexander, N., Ferro, C., Cadena, H., Marín, D., Holford, T. R., Munstermann, L. E., and Ocampo, C. B. (2010). Environmental risk factors for the incidence of American cutaneous leishmaniasis in a sub-Andean zone of Colombia (Chaparral, Tolima). *American Journal of Tropical Medicine and Hygiene*, 82: 243–250.

Vyas, S. and Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21: 459–468.

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

Wen, L.-Y., Yan, X.-L., Sun, F.-H., Fang, Y.-Y., Yang, M.-J., and Lou, L.-J. (2008). A randomized, double-blind, multicenter clinical trial on the efficacy of ivermectin against intestinal nematode infections in China. *Acta Tropica*, 106: 190–194.

Werneck, G. L. and Maguire, J. H. (2002). Spatial modeling using mixed models: an ecologic study of visceral leishmaniasis in Teresina, Piauí state, Brazil. *Cadernos de Saúde Pública*, 18: 633–637.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

WHO (2002a). Prevention and control of schistosomiasis and soil-transmitted helminthiasis. *World Health Organization technical report series*, 912: 1–57.

WHO (2002b). The world health report 2002 - reducing risks, promoting healthy life. World Health Organization, Geneva, Switzerland.

WHO (2006). *Preventive chemotherapy in human helminthiasis: coordinated use of anthelminthic drugs in control interventions: a manual for health professionals and programme managers*. World Health Organization.

WHO (2007). Sixtieth World Health Assembly. World Health Organization, Geneva, Switzerland.

WHO (2010a). Control of the leishmaniases. *World Health Organization technical report series*, 949: 1–152.

WHO (2010b). Soil-transmitted helminthiasis. Number of children treated 2007-2008: update on the 2010 global target. *Weekly Epidemiological Record*, 85: 141–147.

WHO (2012a). Accelerating work to overover the global impact of neglected tropical diseases. World Health Organization, Geneva, Switzerland.

WHO (2012b). African Programme for Onchocerciasis Control: meeting of national onchocerciasis task forces. *Weekly Epidemiological Record*, 87: 494–502.

WHO (2012c). Eliminating soil-transmitted helminthiases as a public health problem in children. progress report 2001-2010 and strategic plan 2011-2020. World Health Organization, Geneva, Switzerland.

WHO (2013a). Estimated number of people covered by preventive chemotherapy: update for 2010 and 2013. *Weekly Epidemiological Record*, 88: 24–28.

WHO (2013b). Rolling out and scaling up integrated preventive chemotherapy for selected neglected tropical diseases. *Weekly Epidemiological Record*, 88: 161–166.

WHO (2013c). Sustaining the drive to overcome the global impact of neglected tropical diseases: second WHO report on neglected tropical diseases. World Health Organization, Geneva, Switzerland.

WHO (2014). Soil-transmitted helminthiases: number of children treated in 2012. *Weekly Epidemiological Record*, 89: 133–139.

WHO and UNICEF (2006). Core questions on drinking-water and sanitation for household surveys. World Health Organization, United Nations Children's Fund, Geneva, Switzerland.

WHO Multicentre Growth Reference Study Group (2006). Assessment of differences in linear growth among populations in the WHO Multicentre Growth Reference Study. *Acta Paediatrica*, 95: 56–65.

Wildlife Conservation WCS, Center for International Earth Science Information Network (CIESIN) (2005). Last of the wild data version 2, 2005 (LTW-2): global human footprint dataset (geographic).

Zenonos, Z. A., Dummler, S. K., Müller-Sienerth, N., Chen, J., Preiser, P. R., Rayner, J. C., and Wright, G. J. (2015). Basigin is a druggable target for host-oriented antimalarial interventions. *The Journal of experimental medicine*, pages jem–20150032.

Ziegelbauer, K., Speich, B., Mäusezahl, D., Bos, R., Keiser, J., and Utzinger, J. (2012). Effect of sanitation on soil-transmitted helminth infection: systematic review and meta-analysis. *PLoS Medicine*, 9: e1001162.