LANCASTER UNIVERSITY

Geostatistical Methods and Applications in Global Health

by

Benjamin Amoah

A thesis submitted for the degree of Doctor of Philosophy

in the

Faculty of Health and Medicine
Lancaster Medical School

September 2018

Declaration of Authorship

I, BENJAMIN AMOAH, declare that this thesis titled, 'GEOSTATISTICAL METH-ODS AND APPLICATIONS IN GLOBAL HEALTH' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

- I have acknowledged all main sources of help.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

"Every one who is seriously involved in the pursuit of science becomes convinced that a spirit is manifest in the laws of the Universe– a spirit vastly superior to that of man, and one in the face of which we with our modest powers must feel humble."

Albert Einstein

Abstract

Sub-saharan Africa shares a high portion of the global disease burden and has attracted the attention of several intervention programmes. Intervention programmes need an in-depth understanding of the spatial and temporal distribution of diseases and the underlying risk factors in order to plan effective control strategies. Geostatistical methods provide a means to map disease outcomes whilst explaining measured and unmeasured underlying risk factors. This thesis, made up of three papers, focuses of developing and applying geostatistical methods to understand the spatial (and temporal) distributions and risk factors of childhood undernutrition, malaria and *Loa loa* in sub-Saharan African countries.

The relationship between the rate of infectious mosquito bites and the prevalence of malaria parasite in human hosts can highlight aspects of malaria epidemiology that are pertinent to malaria control. However, this relationship is poorly understood. In our first paper, we develop geostatistical models to study the spatio-temporal distributions of *Plasmodium falciparum* parasite prevalence and the rate of infectious mosquito bites. We then highlight key aspects of the malaria epidemiology relevant for intervention policies by using mechanistic and empirical statistical models to explore the relationship between infectious bites and parasite prevalence in a rural community in Malawi.

The question of whether or not malaria is associated with growth in children has been studied for years, with different studies reporting contradictory results. However, none of these studies used spatial statistical methods. In the second paper, we develop a geostatistical model to investigate this association using 20 Demographic and Health Survey datasets from 13 sub-Saharan African countries. We then propose novel extensions of the modelling strategy to growth and malaria data collected as a spatial longitudinal study.

Disease prevalence data are often obtained using different diagnostics, but in the absence of spatial statistical methods to jointly analyse such data, most studies report the results of separate analyses on the data from each diagnostic. A joint analysis can explain possible correlations between different diagnostics, which can then be exploited to make more precise and more reliable predictions. In the third paper, we developed a geostatistical framework for combining prevalence data from different diagnostics and apply the novel methodology to map malaria in the highlands of Western Kenya and *Loa loa* in sub-Saharan Africa.

Acknowledgements

Contents

List of Figures

List of Tables

List of Papers

Paper 1. *"On the Relationship Between Plasmodium falciparum Parasite Prevalence and Entomological Inoculation Rate: a Case Study in Rural Malawi"*

Amoah B., McCann R.S., Kabaghe A.N., Paula M., Gowelo S., Mburu M., Chipeta M.G., Tizifa T., van den Berg H., Mzilahowa T., Takken W., van Vugt M., Phiri K.S., Diggle P.J., Terlouw D.J., Giorgi E.

Final draft ready for submission.

Contributions: Lead author, writing of code for statistical analysis, statistical analysis and interpretation of results, writing of first draft.

Paper 2. *"Geostatistical modelling of the association between malaria and child growth in Africa"*

Amoah B., Giorgi E., Hayes D. J., Van Buuren S., Diggle P. J.

Published in *International Journal of Health Geographics*

Contributions: Lead author, writing of code for statistical analysis, retrieving of data, statistical analysis and interpretation of results, writing of the paper.

Paper 3. *"A Geostatistical Framework for Combining Spatially Referenced Disease Prevalence Data from Multiple Diagnostics"*

Amoah B., Giorgi E., Diggle P. J.

Under review in *Biometrics*

Contributions: Lead author, conceptualization, writing of code for statistical analysis, statistical analysis and interpretation of results, writing of first draft.

To Felicia and Christlove

Chapter 1

Introduction

The burden of diseases can vary greatly within defined geographical location, either at the global, continental, national or sub-national levels. Reliable estimates of the extents of disease burden across different geographical regions and over different times, and an understanding of the factors underlying the varying extents, are beneficial to disease control, from the planning of intervention strategies to their evaluation. This thesis focuses on developing geostatistical methods to map malaria, undernutrition and *Loa loa* in sub-Saharan Africa. In the first paper, we develop geostatistical models to map *Plasmodium falciparum* parasite prevalence and entomological inoculation rate. We then model the relationship between the two and use this relationship to highlight key aspects of the malaria epidemiology that are relevant for policy. In the second paper, we develop geostatistical models to map child growth, as measured by height-for-age z-scores in sub-Saharan Africa and investigate the relationship between malaria and child growth in this region. In the third paper, we develop a geostatistical framework to combine disease prevalence data obtained using different diagnostics.

1.1 Fundamentals of model-based geostatistical methods for disease mapping

1.1.1 Geostatistical data

Within the broader context of the analyses of spatially referenced data is a class of statistical models and methods known as *geostatistics*. Geostatistical methods are used to analyse data that consists of a set of unique locations $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^2$ from which data $Y = (y_1, y_2, \ldots, y_n)$ have been sampled correspondingly. Each y_i is a realization of a random variable Y_i, or in the case where there are multiple responses at the same location, a random vector $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^\top$, usually known as the response variable. The distribution of the response variable is dependent on the value at the location x_i of an underlying spatially continuous latent stochastic process $S(x)$. Additionally, we may observe explanatory variables $d(x) = \{d(x_1), d(x_2), \ldots d(x_n)\}$ at the set of locations, where each $d(x_i)$ is a p-dimensional vector of explanatory variables at location x_i. In the context of disease mapping, $d(x_i)$ include remotely sensed data such as land surface temperature, normalized difference vegetation index (NDVI), rainfall, elevation, or other location specific correlates of the disease outcomes such as socio-economic indicators, for example, night-time light, distances to important facilities, road network, urban/rural, population size and composition or other area-level information. One of the most basic geostatistical models is the linear model, which provides a basis for understanding more complex geostatistical models.

1.1.2 The geostatistical linear model

The basics of the linear model are that conditionally on zero-mean mutually independent latent Gaussian variables Z_i and the latent process $S(x_i)$, Y_i are mutually independent Gaussian random variables with mean $\mu(x_i)$ given by

$$\mu(x)_i = d(x_i)^\top \beta + S(x_i) + Z_i, \tag{1.1}$$

where β is a vector of regression parameters corresponding to the explanation variables observed at location x_i. The vector Y then has a multivariate Gaussian distribution with mean $\mu = D\beta$ and covariance matrix Σ, where D is a design matrix of covariates and $\Sigma = \sigma^2 R(\theta) + \tau^2 I_n$, where $[R]_{ij} = \rho(u_{ij}; \theta)$, with the argument u_{ij} of the correlation function ρ being the Euclidean distance between locations x_i and x_j and θ being a vector of parameters.

In this thesis, we model the latent stochastic process $S(x)$ as Gaussian process. Gaussian stochastic processes are commonly used in geostatistical analyses because they are convenient empirical models that can capture a wide-range of spatial variation, rather than their use having any physical justification. However, the compatibility of a specified process with the data for which it is chosen can be checked based on the variogram. We demonstrate this for the models developed in this thesis.

A Gaussian process is completely specified by its mean and covariance functions $E[S(x)]$ and $\mathrm{Cov}(S(x_i), S(x_j))$ respectively, where x_i and x_j are two geographical locations. We assume that $S(x)$ is a zero-mean $(E[S(x)] = 0)$ stationary and isotropic process, that is, the distribution of $S(x)$ remains invariant under rotation and translation. By assuming spatial stationarity, specification of the covariance function reduces to specification of a scaler parameter $\sigma = \mathrm{Var}(S(x))$ and a correlation function $\rho(u_{ij}) = \mathrm{Corr}\{S(x_i), S(x_j)\}$, which depends only on the distance u_{ij} between x_i and x_j.

Any parametric family of functions is a legitimate class of correlation functions of $S(x)$ if and only it is positive definite. However, this condition is not easy to check. Hence, it is useful to work with standard classes of correlation functions known to satisfy the condition and flexible enough to adopt to a wide range of geostatistical or disease mapping problems. Our choice is the family of Matérn (1986) correlation functions given by

$$\rho(u_{ij}) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}\left(\frac{u_{ij}}{\phi}\right)^{\kappa}\mathcal{K}_{\kappa}\left(\frac{u_{ij}}{\phi}\right), \tag{1.2}$$

where $\mathcal{K}_{\kappa}(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$. The parameter κ governs the smoothness of $S(x)$ which is $\lceil \kappa \rceil - 1$ times differentiable, where

$\lceil a \rceil$ is the smallest integer greater than or equal to a. The scale parameter ϕ regulates the rate at which the spatial correlation decays with increasing distance u.

The reason for our choice of the Matérn family of correlation function throughout this thesis is as follows. The integer part of κ determines the mean square differentiability, which affects the smoothness of spatial predictions of the underlying process. Thus, the knowledge of the process can help to inform the choice of κ, which otherwise, is poorly identified from typically datasets. In geostatistics, typical choices of κ values are 0.5, 1.5 and 2.5, which respectively correspond to processes $S(x)$ which are mean-square continuous, once differentiable and twice differentiable, where the smoothness of $S(x)$ increases with κ. For $\kappa = 0.5$, the Matérn correlation function reduces to the exponential, $\rho(u) = \exp(-u/\phi)$, a popular choice for disease mapping applications.

Other choices of correlation functions are the powered exponential, spherical and Cauchy families. The powered exponential family, defined by the correlation function

$$\rho(u) = \exp\{-(u/\phi)^{\kappa}\}, \tag{1.3}$$

like the Matérn family, has a scale parameter $\phi > 0$, a shape parameter κ, where $0 < \kappa \leq 2$. This family generates correlation functions which decrease monotonically with increasing u but are less flexible than the Matérn family.

Gneiting et al. (2001) introduced a three-parameter family of correlation functions know as the Cauchy family, whose members are of the form

$$\rho(u) = \left(1 + (t/\phi)^{\kappa}\right)^{-\beta/\kappa}, \tag{1.4}$$

where $0 < \kappa \leq 2$ and $\phi > 0$ are respectively the smoothness and and scale parameters, and the extra parameter $\beta > 0$ determines the asymptotic power law, $\rho(u) \sim u^{-\beta}$, with which the correlation decays to 0 for increasing distance u.

The spherical family has the correlation function given by

$$
\rho(u) = \begin{cases} 1 - \frac{2}{3}(u/\phi) + \frac{1}{2}(u/3)^3 & : \ 0 \leq u \leq \phi \\ 0 & : \ u \geq \phi, \end{cases} \tag{1.5}
$$

has a single parameter $\phi > 0$ with the dimension of distance and are less flexible than the Matérn and powered exponential and Cauchy families.

The stationarity assumption about $S(x)$ may need relaxation due to non-stationarities in the covariance structure. This could arise because the covariance between two locations x_i and x_j depends on their relative positions, or that the process $S(x)$ may have variable smoothness in the study region, possibly because it depends on underlying environmental variables that are smoother in some subregions than others. In such cases, the assumption of stationary may not hold. One way spatial statistics has dealt with this problem is by mapping the original region to a new space in which stationarity is assumed. Another solution to the problem is to choose from classes of non-stationary covariance functions for Gaussian processes (Higdon et al., 1999, Paciorek and Schervish, 2004)

1.1.3   The generalized linear geostatistical model

In the case of non-Gaussian data, by following the generalized linear modelling framework (McCullagh, 1984, Nelder and Baker, 1972), the geostatistical linear model (1.1) is easily generalized to the case where the data Y_i follow a common distributional family with expected values μ_i. Such a case may arise either because the empirical data deviates substantially from the linear Gaussian assumptions, for example, they show strong skewness, or that we know that the mechanism that generates the data necessitates the use of some other probability distribution. An approach to deal with the problem is to transform the data so that Gaussian assumptions are satisfied. A widely used approach for such transformation is the Box-Cox family of transformations (Box and Cox, 1964)

given by

$$\rho(u) = \begin{cases} (Y^\lambda - 1)/\lambda & : \ \lambda \neq 0 \\ \\ \log Y & : \ \lambda = 0 \end{cases} \tag{1.6}$$

Another approach to specify the mean μ_i by a known link function $h(\cdot)$, where $h(\mu_i) = \eta_i$ is the linear predictor, giving rise to a generalized linear model for geostatistical data (Diggle and Ribeiro, 2007).

In each of the three papers that make up the core of the thesis, we develop models that build on the (generalized) linear geostatistical model. The datasets analysed in this thesis follow the Gaussian, Poisson, Bernoulli or Binomial probability distributions. We apply the log-link function to the Poisson data and the logit-link to the Bernoulli and binomial data.

1.2 Objectives and structure of the thesis

This thesis comprises three sets of research objectives, each of which is studied in a separate chapter.

1.2.1 Objectives of Chapter 2

In Chapter 2 (Paper 1), we analyse three-years spatio-temporal entomological and parasitaemia data from an integrated malaria project carried out in three administrative units, known as focal areas, surrounding the Majete Wildlife Reserve (MWR) perimeter, in the Chikwawa District, Southern Malawi. The specific questions of scientific interest are the following.

I) What are the spatio-temporal distributions of the *plasmodium falciparum* parasite prevalence (*Pf*PR) and *plasmodium falciparum* entomological inoculation rate (*Pf*EIR) in three focal areas?

II) How do the spatio-temporal patterns of *Pf*EIR and *Pf*PR compare?

III) What functional relationship exists between PfPR and PfEIR, and what policy relevant features of the malaria epidemiology in the study region does this relationship highlight?

To answer the first question, we develop spatio-temporal geostatistical models to map PfPR and PfEIR in the study region. We model each as a function of time, location and explanatory variables in a model-based geostatistical framework. We estimate PfEIR as a product of model-based predicted human biting rate, which we estimate as the densities of the various malaria vectors in the study region, and the sporozoite rate, which is the proportion of vectors carrying the sporozoite stages of the parasite. We model the mosquito densities using spatio-temporal geostatistical Poisson log-linear models and the sporozoite rates using generalized linear binomial models. To answer the second question, we identify and compare hot-spots of PfPR to those of PfEIR by mapping the predictive probabilities that PfPR and PfEIR exceed predefined thresholds with high probability. To answer the third question, we model the relationship between PfPR and model-based predicted PfEIR. We use a range of models that can explain the relationship and then choose the model that best explains the data. We then use the parameters of the best model to highlight features of the malaria epidemiology that have relevance for policy.

1.2.2 Objectives of Chapter 3

In Chapter 3 (Paper 2), we investigate the association between the geographical distribution of malaria and linear growth in children under five years old, using 20 Demographic and Health Survey (DHS) datasets from 13 sub-Saharan African countries. This association has been studied for years, yet it is poorly understood since most studies have reported contradictory results. The specific questions we ask in this study are the following.

I) What is the association between malaria and linear growth in children 0-5 years in sub-Saharan Africa?

II) What factors modulate the association between malaria and linear growth in the specified population?

III) What is the spatial distribution of stunted growth based on each DHS?

To answer the first question, we first explore the relationship between malaria exposure and hight-for-age z-score (HAZ) in a univariate analysis, with malaria as the only predictor of HAZ in a linear model. We then modelled HAZ trajectories across different ages for each dataset, using cubic splines and then investigate the variation in the trajectories of HAZ in different malaria exposure levels. Finally, we develop a linear geostatistical model in which malaria is a predictor of HAZ, whilst we adjust for know confounders. We then make inferences on the regression coefficients associated with malaria. To answer the second question, we model the estimated regression coefficients as a function of a number of World Bank African Development Indicators to find out which indicators possibly explain variations in the estimated regression coefficients of malaria for the different DHS. To answer the third question, we map the risk of stunting for each survey by mapping the predictive probability that the average HAZ at each prediction location is less than -2, the threshold that defines stunted growth.

1.2.3 Objectives of Chapter 4

In Chapter 4 (Paper 3), we develop a geostatistical framework that allows the joint analysis of disease prevalence data that have been obtained using different diagnostics. Prevalence data of the same disease are often obtained using different diagnostics for two main reasons: (1) as a means to achieve a required sample size in a cost effective way, and (2) as a means to fully capture the distribution of the disease, since the different diagnostics can provide complementary information about the epidemiology of the disease. Whilst prevalence estimates obtained by different diagnostics are correlated, considerably differences might also occur. Most studies have reported separate analyses for different diagnostics, thus failing to explain and utilise the possible correlations in the data, which can be exploited to benefit statistical inference. The specific questions of scientific interest are the following.

I) How can we predict disease prevalence as defined by a gold-standard diagnostic in geographical regions where data is available only for a low-cost, possibly biased alternative test?

II) How can we more precisely predict disease prevalence as defined by each of several complementary diagnostics by borrowing information across the others?

To answer the first question, we develop a bivariate geostatistical model that exploits the correlation between a gold standard and a low cost diagnostic to help predict prevalence as defined by the gold standard. To answer the second question, we develop a bivariate geostatistical model that exploits the cross-correlation across complementary diagnostics. We apply the model developed for the first objective to a *Loa loa* mapping problem in Central and West Africa where a questionnaire-based diagnostic, RAPLOA, and a microscopy-based parasitology diagnostic were used. We apply the model developed for the second objective to a malaria mapping problem in highlands of Western Kenya, where a rapid diagnostic test and a polymerase chain reaction where used. Finally, we extend the novel methodology to a model that can be used to jointly analyses prevalence data from a gold standard and a number of complementary diagnostics.

In Chapter 5 we conclude by discussing other possible applications and the methodological extensions of the models developed in the preceding chapters.

References

Box, G. E. and Cox, D. R. (1964). "An analysis of transformations." *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

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

Gneiting, T., Sasvári, Z., and Schlather, M. (2001). "Analogies and correspondences between variograms and covariance functions." *Advances in Applied Probability*, 33(3):617–630.

Higdon, D., Swall, J., and Kern, J. (1999). "Non-stationary spatial modeling." *Bayesian statistics*, 6(1):761–768.

Matérn, B. (1986). Spatial variation, volume 36 of lecture notes in statistics.

McCullagh, P. (1984). "Generalized linear models." *European Journal of Operational Research*, 16(3):285–292.

Nelder, J. A. and Baker, R. J. (1972). "Generalized linear models." *Encyclopedia of statistical sciences*.

Paciorek, C. J. and Schervish, M. J. (2004). "Nonstationary covariance functions for Gaussian process regression." In *Advances in neural information processing systems*, pages 273–280.

Chapter 2

Paper 1. On the Relationship Between *Plasmodium falciparum* Parasite Prevalence and Entomological Inoculation Rate: a Case Study in Rural Malawi

Benjamin Amoah [1] [§], Robert S. McCann[2, 3], Alinune N. Kabaghe[3,4], Moraga Paula[1] Steven Gowelo[2, 3], Monicah Mburu[2, 3], Michael G. Chipeta[3, 5, 6], Tinashe Tizifa[3, 4], Henk van den Berg[2] Themba Mzilahowa[3] Willem Takken[2] Michéle van Vugt[4] Kamija S. Phiri[3] Peter J. Diggle[1] Dianne J. Terlouw[3, 5, 6] Emanuele Giorgi[1]

1 Centre for Health Informatics, Computing, and Statistics (CHICAS), Lancaster Medical School, Lancaster University, Lancaster, United Kingdom

2 Entomology Lab, Wageningen University and Research, Wageningen, The Netherlands

3 College of Medicine, University of Malawi, Blantyre, Malawi

4 Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

5 Liverpool School of Tropical Medicine, Liverpool, United Kingdom

6 Malawi-Liverpool Wellcome Trust, Blantyre, Malawi

§b.amoah@lancaster.ac.uk

2.1 Summary

Background

Primarily, infection with a malaria parasite occurs through the bite of an infective mosquito; thus, an unavoidable relationship between the rate of infectious bites, directly measured as the entomological inoculation rate (EIR), and malaria prevalence (PR). This relationship can highlight key features of the malaria epidemiology to inform the effective planning of control strategies; however, it has been scarcely studied, limiting insightful discussions on its nature and implications. We study the *Plasmodium falciparum* EIR–PR relationship and present insights into its possible implications for malaria control.

Methods and findings

Using spatio-temporal data collected over a period of 38 months from rural Malawi, we mapped the *P. falciparum* EIR and PR in children 0.5-5 y/o and in women 15-49 y/o to identify areas of elevated malaria risk and to compare the spatio-temporal patterns of EIR and PR. We then quantified the EIR–PR relationship using a number of statistical models. The predictive maps of EIR and PR showed similar spatio-temporal heterogeneities and seasonal patterns. Hot-spots of EIR and PR mainly overlapped, but with regions that, at certain times, were hot-spots of EIR but not of PR and vice versa. We found that EIR had a one-month delayed effect on PR. Increasing EIR was associated with an initial rapid rise in PR, followed by saturation. We found substantial levels of PR to be associated with even very low EIR. We observed residual hot spots of PR when EIR remained at zero levels. We observed a rise in PR at times when EIR was undetectable. We found a consistently higher PR and children as compared to women.

Conclusions

Our results highlight the following aspects of the EIR-PR relationship: (1) When EIR has been reduced to marginal levels, substantial average levels and residual hot spots of PR are to be expected. (2) At low EIR levels, an increase in EIR could manifest as a rapid rise in PR (3) At high EIR levels, reductions in EIR may not translate into

remarkable reduction in PR. (4) At undetectable levels of EIR, PR levels can still increase (5) Children may play a more active role in the sustenance of malaria transmission than adults. We conclude by emphasising the need for an integrated malaria control, viz, vector and human host management strategies, and the need for the monitoring of both entomological and parasitaemia indices in the fight against malaria.

Keywords: *Plasmodium falciparum*, entomological inoculation rate, malaria prevalence, spatio-temporal, relationship, geostatistics, disease mapping.

Author summary

Why was this study done?

> The relationship between the rate of infectious mosquito bites, the most direct measure of malaria transmission, and malaria prevalence can inform the effective planning of control strategies, but it is poorly understood.

> The study was done to quantify the relationship between the rate at which individuals are bitten by mosquitoes infected with *P. falciparum* and the proportion of the population carrying this parasite.

What did the researchers do and find?

> To identify areas of elevated malaria risk and to compare the spatial and temporal patterns of the rate of infectious bites to that of malaria prevalence, we mapped the rate at which individuals are bitten by mosquitoes carrying *P. falciparum* and the proportions of children 0.5-5 y/o and of women 15-49 y/o infected with the parasite.

> To understand the quantitative relationship between the rate of infectious bites and prevalence, we used a number of statistical models under different assumptions to estimate the relationship.

> We observed a higher malaria prevalence in children as compared to women. Even the lowest malaria transmission levels were associated with substantial malaria prevalence and hot spots of malaria prevalence.

> The relationship we found is such that when malaria transmission is low, an increase in transmission results in a quick rise in prevalence. This is then followed by a slowing rise in prevalence as transmission levels continue to increase.

> Malaria prevalence increased even when malaria transmission could not be detected.

What do these findings mean?

> Vector control strategies are the sole malaria intervention techniques used in some countries. Our study indicates that when transmission has been interrupted through vector control, a large reservoir human hosts could be left, who could refuel transmission whenever conditions permit vector-host contacts. This suggests that if not coupled with detection and treatment of both clinical cases and asymptomatic infections, vector control alone may not be enough to drive malaria to elimination levels. Our study further indicates that the reservoir human host are likely to be concentrated at residual hot spots.

> Our results highlight possible difficulties in eliminating malaria in low transmission settings: a little rise in transmission could translate into an escalation in prevalence, making elimination difficult. We suggest that when malaria transmission has been reduced to the lowest levels, both entomological and parasitaemia indices need to be closely monitored in order to quickly detect and intervene in any outburst in prevalence.

> Our study demonstrates that meaningful public health impact on malaria prevalence may not be immediately achieved through vector control programmes in high transmission settings. However, when the control is persistent, notable impact would eventually be achieved.

> Our results exemplify that when the impact of vector control is monitored through entomological surveillance, if infectious bites can no more be detected, vector control may need to continue for some time since there could still be enough transmission to facilitate a rise in prevalence.

2.2 Introduction

Collaborative efforts of malaria control programmes and all stakeholders led to extensive intervention coverage over the last two decades, leading to large reductions in malaria morbidity and mortality (Bhatt et al., 2015). However, malaria is still a leading global health problem. The previous successes and current challenges have motivated ambitious, although achievable, global and national targets towards achieving malaria elimination. Control strategies to achieve the set targets have been outlined (Patouillard et al., 2017, World Health Organization, 2015), but monitoring epidemiological measures of malaria risk is critical for continued assessment of the effectiveness of control, so as to identify areas where there is the need to re-strategise or increase efforts to be able to push towards elimination.

Three malaria risk measures commonly monitored are the clinical burden, prevalence (PR) and intensity of transmission. Monitoring the clinical burden involves accounting for the number of (symptomatic) clinical cases and deaths through health system records, which are however often incomplete since many sickness are never reported at health facilities, and thus biased. Another means to monitor clinical burden is to follow cohorts, which is expensive. Prevalence is monitored through (repeated/rolling) cross-sectional surveys carried out in communities or at health centres. Monitoring transmission intensity is carried out by measuring the rate of new infections in cohorts or through the entomological inoculation rate (EIR), which is the rate of infectious bites per person per unit time. Alternative serological markers of malaria exposure may also be used, but since these depend on malaria antibodies which can persist in the blood for years after exposure has ceased, they are quite insensitive to changes in transmission (Bousema et al., 2010, Cook et al., 2011, Corran et al., 2007, Drakeley et al., 2005).

EIR is estimated as the product of the average number of mosquito bites per person per unit time, also known as the human biting rate (HBR), and the proportion of mosquitoes carrying the sporozoite stages of malaria parasites, referred to as sporozoite rate (SR). Estimating the HBR is challenging since vectors may not be found when and where they are being looked for and thus mosquito samples may neither represent mosquito populations nor their biting activities. Moreover, SR is less precise when mosquito density is low and can be affected by the average age of adult mosquitoes (Tusting et al., 2014). These limitations make estimating the EIR a difficult problem, especially in low transmission settings. For example, EIR could be zero in places where transmission could still be taking place according to some other indicator, say increasing PR.

Despite the challenges in estimating the EIR, it is an important malaria risk measure, since it can provide information about transmission that other risk measures cannot, for example, the contribution and effectiveness of different vector species in transmitting different parasites species. Moreover, previous studies into its functional relationship with PR (Beier et al., 1999, Smith et al., 2005) and with clinical incidence (Beier et al., 1994) have shown that EIR is the most direct measure of the intensity of malaria transmission. Understanding the functional relationship between EIR and PR is crucial for effective planing, monitoring, and evaluation of malaria control since this can help to better understand policy-relevant aspects of the malaria epidemiology. For example, the levels of reduction in transmission needed to achieve meaningful reductions in PR can be estimated through this relationship. Furthermore, progress made at reducing malaria prevalence through vector control can be tracked and evaluated through this relationship.

In spite of its usefulness, much uncertainty still surrounds the EIR–PR relationship. For example, the following questions have not yet been investigated: 1. Each of EIR and PR has been found to vary in space and time, but how do spatio-temporal patterns of EIR compare to PR? 2. Will geographical regions or times of elevated EIR necessarily correspond to higher PR? 3. Do EIR and PR lead to the identification of the same hot-spots? 4. Does EIR have a lagged effect on PR? 5. Does the EIR-PR relationship

vary across demographic groups which experience a different exposure to malaria, e.g. children between 6 and 60 months and women of reproductive age?

In this paper, we investigate and provide answers to these questions and discuss their implications for malaria control. The data analysis is divided into two main parts. In the first part, we map on a high spatio-temporal resolution the *P. falciparum* entomological inoculation rate (PfEIR), the rate at which individuals are bitten by mosquitoes infected with *P. falciparum*, and *P. falciparum* prevalence (PfPR), the proportion of individuals carrying the parasite in their blood. By this, we pursue two primary objectives: 1. To identify areas of elevated risks, which can be considered for a more targeted intervention 2. To compare spatial heterogeneities, trends, seasonal patterns and hotspots of PfEIR and PfPR. In the second part, we consider several statistical models for the relationship between PfEIR and PfPR that can be distinguished as follows: mechanistic models that are based on different epidemiological assumptions to find the most plausible description(s) of the nature of the PfEIR–PfPR relationship; empirical models where the PfEIR–PfPR relationship is purely informed by the data. Finally, we discuss the possible implications of the estimated relationships.

2.3   Materials and methods

2.3.1   Study area

Surrounding the Majete Wildlife Reserve (MWR) in the Lower Shire Valley in the Chikwawa District in Southern Malawi are 19 community based organizations (CBOs), home to a population of approximately 100,000. In this area, the Majete Malaria Project (MMP) was carried out in order to select three distinct geographical regions, referred to as Focal Areas (FAs) A, B and C (see Figure 2.1), where interventions aimed at reducing the malaria burden have been implemented. The catchment area of the FAs comprises 65 villages with about 6600 households and a population of about 25,000.

Chikwawa experiences highly variable rainfall during its single raining season, which spans November/December to April/May. Temperatures are generally high, with an

average annual maximum temperature of 37.6°C, occurring in December and a minimum of 27.6 °C, occurring in July every year (Joshua et al., 2016). During the raining season, the Shire and Mwanza rivers, which run by MWR, create marshy habitats, paddies, occasional depressions and watering holes, suitable as breeding sites for *A. arabiensis s.s.*, *A. funestus s.s.*, *An. gambiae s.s.* and *A. quadriannulatus* (Spiers et al., 2002).
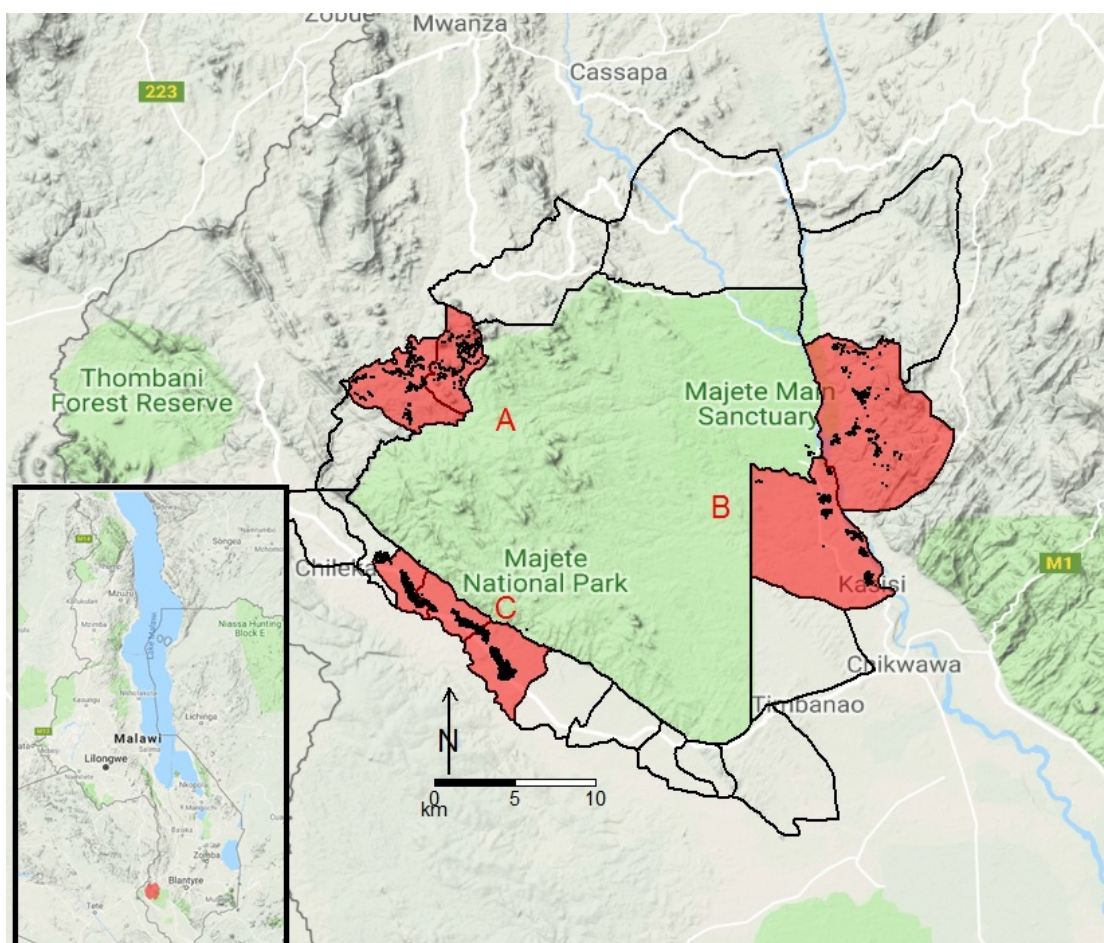
FIGURE 2.1: Map of Malawi (insert) highlighting the Majete Wildlife Reserve and borders of the 19 community-based organisations (CBOs), surrounding the Majete perimeter. Three focal areas (red patches), labelled as A, B, and C, show the communities/households (black points) selected for the parasitaemia and entomological surveys by the Majate Malaria Project (MMP).

2.3.2 Data

As part of the MMP, randomised trials of community-level malaria interventions were rolled out. To quantity their effectiveness, a rolling malaria indicator survey (rMIS) (Roca-Feltrer et al., 2012b) was conducted in conjunction with entomological surveillance. The study protocol for the trial can be found elsewhere (McCann et al., 2017). A

baseline survey was conducted between April 2015 and March 2016, and interventions, including house improvement and larval sources management, were carried out between April 2016 and May 2018.

At each round of the rMIS data collection during the baseline period, an adoptive geo-statistical sampling design (Chipeta et al., 2016), was used for random selection of households. This sampling design allowed locations to be sampled at each surveillance round to depend on the uncertainty in PR estimates from the previous rounds, so as to effectively identify hot-spots with minimal sampling intensity. During the trial period, an inhibitory geostatistical design (Chipeta et al., 2017) was instead used. This design was to ensured that any pair of sampled locations are separated by at least an inhibitory distance of 0.123 km to achieve some uniformity in the sampled locations over a study region (Chipeta et al., 2017). At each round of rMIS data collection in the base-line and trial phases, 75% of the sampled households chosen for the rMIS sampling were then randomly selected for the entomological surveillance.

In each sampled household, children under five (0.5-5 y/o) and women of reproductive age (15-59 y/o) were tested for *P falciparum* using Rapid Diagnostic Test (RDT SD BIOLINE Malaria Ag P.f. HRP-II, Standard Diagnostics, Yongin-si, Republic of Korea).

Mosquitoes were collected using suna traps (Hiscox et al., 2014) with an Mbita (MB5) and Carbon dioxide (CO2) blend odour-bait (Mukabana et al., 2012, Pombi et al., 2014). Traps were hanged 2m above the floor and set in operation from 6pm to 6am. For a selected household in a surveillance round, the trap was fitted either outside the house (outdoor) or in a room (indoor) on the first day and the second day, respectively. Trapped mosquitoes were killed, transported to the lab and sexed using taxonomic keys (Becker et al., 2003, Reinert et al., 2000). Female anophelines were further classified and tested for the presence of *P. faciparum* sporozoites using polymerase chain reaction (PCR).

2.3.3   Environmental and climatic factors

Environmental and climatic factors affect the abundance and suitability of breeding sites that support larvae development (Loetti et al., 2011, Madder et al., 1983), the duration of the immature stages of mosquitoes (Ciota et al., 2014, Craig et al., 1999, Loetti et al., 2011), their host seeking and biting behaviour and, finally, the incubation period and replication of the parasites in mosquitoes (Amek et al., 2011, Rumisha et al., 2014).

Using hourly measurements of temperature and relative humidity (RH) from a weather station in each of the FAs, we computed the average temperature and RH for different range of days prior to the day of data collection. Details of the days over which we computed average temperature and RH are given in Table 2.2 in S1 Appendix.

Spectral indices, namely normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) were computed using remotely sensed multi-spectral imagery from the Landsat Program. These data are freely available from the United State Geological Survey (USGS) Earth Explorer (earthexplorer.usgs.gov) as raster files at a spatial resolution of 30×30m for every 16 days. For our analysis, we averaged each spectral index over five years, from April 2013 to April 2018, while omitting scenes that were dominated by clouds artefacts from the average.

We also extracted raster data of surface elevation and the ASTER 2011 global digital elevation model (DEM) generated using measurements from the advanced spaceborn thermal emission and reflection radiometer. A detailed description of these data has been given elsewhere (Tachikawa et al., 2011) and are freely available for download from the USGS Earth Explorer. Using a flow accumulation map derived from the DEM, a river network map was generated, and distance to small rivers and large rivers (henceforth, DSR and DLR, respectively) were computed and stored as rasters. Details of the method used to derive the accumulated flow from the DEM can be found elsewhere (Tarboton et al., 1991), and an analytic method for determining an appropriate threshold value for river network delineation can also be found elsewhere (Jenson and Domingue, 1988).

2.3.4 Geostatistical Analysis

In order to map PfEIR, we first fit a generalized linear mixed model to each of HBR and *P. falciparum* sporozoite rate (PfSR) and then compute PfEIR as a product of the two quantities. We carry out separate analyses for *A. arabiansis s.s.* and *A. funestus s.s.*, using covariates and random effects structures that we find to be more suitable in each case.

Details of the steps we take to build the HBR, PfSR and PfPR models are given in S1 Appendix. The covariates for different models and mosquito species, although always written as $d(x_i, t_i)$ in each of the geostatistical models following, are not necessarily the same. A Detailed description of the final sets of covariates for *A. arabiansis s.s.* HBR, *A. funestus s.s.* HBR, PfSR and PfPR models later described in this section are respectively given in Tables 2.3, 2.4, 2.5, 2.6 in S1 Appendix. The geostatistical models for the HBR and PfPR data described below are fitted using `PrevMap` (Giorgi and Diggle, 2016), freely available from the Comprehensive R Archive Network (www.r-project.org).

Human biting rate: We assume that the number of mosquitoes trapped by Suna traps estimates HBR as these primarily target host-seeking mosquitoes. Let $Y(x_i, t_i)$ denote counts of mosquitoes trapped at location x_i time $t_i \in \{1, \ldots, 38\}$ (where $t_i = 1$ denotes April 2015). We then model the $Y(x_i, t_i)$ using Poisson mixed models expressed by the following linear predictor

$$\log\{HBR(x_i, t_i)\} = d(x_i, t_i)^\top \beta + f(t_i; \alpha) + S(x_i) + Z_i, \tag{2.1}$$

where: $d(x_i, t_i)$ is a vector of spatio-temporal covariates with associated regression coefficients β; $f(t_i, \alpha)$ linear combination of sine and cosine functions with a period of 12 months and linear terms, used to capture seasonal variation and trends; the Z_i are independent and identically distributed Gaussian random variables with variance τ^2; the $S(x)$ are spatial random effects that we model as a zero-mean stationary and isotropic Gaussian process with variance σ^2 and the exponential correlation function $\rho(u) = \exp(-u/\phi)$, where ϕ regulates the pace at which the spatial correlation decays for increasing distance u between any two locations.

Plasmodium falciparum sporozoite rate Let $Y^*(x_i, t_i)$ be the number of mosquitoes that tested positive for the presence of *P. falciparum* sporozoites. We then assume that the $Y^*(x_i, t_i)$ follow a Binomial mixed model with number of trials $N^*(x_i, t_i)$, i.e. the total number of captured mosquitoes, and probability of testing positive $PfSR(x_i, t_i)$. We model the latter as a logit-linear regression given by

$$\log\left(\frac{PfSR(x_i, t_i)}{1 - PfSR(x_i, t_i)}\right) = d(x_i, t_i)^\top \beta^* + f^*(t_i; \alpha^*) + Z_i^*, \tag{2.2}$$

where each term in the equation above has an analogous interpretation to those of equation (2.1).

Estimating the Plasmodium falciparum entomological inoculation rate: Let $PfEIR_f(x, t)$ and $PfEIR_a(x, t)$ denote the PfEIR for *A. funestus* and *A. arabiensis* at a given location x and month t. We estimate each of these two as

$$PfEIR_f(x, t) = HBR_f(x, t)PfSR_f(x, t)l(t)$$
$$PfEIR_a(x, t) = HBR_a(x, t)PfSR_a(x, t)l(t)$$

where $l(t)$ is the number of days in month t. Finally we estimate the overall PfEIR as

$$PfEIR(x, t) = PfEIR_f(x, t) + PfEIR_a(x, t). \tag{2.3}$$

We map the estimates for PfEIR as in equation (2.3) above over a 30 by 30 m regular grid covering the whole of the study area for each of $t = 1, \ldots, 36$.

Plasmodium falciparum prevalence: We map PfPR in women and in children by fitting geostatistical models for the two groups. More specifically, let $I(x_i, t_i)$ denote the number of RDT positives out of N_i sampled individuals at a location x_i and month t_i. We then assume that the $I(x_i, t_i)$ follow a Binomial mixed models with probability of a positive RDT $PfPR(x_i, t_i)$, such that

$$\log\left\{\frac{PfPR(x_i, t_i))}{1 - PfPR(x_i, t_i)}\right\} = d(x_i, t_i)^\top \varphi + g(t_i, \varrho) + T(x_i) + U_i, \tag{2.4}$$

where $T(x_i)$ is a stationary and isotropic Gaussian process with exponential correlation function and U_i is Gaussian noise, $g(t_i, \varrho)$ is a function of time accounting for trends and seasonality, and φ and ϱ are regression parameters.

Hot-spots detection using PfEIR and PfPR: To identify hot-spots of PfEIR and PfPR, we map the respective predictive probabilities that PfEIR and PfPR exceeded predetermined threshold values. For PfEIR, we choose the threshold of 0.1 ib/person-/month. For PfPR, we take a threshold of 30% for children and 16% for women. The PfPR thresholds are taken to correspond to the PfEIR threshold based on the best of functional relationships between PfEIR and PfPR as described in the next section.

2.3.5 Modelling the relationship between PfEIR and PfPR

In this section, we describe statistical methods we use to model the relationship between PfEIR and PfPR. Since PfEIR may have a delayed effect on PfPR due to the time taken for *P. falciparum* to develop in the human host, we consider that current PfPR depends on PfEIR some l months prior. In particular, we consider $l = 0, 1, 2$. Hence, we assume that the number of RDT positive individuals, $I(x_i, t_i)$, follow independent Binomial distributions such that

$$PfPR(x_i, t_i) = h\{Pf\hat{E}IR(x_i, t_i - l)\}. \tag{2.5}$$

where $h(\cdot)$ is a function with image $[0, 1]$ and governs the relationship between PfPR and PfEIR, and $Pf\hat{E}IR(x_i, t_i - l)$ is the estimated PfEIR in (2.3). Each of the six models considered provides a different specification for $h(\cdot)$. To account for the uncertainty in $Pf\hat{E}IR(x_i, t_i - l)$, we first obtain predictive samples for $PfEIR(x, t - l)$ using the model in (2.3) and, for each of these, we then fit the model given by (2.5).

Finally, we point out that, in this section, while models from 1 to 4 make explicit assumptions on the underlying mechanism of transmission, models 5 and 6 do not, but describe the functional relationship between PfEIR and PfPR through regression methods.

Model 1: The SIS.

Let b be the probability that an infectious mosquito bite results in an infection, that is, the transmission efficiency. Then infections at $(x_i, t_i - l)$ occur at a rate $bPfEIR(x_i, t_i - l)$. We assume that each infection clears independently, over a duration $1/r$. Then the ratio $\gamma = b/r$ represents the time taken to clear infection per infectious bite. We assume that infections clear independently and the relationship between PfEIR and PfPR is the same throughout the study region. If $PfEIR(x, t - l)$ is constant, then the relationship between $PfEIR(x, t - l)$ and $PfPR(x, t)$ can be described by the Ross Model (Ross, 1911)

$$\frac{\partial PfPR(x,t)}{\partial t} = bPfEIR(x, t - l)(1 - PfPR(x,t)) - rPfPR(x,t). \quad (2.6)$$

We obtain our first model as the non-zero equilibrium solution of (2.6), given by

$$PfPR(x,t) = \frac{\gamma PfEIR(x, t - l)}{\gamma PfEIR(x, t - l) + 1}. \quad (2.7)$$

Model 2: the SIS model with different infection/recovery rates (D.I/R). Model 1 assumes that women and children get infected and recover at the same rate. However, the transmission and recovery rates in children may differ from those in women. We modify Model 1 by allowing different b and r for each category. Let $\xi_{1,it}$, $\xi_{2,it}$ respectively be the proportion of children and women sampled at (x_i, t_i) and $\gamma_k = b_k/r_k$, where $k = 1$ denotes children and $k = 2$ denotes women. Inculcating this heterogeneity in infection/recovery in Model 1 gives Model 2, as

$$PfPR(x,t) = \sum_{k=1}^{2} \xi_{k,it} \frac{\gamma_k PfEIR(x, t - l)}{\gamma_k PfEIR(x, t - l) + 1}. \quad (2.8)$$

Model 3: the SIS model with super infection (S.I.). If individuals are super-infected with *P. falciparum*, then the rate at which infections clear depends on the infection rate, with clearance being faster when infection rate is low, and slow when infection rate is high. We then model infection clearance rate as $g(\vartheta, r) = \vartheta/(e^{\vartheta/r} - 1)$, where $\vartheta = b \times PfEIR$ (Aron and May, 1982, Dletz et al., 1974, Smith et al., 2005,

Walton, 1947). The resulting $PfPR(x,t)$ is then

$$PfPR(x,t) = 1 - e^{-\gamma PfEIR(x,t-l)} \tag{2.9}$$

Model 4: the SIS model with S.I and D.I/R. We modify Model 1 by inculcating both the assumptions of heterogeneous infection/recovery rates, as in Model 2 and super infection, as in Model 3. The resulting model is

$$PfPR(x,t) = \sum_{k=1}^{2} \xi_{k,it}\big(1 - e^{-\gamma_k PfEIR(x,t-l)}\big). \tag{2.10}$$

Diagrammatic representation of Models 1 to 4 are given in Figure 2.2.
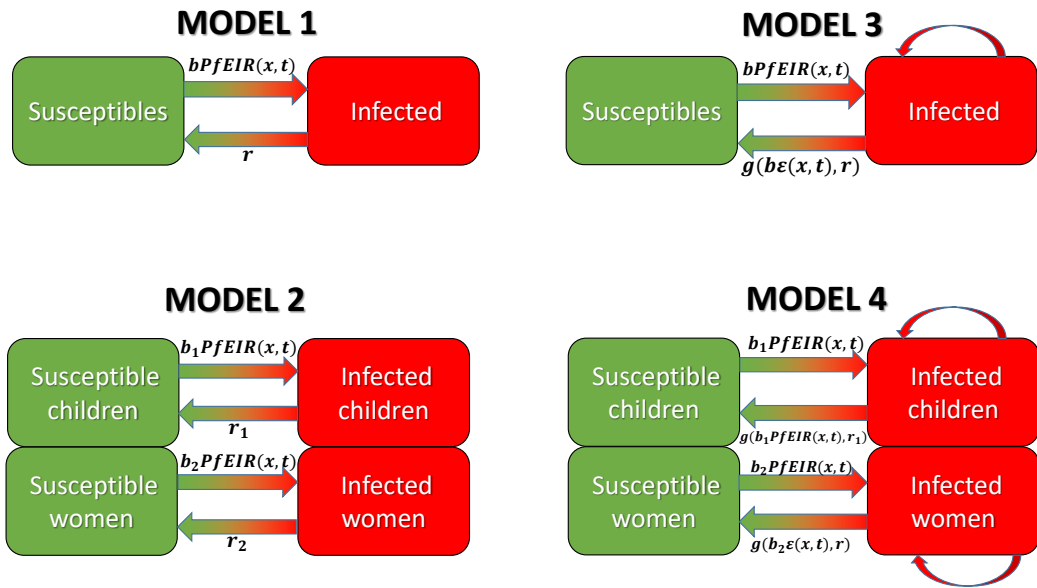
FIGURE 2.2: Diagrammatic representation of the mechanistic models for the relationship between PfEIR and PfPR.

We obtain the parameters of each SIS model by minimizing the log-likelihood function

$$\sum_{t_i}\sum_{x_i} I(x_i,t_i)\log(PfPR(x_i,t_i)) + (N_{it} - I(x_i,t_i))\log(1 - PfPR(x_i,t_i)). \tag{2.11}$$

Model 5: The log-linear model. Beier et al. (1999) assumed that the log of PfEIR is linearly related to PfPR, and fitted the regression model analogous to

$$PfPR(x,t) = a + b\log(PfEIR(x, t - l)), \tag{2.12}$$

the so called "log-linear model".

Model 6: The logit-linear model. The log-linear model has the limitation that PfPR approaches $-\infty$ as PfEIR goes to 0 and to ∞ as PfPR goes approaches ∞, something that is not interpretable epidemiologically. Instead, if PfEIR goes to 0 then sooner or later, we would expect PfPR to go to 0 as well, and if PfEIR goes to ∞, then we will expect PfPR to approach 1. To capture this behaviour, we apply the logit-link function to PfPR to give the model

$$\log\left(\frac{PfPR(x,t)}{1 - PfPR(x,t)}\right) = a + b\log(PfEIR(x, t - l)). \tag{2.13}$$

To fit each of the six models, we obtain 10,000 bootstrapped data sets of predicted PfEIR as in (2.3) at the set of all space-time locations sampled for the rMIS. We do this for two reasons: (1) To obtain EIR data at locations (x_i, t_i) that were sampled for rMIS but not for the entomological surveillance. (2) To account for the uncertainty in EIR. By fitting each model to each of the 10,000 datasets, we obtain 10,000 boostrapped samples $\{\hat{\theta}_1, \ldots, \hat{\theta}_{10000}\}$ for the vector of parameters θ of each candidate model and summarize the parameter estimates by the mean and 95% confidence interval. We do this process for $l = 0, 1, 2$.

2.4   Results

2.4.1   rMIS and mosquito sampling

From April 2015 to May 2018 a total of 5685 individuals were tested, of which 19.0% resulted positive. Among the 2401 tested children between 6 and 60 months, 25.5% were positive, while 14.3% of the 3284 tested women were positive.

A total of 6870 traps were placed in 2425 households resulting in the collection of 623 female Anopheles mosquitoes. Table 2.1 gives details of the number of mosquitoes, sporozoite rates and PfEIR by species. We note that despite the relatively low abundance of *A. funestus s.s.*, it has a higher sporozoite rate, making its PfEIR almost equivalent to that of *A. arabiensis s.s.* (See Table 2.1). The total PfEIR for the 38 months' period is 7.57 ib/person, equivalent to 2.39 ib/person/year.

TABLE 2.1: Details of Anopheles female mosquitoes collected. The table shows the observed number, HBR, PfSE and PfPR for the anopheles species sampled.

Species	Number Collected	Emp. HBR	Emp. PfSR	Emp. PfEIR
A. arabiensis s.s.	438	73.64	5.48%	4.04
A. funestus s.s.	170	28.58	11.17%	3.19
A. gambiae s.s.	11	1.85	18.18%	0.34
A. quadriannulatus	4	0.67	0.00%	0.00
TOTAL	623	104.74		7.57

For estimates of the parameters of the HBR model fitted to *A. arabiensis*, and *A. funestus*, and those of the PfSR and PfPR, see Tables 2.7, 2.8, 2.9 and 2.10 in S1 Appendix respectively.

2.4.2 Hot-spots detection using PfEIR and PfPR

An objective of this study was to map PfEIR and PfPR to compare their spatio-temporal patterns. In S2 Appendix, we show spatio-temporal maps of PfEIR and PfPR in children and in women on a fine spatial resolution of 30×30m and a temporal resolution of 1 month, from April 2015 to March 2018. The maps of PfEIR show spatio-temporal heterogeneities (PfEIR ranging between 0-5ib/person/month, corresponding to cyan to red areas), showing areas of high risk until August 2017, after which PfEIR appears homogeneous and remains at zero-levels everywhere in the study region.

The PfPR maps for both children and women, show high spatio-temporal heterogeneities, especially in Focal Area B. The maps of the two groups show similar spatio-temporal patterns but PfPR in children is generally higher than PfPR in women throughout the study region over the whole period.

Focal Area A generally shows the lowest PfEIR and PfPR. Focal Area C shows higher levels of PfEIR but similar levels of PfPR as does Focal Area B. We note that, although PfEIR remains at zero levels throughout the study region from August 2017 to March 2018, there are substantial levels of PfPR in all Focal Areas.

We summarise the predicted PfEIR maps by the median PfEIR and the 95% confidence intervals over the whole study region for each month, as shown in Figure 2.3 (a). Similarly, we summarise the predicted PfPR maps by the mean over the study region and the 95% confidence intervals for children in Figure 2.3 (b) and for women in Figure 2.3 (c). The predicted PfEIRs and PfPRs (shown as triangular points) are similar to the observed (shown as round points). Children show a higher level of PfPR consistently throughout the study period. Each of PfPR and PfEIR exhibits seasonal patterns with single peaks, which are almost concurrent. PfEIR increases from November to a peak in May and decreases to a trough in November. PfPR appears to start increasing from December to a peak around July, after which it starts to decrease down to a trough between November and December. An interesting observation about Figure 2.3 is that PfEIR is zero after August 2017; however, substantial levels of PfPR persist throughout the same period. The single most striking feature of Figure 2.3 is that, PfPR even increases in both children and women between November 2017 and April 2018, despite PfEIR being zero throughout this period.

This study aimed to compare hot spots of PfEIR and PfPR. We mapped hot spots of PfEIR and PfPR by mapping the probabilities that PfEIR and PfPR exceed defined thresholds with high (0.9) probability. For PfEIR, we chose a threshold of 0.1 ib/person-/month, which corresponds to PfPR thresholds of 30% and 16% respectively in children and women, based on the best fitted relationship between PfEIR and PfPR (given in the next section). S3 Appendix shows spatio-temporal maps comparing the hot-spots of PfPR and PfEIR. We note that the hot-spots of PfEIR and PfPR mainly overlap. However, there are hot-spots of PfEIR that are not necessarily hot-spots of PfPR, and vice versa. We also note that after August 2017, hot spots of PfEIR disappear, but several residual hotspots of PfPR remain in the study region. We note that the hot-spots
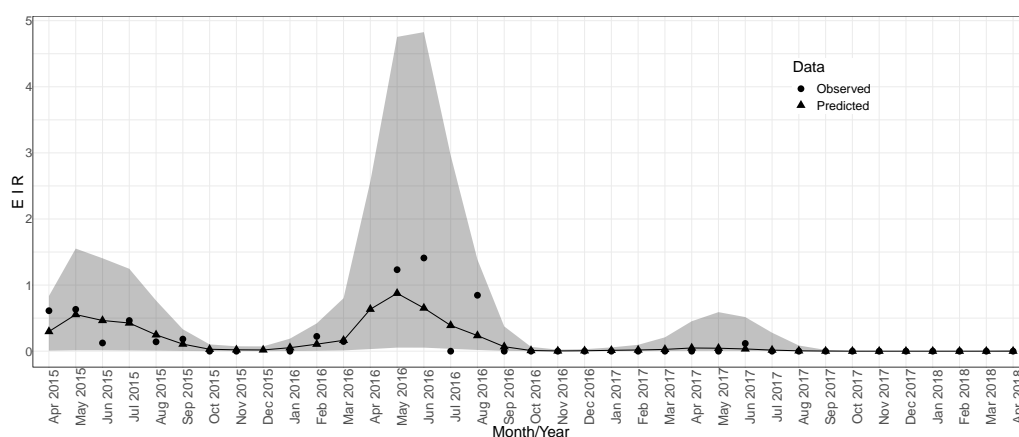
(a)
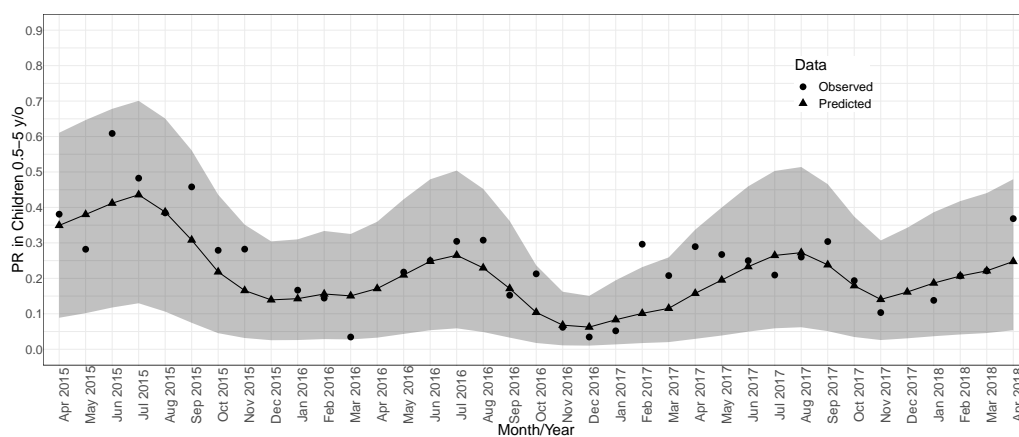
(b)
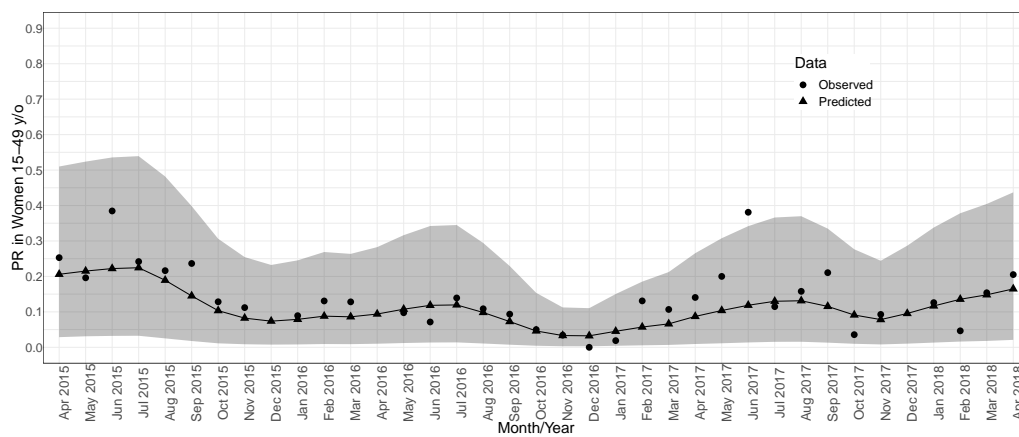
(c)

FIGURE 2.3: Summaries of monthly PfEIR and PfPR. The plot shows monthly median PfEIR (a), mean PfPR in children 0.5-5 y/o (b) and mean PfPR in women 15-49 y/o (c), over the study region. The round points are the observed data and the triagular points are the predictictions from our models. The shaded regions represent the coresponding 95% confidence interval of the predicted values.

of PfEIR and PfPR are mostly unstable, except a few hot-spots of PfPR that persist throughout the study period.

2.4.3 The relationship between PfEIR and PfPR

We investigated if PfEIR has a lagged effect on PfPR. We therefore considered a lag of 1 and 2 months. For each of the six models, the model with the lag effect of 1 month was better (AIC difference > 9) than the corresponding models with no lag effect and a two-month a lagged effect. Table 2.11 in S1 Appendix shows the results of fitting the six models relating PfPR to the PfEIR one month prior.

The logit-linear and log-linear models showed the least AIC as compared to the four SIS models. However, we choose the logit-linear model as the overall best model since it shows an AIC 8 lower than that of the log-linear model (see Table 2.11 in S1 Appendix), and has an asymptotic behaviour more consistent with the epidemiology of malaria. Figure 2.17 in S4 Appendix shows the linear relationship between the logit of PfPR and the log of PfEIR on a scatter plot for children and women combined. Figure 2.4 shows a plot of the fitted logit-linear model with its 95% confidence region for children and women combined. At very low PfEIR, PfPR rises quickly with increasing PfEIR, followed by a flattening off or saturation. Of all 10,000 fitted logit-linear models, PfEIR was a significant predictor of PfPR in 94.7% of the cases. However, the R^2 values for the 10,000 fitted models were low, ranging between 0.13 and 0.39, with an average of 0.27. From the estimated relationship for women and children combined, a decrease in PfEIR from 1 ib/person/month to a very low PfEIR of 0.0001 ib/person/month is associated with a reduction in PfPR from 28.7% to 12.6% (i.e., a 51.3% decrease in PfPR) on average. We note that even when transmission has been driven to an almost zero level, there is still a high PfPR, despite the notable decrease.

An objective of this paper was to investigate if the relationship between PfEIR and PfPR differ between women and children. An indication of differences in the PfEIR–PfPR relationship between children and women lies in the logit-linear model fitted to children and women separately. See the relationships shown in Figure 2.4(b). The average trajectory of PfPR and corresponding 95% confidence interval with varying

PfEIR are distinct for women and children. PfPR in children tend to show a steeper rise with increasing PfEIR than in women. From the estimated relationship for children, a decrease in PfEIR from 1 ib/person/month to 0.0001 ib/person/month is associated with a reduction in PfPR from 38.3% to 15.7% (i.e., a 59.0% decrease in PfPR) on average. From the estimated relationship for women, the same decrease in PfEIR is associated with a reduction in PfPR from 19.4% to 10.2% (i.e., a 47.4% decrease in PfPR) on average. We note two things: with decreasing PfEIR, (1) the percentage reduction in PfPR to be expected in children tends to be higher than in women; (2) when transmission has been driven to an almost zero-level, PfEIR in children remains higher.

2.5 Discussion

The main objective of this paper was to investigate the relationship between PfEIR and PfPR. Using spatio-temporal data from rural Malawi, we have mapped PfEIR and PfPR, and identified hot-spots using exceedance probabilities. We have investigated and quantified the relationship between the PfEIR and PfPR. We found a logit-linear model to best explain the relationship. The geographical regions and times of elevated PfEIR and PfPR we have identified serve as targets for interventions.

We found a vector composition of 70.3% *A. arabiensis s.s.*, 27.3% *A. funestus s.s.*, 1.8% *A. gambiae s.s.*, and 0.5% *A. quadriannulatus.* A study by Mzilahowa et al. (2012) conducted in the same District 13 years earlier found a vector composition of 36.3%, 51.6%, 11.8% and 0.3% respectively. The difference in vector composition are remarkable and could indicate a shift in vector composition in the study area. Shifts in vector compositions have been observed elsewhere in Kenyan and Tanzania (Derua et al., 2012, Kitau et al., 2012, Mwangangi et al., 2013) and can impact on the transmission dynamics of a region, since different vectors have different host seeking behaviours and transmission capacities. A possible reason for the increased percentage of *A. arabiensis s.s.* and the reduced percentage of *A. funestus s.s.* and *A. gambiae s.s.* is the use of conventional and long-lasting insecticide treated nets (LLINs), which a study (Mwangangi et al.,
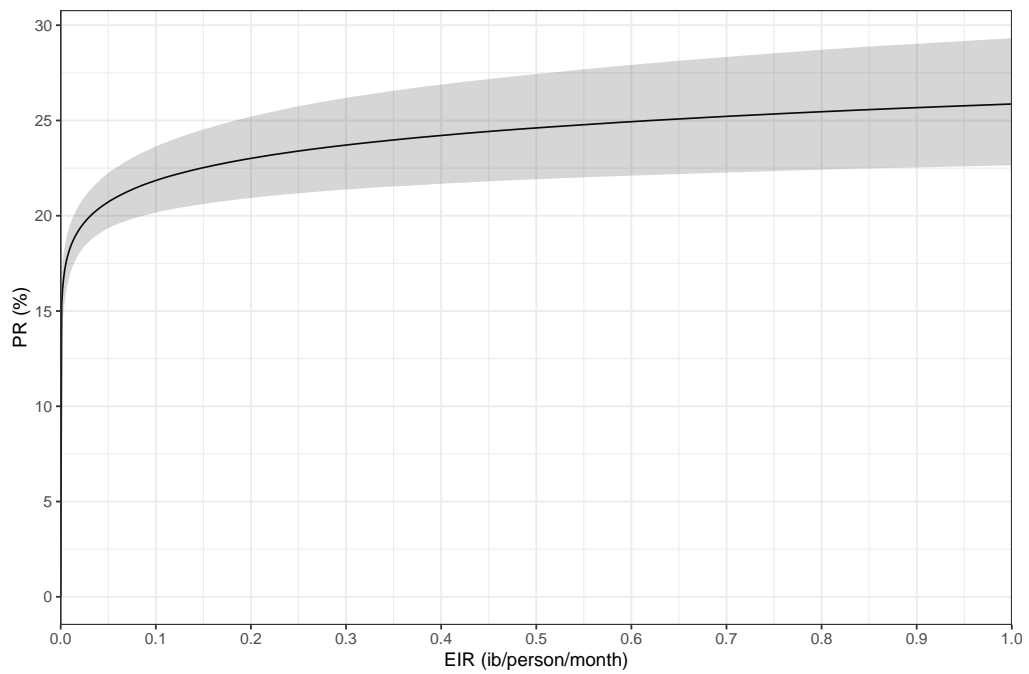
(a)
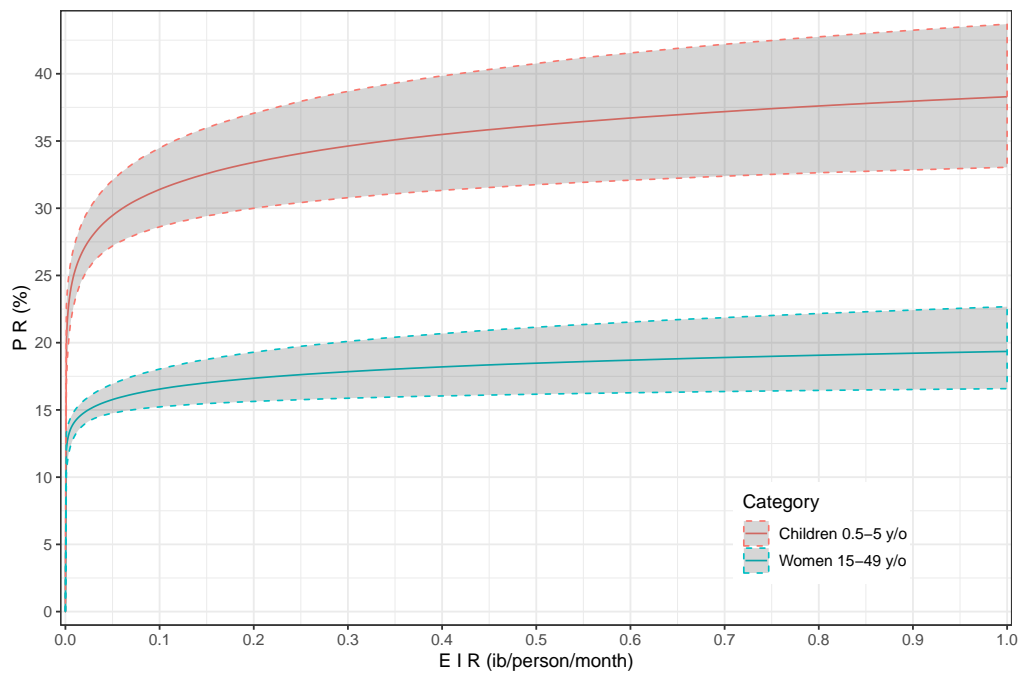
(b)

FIGURE 2.4: A plot of the estimated logit-linear relationship between PfPR and PfEIR. The solid lines are the estimated relationships and the shaded areas are the associated 95% confidence region for children and women combined (a) and for children and women separated (b)

2013) found to be more effective at killing *A. funestus s.s.* and *A. gambiae s.s.* than *A. arabiensis s.s.*. The use of ITNs (in children under 5 y/o) has increased in Malawi over the years: from 7.6% in 2000 to 20.2 in 2004, 59.6 in 2010 (Mathanga et al., 2012), and 73.0% in 2014 (Zamawe et al., 2016). Meanwhile, Mzilahowa et al. (2012) reported an EIR of 15 ib/person/month, which when compared our reported 2.39 ib/person/month could indicate a reduction in transmission, as has been observed throughout the country over the years, and has been attributed partly to the countries advances in vector control (Zamawe et al., 2016). Our statistical analysis did not include *A. gambiae s.s.*, and *A. quadriannulatus* due to their very low abundance in the study region throughout the study period.

Our study found, as expected (Mwandagalirwa et al., 2017, Zhou et al., 2016), a consistently higher PfPR in children 0.5-5 y/o than in women 15-49 y/o thoughout the study region and throught the study period. When we investigate the hypothesis that households with children had a different HBR as compared to those that did not, we did not find evidence in support of the hypothesis, which is in line with some studies (Clyde and Shute, 1958, Smith et al., 1956) that found that children and adults are not bitten at different rates, but contrary to others studies (Boreham et al., 1978, Bryan and Smalley, 1978, Muirhead-Thomson, 1951, Thomas, 1951) that have suggested differences. If children are bitten as often as women, then why is PfPR higher in children? On this question, our data better supported the assumption that either transmission efficiency is higher and/or time to clear a *P. falciparum* infection is longer in children than in women (Model 2, the best of the four SIS models), as compared to an assumption of equal transmission efficiency and infection clearance time (Model 1, AIC difference of 914). Both transmission efficiency and time to clear infection could be modulated by immunity. Active acquired immunity to blood stages of the parasites, which generally increases with age (Baird, 1995, Ladeia-Andrade et al., 2009), can cause inefficient transmissions, rapid clearance of parasites, or may keep parasite density too low to be detected (Doolan et al., 2009, John et al., 2005).

A question investigated in this research was whether or not PfEIR has a delayed effect on PfPR. Our data better support a month's delayed effect of PfEIR on PfPR as compared

to no delay effect and two months' delayed effect. The one month's lagged time could be due to the incubation periods of parasites within the human and in the mosquito. Prolonging these incubation periods (either through medicine or control measures) has been demonstrated to have the potential of reducing prevalence (Ruan et al., 2008).

On the nature of the PfEIR-PfPR relationship, we found that a logit-linear model, among other candidates, best explained the relationship in the study region, unlike Smith et al. (2005) who found an SIS model that assumes both heterogeneous infection rates and superinfection, a model analogous to our Model 4, to be the best model. Our results are however similar to the results of Beier et al. (1999), who modelled the same relationship with a log-linear model. The estimated relationship is such that every detectable PfEIR is associated with PfPR and even very low PfEIR levels are associated with substantial PfPR, similar to the findings of Beier et al. (1999). The association of very low (or undetectable) PfEIR with substential levels of PfPR has been observed before (Kabiru, 1994, Mbogo et al., 1995). This finding suggests that in contrast with other studies that have suggested that vector control programmes could be effective for controlling and even eliminating malaria in certain areas where transmission levels are marginal, our results suggest that although substantial reductions in PfPR may be expected with decreasing PfEIR through vector control, they may not be enough to drive malaria to elimination levels if they are not coupled with the detection and treatment of asymptomatic cases. This is manifested in our spatio-temporal maps for the months of August 2017 to March 2018, where PfEIR is zero everywhere in the study region, yet PfPR remains appreciably high ($> 10\%$) in most areas. The PfPR levels after a prolonged zero PfEIR correspond to a residual human host to be expected when transmission has been interrupted (see the spatio-temporal maps in S2 Appendix). Moreover, this residual hosts are more likely to be children (as evidenced in or spatio-temporal maps showing that at the end of the high transmission season, PfPR in consistently higher in children throughout the study region). This finding, coupled with the finding that transmission may be more efficient in children and they could take a longer time to clear infection, suggests that children may play a more active role in transmission during a high transmission period and in the resurgence of transmission after a period of low transmission, when vector-host contacts

are enhanced, similar to what other studies have suggested (Lin Ouédraogo et al., 2015, Walldorf et al., 2015).

The saturation in PfPR with increasing PfEIR at high transmission levels may be explained by a number of factors: (1) the same infected individuals are bitten by an increasing number of mosquitoes; (2) some individuals are well protected against vector contacts and do not get infected; (3) some individuals have acquired an immunity such that parasites densities in their blood are kept at levels lower than can be detected by RDT. Whatever be the reason, the saturation has practical importance on what results to expect in vector control programmes. It could mean that reduction in PfEIR may not be met with an immediate appreciable reduction in PfPR, if PfEIR is high. Additionally, when transmission in low, the quick rise in PfPR with increasing PfEIR suggest that nearing elimination, a little increase in PfEIR could be met with a quick rise in PfPR, implying extra difficulty in eliminating the disease when substantial reduction have already been achieved.

We found PfPR to increase during the months spanning November 2017 to April 2018, whilst PfEIR remained at zero-levels over the same period. What remains unclear is whether the rise in PfPR during the said period is due to new (undetected) infections or that parasite density in already-infected individuals increased to detectable levels. Either ways, this results exemplifies that a rise a prevalence may be observed even when transmission cannot be detected; thus the need for monitoring both parasitaemia and entomological indices in malaria surveillance.

The previous studies on the EIR-PR relationship (Beier et al., 1999, Smith et al., 2005) relied on data reported by several isolated surveys conducted in different geographical regions throughout Africa and at different time points. We optimized our sampling framework by sampling in same geographic region, across different transmission seasons and over a higher temporal scale of a month. However, we found relatively weaker association between PfEIR and PfPR (evidenced by low r^2 values) as compared to Beier et al. (1999). Indeed, in low EIR settings like our study region, the EIR–PR relationship can be noisy (see for example the low EIR region of Figure 1 of Beier et al. (1999) and

Figure 1 of Smith et al. (2005)). This limitation might be overcome by more intensive spatio-temporal sampling.

We note that the regression parameters for NDVI in the *A. funestus* model was not statistically significant (Table 2.8). Greenness of the vegetation is however known to affect the abundance of mosquitoes due to the availability of breeding sites in green areas and times. Greenness is also seasonal and may be related to humidity and temperature, which are all adjusted for in the model. Thus, the lack of statistical significance may not be interpreted as an actual lack of effect. We also note that in the same model, one of the temperature variables was significant whereas the other was not. However, these all helped to improve the overall fit of the model. Since temperature is seasonal, and having included sinusoidal terms to account for seasonality, it is difficult to explain the lack of significance of this variable due to the complexity of the model. Nevertheless, the main goal of the analysis was predictive rather than explanatory.

A concern that may be raised about our analyses is that in modelling the relationship between PfEIR and PfPR, we did not accounted for the sensitivity and specificity of RDT, the diagnostic used in testing for *P falciparum*. If the sensitivity α and specificity β were known, we could account for them by setting $PfPR(x,t)$ as used in our analysis to $\alpha PfPR(x,t) + (1-\beta)(1 - PfPR(x,t))$. Thus, to be strict with terminology, what we have called PfPR should be interpreted as the probability of testing positive for *P falciparum* using RDT.

Malaria transmission in Malawi is known to be highly seasonal (Roca-Feltrer et al., 2012a, Rogerson et al., 2000) (as also evident in our results), with a single peak in the rainy months (December to May) and a trough in the dry season (June to November), which is mainly climate driven (Lowe et al., 2013). In the light of this background knowledge and inspection of our empirical data, we accounted for seasonality using sinusoidal functions with a period of 12 months.

Interpreting the relationship between EIR and PR is indeed a much more complex problem than presented in this and other papers that have considered the subject. First, there needs to be a transmission before an individual acquires malaria. This makes prevalence, at least in theory, a function of EIR. However prevalence at any point in time

also depends on the duration of infection, which has a different transition probability than EIR. Second, transmission will take place only when there are infected people. Furthermore, transmission may occur at a higher rate when prevalence is higher. Thus, in turn, EIR may also depend on prevalence. Our modelling therefore is a construct of a rather complex process that is still not fully understood.

Although the PR data were available at individual level, we fitted the model at location level since after fitting a separate model to women and children, no individual level variable was significant.

Hurdle or zero-inflated models are sometimes chosen for count data when their counterparts without zero-inflation are unable to predict equivalent rates of zeros as are in the data. However, our predictions of EIR are similar to the empirical data; thus our HBR and SR models, although without zero-inflation, are parsimonious for the data.

Potential overdispersion in count data may be accounted for by using negative binomial models. However, a geostatistical model already adjusts for possible overdispersion. Diggle and Giorgi (2019) discuss this in detail detail.

Our model building strategy was an incremental approach, where we start with a simple generalized linear model and increase complexity by adding random effects, spatially correlated random effects, and spatio-temporal random effects subject to the current best model not producing satisfactory validation based on the variogram validation approach developed by Giorgi et al. (2018). All models were satisfactory after adding spatially correlated random effects except for the SR models that didn't need spatially correlated random effects.

Another approach we could have used to map EIR is a marked point process to model empirical EIR. However, both HBR and PfSR are noisy data and therefore empirical $EIR = HBR \times PfSR$ is even noisier. By modelling HBR and PfSR separately using appropriate covariates, the signal in each could then be extracted, reducing the noise in predicted EIR.

In obtaining parameters of each of the models relating EIR and PR, we drew 10,000 Monte Carlo samples from the predictive distribution of EIR, taking into account the

uncertainties in the parameters of its predictive models. Thus, the resulting samples from the set of parameters of each model relating EIR and PR can be seen as random samples from its distribution.

Human landing catches (HLC) are the gold standard for providing reliable estimates of HBR by anophelines (Hawkes et al., 2017, Organization., 1975), although arguements have been made that it may result in overestimation (Gao et al., 2018) or underestimate (Meza et al., 2019). The main disadvantages of HLC is that it exposes human catchers to several infection transmited by the vector and has baises resulting from differential skill of the catcher. Our study used Suna traps, which has been found to yeild HBR estiamtes that are consistent with with HLC (Mburu et al., 2019). We therefore estimated HBR using the number of mosquitoes traped by Suna traps, without multiplying by any factor.

Distributed lag non-linear models (Gasparrini et al., 2010) may be used to account for the effect of environmental factors such as temperature and relative humidity, which often have delayed effects of on the outcome varibles. Some studies on malaria (Zhao et al., 2014) and mosquitoes (Chen et al., 2010) have used this approach. However, these models often require the estimation of several parameters, more than the amount of data can support. In our analysis, we used a series of accumulated temperature and relative humidity, since it is important to aggregate climatic variables to detect biological relationships (Lozano-Fuentes et al., 2012, Roiz et al., 2014, Shone et al., 2001).

The splines used to adjust for time trends and approximated non-linear relationships between outcome variable and explanatory variables were based on epidemiological knowledge and visual inspection of trends. Temperature and RH are known to be non-linearly related to each of our outcomes. For each linear spline, we considered a range of values for each knot position based on visual inspection and/or epidemiological knowledge, and chose the best knot position based on the using AIC. We used the AIC because models were nested, and their overall fit were of more concern than error of prediction.

An extension of the analysis presented in this paper is a joint model for PfEIR and PfPR. The joint model might then allow for the borrowing of information from the entomological data in order to predict parasitaemia prevalence more precisely. An exploratory analysis for the need and plausibility of such a joint analysis will to be check if the

residual errors of PfEIR are correlated with those of PfPR. Correlations of the errors of the two outcomes will suggest that predictions of prevalence might benefit from the joint modelling of entomological and parasitaemia data. We propose a model for such joint analysis in concluding chapter of the thesis. Such a joint analysis opens the question of mediation variables since some variables may have and effect on PfPR through PfSR or HBR or both.

Conclusion

We have mapped PfEIR and PfPR, compared their spatio-temporal patterns, and identified their respective hot spots. We found a clear relationship between PfPR and PfEIR, emphasising that PfEIR is a direct measure of malaria transmission intensity. First, our results suggest that at low transmission settings, a rapid increase in prevalence is to be expect with the slightest increase in transmission. Second, the results suggest that an immediate reduction in prevalence may not be expected through vector control programmes in high transmission settings. Third, the results suggest that it may not be possible to eliminate malaria or drive it to elimination levels by using only vector control programmes. Intervention programmes may therefore need to integrate vector control with the detection and treatment of asymptomatic individuals (who are likely to be found in residual hot-spots) to drive malaria to elimination levels. Our results are likely to hold in other geographical regions with marginal *P falciparum* transmission or where there are seasonal effects in malaria transmission.

S1 Appendix: Details of the models and estimates of their parameters.

Lagged averages of temperature and relative humidity

Let $Avg(\text{Temp}(x_i, t_i), s_1, s_2)$ and $Avg(\text{RH}(x_i, t_i), s_1, s_2)$ respectively denote the average temperature and relative humidity taken over s_1 to s_2 days prior to the data collection. Table 2.2 shows the s_1 and s_2 values over which average temperature and relative humidity were considered for the analysis.

TABLE 2.2: Range of days prior to data collections over which temperature and relative humidity were averaged

To (s_2)	0	3	5	7	14	21	28	35	42
From (s_1)									
0	✓[a]	✓	✓	✓	✓	✓	✓	✓	✓
3			✓	✓	✓	✓	✓	✓	✓
5				✓	✓	✓	✓	✓	✓
7					✓	✓	✓	✓	✓
14						✓	✓	✓	✓
21							✓	✓	✓
28								✓	✓
35									✓

[a] The check marks indicate the days from/to which average temperature and relative humidity were taken.

Procedure for building the HBR, PfSR and PfEIR models

We selected the best combination of fixed and random effects that best explain HBR, PfSR and PfPR using the following procedure.

1. We first built a generalized linear model in which temperature and RH are considered together with time trends and sine and cosine functions for seasonality. For $Avg(\text{Temp}(x_i, t_i), s_1, s_2)$, $Avg(\text{RH}(x_i, t_i), s_1, s_2)$, the choice of s_1 and s_2, as illustrated by Table 2.2, was based on the deviance profile of the variable involved, i.e.

either temperature or RH. Piecewise-linear transformations of temperature and RH were considered based on visual inspection and epidemiological knowledge.

2. Potential confounding between seasonal sinusoids, temperature and RH were checked, and if confounding existed, the variable with the highest variance inflation factor was excluded from the model. Furthermore, variables that did not improve the model fit as judged by the AIC were excluded. Sin-cosine terms were always considered together as if they were one covariate.

3. With the current model as a basic model we include other available explanatory variables based on forward selection.

4. When no more explanatory variables significantly improve the model fit, we fit a generalized linear mixed model with a random effect for each unique space-time location.

5. We then check for the presence of residual spatial, temporal, and spatio-temporal correlations using the algorithm described in Giorgi et al. (2018), and then include the random effect terms that improve the model fit.

The selected fixed effects for the HBR, PfSR and PfEIR models

TABLE 2.3: Details of the fixed effects for the *A. arabiensis* human biting rate model.

Effect	Details
Covariates $(d(x_i, t_i)^\top \beta)$	
$\beta_1 \mathbf{1}(x_i \in \mathcal{A})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area A and 0 otherwise.
$\beta_2 \mathbf{1}(x_i \in \mathcal{B})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area B and 0 otherwise.
$\beta_3 \mathbf{1}(x_i \in \mathcal{C})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area C and 0 otherwise.
$\beta_4 \mathbf{1}(\text{Indoor})$	A binary indicator taking the value 1 if location the mosquito trap was set indoors and 0 otherwise.
$\beta_5 \text{DSR}(x_i)$	Distance from location x_i to the closest small river
$\beta_6 Avg(\text{RH}(x_i, t_i), 14, 35)$	Average relative humidity 14 to 35 days prior to the data collection.
$\beta_7 \min\{Avg(\text{Temp}(x_i, t_i), 7, 14), 22.9\}$	The effect of temperature when temperature is below $22.9°C$.
$\beta_8 \max\{Avg(\text{Temp}(x_i, t_i), 7, 14) - 22.9, 0\}$	The effect of temperature when temperature is $22.9°C$ or more.
Seasonality and Trends $(f(t_i, \alpha))$	
$\alpha_1 \sin(2\pi t/12)/t$	Seasonal and trend effect
$\alpha_2 \cos(2\pi t/12)/t$	Seasonal and trend effect

TABLE 2.4: Details of the fixed effects for the *A. funestus* human biting rate model.

Effect	Details
Covariates $(d(x_i, t_i)^\top \beta)$	
β_0	Intercept
$\beta_1 \text{Elevation}(x_i)$	Elevation of the location x_i.
$\beta_2 \text{DSR}(x_i)$	Distance from location x_i to the nearest small river.
$\beta_3 \text{NDVI}(x_i)$	Normalized difference vegetation index at location x_i.
$\beta_4 Avg(\text{Temp}(x_i, t_i), 0, 7)$	Average temperature one week prior to data collection.
$\beta_5 Avg(\text{Temp}(x_i, t_i), 7, 14)$	Average temperature 7 to 14 days prior to data collection.
$\beta_6 Avg(\text{RH}(x_i, t_i), 14, 21)$	Average relative humidity 14 to 21 days prior to data collection.
Seasonality and Trends $(f(t_i, \alpha))$	
$\alpha_1 \sin(2\pi t_i / 12)$	Seasonal effect
$\alpha_2 \cos(2\pi t_i / 12)$	Seasonal effect
$\alpha_3 \min\{t_i, 12\}$	Trend term
$\alpha_4 \max\{t_i - 12, 0\}$	Trend term

TABLE 2.5: Details of the fixed effects for the sporozoite rate model.

Effect	Details
Covariates $(d(x_i, t_i)^\top \beta^*)$	
β_0^*	Intercept
$\beta_1^* \mathrm{DSR}(x_i)$	Distance from location x_i to the nearest small river.
$\beta_2^* \mathrm{DLR}(x_i)$	Distance from location x_i to the nearest large river.
$\beta_3^* \mathrm{Elevation}(x_i)$	Elevation of location x_i
$\beta_4^* \mathrm{EVI}(x_i)$	Enhanced vegetation index of location x_i
Seasonality and Trends $(f^*(t_i, \alpha^*))$	
$\alpha_1^* \sin(2\pi t_i/12)$	Seasonal effect
$\alpha_2^* \cos(2\pi t_i/12)$	Seasonal effect
$\alpha_3^* \min\{t_i, 12\}$	Trend effect
$\alpha_4^* \max\{t_i - 12, 0\}$	Trend effect

TABLE 2.6: Details of the fixed effects for the *P. falciparum* prevalence model.

Effect	Details
Covariates $(d(x_i, t_i)^\top \varphi)$	
$\varphi_1 \mathbf{1}(x_i \in \mathcal{A})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area A and 0 otherwise.
$\varphi_2 \mathbf{1}(x_i \in \mathcal{B})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area B and 0 otherwise.
$\varphi_3 \mathbf{1}(x_i \in \mathcal{C})$	A binary indicator taking the value 1 if location x_i belongs to Focal Area C and 0 otherwise.
$\varphi_4 \text{Elevation}(x_i)$	Elevation of the location x_i.
$\varphi_5 \text{DLR}(x_i)$	Distance from location x_i to the nearest large river.
$\varphi_6 Avg(\text{Temp}(x_i, t_i), 14, 42)$	Average temperature 14 to 42 days prior to data collection.
$\varphi_7 \text{NDVI}(x_i)$	Normalized difference vegetation index at location x_i.
$\varphi_8 \text{Wealth}(x_i)$	Wealth index of the i-th household.
Seasonality and Trends $(g(t_i, \varrho))$	
$\varrho_1 \sin(2\pi t_i / 12)$	Seasonal effect
$\varrho_2 \cos(2\pi t_i / 12)$	Seasonal effect
$\varrho_3 \min\{t_i, 21\}$	Trend term
$\varrho_4 \max\{t_i - 21, 0\}$	Trend term

TABLE 2.7: Regression table for the *A. arabiensis* human biting rate model

Variable	Parameter	Point Estimate
Covariates		
$\mathbf{1}(x_i \in \mathcal{A})$	β_1	-13.525
		$(-16.217, -10.833)^a$
$\mathbf{1}(x_i \in \mathcal{B})$	β_2	-9.995
		$(-12.656, -7.333)$
$\mathbf{1}(x_i \in \mathcal{C})$	β_3	-10.848
		$(-13.514, -8.182)$
$\mathbf{1}(\text{Indoor})$	β_4	0.456
		$(0.264, 0.647)$
$\text{DSR}(x_i)$	β_5	0.631×10^{-3}
		$(0.143, 1.120\ ) \times 10^{-3}$
$Avg(\text{RH}(x_i, t_i), 14, 35)$	β_6	0.056
		$(0.038, 0.073)$
$\min\{Avg(\text{Temp}(x_i, t_i), 7, 14), 22.9\}$	β_7	0.180
		$(0.072, 0.289)$
$\max\{Avg(\text{Temp}(x_i, t_i), 7, 14) - 22.9, 0\}$	β_8	-0.132
		$(-0.22, -0.044)$
Seasonality and Trends		
$\sin(2\pi t_i/12)/t$	α_1	-0.291
		$(-0.907, 0.325)$
$\cos(2\pi t_i/12)/t$	α_2	1.092
		$(-0.759, 2.943)$
Spatial Correlation		
Signal variance	σ^2	4.114
		$(3.262, 5.189)$
Scale (km)	ϕ	0.649
		$(0.492, 0.856)$
Nugget variance	τ^2	0.162
		$(0.124, 0.21)$

Dependent Variable: log of *A. arabiensis* Mosquito Density

[a] 95% confidence intervals are in brackets.

TABLE 2.8: Regression table for the *A. funestus* human biting rate model

Variable	Parameter	Point Estimate
Covariates		
Intercept	β_0	2.523
		$(-3.209,\ 8.256)^a$
Elevation(x_i)	β_1	-5.583×10^{-3}
		$(-7.896,\ -3.271) \times 10^{-3}$
DSR(x_i)	β_2	2.993×10^{-3}
		$(2.329,\ 3.658) \times 10^{-3}$
NDVI(x_i)	β_3	1.392
		$(-1.251,\ 4.035)$
$Avg(\text{Temp}(x_i, t_i), 0, 7)$	β_4	-0.154
		$(-0.279,\ -0.028)$
$Avg(\text{Temp}(x_i, t_i), 7, 14)$	β_5	-0.116
		$(-0.295,\ 0.064)$
$Avg(\text{RH}(x_i, t_i), 14, 21)$	β_6	-0.043
		$(-0.078,\ -0.008)$
Seasonality and Trends		
$\sin(2\pi t_i/12)$	α_1	-0.291
		$(-0.907,\ 0.325)$
$\cos(2\pi t_i/12)$	α_2	1.092
		$(-0.759,\ 2.943)$
$\min\{t_i, 12\}$	α_3	-0.291
		$(-0.907,\ 0.325)$
$\max\{t_i - 12, 0\}$	α_4	1.092
		$(-0.759,\ 2.943)$
Spatial Correlation		
Signal variance	σ^2	4.456
		$(3.379,\ 5.876)$
Scale (km)	ϕ	0.906
		$(0.66,\ 1.245)$
Nugget variance	τ^2	0.142
		$(0.105,\ 0.191)$

Dependent Variable: log of *A. funestus* Mosquito Density

[a] 95% confidence intervals are in brackets.

TABLE 2.9: Regression table from fitting the *P. falciparum* sporozoite rate models.

Variable	Parameter	*A. funestus s.s.*	*A. arabiensis s. s.*
Covariates			
Intercept	β_0^*	0.139	-3.392
		$(-7.793, 8.071)^a$	(-4.772, -2.125)
DLR(x_i)	β_1^*	-1.945×10^{-3}	—
		$(-3.345, -0.545)\times10^{-3}$	
DSR(x_i)	β_2^*	-4.309×10^{-3}	—
		(-4.309, -1.119)	
Elevation(x_i)	β_3^*	7.786×10^{-3}	—
		$(5.819, 9.752)\times10^{-3}$	
EVI(x_i)	β_4^*	-36.648	—
		(-65.090, -8.206)	
Seasonality and Trends			
$\sin(2\pi t_i/12)$	α_1^*	-0.378	-0.253
		(-0.565, -0.19)	(-0.882, 0.375)
$\cos(2\pi t_i/12)$	α_2^*	-0.722	-0.867
		(-0.954, -0.489)	(-1.864, 0.13)
$\min\{t_i, 12\}$	α_3^*	-0.056	0.027
		(-0.072, -0.041)	(-0.086, 0.140)
$\max\{t_i - 12, 0\}$	α_4^*	0.061	-0.089
		(0.039, 0.084)	(-0.305, 0.127)

Dependent Variables: logits of the probability of a positive test from children under 5 y/o and for women 15-49 y/o.

[a] 95% confidence intervals are in brackets.

TABLE 2.10: Regression table for the *P. falciparum* parasite rate model.

Variable	Parameter	Children under 5 y/o	Women 15-49 y/o
Covariates			
$\mathbf{1}(x_i \in \mathcal{A})$	φ_1	0.685	-0.506
		(-1.877 , 3.247)	(-3.166 , 2.155)
$\mathbf{1}(x_i \in \mathcal{B})$	φ_2	2.829	2.568
		(0.41 , 5.248)	(0.134 , 5.002)
$\mathbf{1}(x_i \in \mathcal{C})$	φ_3	3.192	2.641
		(0.806 , 5.577)	(0.224 , 5.058)
Elevation(x_i)	φ_4	5.165×10^{-3}	5.920×10^{-3}
		(2.322 , 8.008)$\times 10^{-3}$	(3.039 , 8.800)$\times 10^{-3}$
DLR(x_i)	φ_5	-0.372×10^{-3}	-0.181×10^{-3}
		(-0.522 , -0.222) $\times 10^{-3}$	(-0.353, -0.009) $\times 10^{-3}$
$Avg(\text{Temp}(x_i, t_i), 14, 42)$	φ_6	-0.112	-0.096
		(-0.201 , -0.023)	(-0.187 , -0.005)
NDVI(x_i)	φ_7	-2.424	-5.556
		(-4.703 , -0.144)	(-7.63 , -3.482)
Wealth(x_i)	φ_8	-0.212	-0.159
		(-0.283 , -0.141)	(-0.215 , -0.102)
Seasonality and Trends			
$\min\{t_i, 21\}$	ϱ_1	-0.079	-0.079
		(-0.098 , -0.06)	(-0.1 , -0.059)
$\max\{t_i - 21, 0\}$	ϱ_2	0.072	0.086
		(0.042 , 0.102)	(0.056 , 0.117)
$\cos(2\pi t_i/12)$	ϱ_3	-0.045	0.101
		(-0.265 , 0.175)	(-0.123 , 0.324)
$\sin(2\pi t_i/12)$	ϱ_4	0.209	0.175
		(-0.138 , 0.556)	(-0.173 , 0.523)
Spatial Correlation			
Signal variance	σ^2	0.347	0.602
		(0.222 , 0.542)	(0.416 , 0.872)
Scale (km)	ϕ	1.175	1.055
		(0.617 , 2.238)	(0.631 , 1.765)
Nugget variance	τ^2	1.546	1.368
		(0.956 , 2.500)	(0.932 , 2.007)

Dependent Variables: logits of the probability of a positive test from children under 5 y/o and for women 15-49 y/o.

[a] 95% confidence intervals are in brackets.

TABLE 2.11: Parameter estimates from the models for the relationship between PfEIR and PfPR. The models' goodness of fit are assessed by the AIC and their predictive abilities by the root-mean-square error (RMSE) and bias.

Model	$p(x,t)$	γ	γ_1	γ_2	AIC	RMSE	Bias
1. SIS	$\frac{\gamma PfEIR(x,t-1)}{\gamma PfEIR(x,t-1)+1}$	7.02			7633	0.361	7.597×10^{-3}
		(3.906, 12.284)					
2. SIS with D.I/R	$\sum_{k=1}^{2} \xi_{k,it} \frac{\gamma_k PfEIR(x,t-1)}{\gamma_k PfEIR(x,t-1)+1}$		107.208	0.762	6719	0.353	27.386×10^{-3}
			(0.088, 381.139)	(0.485, 24.344)			
3. SIS with S.I.	$1 - e^{-\gamma PfEIR(x,t-1)}$	1.728			9231	0.351	78.301×10^{-3}
		(0.638, 3.087)					
4. SIS with S.I. and D.I/R	$\sum_{k=1}^{2} \xi_{k,it}(1 - e^{-\gamma_k PfEIR(x,t-1)})$		22.603	0.471	7677	0.392	99.390×10^{-3}
			(0.128, 67.048)	(0.234, 7.02)			
			a	b			
5. Beier	$a + b\log(PfEIR(x,t-1))$		0.253	0.013	4628	0.328	5.376×10^{-3}
			(0.232, 0.283)	(0.009, 0.021)			
6. Logit-linear	$\frac{PfEIR(x,t-1)^b}{PfEIR(x,t-1)^b+\exp(-a)}$		-0.986	0.100	4620	0.327	4.874×10^{-3}
			(-1.160, -0.804)	(0.062, 0.147)			
Logit-linear for children only			-0.523	0.119			
			(-0.742, -0.296)	(0.073, 0.174)			
Logit-linear for women only			-1.427	0.083			
			(-1.575, -1.218)	(0.046, 0.133)			

S.I. denotes supper infection and D.I/R denotes different infection/recovery rates for children and women. 95% confidence intervals are in brackets. AIC is the median AIC from 10,000 Simulations. RMSE is the root-mean-square error.

S2 Appendix: Maps of predicted PfEIR and PfPR.
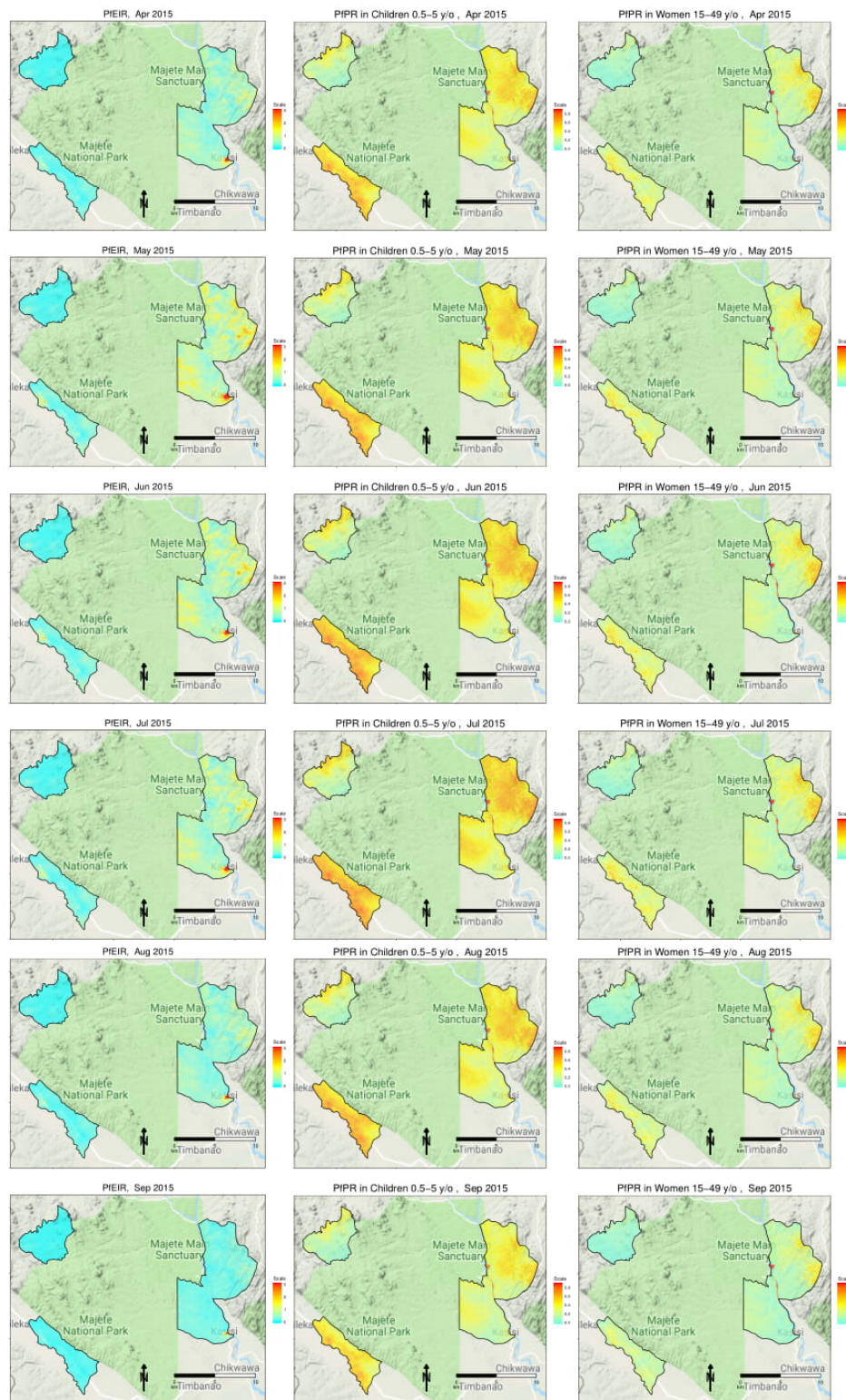
FIGURE 2.5: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2015 to September 2015.
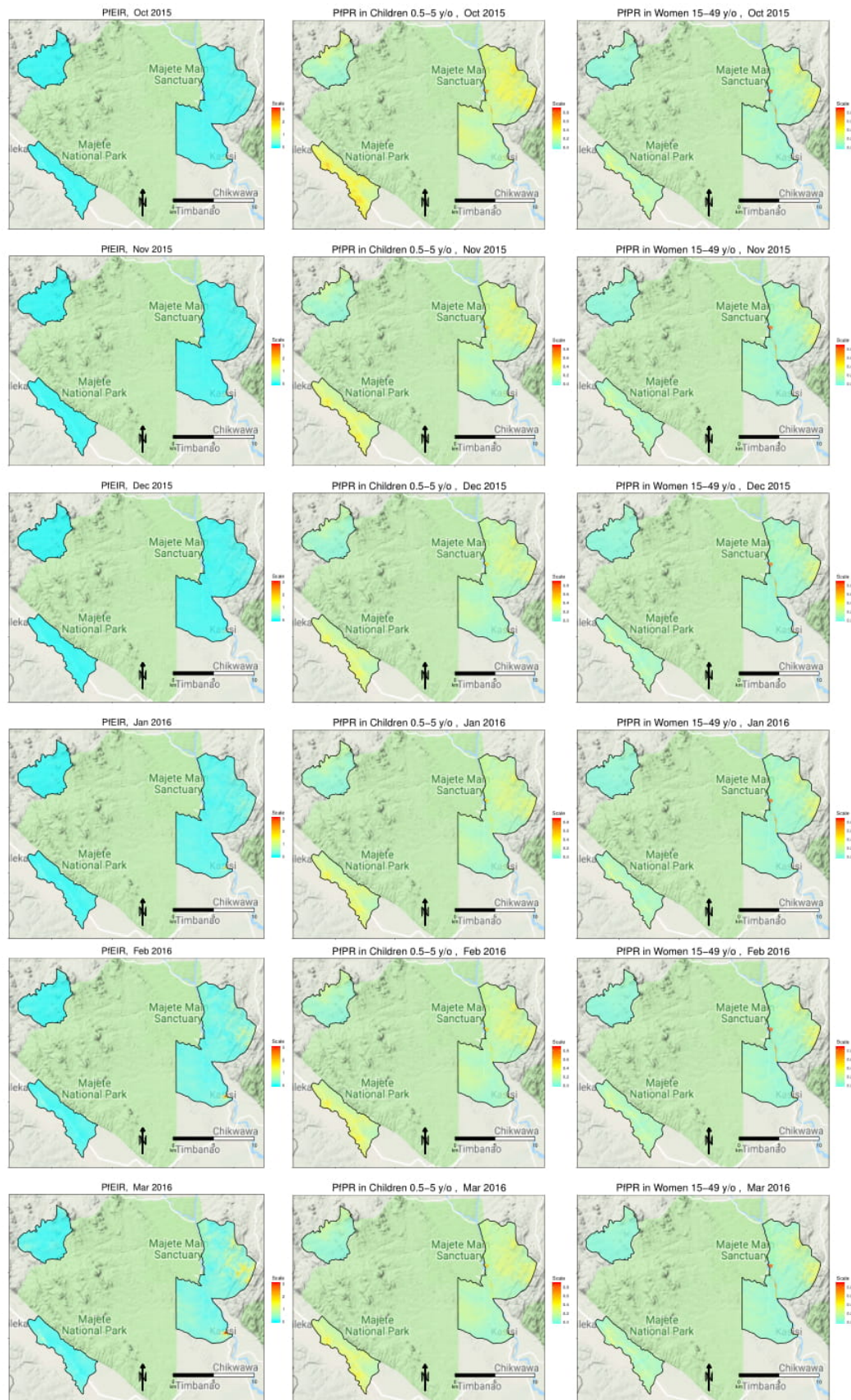
FIGURE 2.6: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2015 to March 2016.
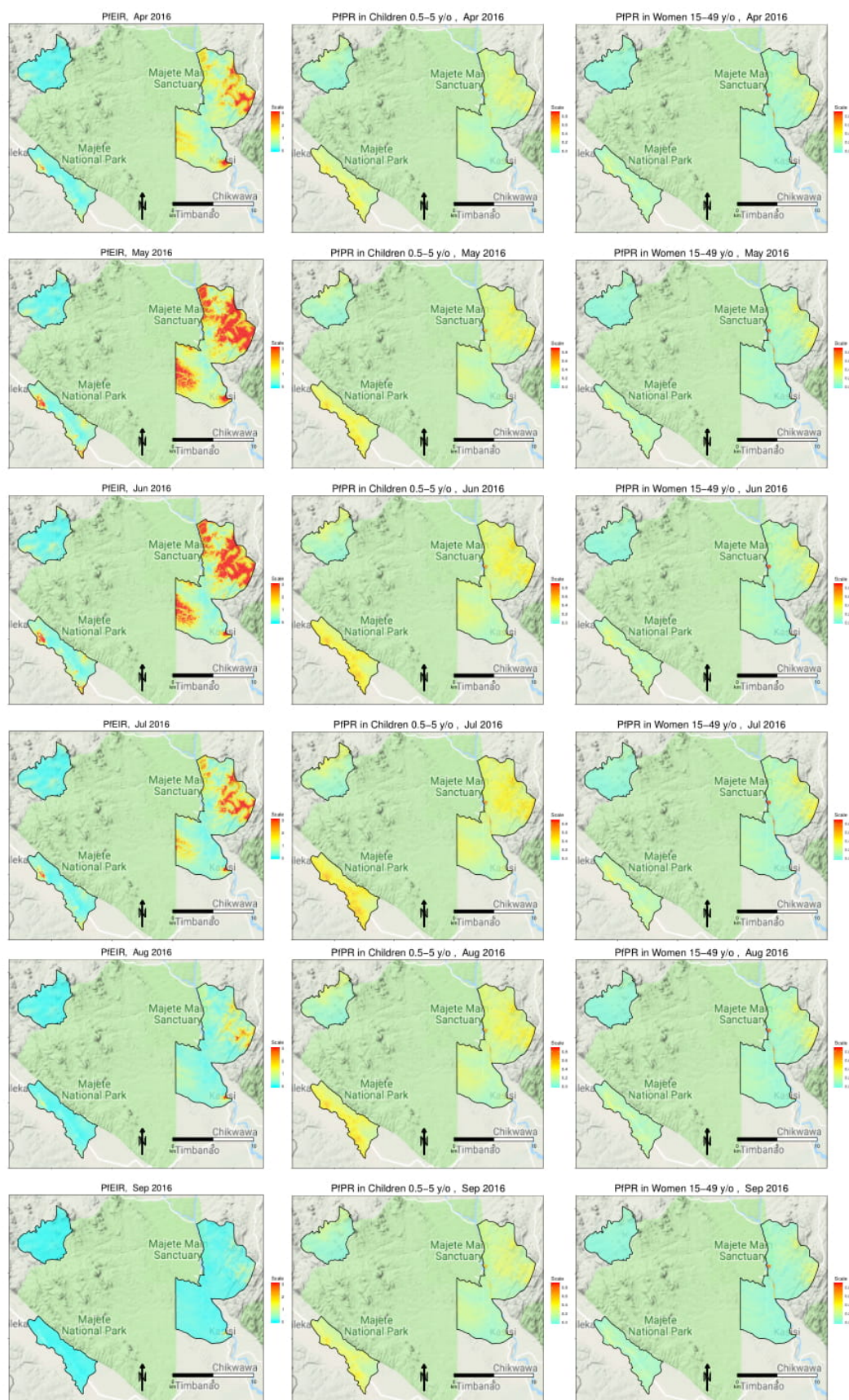
FIGURE 2.7: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2016 to September 2016.
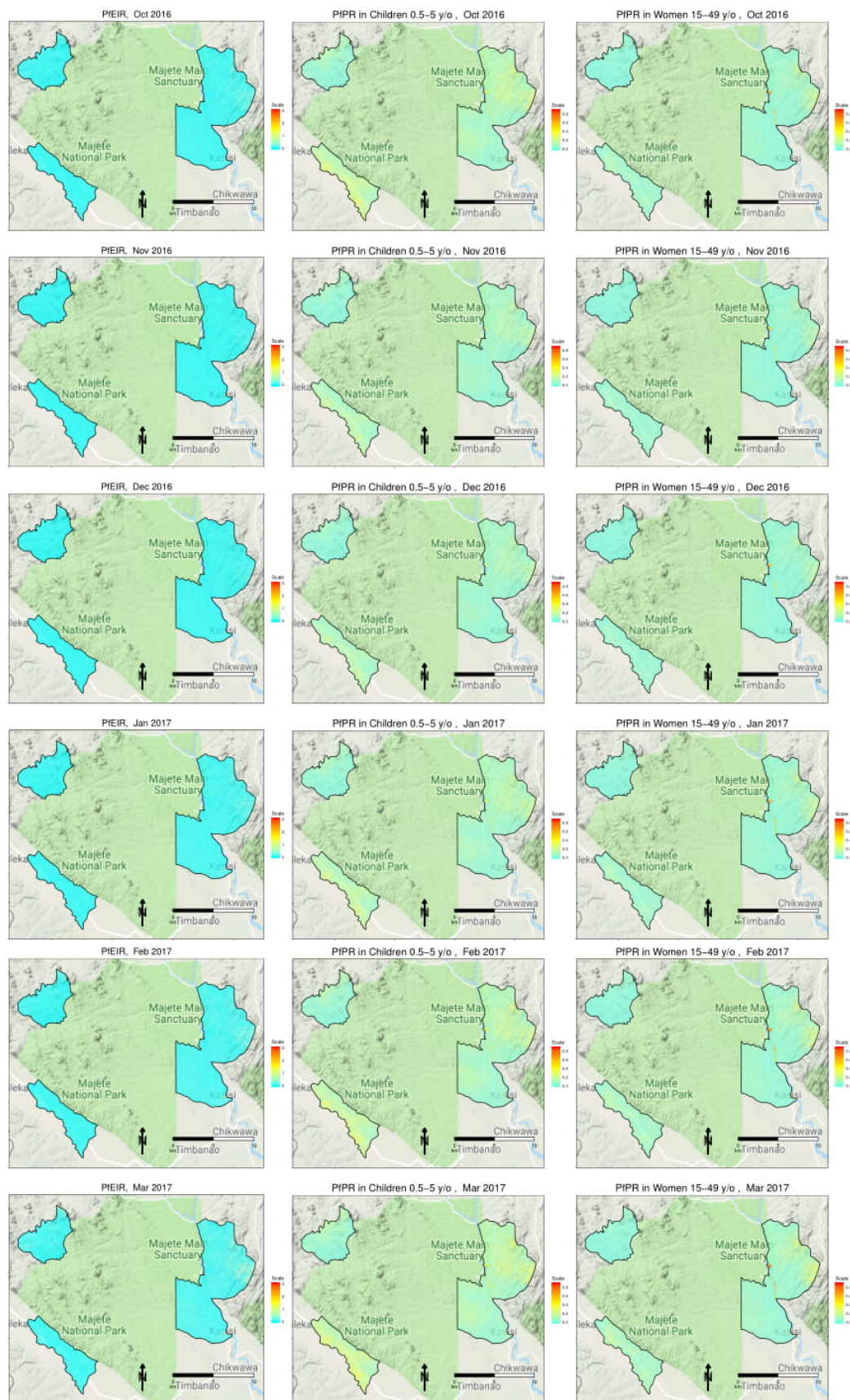
FIGURE 2.8: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2016 to March 2017.
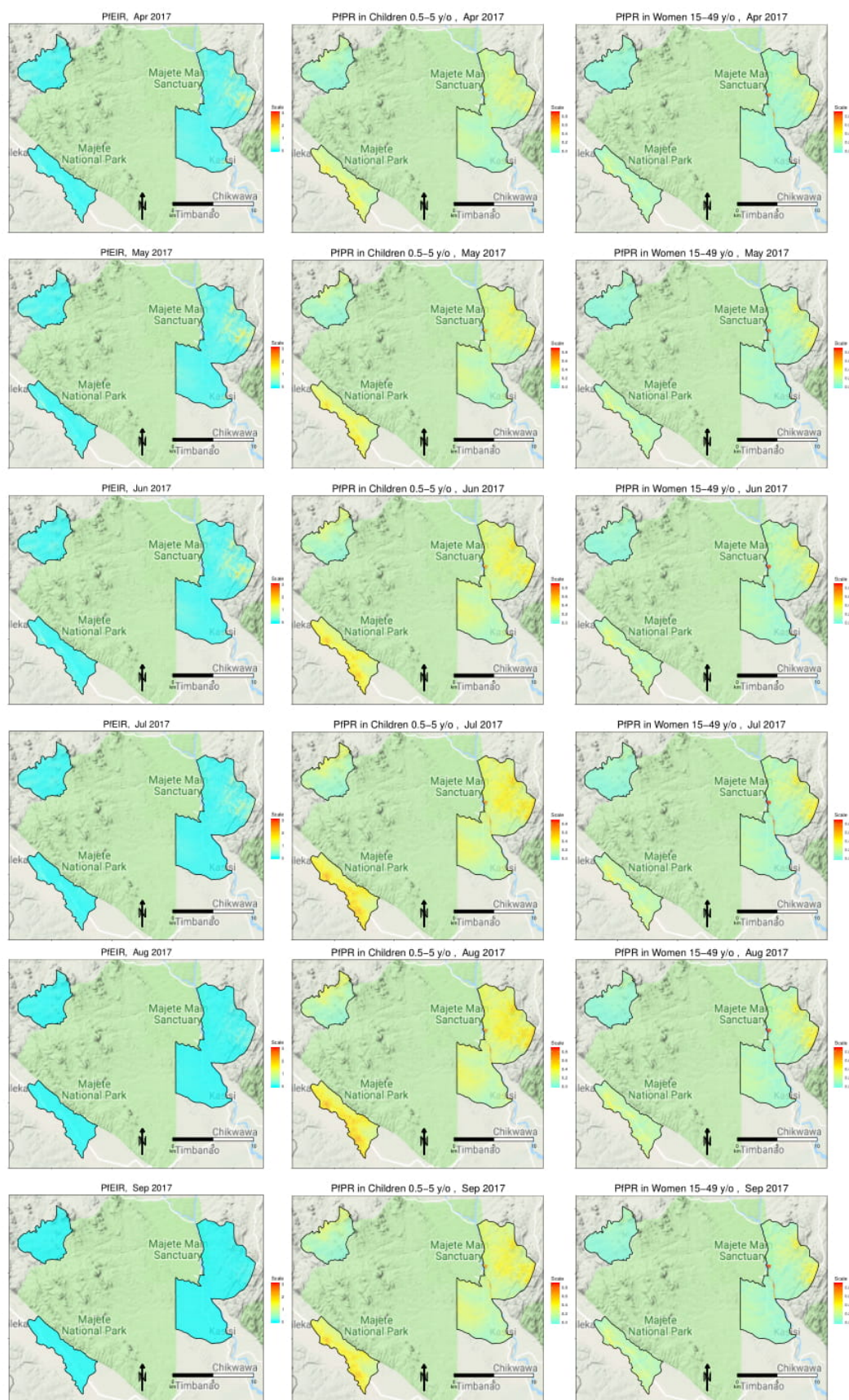
FIGURE 2.9: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2017 to September 2017.
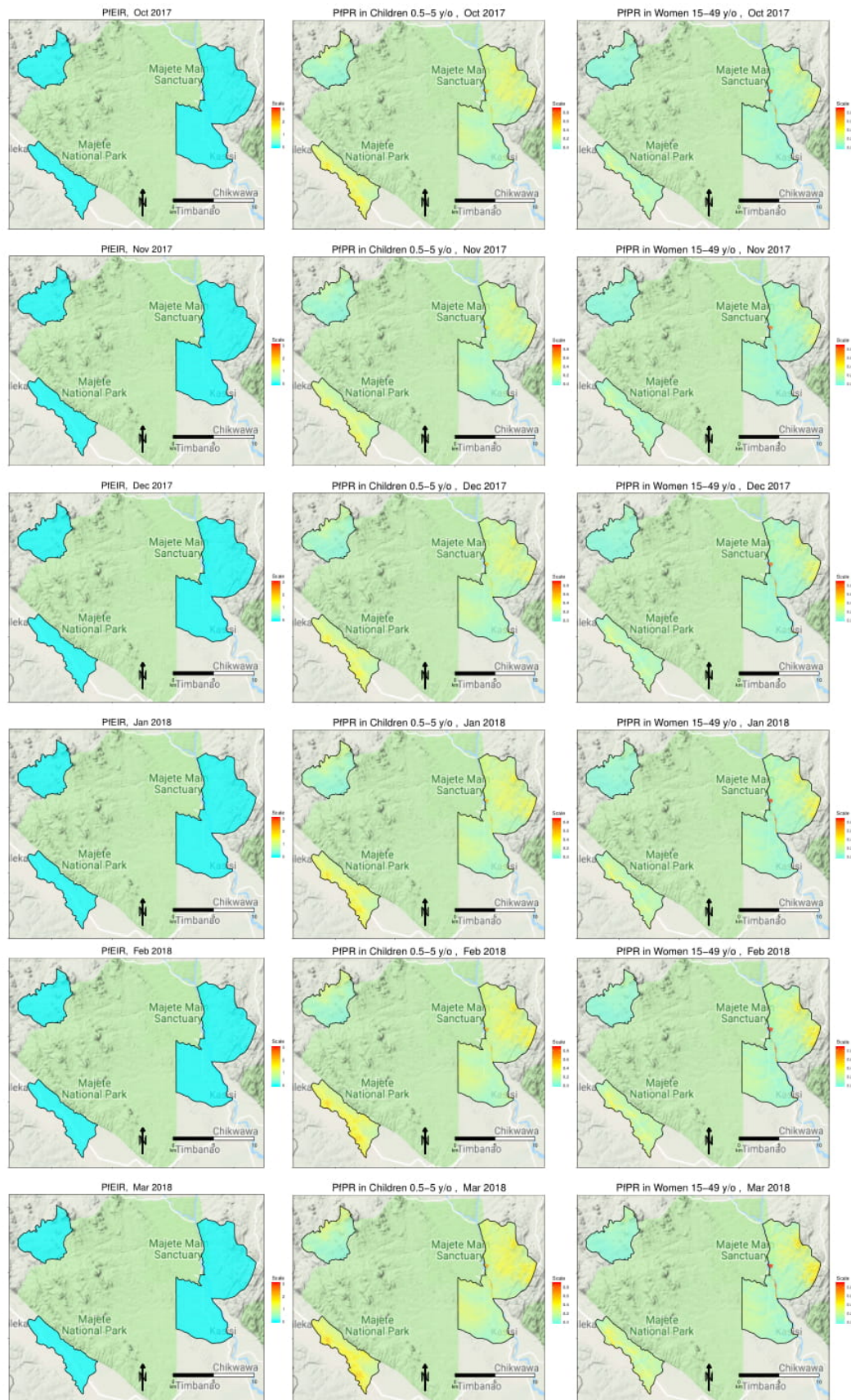
FIGURE 2.10: Predicted PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2017 to March 2018.

S3 Appendix: Maps of Exceedance probabilities and Hotspots of PfEIR and PfPR.
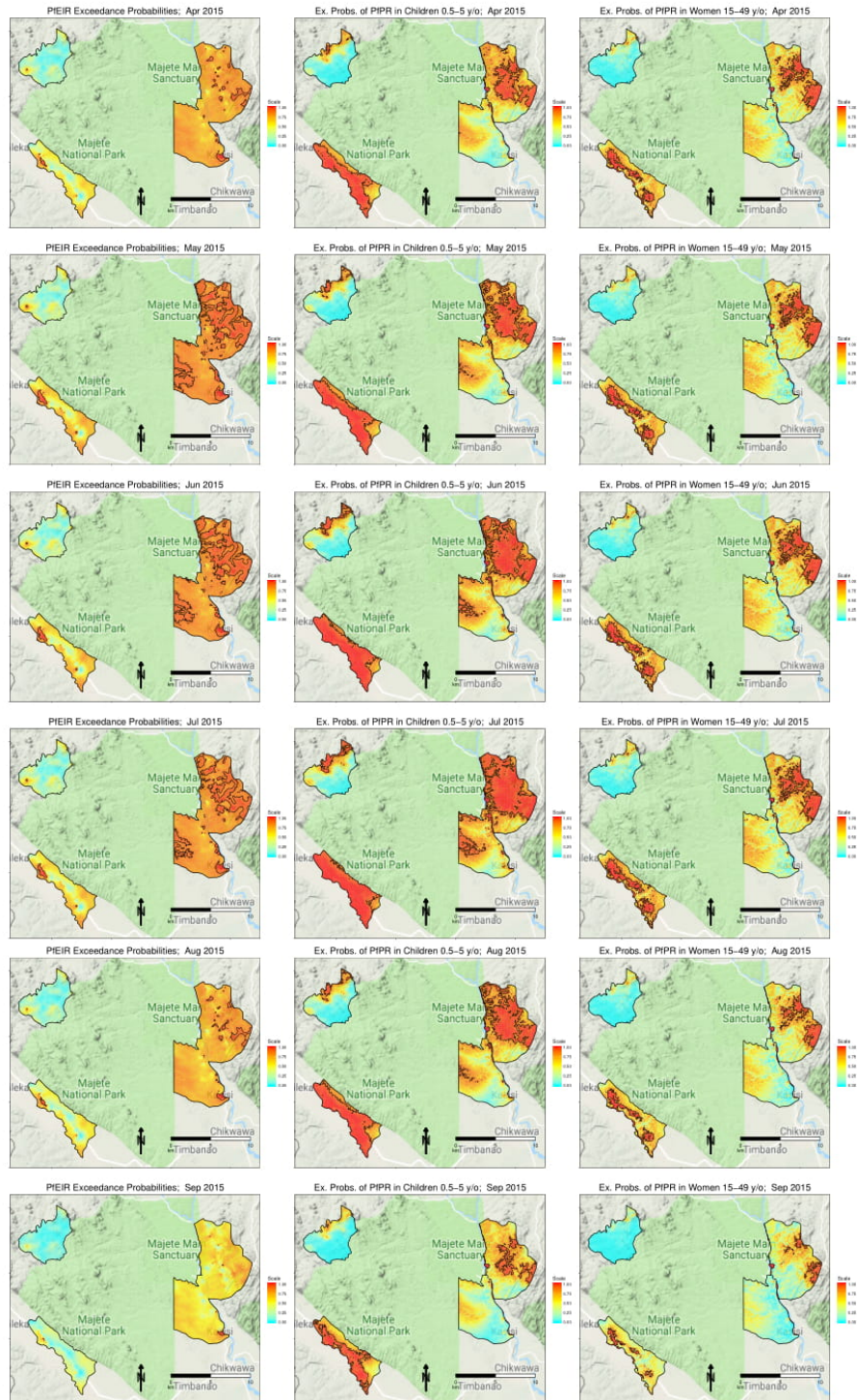
FIGURE 2.11: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2015 to September 2015.
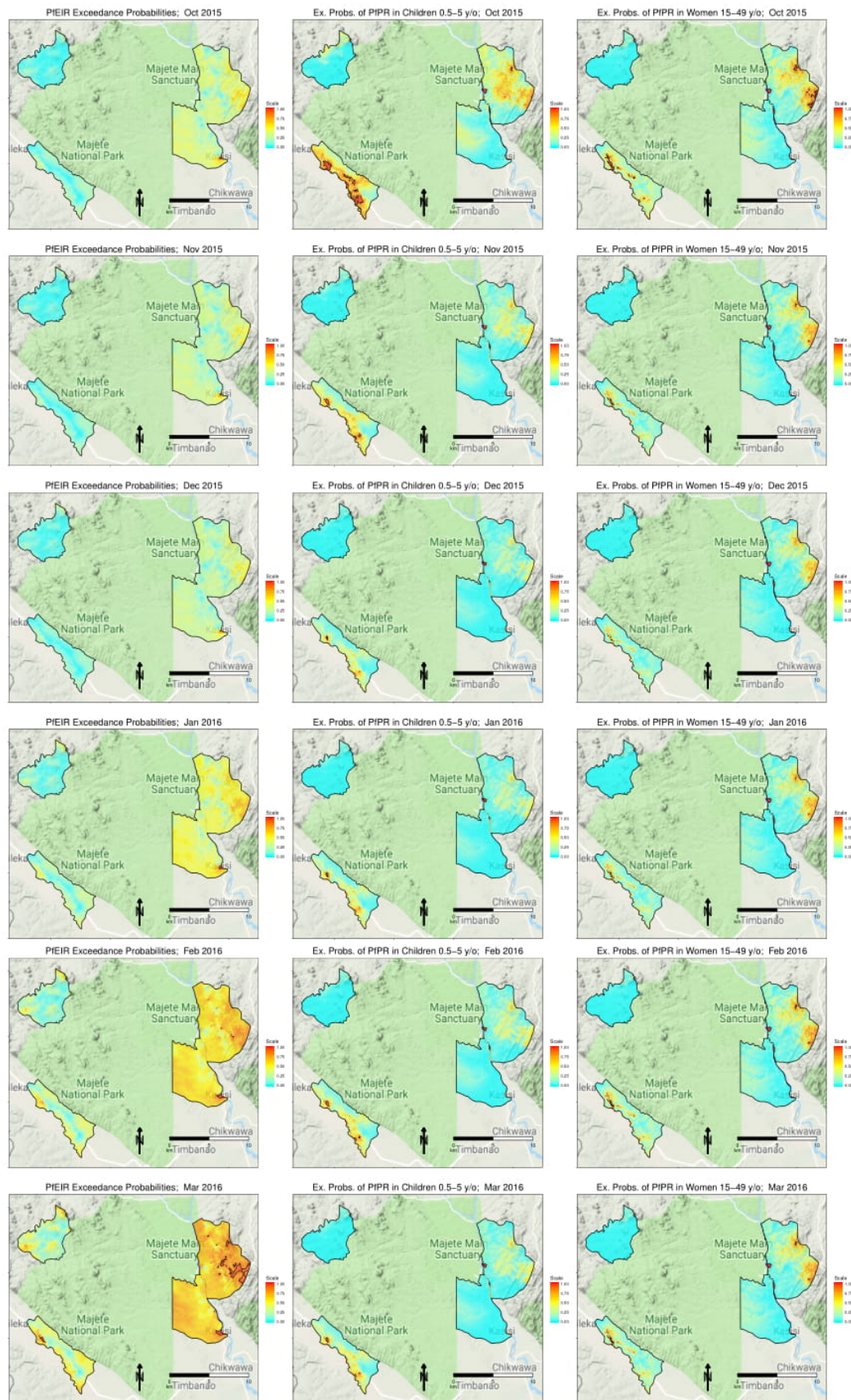
FIGURE 2.12: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2015 to March 2016.
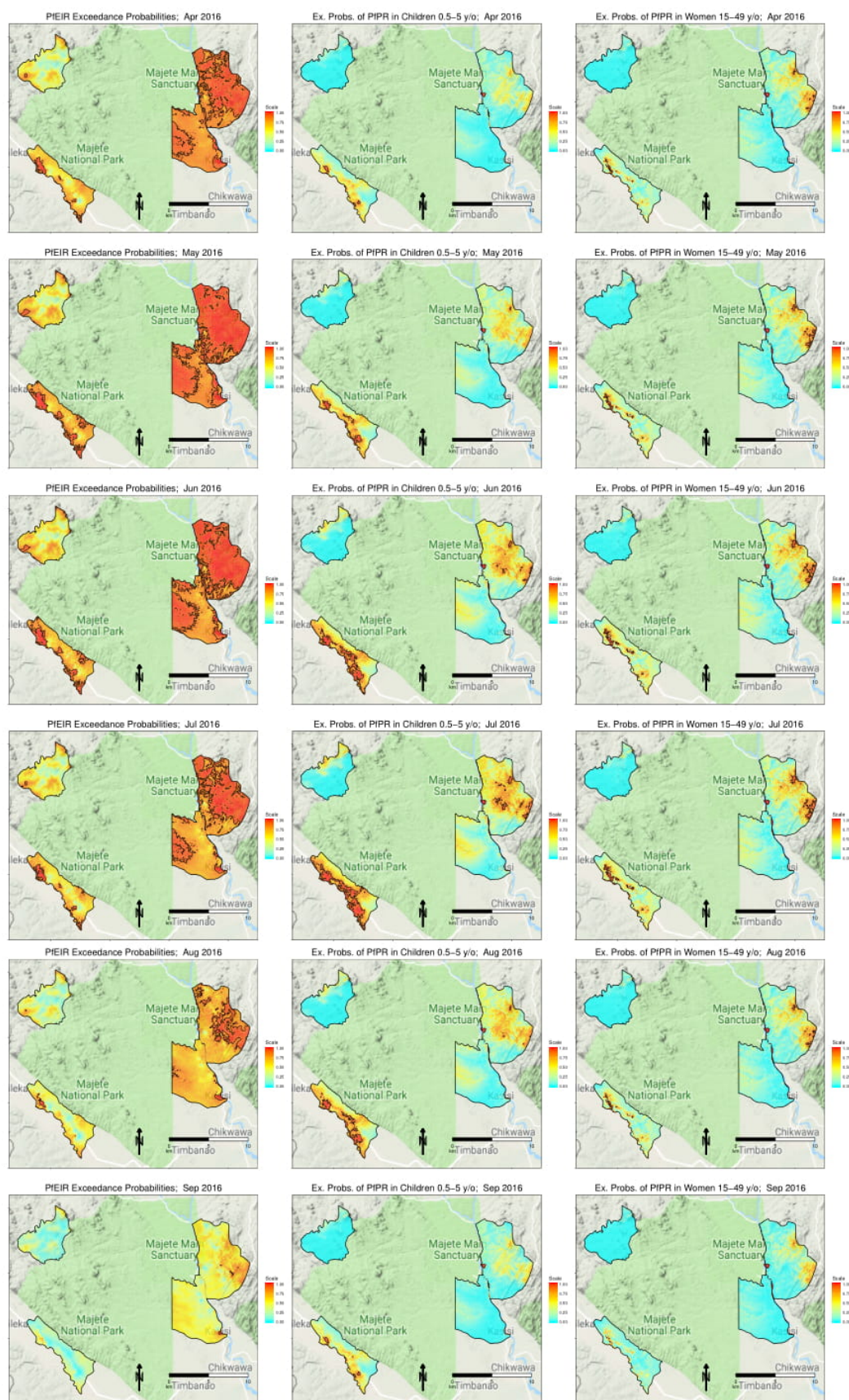
FIGURE 2.13: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2016 to September 2016.
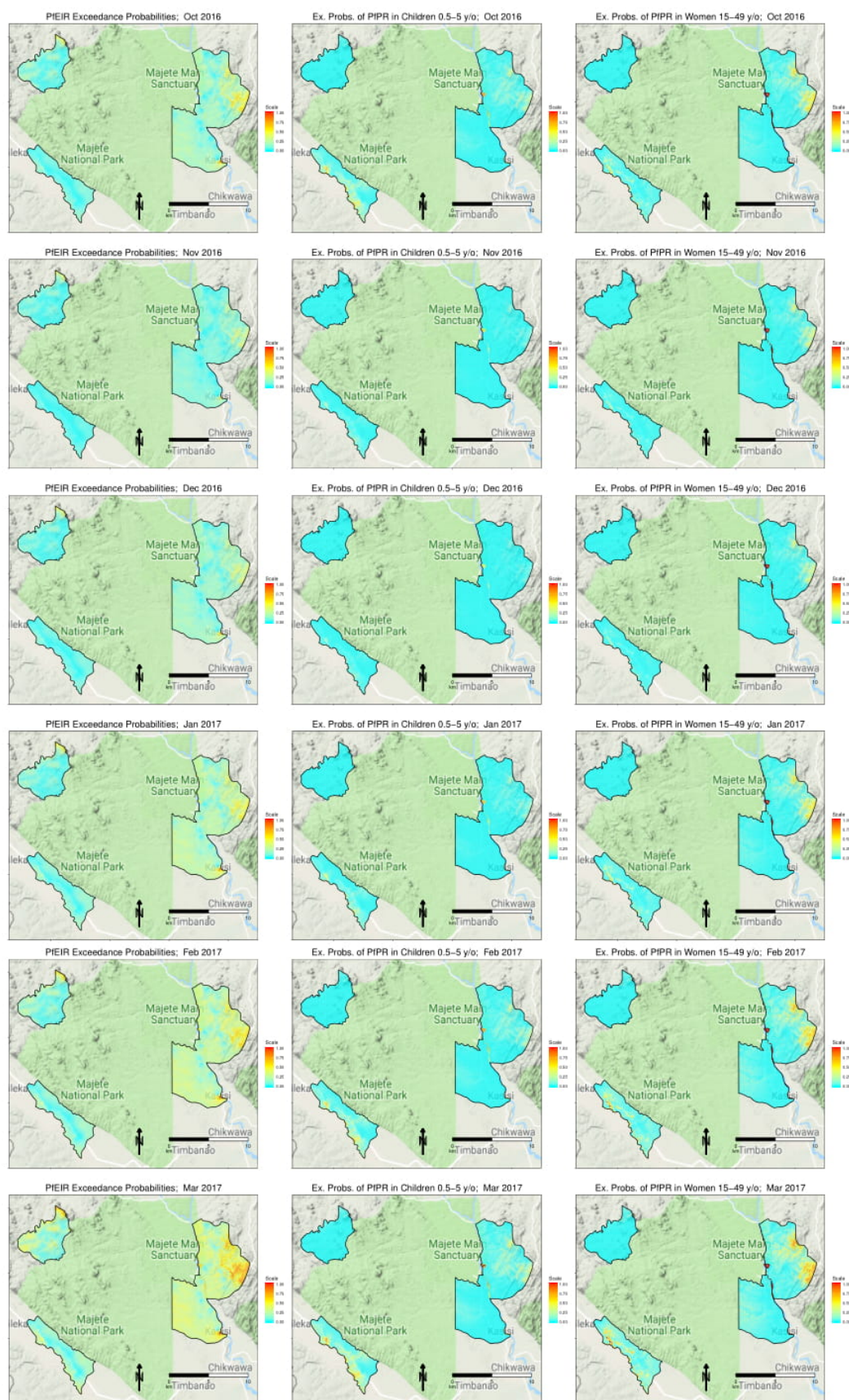
FIGURE 2.14: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2016 to March 2017.
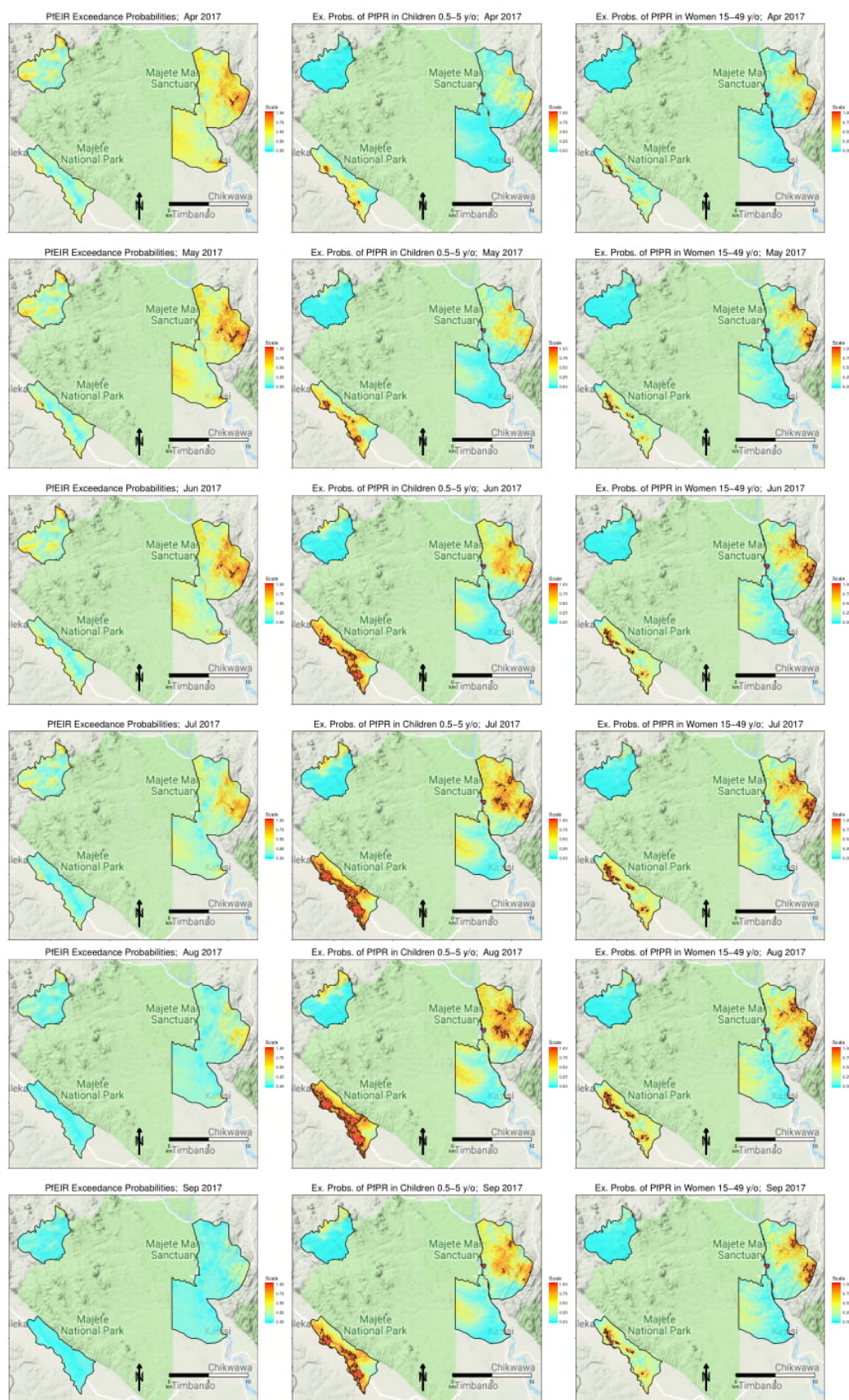
FIGURE 2.15: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from April 2017 to September 2017.
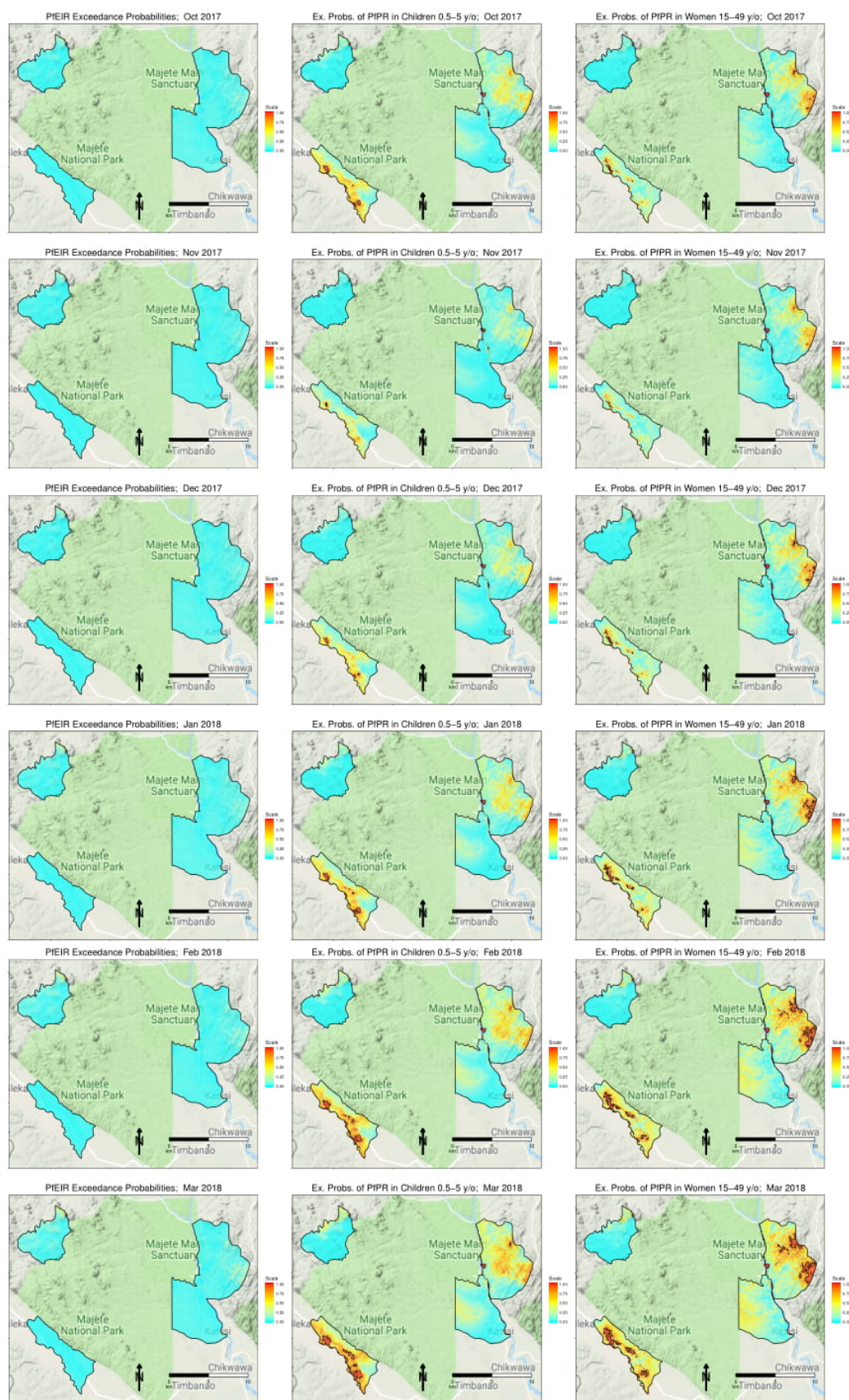
FIGURE 2.16: Exceedance probabilities of PfEIR (left panel), PfPR in children 0.5-5 y/o (middle panel) and women 15-49 y/o (right panel) from October 2017 to March 2018.

S4 Appendix: Additional figures.
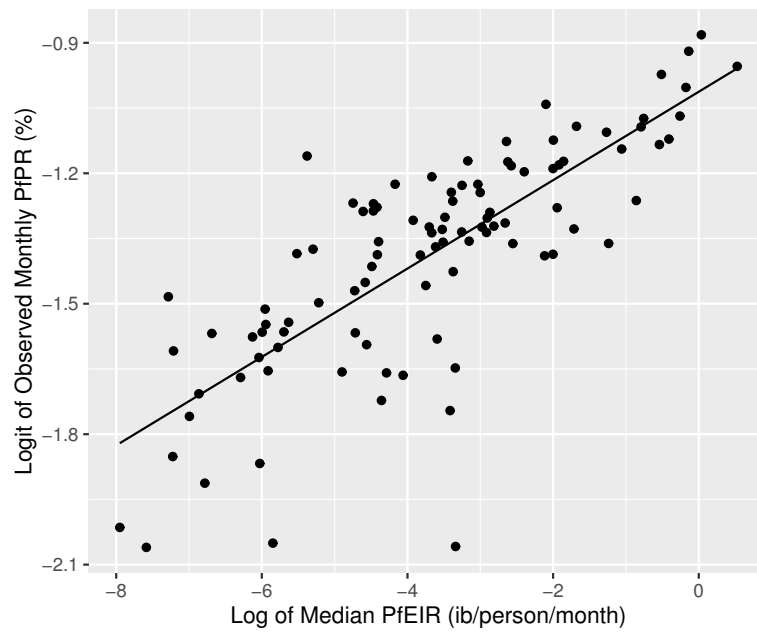
FIGURE 2.17: Scatter plots of the linear relationship between the logit of PfPR and the log of PfEIR.

References

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

Aron, J. L. and May, R. M. (1982). "The population dynamics of malaria." In *The population dynamics of infectious diseases: theory and applications*, pages 139–179. Springer.

Baird, J. K. (1995). "Host age as a determinant of naturally acquired immunity to Plasmodium falciparum." *Parasitology today*, 11(3):105–111.

Becker, N., Petrić, D., Boase, C., Lane, J., Zgomba, M., Dahl, C., and Kaiser, A. (2003). "Mosquitoes and their control." 2.

Beier, J. C., Killeen, G. F., and Githure, J. I. (1999). "Short report: entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa." *The American Journal of Tropical Medicine and Hygiene*, 61(1):109–113.

Beier, J. C., Oster, C. N., Onyango, F. K., Bales, J. D., Sherwood, J. A., Perkins, P. V., Chumo, D. K., Koech, D. V., Whitmire, R. E., Roberts, C. R., et al. (1994). "Plasmodium falciparum incidence relative to entomologic inoculation rates at a site proposed for testing malaria vaccines in western Kenya." *The American journal of tropical medicine and hygiene*, 50(5):529–536.

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(7572):207.

Boreham, P., Chandler, J., and Jolly, J. (1978). "The incidence of mosquitoes feeding on mothers and babies at Kisumu, Kenya." *The Journal of tropical medicine and hygiene*, 81(4):63–67.

Bousema, T., Youssef, R. M., Cook, J., Cox, J., Alegana, V. A., Amran, J., Noor, A. M., Snow, R. W., and Drakeley, C. (2010). "Serologic markers for detecting malaria in areas of low endemicity, Somalia, 2008." *Emerging infectious diseases*, 16(3):392.

Bryan, J. H. and Smalley, M. (1978). "The use of ABO blood groups as markers for mosquito biting studies." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 72(4):357–360.

Chen, S.-C., Liao, C.-M., Chio, C.-P., Chou, H.-H., You, S.-H., and Cheng, Y.-H. (2010). "Lagged temperature effect with mosquito transmission potential explains dengue variability in southern Taiwan: insights from a statistical analysis." *Science of the total environment*, 408(19):4069–4075.

Chipeta, M., Terlouw, D., Phiri, K., and Diggle, P. (2017). "Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure." *Environmetrics*, 28(1):e2425.

Chipeta, M. G., Terlouw, D. J., Phiri, K. S., and Diggle, P. J. (2016). "Adaptive geostatistical design and analysis for prevalence surveys." *Spatial Statistics*, 15:70–84.

Ciota, A. T., Matacchiero, A. C., Kilpatrick, A. M., and Kramer, L. D. (2014). "The effect of temperature on life history traits of Culex mosquitoes." *Journal of medical entomology*, 51(1):55–62.

Clyde, D. and Shute, G. (1958). "Selective feeding habits of anophelines amongst Africans of different ages." *The American journal of tropical medicine and hygiene*, 7(5):543–545.

Cook, J., Kleinschmidt, I., Schwabe, C., Nseng, G., Bousema, T., Corran, P. H., Riley, E. M., and Drakeley, C. J. (2011). "Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, equatorial Guinea." *PloS one*, 6(9):e25137.

Corran, P., Coleman, P., Riley, E., and Drakeley, C. (2007). "Serology: a robust indicator of malaria transmission intensity?" *Trends in parasitology*, 23(12):575–582.

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

Derua, Y. A., Alifrangis, M., Hosea, K. M., Meyrowitsch, D. W., Magesa, S. M., Pedersen, E. M., and Simonsen, P. E. (2012). "Change in composition of the Anopheles gambiae complex and its possible implications for the transmission of malaria and lymphatic filariasis in north-eastern Tanzania." *Malaria journal*, 11(1):188.

Diggle, P. J. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications.* Chapman and Hall/CRC.

Dletz, K. et al. (1974). "A malaria model tested in the African savanna." *Bulletin of the World Health Organization*, 50:347–357.

Doolan, D. L., Dobaño, C., and Baird, J. K. (2009). "Acquired immunity to malaria." *Clinical microbiology reviews*, 22(1):13–36.

Drakeley, C., Corran, P., Coleman, P., Tongren, J., McDonald, S., Carneiro, I., Malima, R., Lusingu, J., Manjurano, A., Nkya, W., et al. (2005). "Estimating medium-and long-term trends in malaria transmission by using serological markers of malaria exposure." *Proceedings of the National Academy of Sciences*, 102(14):5108–5113.

Gao, Q., Wang, F., Lv, X., Cao, H., Zhou, J., Su, F., Xiong, C., and Leng, P. (2018). "Comparison of the human-baited double net trap with the human landing catch for Aedes albopictus monitoring in Shanghai, China." *Parasites & vectors*, 11(1):483.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010). "Distributed lag non-linear models." *Statistics in medicine*, 29(21):2224–2234.

Giorgi, E. and Diggle, P. (2016). Prevmap: an r package for prevalence mapping.

Giorgi, E., Diggle, P. J., Snow, R. W., and Noor, A. M. (2018). "Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys." *International Statistical Review.*

Hawkes, F., Manin, B. O., Ng, S. H., Torr, S. J., Drakeley, C., Chua, T. H., and Ferguson, H. M. (2017). "Evaluation of electric nets as means to sample mosquito vectors host-seeking on humans and primates." *Parasites & vectors*, 10(1):338.

Hiscox, A., Otieno, B., Kibet, A., Mweresa, C. K., Omusula, P., Geier, M., Rose, A., Mukabana, W. R., and Takken, W. (2014). "Development and optimization of the Suna trap as a tool for mosquito monitoring and control." *Malaria journal*, 13(1):257.

Jenson, S. K. and Domingue, J. O. (1988). "Extracting topographic structure from digital elevation data for geographic information system analysis." *Photogrammetric engineering and remote sensing*, 54(11):1593–1600.

John, C. C., Moormann, A. M., Pregibon, D. C., Sumba, P. O., McHugh, M. M., Narum, D. L., Lanar, D. E., Schluchter, M. D., and Kazura, J. W. (2005). "Correlation of high levels of antibodies to multiple pre-erythrocytic Plasmodium falciparum antigens and protection from infection." *The American journal of tropical medicine and hygiene*, 73(1):222–228.

Joshua, M. K., Ngongondo, C., Monjerezi, M., Chipungu, F., Liwenga, E., Majule, A. E., Stathers, T., and Lamboll, R. (2016). "Climate change in semi-arid Malawi: Perceptions, adaptation strategies and water governance." *Jàmbá: Journal of Disaster Risk Studies*, 8(3):1–10.

Kabiru, E. (1994). "Sporozoite challenge and transmission patterns as determinants of occurrence of severe malaria in residents of Kilifi district, Kenya." *Nairobi: University of Nairobi*.

Kitau, J., Oxborough, R. M., Tungu, P. K., Matowo, J., Malima, R. C., Magesa, S. M., Bruce, J., Mosha, F. W., and Rowland, M. W. (2012). "Species shifts in the Anopheles gambiae complex: do LLINs successfully control Anopheles arabiensis?" *PLoS one*, 7(3):e31481.

Ladeia-Andrade, S., Ferreira, M. U., de Carvalho, M. E., Curado, I., and Coura, J. R. (2009). "Age-dependent acquisition of protective immunity to malaria in riverine populations of the Amazon Basin of Brazil." *The American journal of tropical medicine and hygiene*, 80(3):452–459.

Lin Ouédraogo, A., Gonçalves, B. P., Gnémé, A., Wenger, E. A., Guelbeogo, M. W., Ouédraogo, A., Gerardin, J., Bever, C. A., Lyons, H., Pitroipa, X., et al. (2015). "Dynamics of the human infectious reservoir for malaria determined by mosquito

feeding assays and ultrasensitive malaria diagnosis in Burkina Faso." *The Journal of infectious diseases*, 213(1):90–99.

Loetti, V., Schweigmann, N., and Burroni, N. (2011). "Development rates, larval survivorship and wing length of Culex pipiens (Diptera: Culicidae) at constant temperatures." *Journal of Natural history*, 45(35-36):2203–2213.

Lowe, R., Chirombo, J., and Tompkins, A. M. (2013). "Relative importance of climatic, geographic and socio-economic determinants of malaria in Malawi." *Malaria journal*, 12(1):416.

Lozano-Fuentes, S., Hayden, M. H., Welsh-Rodriguez, C., Ochoa-Martinez, C., Tapia-Santos, B., Kobylinski, K. C., Uejio, C. K., Zielinski-Gutierrez, E., Delle Monache, L., Monaghan, A. J., et al. (2012). "The dengue virus mosquito vector Aedes aegypti at high elevation in Mexico." *The American journal of tropical medicine and hygiene*, 87(5):902–909.

Madder, D., Surgeoner, G., and Helson, B. (1983). "Number of generations, egg production, and developmental time of Culex pipiens and Culex restuans (Diptera: Culicidae) in southern Ontario." *Journal of medical entomology*, 20(3):275–287.

Mathanga, D. P., Walker, E. D., Wilson, M. L., Ali, D., Taylor, T. E., and Laufer, M. K. (2012). "Malaria control in Malawi: current status and directions for the future." *Acta tropica*, 121(3):212–217.

Mbogo, C. N., Snow, R. W., Khamala, C. P., Kabiru, E. W., Ouma, J. H., Githure, J. I., Marsh, K., and Beier, J. C. (1995). "Relationships between Plasmodium falciparum transmission by vector populations and the incidence of severe disease at nine sites on the Kenyan coast." *The American journal of tropical medicine and hygiene*, 52(3):201–206.

Mburu, M. M., Zembere, K., Hiscox, A., Banda, J., Phiri, K. S., Van Den Berg, H., Mzilahowa, T., Takken, W., and McCann, R. S. (2019). "Assessment of the Suna trap for sampling mosquitoes indoors and outdoors." *Malaria journal*, 18(1):51.

McCann, R. S., van den Berg, H., Diggle, P. J., van Vugt, M., Terlouw, D. J., Phiri, K. S., Di Pasquale, A., Maire, N., Gowelo, S., Mburu, M. M., et al. (2017). "Assessment of the effect of larval source management and house improvement on malaria transmission when added to standard malaria control strategies in southern Malawi: study protocol for a cluster-randomised controlled trial." *BMC infectious diseases*, 17(1):639.

Meza, F. C., Kreppel, K. S., Maliti, D. F., Mlwale, A. T., Mirzai, N., Killeen, G. F., Ferguson, H. M., and Govella, N. J. (2019). "Mosquito electrocuting traps for directly measuring biting rates and host-preferences of Anopheles arabiensis and Anopheles funestus outdoors." *Malaria journal*, 18(1):83.

Muirhead-Thomson, R. (1951). "Distribution of anopheline mosquito bites among different age groups." *British Medical Journal*, 1(4715):1114.

Mukabana, W. R., Mweresa, C. K., Otieno, B., Omusula, P., Smallegange, R. C., Van Loon, J. J., and Takken, W. (2012). "A novel synthetic odorant blend for trapping of malaria and other African mosquito species." *Journal of chemical ecology*, 38(3):235–244.

Mwandagalirwa, M. K., Levitz, L., Thwai, K. L., Parr, J. B., Goel, V., Janko, M., Tshefu, A., Emch, M., Meshnick, S. R., and Carrel, M. (2017). "Individual and household characteristics of persons with Plasmodium falciparum malaria in sites with varying endemicities in Kinshasa Province, Democratic Republic of the Congo." *Malaria journal*, 16(1):456.

Mwangangi, J. M., Mbogo, C. M., Orindi, B. O., Muturi, E. J., Midega, J. T., Nzovu, J., Gatakaa, H., Githure, J., Borgemeister, C., Keating, J., et al. (2013). "Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years." *Malaria journal*, 12(1):13.

Mzilahowa, T., Hastings, I. M., Molyneux, M. E., and McCall, P. J. (2012). "Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi." *Malaria journal*, 11(1):380.

Organization., W. W. H. (1975). "Manual on practical entomology in malaria. Part II. Methods and techniques." *WHO Offset Publ. 13.*

Patouillard, E., Griffin, J., Bhatt, S., Ghani, A., and Cibulskis, R. (2017). "Global investment targets for malaria control and elimination between 2016 and 2030." *BMJ global health*, 2(2):e000176.

Pombi, M., Jacobs, F., Verhulst, N. O., Caputo, B., Della Torre, A., and Takken, W. (2014). "Field evaluation of a novel synthetic odour blend and of the synergistic role of carbon dioxide for sampling host-seeking Aedes albopictus adults in Rome, Italy." *Parasites & vectors*, 7(1):580.

Reinert, J. F. et al. (2000). "New classification for the composite genus Aedes (Diptera: Culicidae: Aedini), elevation of subgenus Ochlerotatus to generic rank, reclassification of the other subgenera, and notes on certain subgenera and species." *JOURNAL-AMERICAN MOSQUITO CONTROL ASSOCIATION*, 16(3):175–188.

Roca-Feltrer, A., Kwizombe, C. J., Sanjoaquin, M. A., Sesay, S. S., Faragher, B., Harrison, J., Geukers, K., Kabuluzi, S., Mathanga, D. P., Molyneux, E., et al. (2012a). "Lack of decline in childhood malaria, Malawi, 2001–2010." *Emerging infectious diseases*, 18(2):272.

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

Rogerson, S. J., Van den Broek, N., Chaluluka, E., Qongwane, C., Mhango, C., and Molyneux, M. E. (2000). "Malaria and anemia in antenatal women in Blantyre, Malawi: a twelve-month survey." *The American journal of tropical medicine and hygiene*, 62(3):335–340.

Roiz, D., Ruiz, S., Soriguer, R., and Figuerola, J. (2014). "Climatic effects on mosquito abundance in Mediterranean wetlands." *Parasites & vectors*, 7(1):333.

Ross, R. (1911). *The prevention of malaria.* John Murray; London.

Ruan, S., Xiao, D., and Beier, J. C. (2008). "On the delayed Ross–Macdonald model for malaria transmission." *Bulletin of mathematical biology*, 70(4):1098–1114.

Rumisha, S. F., Smith, T., Abdulla, S., Masanja, H., and Vounatsou, P. (2014). "Modelling heterogeneity in malaria transmission using large sparse spatio-temporal entomological data." *Global health action*, 7.

Shone, S. M., Ferrao, P. N., Lesser, C. R., Norris, D. E., and Glass, G. E. (2001). "Analysis of mosquito vector species abundances in Maryland using geographic information systems." *Annals of the New York Academy of Sciences*, 951(1):364–368.

Smith, A. et al. (1956). "The attractiveness of an adult and child to A. gambiae." *East African medical journal*, 33(10).

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

Spiers, A., Mzilahowa, T., Atkinson, D., and McCall, P. (2002). "The malaria vectors of the lower Shire Valley, Malawi." *Malawi Medical Journal*, 14(1):4–7.

Tachikawa, T., Hato, M., Kaku, M., and Iwasaki, A. (2011). "Characteristics of ASTER GDEM version 2." In *Geoscience and remote sensing symposium (IGARSS), 2011 IEEE international*, pages 3657–3660. IEEE.

Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I. (1991). "On the extraction of channel networks from digital elevation data." *Hydrological processes*, 5(1):81–100.

Thomas, T. (1951). "Biting activity of Anopheles gambiae." *British Medical Journal*, 2(4744):1402.

Tusting, L. S., Bousema, T., Smith, D. L., and Drakeley, C. (2014). "Measuring changes in Plasmodium falciparum transmission: precision, accuracy and costs of metrics." In *Advances in parasitology*, volume 84, pages 151–208. Elsevier.

Walldorf, J. A., Cohee, L. M., Coalson, J. E., Bauleni, A., Nkanaunena, K., Kapito-Tembo, A., Seydel, K. B., Ali, D., Mathanga, D., Taylor, T. E., et al. (2015). "School-age children are a reservoir of malaria infection in Malawi." *PLoS One*, 10(7):e0134061.

Walton, G. (1947). "On the Control of Malaria in Freetown, Sierra Leon: I.—Plasmodium Falciparum and Anopheles Gambiae in Relation to Malaria Occurring in Infants." *Annals of Tropical Medicine & Parasitology*, 41(3-4):380–407.

World Health Organization (2015). *Global technical strategy for malaria 2016-2030*. World Health Organization.

Zamawe, C. O., Nakamura, K., Shibanuma, A., and Jimba, M. (2016). "The effectiveness of a nationwide universal coverage campaign of insecticide-treated bed nets on childhood malaria in Malawi." *Malaria journal*, 15(1):505.

Zhao, X., Chen, F., Feng, Z., Li, X., and Zhou, X.-H. (2014). "Characterizing the effect of temperature fluctuation on the incidence of malaria: an epidemiological study in south-west China using the varying coefficient distributed lag non-linear model." *Malaria journal*, 13(1):192.

Zhou, Z., Mitchell, R. M., Kariuki, S., Odero, C., Otieno, P., Otieno, K., Onyona, P., Were, V., Wiegand, R. E., Gimnig, J. E., et al. (2016). "Assessment of submicroscopic infections and gametocyte carriage of Plasmodium falciparum during peak malaria transmission season in a community-based cross-sectional survey in western Kenya, 2012." *Malaria journal*, 15(1):421.

Chapter 3

Paper 2. Geostatistical Modelling of the Association between Malaria and Child Growth in Africa

Benjamin Amoah[1]*, Emanuele Giorgi[1], Daniel J. Hayes[2], Stef van Buuren[3], Peter J. Diggle[1]

*Correspondence: Benjamin Amoah, CHICAS, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, LA1 4YB, UK.

E-mail: b.amoah@lancaster.ac.uk

3.1 Summary

Background

Undernutrition among children under five years of age continues to be a public health challenge in many low- and middle-income countries (LMICs) and can lead to growth stunting. Infectious diseases may also affect child growth, however, the association between malaria and stunting remains unclear. In this paper, we study this association using data from 20 Demographic and Health Surveys (DHS) conducted in 13 African countries. Our objective is to make inference on the association between malaria incidence during the first year of life and height-for-age Z-scores (HAZs).

Methods

We develop a geostatistical model for HAZs as a function of both measured and unmeasured child-specific and spatial risk factors. We visualize stunting risk in each of the 20 analysed surveys by mapping the predictive probability that HAZ is below -2. Finally, we carry out a meta-analysis by modelling the estimated effects of malaria incidence on HAZ from each DHS as a linear regression on national development indicators from the World Bank.

Results

A non-spatial univariate linear regression of HAZ on malaria incidence showed a negative association in 18 out of 20 surveys. However, after adjusting for spatial risk factors and controlling for confounding effects, we found a weaker association between HAZ and malaria, with a mix of positive and negative estimates, of which 3 out of 20 are significantly different from zero at the conventional 5% level. The meta-analysis showed that this variation in the estimated effect of malaria incidence on HAZ is significantly associated with the amount of arable land.

Conclusion

Confounding effects on the association between malaria and stunting vary both by country and over time. Geostatistical analysis provides a useful framework that allows to account for unmeasured spatial confounders. Establishing whether the association between malaria and stunting is causal would require longitudinal follow-up data on individual children.

Keywords

Child growth; exceedance probability; geostatistics; malaria; stunting.

Author summary

Why was this study done?

> Many studies have investigated the association between malaria and child growth is, but have reported contrary results, namely positive, negative and no associations. However, possible reasons for the mix of positive, negative and no associations have not been investigated.

> Child growth may be spatially structured, and failure to account for any spatial correlations present may lead to unreliable inferences on any regression relationships.

What did the researchers do and find?

> We used 20 DHS datasets from 13 African countries to model height-for-age Z-scores as a function of malaria and other measured and unmeasured (spatial) risk factors.

> We found that the regression coefficients corresponding to malaria were a mix of significant positives and negatives, and no associations.

> We modelled the regression coefficients corresponding to malaria as a function of World Bank National Development Indicators and found that arable land significantly explained 26% of the variability in the estimated coefficients.

What do these findings mean?

> Our results suggest that geo-political difference in the economy, agriculture and environment of countries could account for differences in the signs and magnitude of the association between malaria and child growth, and should be controlled for when investigating this association.

> Our study also presents stunting risk maps based on each of the 20 DHS datasets, which can be used for spatial targeting of stunting interventions.

3.2 Background

Undernutrition underlies 45% of all child deaths among children under five years (Black et al., 2013). A very low height-for-age, usually referred to as stunting, is an important indicator that reflects the cumulative effects of undernutrition and disease infections (UNICEF et al., 2013). Stunted children are more prone to illness and premature death. Stunting among children is known to be associated with poor cognitive development (Daniels and Adair, 2004, Walker et al., 2005). Long-term consequences of stunting include lower adult economic productivity, higher risks of ill-health and, among women with short stature, an increased risk of death during delivery (Cunha and Heckman, 2007, Currie, 2000, 2008, Heckman et al., 2006). Globally, the rate of stunting in children under five years reduced from 32.7% (198 million) in year 2000 to 23.2% (156 million) in year 2015 (UNICEF et al., 2016). In Africa however, the rates reduced from 38% in 2000 to 32% in 2015, representing more limited progress than in Asia, Latin America and the Caribbean where stunting rates dropped by more than one third over the same period (UNICEF et al., 2016). In many low- and middle-income countries (LMICs), over 50% of 12-23 months old children are stunted (Marriott et al., 2012, Stevens et al., 2012, Victora et al., 2010). In 2014, less than half of all children under five years lived in LMICs, yet

these countries accounted for two-thirds of all stunted children globally (UNICEF et al., 2015). Although the main risk factor for stunting is inadequate nutrition, exposure to infectious diseases may also lead to an increase in stunting risk (Custodio et al., 2009, Verhoef et al., 2002). However, there are indirect effects of malaria not fully understood (Holding and Kitsao-Wekulo, 2004, Shanks et al., 2008), and it is unclear if part of the stunting burden can be attributed to malaria.

Malaria is still a public health threat, although the ongoing global fight against it has resulted in 50% decrease in the infection prevalence and 40% decrease in the clinical incidence in the endemic region of Africa between 2000 and 2015 (Bhatt et al., 2015). In 2015, there were an estimated 214 million malaria cases and 438 thousand deaths from malaria worldwide, of which 88% occurred in sub-Saharan Africa and 70% in children under the age of 5 years, with 10% of all deaths in children under the age of 5 years due to malaria (World Health Organization, 2016). In 2017, similar global estimates were still reported: 216 million malaria cases and 445 thousand malaria deaths, of which 91% occurred in sub-Saharan Africa, with most of the deaths still occuring in children under 5 years (World Health Organization, 2017). The association between malaria and stunting is unclear and still a matter of debate, with studies showing contrasting results. For example, maternal malaria has been found to impact on child growth (Kalanda et al., 2005), with infants born to women who experienced malaria during pregnancy having an increased risk of impaired height and weight gain (De Beaudrap et al., 2016, Guyatt and Snow, 2004, McGregor et al., 1983, Uddenfeldt Wort et al., 2004). The risk of stunting has been found to increase for every malaria episode (Kang et al., 2013). On the other hand, some studies suggest that stunting may modulate susceptibility to malaria, especially during the first two years of life (Nyakeriga et al., 2004, Olney et al., 2009). Whilst some studies suggest that stunted children may be at higher risk of developing malaria episodes (Deen et al., 2002), others report that stunting may have a protective effect against malaria (Genton et al., 1998, Murray et al., 1978). In other studies, instead, no association is found (Müller et al., 2003, Snow et al., 1991). More recently, Fink et al. (2013) found a significant effect of malaria exposure on cognitive development and socio-emotional development, but not on height, for which they report an estimated effect of about 3.000 and associated 95% confidence interval (-11.350, 4.606).

The height-for-age Z-score (HAZ) measures the deviation from heights based on the World Health Organization (WHO) growth standards (Onis, 2006, World Health Organization et al., 2006) and are comparable across ages and gender. Values of HAZ below -2 are used as an indicator of stunted growth. In this paper, we analyse data from 20 Demographic and Health Surveys (DHS) conducted in Senegal, Mozambique, Ghana, Burkina Faso, Zambia, Malawi, Rwanda, Cote d'Ivoire, Burundi, Liberia, Namibia, Togo and Tanzania to pursue the following objectives: (1) to investigate the association between malaria and HAZ by developing a geostatistical framework that accounts for both measured and unmeasured risk factors for stunting; (2) to understand how such association varies across the African countries considered in this study; (3) to map the risk of stunting. We also discuss the limitations of this study and provide a detailed description on how the proposed modelling framework could be further extended to a longitudinal setting. To the best of our knowledge, this is the first study that investigates the association between the geographical distribution of malaria and HAZ using a model-based geostatistical approach.

3.3 Methods

3.3.1 Data

DHS are nationally representative household surveys that are generally repeated every five years and provide information on a range of health and population indicators, including anthropometric information. The DHS methodology is usually based on a stratified two-stage cluster design. At the first stage, enumeration areas are drawn from census files. At the second stage, for each enumeration area selected, samples of households are drawn from an updated list of households to form groups of households known as sampling clusters. The GPS location of the center of each sampling cluster is taken as the cluster location. Each child is allocated to a spatially-referenced sampling cluster. We analyse data from 20 DHS conducted between 2003 and 2014 (DHS Surveys, 2014). Table 3.1 shows the number of clusters and individuals for each survey. The average

number of children per cluster varies from one survey to another, with the highest value of about 21.7 in Burkia Faso in 2003 and the lowest of about 5.7 in Malawi in 2010.

TABLE 3.1: Sample size summaries for the analysed DHS data indicating the country, year of survey, number of children, number of sampled clusters, and average number of children per cluster.

Country	Year	No. of children	No. of clusters	Average no. of children per cluster
Senegal	2005	2710	355	7.6
Senegal	2011	3694	384	9.6
Mozambique	2011	9595	609	15.8
Ghana	2003	3010	393	7.7
Ghana	2008	2350	393	6.0
Ghana	2014	2671	410	6.5
Burkina Faso	2003	8581	396	21.7
Burkina Faso	2010	6290	540	11.6
Zambia	2007	5243	317	16.5
Zambia	2014	4635	303	15.3
Malawi	2004	6238	386	16.2
Malawi	2010	4623	811	5.7
Rwanda	2005	3692	455	8.1
Cote d'Ivoire	2007	3305	288	11.5
Burundi	2010	3449	376	9.2
Liberia	2007	4197	270	15.5
Liberia	2013	3206	319	10.1
Namibia	2007	3669	484	7.6
Togo	2014	3209	328	9.8
Tanzania	2010	6581	453	14.5

The variables used in the analysis are the following.

Child-specific variables. Data on a child's height, age and gender, family's wealth index and mother's education level were obtained from the DHS for all sampled children aged less than 5 years. Families' wealth indices are constructed using principal component analysis (Rutstein et al., 2004) on household's ownership of television, radio, watch, vehicles and agricultural land, type and number of animals owned, bank account, materials used for housing construction, type of water access and sanitation facilities (Rutstein et al., 2004).

Urban extent indicator. We use information on urban extents, available as raster data at a spatial resolution of 1 km by 1 km, from the Global Rural-Urban Mapping Project (Center for International Earth Science Information Network – CIESIN – Columbia University et al., 2011). This variable is a binary indicator that classifies each spatial

grid cell as urban or rural, based on a combination of population counts, settlement points, and presence of night-time lights.

Estimated malaria incidence rates. We use raster data on estimated *Plasmodium falciparum* incidence as obtained from a Bayesian spatio-temporal model implemented by the Malaria Atlas Project (Bhatt et al., 2015). The data are available at a temporal resolution of one year, from 2000 to 2015, and a spatial resolution of $0.05° \times 0.05°$. More specifically, the estimated *Plasmodium falciparum* malaria incidence at pixel-level is the predicted average clinical incidence rate per child per year in the age cohorts 0-5 years. A clinical malaria episode is an attributable febrile episode with a body temperature in excess of $37.5°$C. Multiple bouts of symptoms occurring within a 30-day period are counted as a single episode. These data did not included uncertainties in the malaria incidence estimates. Bhatt et al. (2015) provided these data specifically for the children under five years to the authors.

3.3.2 Model formulation and spatial prediction

Accounting for spatial effects is crucial in order to deliver valid inferences on the regression coefficients (Thomson et al., 1999). Model-based geostatistics allows us to incorporate both explained and unexplained (residual) spatial variation in HAZ and to predict the risk of stunting throughout a geographical area of interest.

Let Y_{ij} denote the HAZ for the jth sampled child at the cluster location x_i. Visual checks of histograms of Y_{ij} for each DHS showed satisfactory Gaussian distribution. We distinguish between two sources of variation in HAZ: between-cluster variation, induced by spatially varying risk factors; and within-cluster variation due to child-specific characteristics. Each of these components depends on both measured and unmeasured risk factors. In order to account for the latter, we define a hierarchical linear model as follows. Let $S(x_i)$ denote a stationary Gaussian process and U_i represent mutually independent zero-mean Gaussian variables with common variance τ^2. We assume that, conditionally on $S(x_i)$ and U_i, the Y_{ij} are Gaussian variables with means $\mu_j(x_i)$ and

variance ω^2, where

$$\mu_j(x_i) = e_{ij}^\top \gamma + d(x_i)\beta + \delta \mathcal{M}_{ij} + f(\mathcal{A}_{ij}) + U_i + S(x_i), \text{ for } \quad i = 1, \ldots, n \quad \text{(3.1)}$$
$$j = 1, \ldots, m_i.$$

In (3.1), n is the number of cluster locations and m_i is the number of individuals at cluster location x_i. In (3.1) we also distinguish between three types of explanatory variables: e_{ij}, a vector of child-specific explanatory variables, including sex, family's wealth index and mother's education level; $d(x_i)$, a spatial indicator variable which takes values 1, if location x_i is classified as urban and 0 if rural; $\mathcal{M}_{ij}$, the estimated malaria incidence at location x_i during the first year of life of the j-th child. The parameters γ, β and δ are the regression parameters associated with each of the three types of explanatory variables, whilst $f(\mathcal{A})$ is a cubic spline function of age, $\mathcal{A}$, with knots at 12 and 24 months.

Our objective is to make inference on the parameter δ, which quantifies the effect of malaria incidence in the first year of life on HAZ. Our assumption is that malaria has a lagged effect on height and, therefore, we use the incidence of malaria during the first year of life to determine the strength of this association. In the remainder of the paper, we shall refer to the parameter δ and the variable $\mathcal{M}_{ij}$ in (3.1) as the effect of malaria on HAZ and malaria incidence, respectively.

In (3.1), the unstructured random effect U_i conflates two sources of residual variation: spatial variation on a scale smaller than the minimum observed distance between clusters; and unexplained unstructured variation at cluster level.

The spatially structured residuals $S(x)$ are modelled as a zero-mean stationary and isotropic Gaussian process with variance σ^2 and exponential correlation function given by

$$\rho(u; \phi) = \exp(-u/\phi), \quad \text{(3.2)}$$

where u is the Euclidean distance between any two locations. The scale parameter ϕ regulates the rate at which the spatial correlation decays with increasing distance u.

We map the risk of stunting for male children, 24 months old, using the predictive probability that HAZ is below -2 over a $0.05° \times 0.05°$ grid. We integrate out the effect of maternal education and wealth index using the following Monte Carlo approach. We generate 10,000 samples from the joint distribution of these two variables and, conditionally on these, we then simulate values of HAZ. The stunting risk is then computed by taking the proportion of simulated HAZ samples that are below -2.

More details on the computational implementation and on the mapping of stunting risk are given in Additional File 1.

3.3.3 Model Validation

To check the validity of the adopted spatial correlation structure for the data, we carry out the following Monte Carlo procedure. We simulate 1,000 empirical variograms under the fitted model and then use these to compute 95% confidence intervals at any given spatial distance of the variogram. If the empirical variogram obtained from the data falls within the 95% tolerance bandwidth, we conclude that the adopted spatial correlation function is compatible with the data. If, instead, that falls outside the 95% tolerance bandwidth, then the data show evidence against the fitted model. More details are provided in Additional File 1.

3.3.4 Understanding the variation in the effect of malaria on HAZ

We carry out a meta-analysis in order to understand the variation in the estimates of the parameter of interest δ, from all the 20 DHS. Let $\hat{\delta}_k$ and s_k denote the maximum likelihood estimate of δ and its standard error, respectively, for $k = 1, \ldots, 20$. We then model $\hat{\delta}_k$ using a weighted least squares fit to the regression model

$$\hat{\delta}_k = \alpha_0 + \alpha_1 v_k + Z_k, \tag{3.3}$$

where v_k is a World Bank African development indicator (World Bank Indicators, 2014) associated with the country and year of the k-th survey, and the Z_k are independent

Gaussian variables with mean zero and variance s_k^2. We select eleven development indicators belonging to the categories of "Agriculture and rural development", "Climate change", "Economy and growth", "Education" and "Environment". A full list of the indicators is given in Additional File 2'.

3.4 Results

3.4.1 Non-spatial analysis

Figure 3.1 shows box-plots of HAZ by categories of family's wealth indices and mother's education level for all surveys combined. We assign integer scores 1 to 5 to the five levels of family wealth from very poor to very wealthy; and scores 1 to 6 to the six levels of mothers education, from no education to higher education. As expected, the box-plots show that the median HAZ tends to increase with increasing levels of wealth and education.

We then investigate the marginal association between malaria incidence and HAZ. Figure 3.2 shows the observed HAZ against malaria incidence, where the solid line is obtained from the least squares fit of a univariate linear model. The dashed horizontal lines indicate HAZ levels of 2, 0 and -2. The dashed vertical lines separate $\mathcal{M}$ into terciles. We see that Malaria incidence takes a maximum value of about 1.5 for all surveys, except Namibia in 2007, where this is about 0.7. We also note that for the surveys in Senegal in 2005, Mozambique in 2011, Ghana in 2003-2008-2014 and Zambia in 2007, the variation in $\mathcal{M}$ is evenly distributed, whereas it is more skewed for Senegal in 2011, Burkina Faso in 2003-2010, Malawi in 2004 and Namibia in 2007. Except for Rwanda in 2005, Zambia in 2014 and Malawi in 2010, in all the remaining 17 surveys we observe that HAZ decreases with increasing values of $\mathcal{M}$. Figure 3.3 shows the least squares estimates and the corresponding 95% confidence intervals. The estimated regression coefficients are negative in 18 surveys, of which 16 are significantly different from zero at 5% level.
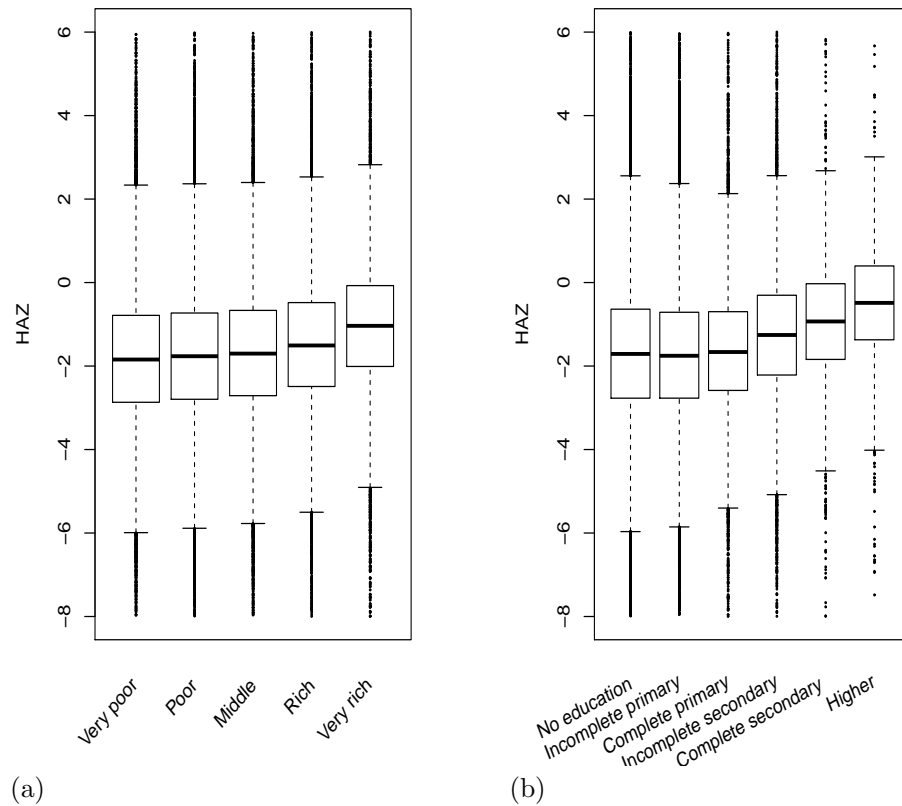
FIGURE 3.1: Box plots of height-for-age z-scores by family's wealth (a) and mother's level of education (b), pooled over all 20 surveys.

Figure 3.4 shows HAZ curves as functions of age, within each of the terciles groups of $\mathcal{M}$, as indicated in Figure 3.2. The fitted curves reflect the typical age-related pattern of HAZ in LMICs: after a decrease in HAZ during the first two years of life, child-growth slowly recovers but never reaches zero. This phenomenon, known as "growth faltering", has been widely observed; see, for example, (Allen, 1994, Rieger and Trommlerová, 2016, Stevens et al., 2012, Victora et al., 2010). We also observe that in Burkina Faso in 2003, Ghana in 2008, Malawi in 2004-2010 and Rwanda in 2005, HAZ curves by terciles groups of $\mathcal{M}$ are partly overlapping, whereas in the remaining 15 surveys, children in the first tercile of $\mathcal{M}$ have the highest levels of HAZ and children in the third tercile with the lowest levels of HAZ, irrespective of age. We further notice that in Burkina Faso in 2003, Burundi in 2010, Rwanda in 2005, Cote d'Ivoire in 2007 and Malawi in 2004, where median HAZ curves fall below the -2 threshold at about 24 months of age, the
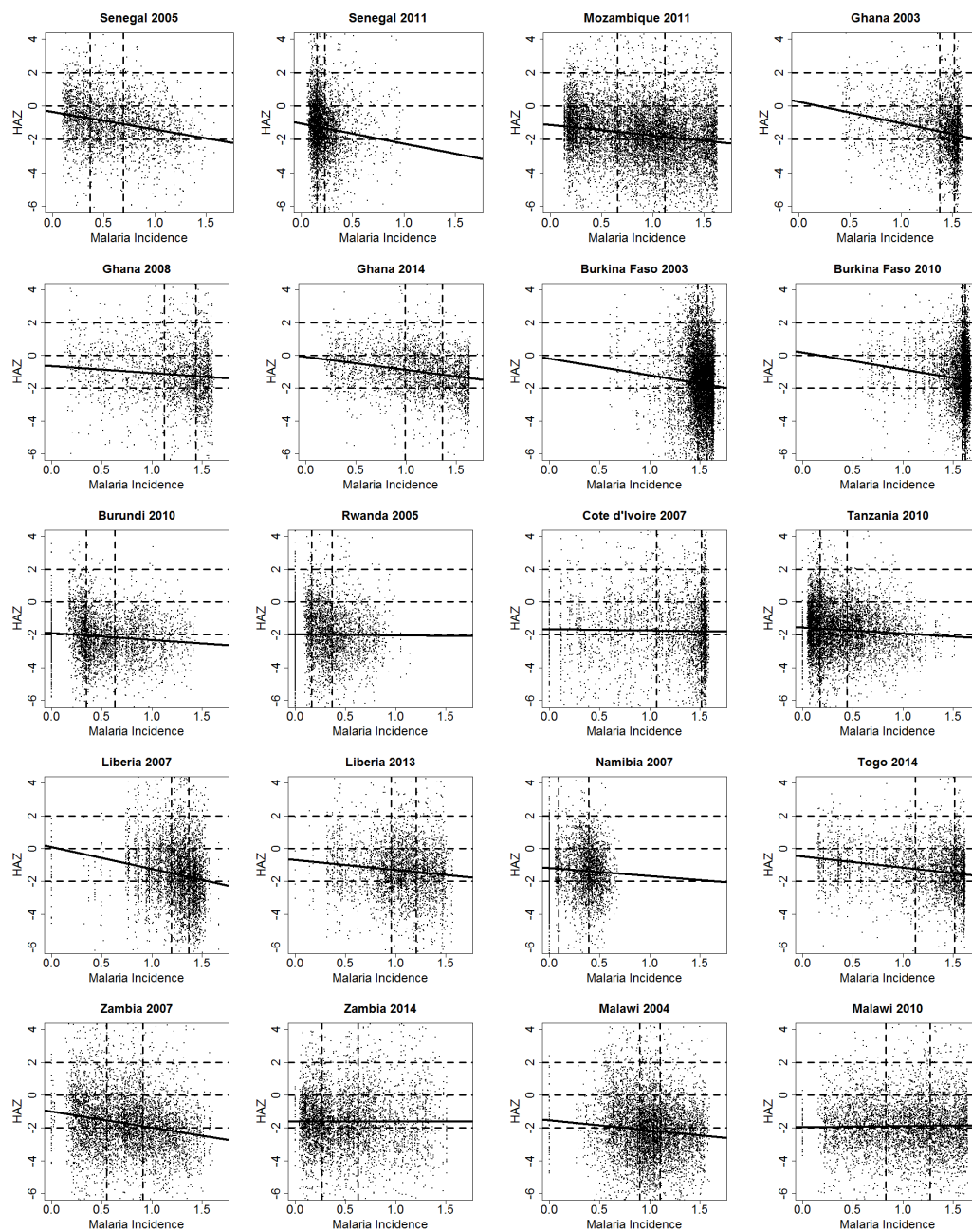
FIGURE 3.2: Scatterplots of height-for-age z-scores (HAZ) against expected malaria incidence in the first year of life ($\mathcal{M}$). The solid line shows the univariate linear model with malaria incidence as the predictor of HAZ. The dashed horizontal lines show HAZ levels of 2, 0 and -2, whilst the dashed horizontal lines separates $\mathcal{M}$ into terciles.
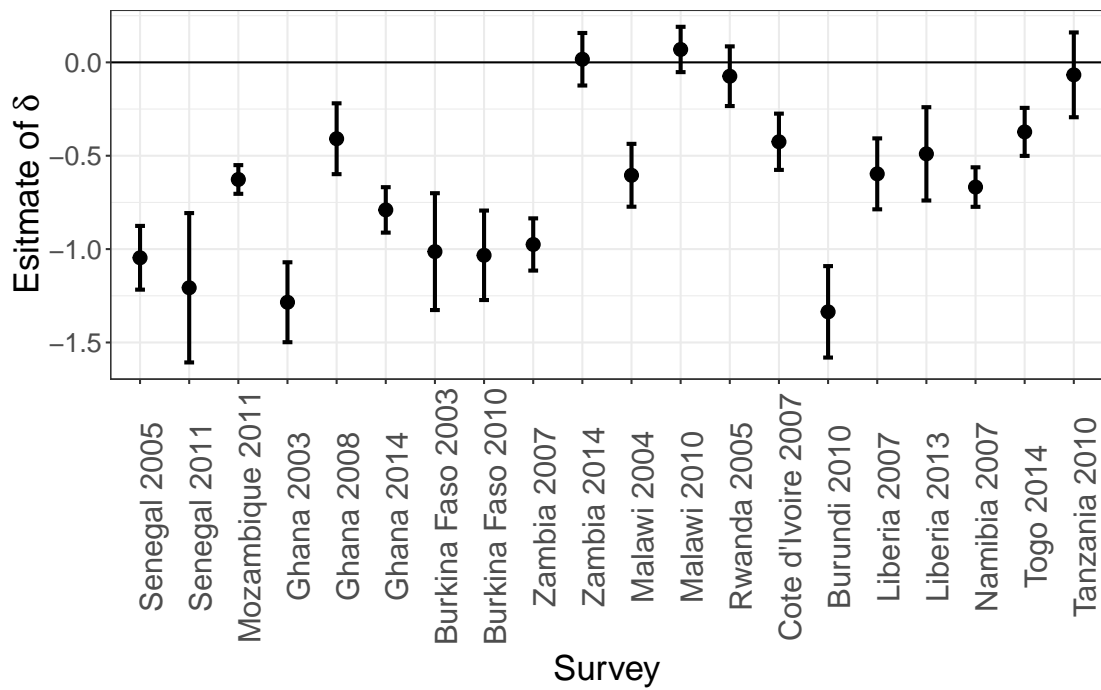
FIGURE 3.3: Plot of estimates of the malaria effect on HAZ with associated 95% confidence intervals obtained from a univariate linear model for each survey.

curves still remain below the -2 threshold in later years.

3.4.2 Geostatistical analysis

Figure 3.5 shows estimates, with associated 95% confidence intervals, of the malaria parameter δ from the fitted geostatistical model in (3.1). The point estimate of δ is negative in 7 surveys with Ghana in 2014 and Liberia in 2007 being significant at the 5% level. Positive values are estimated for the remaining 13 surveys, with only Namibia in 2007 being significant. We note that, after accounting for residual spatial variation and measured potential confounders, the magnitude of the association between malaria incidence and HAZs is smaller than for the marginal association shown in Figure 3.3.

Point estimates of the covariance parameters of (3.1) with associated standard errors are reported in Additional File 3. We see that, for each survey, the variance corresponding to the child-specific variation is consistently larger than both the variance of the spatial process and the nugget variance.
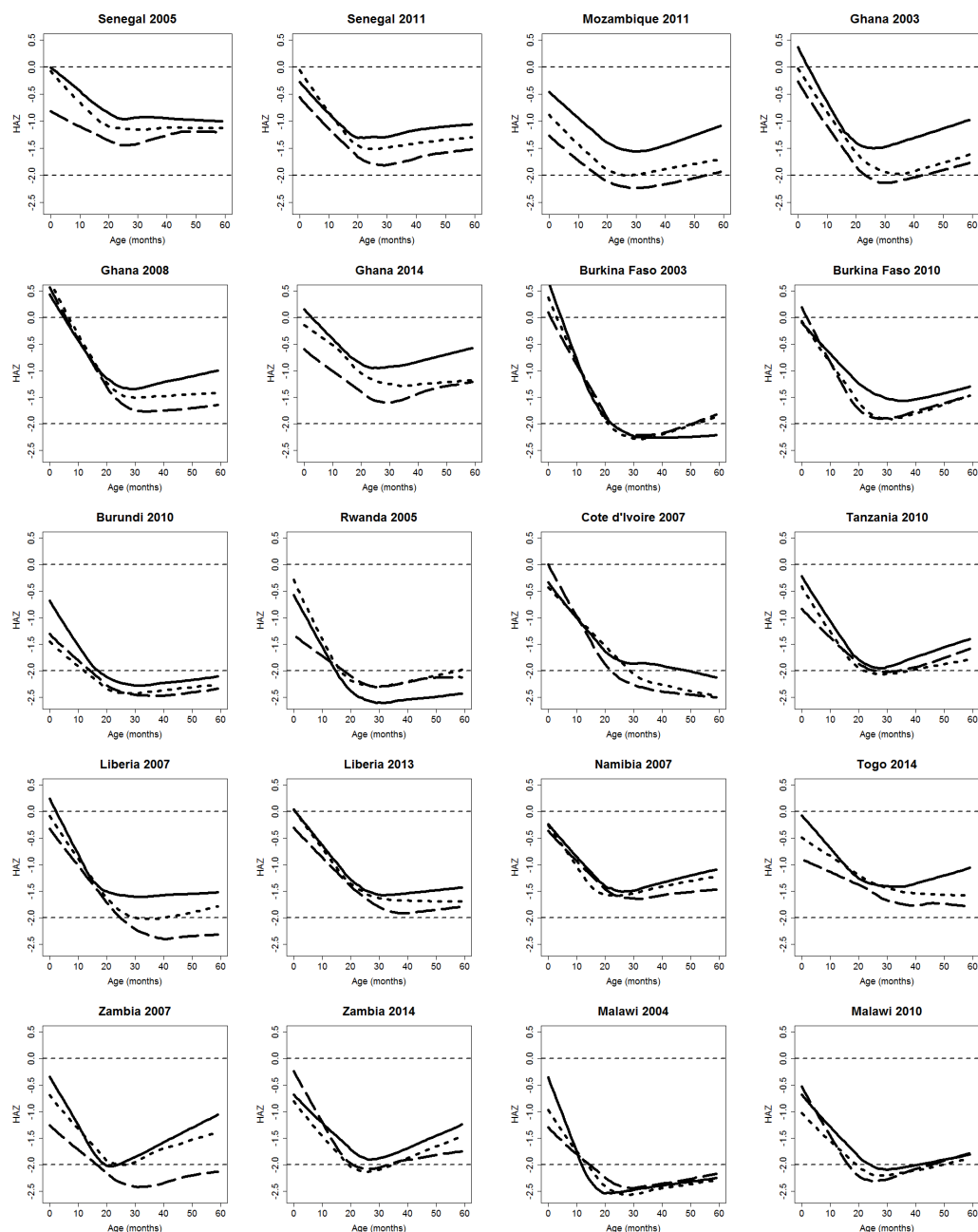
FIGURE 3.4: Estimated trajectories of height-for-age z-scores (HAZ) as a function of age, stratified by malaria incidence ($\mathcal{M}$). Each panel shows three curves. Each curve is a piecewise cubic spline with knots at 12 and 24 months and corresponds to a tercile group of $\mathcal{M}$. The solid, dotted and dashed curves respectively correspond to the first, second and third terciles of $\mathcal{M}$, as indicated in Figure 3.2. The horizontal lines are the HAZ levels of 0 and -2.
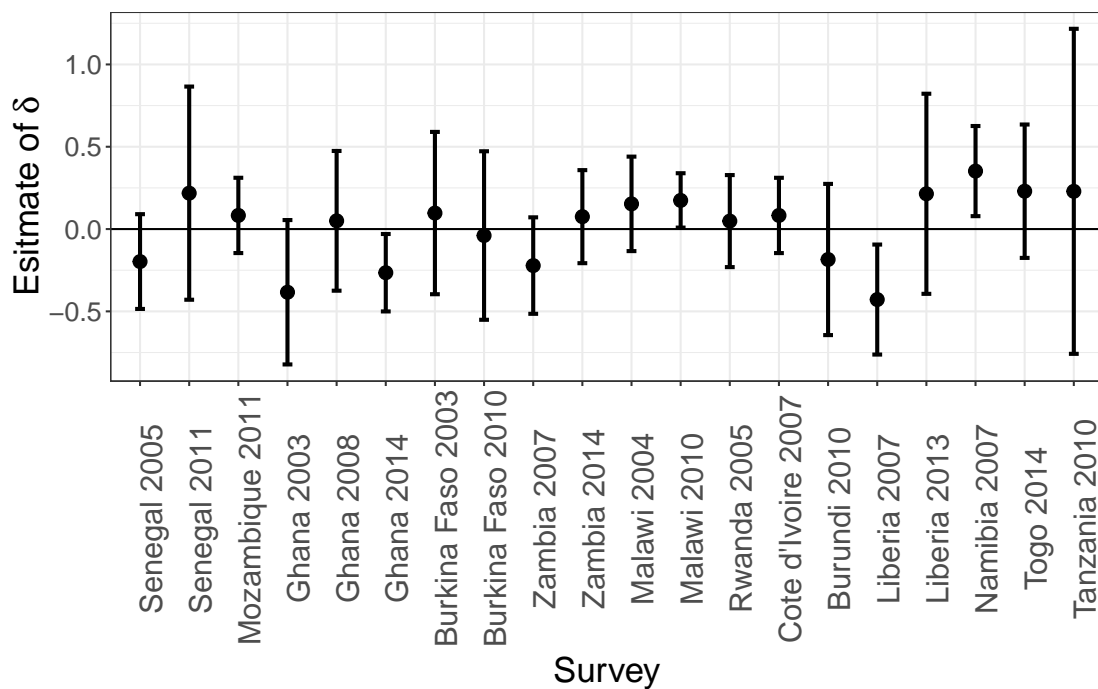
FIGURE 3.5: Plot of estimates of the malaria effect on HAZ with associated 95% confidence intervals, obtained from the geostatistical model in (3.1) for each survey.

The results from the model validation (Additional File 4) show that the fitted geostatistical models are compatible with the data for each of the 20 surveys analysed. We also point out that, although the variograms based on the residuals from the standard linear regression are relatively flat, we still find evidence of non-negligible residual spatial variation in HAZ as indicated by the interval estimates of the parameter of the scale of the spatial correlation in Additional File 3.

3.4.3   Mapping of Stunting Risk

In Figure 3.6, we report the predictive maps of stunting risk for Ghana, Burkina Faso and Mozambique for boys, aged 24 months. In Ghana in 2003-2008-2014, the maps show a remarkable decrease in stunting over time, that is observed almost everywhere within the country. Similarly, in Burkina Faso, we observe a decrease in stunting risk from 2003 to 2010. Mozambique in 2011 shows high spatial heterogeneity in stunting risk, with values ranging from 0.1 to 0.9. Risk maps for the remaining surveys are shown in Additional File 5. In these maps, we observe overall higher levels of stunting risk in

Burundi in 2010 and Malawi in 2004, and lower levels in Senegal in 2008 and Togo in 2014.

3.4.4 Variation in the effect of malaria on HAZ

The amount of arable land (defined as percentage of land under temporary crops, meadows for mowing or for pasture, market or kitchen gardens, and land temporarily fallow) in the country and year of survey is the only World Bank indicator to be significant at 5% level, with a p-value of about 0.013, explaining 26% of the total variation in the estimated effects of the malaria on HAZ. More specifically, we estimate that an increase of 1% in arable land leads to a 0.008 increase in the value of the estimated malaria effect, on average. See "Additional Table 2" in Additional File 2 for details of the estimates from the meta analysis.

We assessed possible spatial concurvity in the generalized additive model framework modelling HAZ as a smooth function of malaria and using three indices namely worst, observed and estimate, all of which are bounded between 0 and 1, with 0 indicating no problem of concurvity, and 1 indicating total lack of identifiability. Each index is based the square of $||g||/||f||$, where it is assumed that a smooth term of HAZ, f, in the model can be decomposed into a part, g, that lies entirely in the space of malaria, and a remainder part that is completely within the HAZ space. If g makes up a large part of f then there is a concurvity problem. For details, see the documentation for the R package *mgcv* (Wood and Wood, 2015), which was used for the concurvity analysis. For all the 20 datasets, each index was less than 1.0×10^{-16}, indicating no concurvity between HAZ and malaria (Table 3.2).

3.5 Discussion

The objective of our study was to model and quantify the association between malaria and HAZs in children aged less than 5 years. Using DHS data from 20 surveys in 13 African countries between 2003 and 2014, we have developed a geostatistical framework to model HAZ as a function of both child-specific and spatial risk factors. As a proxy
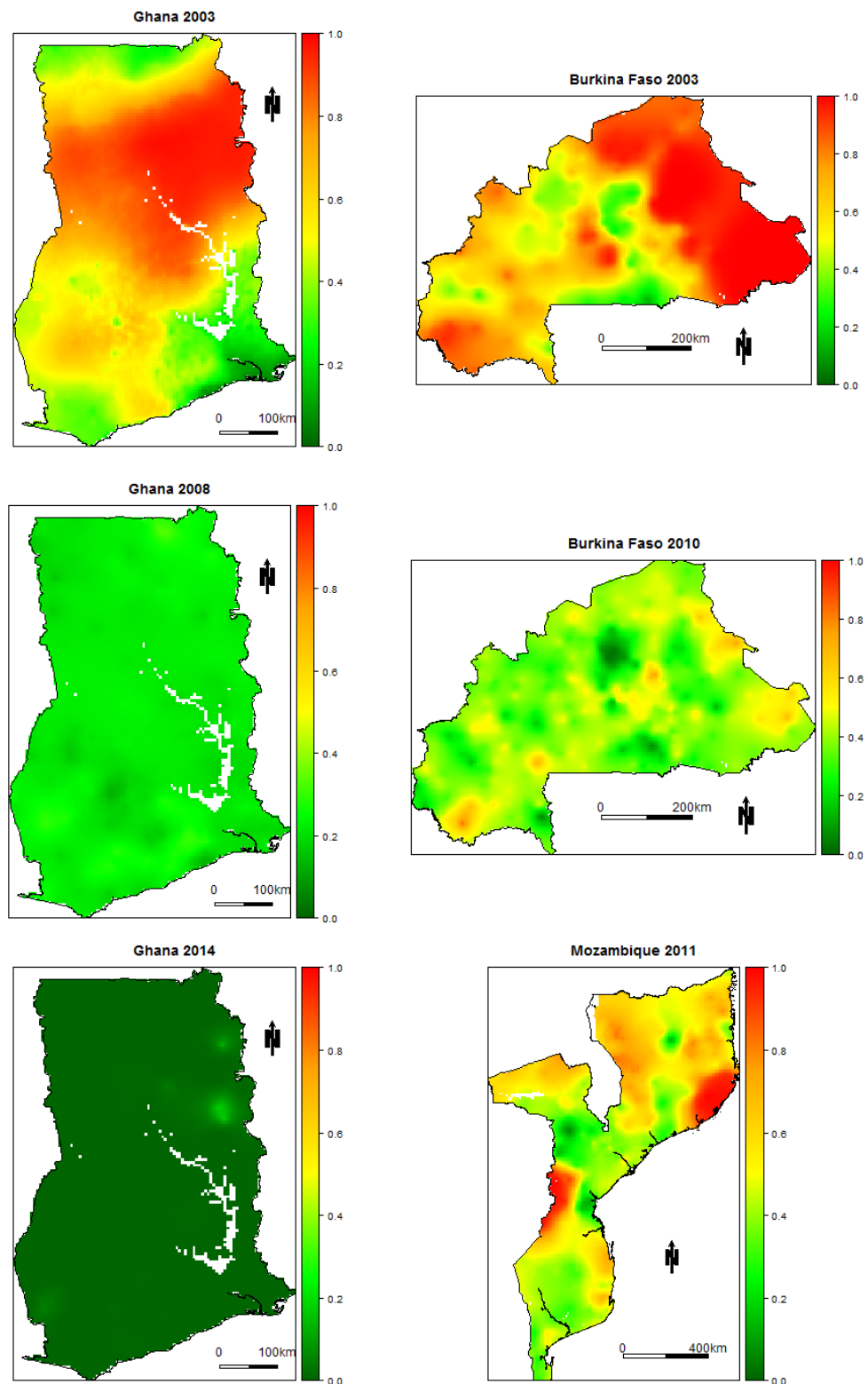
FIGURE 3.6: Predicted stunting risk maps for Ghana, Burkina Faso and Mozambique. The colour scale ranges from green to red with red areas being high risk areas and green areas being low risk areas.

TABLE 3.2: Details of the spatial concurvity analysis.

	Dataset	Worst	Observed	Estimate
1	Senegal 2005	3.3945E-23	3.3945E-23	3.3945E-23
2	Senegal 2011	4.5431E-23	4.5431E-23	4.5431E-23
3	Mozambique 2011	1.4762E-23	1.4762E-23	1.4762E-23
4	Ghana 2003	4.6251E-26	4.6251E-26	4.6251E-26
5	Ghana 2008	9.4906E-23	9.4906E-23	9.4906E-23
6	Ghana 2014	7.5096E-23	7.5096E-23	7.5096E-23
7	Burkina Faso 2003	4.8769E-25	4.8769E-25	4.8769E-25
8	Burkina Faso 2010	1.4972E-24	1.4972E-24	1.4972E-24
9	Zambia 2007	8.2974E-21	8.2974E-21	8.2974E-21
10	Zambia 2014	7.6379E-26	7.6379E-26	7.6379E-26
11	Malawi 2004	6.6261E-24	6.6261E-24	6.6261E-24
12	Malawi 2010	6.0444E-28	6.0444E-28	6.0444E-28
13	Cote d'Ivoire 2007	7.1410E-22	7.1410E-22	7.1410E-22
14	Burundi 2010	1.5801E-21	1.5801E-21	1.5801E-21
15	Liberia 2007	6.6619E-22	6.6619E-22	6.6619E-22
16	Liberia 2013	5.2867E-23	5.2867E-23	5.2867E-23
17	Namibia 2007	3.3047E-18	3.3047E-18	3.3047E-18
18	Togo 2014	2.4733E-24	2.4733E-24	2.4733E-24
19	Tanzania 2010	1.7239E-24	1.7239E-24	1.7239E-24
20	Rwanda 2005	1.8620E-20	1.8620E-20	1.8620E-20

for malaria exposure, we used estimates of malaria incidence in the first year of life from the Malaria Atlas Project. A non-spatial univariate linear regression showed a negative effect of malaria incidence on HAZs. However, after controlling for confounding and residual spatial effects, the estimated effect of malaria on HAZ was weaker and not significant in 17 out of the 20 surveys considered.

One of the main challenges in modelling the association between malaria and HAZ is the need to take account of confounding effects. Among these, socio-economic status has been shown to be one of the most important (Gallup and Sachs, 2001, Sachs and Malaney, 2002, Somi et al., 2007, Teklehaimanot and Paola Mejia, 2008). Education is another important factor that affects both malaria exposure and risk of stunting (Fink et al., 2013, Kere et al., 1993, Thuilliez et al., 2010). Higher levels of education are associated with improved knowledge and practice about the appropriate strategies for the prevention and treatment of malaria (Dike et al., 2006), and about healthy practices in breastfeeding and child nutrition (Abuya et al., 2012). Our results are consistent with these findings in all of the 20 surveys here analysed.

We observed that in surveys where HAZ curves fall off below the -2 threshold in early childhood, the curves never really rise above the -2 threshold in later years. This finding suggest that HAZ recovery after 2 years of age may be very difficult, if at all possible,

when HAZ fall-off in early childhood is severe. This is in line with findings from a study (Crookston et al., 2010) that showed that recovery from stunting is associated with severity of stunting in early years. Other factors that have been found to favour recovery from low HAZ are good nutrition (Lopriore et al., 2004) and higher levels of mother's education (Vella et al., 1994).

In our analysis, we found a mix of positive and negative point estimates of the association between malaria incidence and HAZ among the different surveys. However, findings from previous studies have shown contrasting results, with some reporting statistically significant negative associations between malaria and stunting (Arinaitwe et al., 2012, Deen et al., 2002, Ehrhardt et al., 2006, Kang et al., 2013), and others reporting positive associations (Genton et al., 1998, Murray et al., 1978). To understand such variation in the magnitude and direction of the estimated parameters that quantify the malaria effect, we carried out a meta-analysis by considering several indicators of national development from the World Bank. Among these, the amount of arable land was the only one to show a significant association. Arable land might in fact modulate the association between malaria and HAZ, with a larger surface of arable land leading to a fall in poverty and malnutrition, especially in rural areas (Webb and Block, 2012), but also to a larger number of breeding sites for mosquitoes (Sovi et al., 2013). This suggests that geo-political differences among countries should also be considered, since the implementation of policies aiming to reduce malnutrition can also impact on the epidemiology of malaria. Arable land could be indeed associated with agricultural, economic and environmental factors that are common to both malaria and stunting (Matariya et al., 2016, Smith and Haddad, 2015).

Our Model treated malaria as a risk factor of stunting. Since the malaria data used for the analysis are not individual specific but location level estimates, we cannot use these data to establish or explain causality. In a causal study, if HAZ modulates malaria as suggested by some studies (Genton et al., 1998, Murray et al., 1978, Nyakeriga et al., 2004, Olney et al., 2009), then including malaria in the model can induce a collider bias, and risk factors for malaria can erroneously appear associated with HAZ. In this study instead, neither the exposure (malaria) nor the outcome (HAZ) drove inclusion of

participants in a study. Indeed DHS are nationally representative surveys whose design are not determined by malaria or HAZ, hence, at least in part, avoiding collider bais in our analysis.

We have quantified stunting risk by mapping the predictive probability that HAZ is below a threshold of -2. For countries with repeated surveys, our risk maps showed reductions over time in the risk of stunting. The main factors that might be driving such reductions are improvements in health environments through increasing access to safe water and sanitation, improvements in the quality of caring practices for children through increasing women's education and promoting gender equality, including women's empowerment; and increase in food security by ensuring adequate availability of food at the national level and sufficient nutritional quality of that food (Headey, 2013, Ruel et al., 2013, Smith and Haddad, 2015). Our risk maps showed remarkable spatial heterogeneity in the risk of stunting, identifying geographic areas with high risk that could be considered for a more targeted intervention.

It has been widely observed that HAZ undergoes a rapid decrease in the first 24 months and an increase thereafter (Allen, 1994, Stevens et al., 2012, Victora et al., 2010). For this reason we used cubic splines with knots at 12 and 24 months in order to better capture the non-linear trajectory that we observed across the five years of age.

3.5.1 Limitations of the study

The main limitation of our study is that the information available to us on malaria and HAZ is cross-sectional, rather than longitudinal, in nature. This prevents us from establishing whether our observed associations can be given a causal interpretation. A second limitation is that we have no information on the uncertainty associated with the estimates of malaria incidence. We have assumed the first year of life to be the most important in determining the strength of the association between malaria and child growth. To investigate whether exposure to malaria in other years of childhood could also have an impact on growth would require the fitting of a distributed lag-model.

In Additional File 6, we give methodological details on how to account for uncertainty in malaria incidence in a cross-sectional geostatistical setting.

3.5.2 Novel extensions to longitudinal geostatistical data

To assess the cumulative effect of malaria on child-growth at different developmental stages, we would need longitudinal, individual-level data on children's actual malaria status over the first five years of life. We would then extend our current methodology as follows.

To simplify the notation used in this chapter and without loss of generality, we assume that all the sampled children have identical follow up times. Then, let Y_{ijt} and W_{ijt} denote the HAZ and number of malaria episodes for the j-th child at location x_i and time t, respectively. Also, let $\tilde{S}(x,t)$ denote a latent spatio-temporal Gaussian process. Given $\tilde{S}(x,t)$, we model the W_{ijt} as a set of mutually independent Poisson variables with mean $\mathcal{M}_{ijt}$ such that

$$\log\{\mathcal{M}_{ijt}\} = \tilde{e}_{ijt}^{\top}\tilde{\gamma} + \tilde{d}(x_i)^{\top}\tilde{\beta} + \tilde{S}(x_i,t),$$

where $\tilde{e}_{ijt}$ are child-specific explanatory variables that might vary over time. We then assume that Y_{ijt}, conditionally on $\mathcal{M}_{ijt}$, a spatio-temporal Gaussian process $S(x,t)$ and random effects U_{it} and V_{ij}, are independent Gaussian variables with mean

$$\mu_j(x_i,t) = e_{ijt}^{\top}\gamma + d(x_i,t)\beta + \sum_{h=0}^{t-1}\delta_{t-h}\mathcal{M}_{ij(t-h)} + f(\mathcal{A}_{ijt}) + V_{ij} + S(x_i,t) + U_{it} \quad (3.4)$$

In (3.4), U_{it} is unstructured unexplained variation at location x_i and time t, V_{ij} is unexplained child-specific variation and the lagged parameters δ_{t-h}, for $h = 0,\ldots,t=1$, represents the effect of malaria incidence during the h-th year of life on HAZ. To make the model more parsimonious, the parameters δ_{t-h} can be constrained using a parametric specification, i.e. $\delta_{t-h} = g(t-h;\theta)$ where $g(\cdot;\theta)$ is a known function indexed by the vector of parameters θ.

This modelling framework would allow us to better understand the cumulative effect of malaria on HAZ at different developmental stages by overcoming the current limitation of our study where we assume that $\delta_{t-h} = 0$ for $0 \leq h \leq t - 2$.

The proposed model can more generally be used to investigate the age-specific and cumulative effects of episode of diseases on other indicators of malnutrition in children (??).

3.6 Conclusion

Geostatistical methods provide a useful framework to account for spatially structured confounding effects that modulate the association between malaria and HAZ. This study also highlights that one of the main challenges in modelling this association is that confounding effects vary by country, as well as in time. This can change both the direction and magnitude of the effect of malaria on HAZ, making a generalization on the effect of malaria on HAZ almost impossible using only currently available data. Establishing whether the association between malaria and stunting is causal would require longitudinal follow-up data on individual children.

3.7 List of Abbreviations

DHS: Demographic and Health Surveys

GDP: Gross domestic product

HAZs: Height-for-age z-scores

LMICs: Low- and middle-income countries

WHO: World Health Organization

3.8 Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable (this manuscript does not included reporting on individual participant's data).

Availability of data and material

The statistical methods presented in this manuscript have been implemented in the R package `PrevMap` (Giorgi and Diggle, 2017) which can be freely downloaded from the Comprehensive R Archive Network (https://www.r-project.org). The datasets supporting the conclusions of this manuscript are the following. The DHS data are available on request from the Demographic and Health Surveys repository (http://dhsprogram.com). The urban extent indicator data are available in the Socioeconomic Data and Applications Center repository http://sedac.ciesin.columbia.edu/data/collection/grump-v1. The malaria incidence raster data for the age group 0-5 years are available on request from the Malaria Atlas Project (http://www.map.ox.ac.uk/). The World Bank indicators data are publicly available in the World Bank database (http://data.worldbank.org/indicator).

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

BA wrote the first draft of the manuscript. BA, EG and PJD conducted the statistical analysis. SvB, BA, and DJH retrieved the data. BA and EG developed the code for the analysis. All authors helped to draft the manuscript. All authors read and approved the final manuscript.

Authors' information

[1]CHICAS Research Group, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK. [2] Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK. [3]Department of Child Health, Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands. [4]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands.

Additional File 1: Computational details

Parameter estimation

Let $Y^\top = (Y_1^\top, \ldots, Y_n^\top)$, where Y_i is the vector of observed HAZs at the ith cluster, and Let $N = \sum_{i=1}^{n} m_i$. The vector Y has a multivariate Gaussian distribution with mean $\mu = D\xi$ and covariance matrix Σ, where D is a matrix of covariates including the cubic spline bases, with vector of regression coefficients $\xi^\top = (\gamma^{*\top}, \beta, \delta)$, where γ^* consists of coefficients of child-specific variables, spatial variables and those of the spline bases, and $\Sigma = C(\sigma^2 R + \tau^2 I_n)C^\top + \omega^2 I_N$, where $[R]_{ij} = \rho(u_{ij}; \phi)$ and

$$[C]_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ child has been sampled at location } x_i, \\ 0 & \text{otherwise} \end{cases}.$$

Let $\theta^\top = (\xi^\top, \sigma^2, \tau^2, \omega^2, \phi)$ denote the vector of model parameters; the log-likelihood for θ is given by

$$L(\theta) = -\frac{1}{2}\big\{ \log|\Sigma| + (y - \mu)^T \Sigma^{-1}(y - \mu) \big\}. \tag{3.5}$$

Inversion of the covariance matrix Σ can be simplified using the Woodbury matrix identity to give

$$\Sigma^{-1} = (\tilde{\omega}^{-2} I_N - \tilde{\omega}^{-4} C\big[(R + \tilde{\tau}^2 I_n)^{-1} + \tilde{\tau}^{-2} C^T C\big]^{-1} C^T)/\sigma^2. \tag{3.6}$$

where $\tilde{\omega}^2 = \omega^2/\sigma^2$ and $\tilde{\tau}^2 = \tau^2/\sigma^2$. Using Sylvester's determinant identity, we can also write

$$|\Sigma| = \omega^{2N} |\tilde{\omega}^{-2} C^\top C(R + \tilde{\tau}^2 I_n) + I_n|, \tag{3.7}$$

hence, as in (3.6), computations are carried out on an n by n matrix.

To maximize $L(\theta)$, we then use the profile likelihood for given $\psi = (\tilde{\tau}^2, \tilde{\omega}^2, \phi)$. The profile estimates for ξ and σ^2 are respectively given by

$$\hat{\xi}(\psi) = (D^T Q D)^{-1} D^T Q^{-1} y$$

and

$$\hat{\sigma}^2(\psi) = \frac{1}{N}(y - D\hat{\xi}(\psi))^T Q^{-1}(y - D\hat{\xi}(\psi))$$

where $\sigma^2 Q = \Sigma$. By plugging $\hat{\xi}(\theta)$ and $\hat{\sigma}^2(\theta)$ into (3.5), we obtain

$$L_p(\psi) = -\frac{1}{2}\left\{N \log \hat{\sigma}^2(\psi) + \log|Q|\right\}. \tag{3.8}$$

Finally, numerical optimization can be used to maximize $L_p(\psi)$ with respect to ψ.

Model validation

We carry out model validation to test the validity of the adopted spatial covariance function as follows.

Let $W_j(x_i) = S(x_i) + U_i + V_i$ denote the residual variation in HAZ for j-th children at location x_i, where $V_i \sim \mathcal{N}(0, \omega^2)$ and $S(x_i)$ and $U(x_i)$ are as defined in the geostatistical model for HAZ in equation (1) of the manuscript. The theoretical variogram of the random effects is

$$\gamma(u_{hk}) = \omega^2 + \tau^2 + \sigma^2(1 - \exp\{u_{hk}\}) \tag{3.9}$$

where u_{hk} is the Euclidean distance between location x_h and x_k.

Denote by $\tilde{W}_j(x_i)$ the estimated residuals from a standard linear regression for the j-th child at location x_i. Let $N(u) = \{(h,k) : ||x_h - x_k|| = u_{hk}\}$, i.e. the set of all data-points such that their distance is u_{hk}. The empirical variogram is the defined as

$$\tilde{\gamma}(u_{hk}) = \frac{1}{2|N(u_{hk})|} \sum_{(h,k) \in N(u)} \left(\tilde{W}_j(x_h) - \tilde{W}_{j'}(x_k)\right)^2, \tag{3.10}$$

where $|N(u_{hk})|$ is the number of observations in $N(u_{hk})$.

To generate a 95% bandwidth of the empirical variogram under the fitted model, we first simulate $W_j(x_i)$, at observed locations x_i, from its marginal multivariate Gaussian distribution, as defined by the geostatistical model. Conditionally on the simulated values of $W_j(x_i)$, we simulate HAZ from the conditional model in equation (1) of the manuscript. We then compute the empirical variogram in (3.10) obtained from the simulated data. We repeat this process 1,000 times. Finally, we generate 95% tolerance intervals at each of pre-defined spatial distances of the variogram.

Spatial prediction

Let $T^\top = (T(x_1^*), \ldots, T(x_q^*))$ denote our target of prediction, where x_i^* are q prediction locations. The conditional distribution of T given the data $Y = y$ and all the explanatory variables at each of the prediction locations x_i^*, is a multivariate Gaussian with mean

$$D^*\xi + P\Sigma^{-1}(y - D\xi), \tag{3.11}$$

where D^* is the matrix of explanatory variables at the prediction locations, P is the cross-covariance matrix and ξ the vector of all the regression coefficients reported in equation (1) in the manuscript; the covariance matrix is

$$\sigma^2(R^* + \tilde{\tau}^2 I) - P\Sigma^{-1}P^\top, \tag{3.12}$$

where $[R^*]_{ij} = \exp\{-u_{ij}^*/\phi\}$ and u_{ij}^* is the Euclidean distance between any two prediction locations x_i^* and x_j^*. When carrying out predictions, we plug-in the maximum likelihood estimates for each of the model parameters.

In order to quantify the risk of stunting at a location x, we map

$$\text{Prob}(T(x_i^*) < -2|y), i = 1, \ldots, q. \tag{3.13}$$

In the above equation we fix age at 24 months and gender to male, whilst we integrate out maternal education and wealth index as follows. Let $[\cdot]$ be a shorthand notation for

"the distribution of $\cdot$". The predictive distribution of the target $T(x)$ is then given by

$$[T(x_i^*)|y] \quad = \quad \int [\mathcal{D}][T(x_i^*)|y, \mathcal{D}]d\mathcal{D}, i = 1, \ldots, q \qquad (3.14)$$

where $\mathcal{D} = (\mathcal{W}, \mathcal{E})$, with $\mathcal{E}$ corresponding to maternal education and $\mathcal{W}$ to wealth index, and $[T(x_i)|y, \mathcal{D}]$ is the i-th component of the multivariate Guassian distribution with mean and covariance matrix given by (3.11) and (3.12), respectively. We model the joint distribution of $\mathcal{D}$ as

$$[\mathcal{D}] = [\mathcal{E}][\mathcal{W}|\mathcal{E}]$$

where $[\mathcal{E}]$ is estimated using the empirical distribution obtained from the data of a given survey, and $[\mathcal{W}|\mathcal{E}]$ is a proportional odds cumulative probit model Agresti (1996).

To compute (3.14), we then generate 10,000 samples from $[T(x_i^*)|y]$ by simulating sequentially from $[\mathcal{E}]$, $[\mathcal{W}|\mathcal{E}]$ and $[T(x_i^*)|y, \mathcal{D}]$. Finally, we obtain (3.13) by computing the proportions of simulated samples from $[T(x_i^*)|y]$ that lie below -2, for $i = 1, \ldots, q$.

Additional File 2: Details of the World Bank development indicators

TABLE 3.3: World Bank Development Indicators.

Indicator	Description
Land under cereal production	Land under cereal production refers to harvested area, although some countries report only sown or cultivated area. Cereals include wheat, rice, maize, barley, oats, rye, millet, sorghum, buckwheat, and mixed grains.
Arable land	Arable land includes land defined as land under temporary crops (double-cropped areas are counted once).
Livestock production index	Livestock production index includes meat and milk from all sources, dairy products such as cheese, and eggs, honey, raw silk, wool, and hides and skins.
GDP per capita	GDP per capita based on purchasing power parity.
Primary completion rate, male	Primary completion rate is the number of new entrants in the last grade of primary education, regardless of age, divided by the population at the entrance age for the last grade of primary education.
Population density	Population density is midyear population divided by land area in square kilometers.
Progression to secondary school, female	Progression to secondary school refers to the number of new entrants to the first grade of secondary school in a given year as a percentage of the number of students enrolled in the final grade of primary school in the previous year (minus the number of repeaters from the last grade of primary education in the given year).
Prevalence of anemia among children	Prevalence of anemia, children under age 5, is the percentage of children under age 5 whose hemoglobin level is less than 110 grams per liter at sea level.
Improved water source	Access to an improved water source refers to the percentage of the population using an improved drinking water sources which include piped water on premises and other improved drinking water sources such as public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection.
Improved sanitation facilities, rural/urban	Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities include flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.
Unemployment	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.

TABLE 3.4: Regression table for the meta analysis

World Bank Development Indicator	Point Estimate	Standard Error	p-value
Land under cereal production	5.0388×10^{-8}	4.3233×10^{-8}	0.2590
Arable land	7.9548×10^{-3} *	2.8915×10^{-3}	0.0131
Livestock production index	5.7663×10^{-4}	1.6313×10^{-3}	0.7278
GDP per capita	-4.1747×10^{-5}	3.5107×10^{-5}	0.2498
Primary completion rate, male	-2.3216×10^{-3}	2.9019×10^{-3}	0.4341
Population density	5.5419×10^{-4}	4.8843×10^{-4}	0.2714
Progression to secondary school, female	-3.8597×10^{-3}	3.6859×10^{-3}	0.3089
Prevalence of anaemia among children	-4.6101×10^{-3}	4.7033×10^{-3}	0.3400
Improved water source	-5.1427×10^{-3}	4.3177×10^{-3}	0.2491
Improved sanitation facilities, urban	2.7747×10^{-3}	3.5442×10^{-3}	0.4439
Unemployment	5.0836×10^{-3}	8.4332×10^{-3}	0.5542

Dependent Variable: Estimate of malaria-HAZ association parameter
Significance levels: 0 '***' 0.001 '**' 0.01 '*' 0.05 '·' 0.1 ' ' 1

Additional File 3: Estimates of covariance parameters

TABLE 3.5: Estimates of Covariance Parameters

Survey	$\log(\sigma^2)$	$\log(\phi)$	$\log(\tau^2)$	$\log(\omega^2)$
Senegal 2005	-4.130 (0.981)	4.520 (0.983)	-2.055 (2.042)	0.659 (1.963)
Senegal 2011	-2.915 (0.522)	4.311 (0.724)	-2.180 (1.171)	0.951 (1.045)
Mozambique 2011	-1.802 (0.333)	5.001 (0.457)	-2.197 (0.754)	0.959 (0.666)
Ghana 2003	-2.388 (0.493)	4.780 (0.705)	-2.565 (1.136)	0.687 (0.987)
Ghana 2008	-2.995 (1.552)	3.056 (1.488)	-1.290 (3.355)	0.930 (3.105)
Ghana 2014	-3.154 (0.494)	4.551 (0.786)	-3.601 (1.407)	0.323 (0.988)
Burkina Faso 2003	-0.696 (0.530)	5.328 (0.693)	-2.289 (1.068)	1.105 (1.059)
Burkina Faso 2010	-2.426 (0.394)	3.500 (0.503)	-1.981 (0.971)	0.833 (0.789)
Zambia 2007	-1.886 (0.651)	6.025 (0.854)	-2.755 (1.377)	1.115 (1.303)
Zambia 2014	-2.713 (0.487)	3.600 (1.124)	-1.830 (1.138)	1.041 (0.975)
Malawi 2004	-2.905 (0.674)	4.975 (1.225)	-2.678 (1.441)	1.117 (1.348)
Malawi 2010	-3.329 (0.682)	3.086 (0.634)	-2.443 (1.628)	1.015 (1.365)
Rwanda 2005	-3.0599 (0.549)	3.019 (0.731)	-1.970 (1.221)	0.812 (1.099)
Cote d'Ivoire 2007	-1.155 (0.678)	1.798 (0.869)	-2.186 (3.046)	1.375 (1.357)
Burundi 2010	-2.909 (0.424)	2.122 (0.527)	-2.816 (1.175)	0.502 (0.848)
Liberia 2007	-2.885 (0.745)	2.955 (0.648)	-2.665 (1.966)	1.111 (1.491)
Liberia 2013	-3.324 (0.700)	3.563 (0.693)	-2.821 (1.741)	0.874 (1.400)
Namibia 2007	-2.255 (0.431)	5.011 (0.522)	-3.714 (1.447)	0.743 (0.862)
Togo 2014	-3.335 (1.454)	1.674 (0.714)	-3.847 (5.168)	0.540 (2.908)
Tanzania 2010	-2.313 (0.304)	4.860 (0.489)	-3.175 (0.853)	0.670 (0.607)

Standard errors in brackets

Additional File 4: Results from the model validation
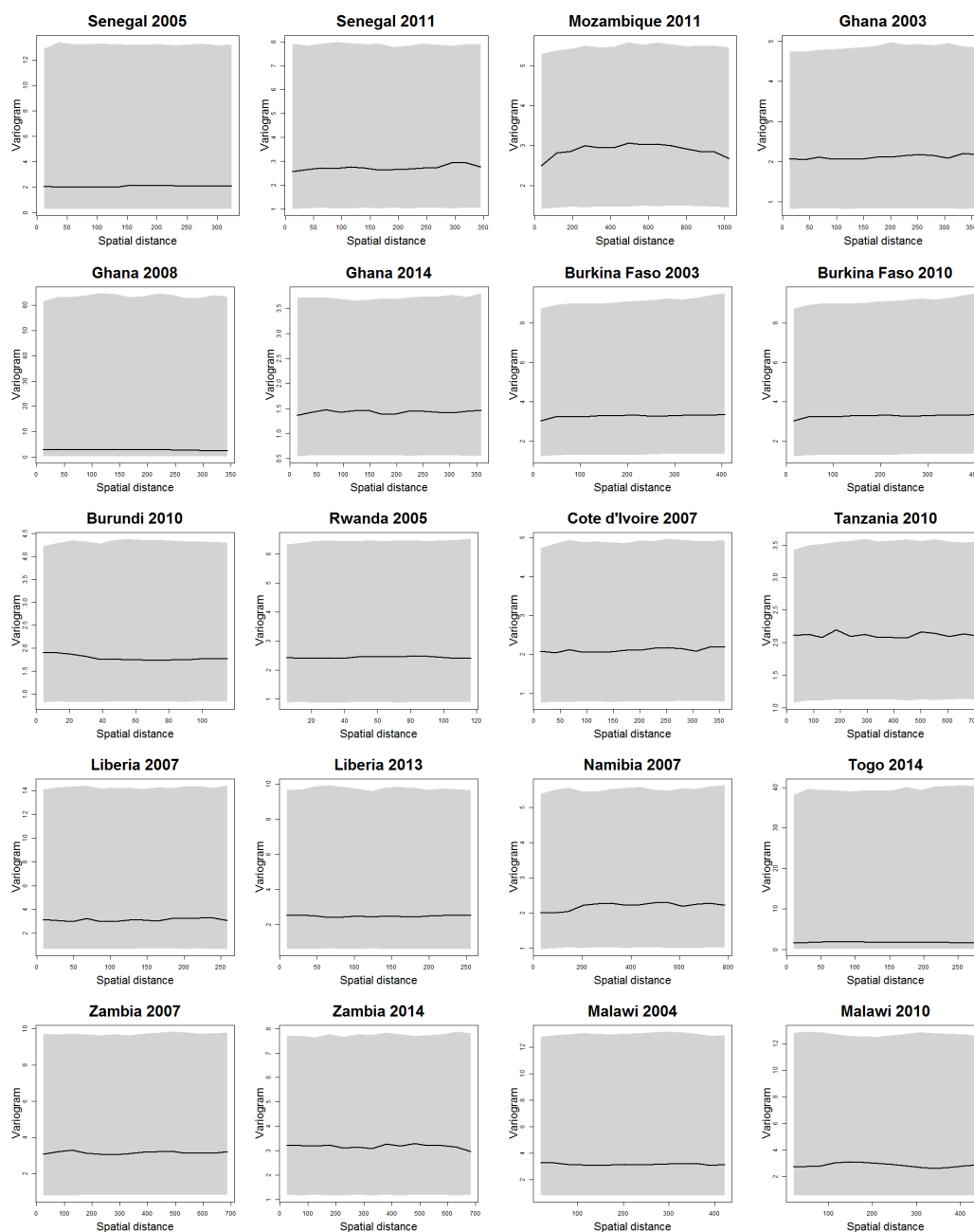
ADDITIONAL FIGURE 3.7: The solid line corresponds the empirical variogram of the residuals from a standard linear regression analysis. The shaded area is the 95% tolerance bandwidth generated under the hypothesis that the adopted correlation function is the true model.

Additional File 5: Maps of stunting risk
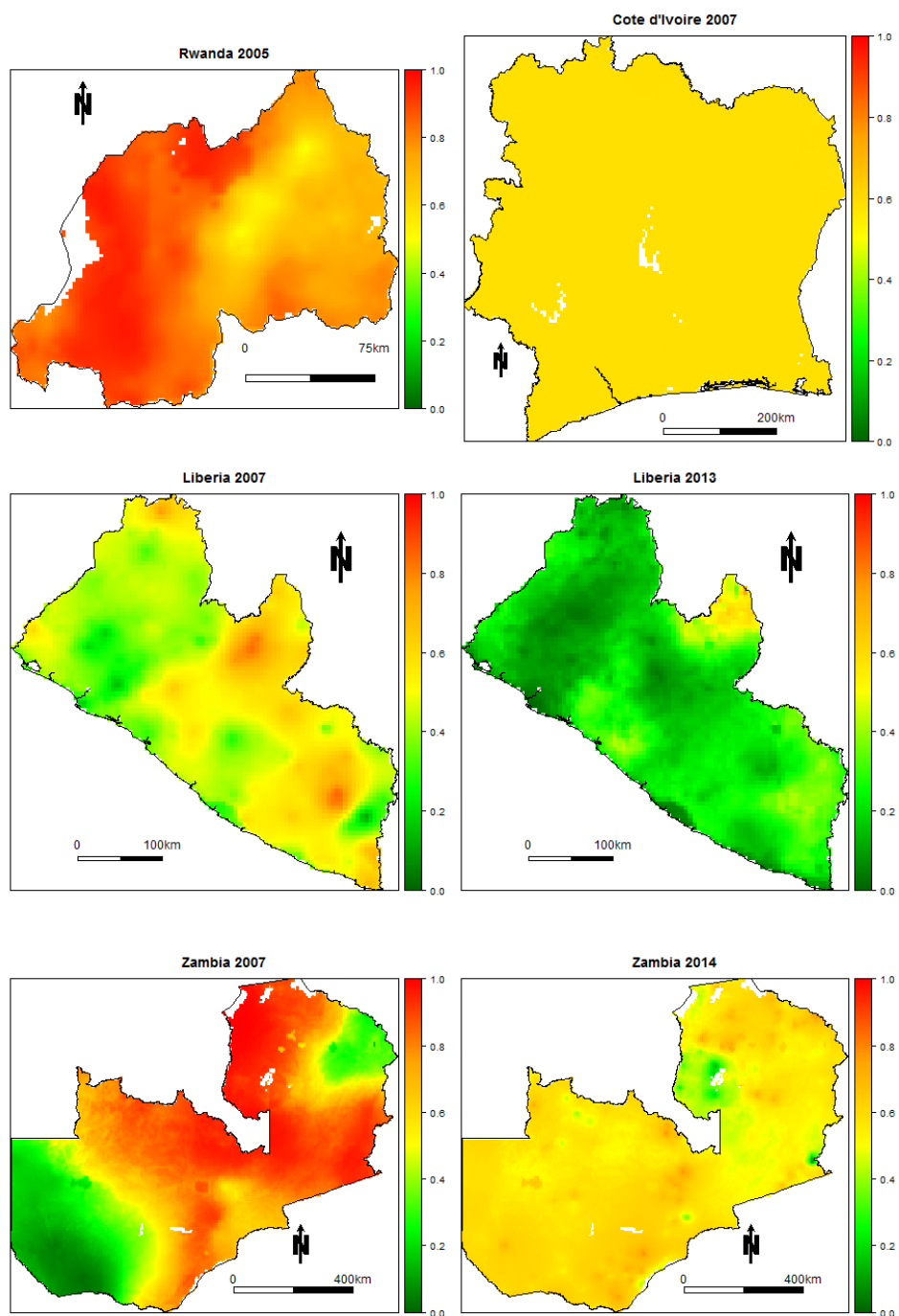
ADDITIONAL FIGURE 3.8: Predicted stunting risk maps for Rwanda, Cote d'voire, Liberia and Zambia.
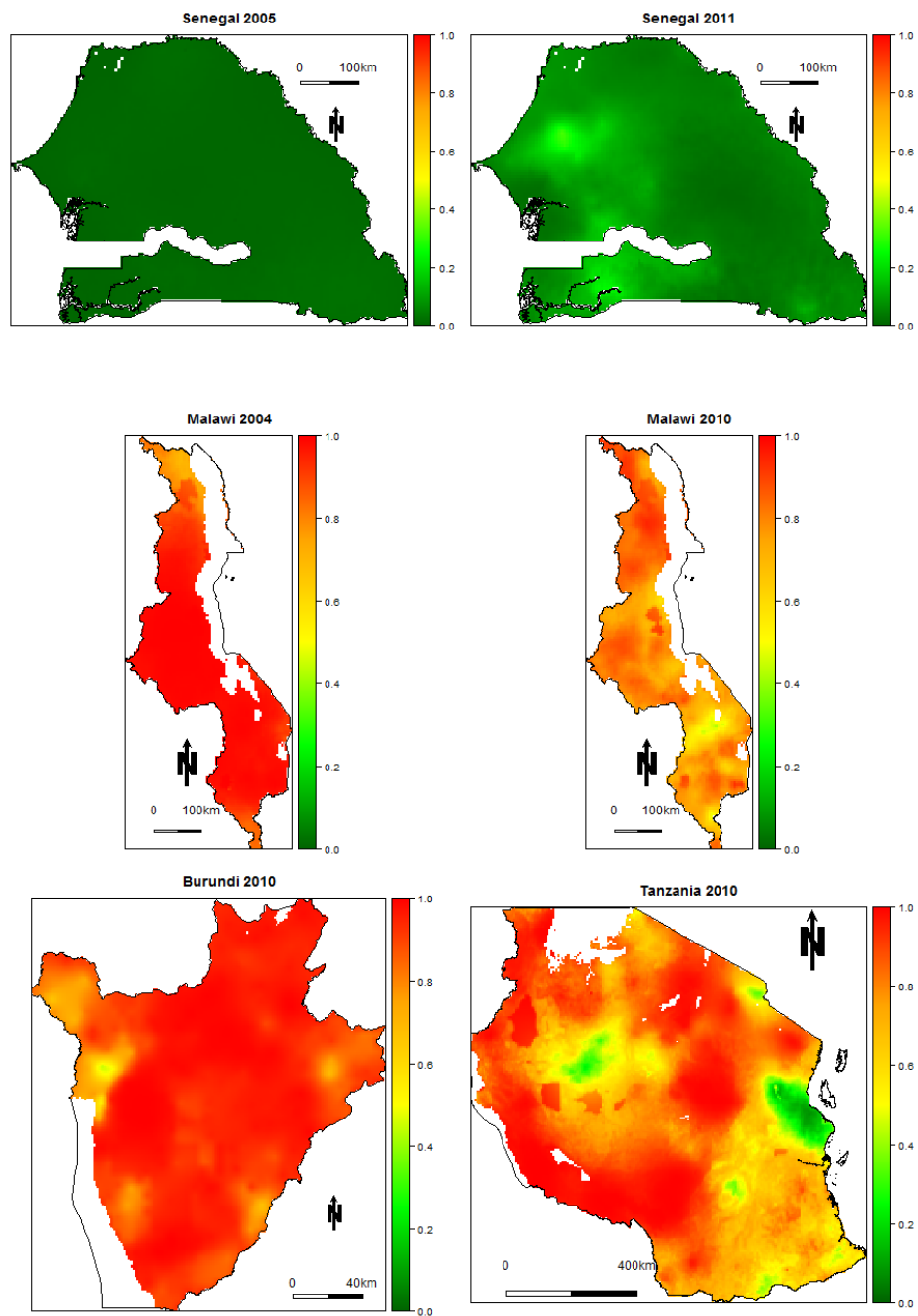
ADDITIONAL FIGURE 3.9: Predicted stunting risk maps for Senegal, Malawi, Burundi and Tanzania.
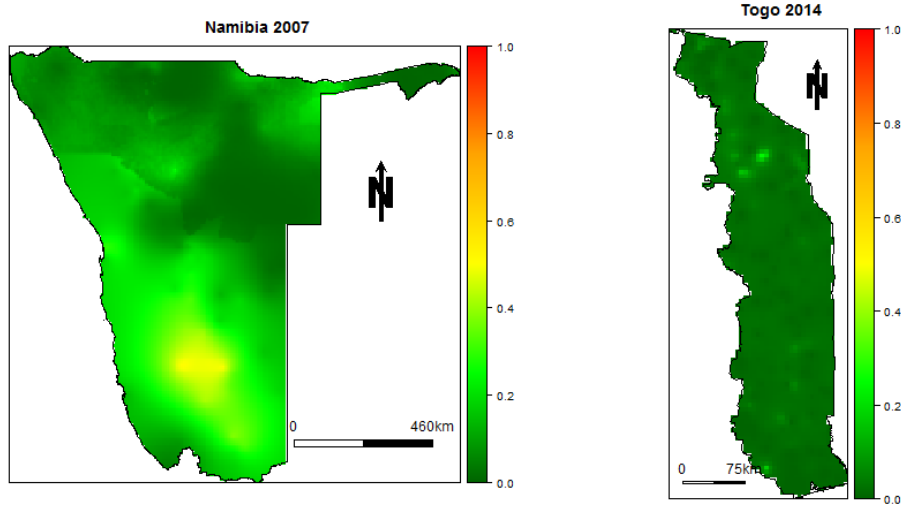
ADDITIONAL FIGURE 3.10: Predicted stunting risk maps for Namibia and Togo.

Additional File 6: Accounting for the uncertainty in malaria incidence

To account for the uncertainty in malaria incidence, we consider two possible scenarios: (A) predictive samples or measures of uncertainty from the model used to estimate malaria incidence are available; (B) the data on malaria incidence are available. In what follows, we shall use $\mathcal{M}_{ij}$ to denote the mean of the predictive distribution of malaria incidence during the first year of life of the j-th child at location x_i.

Scenario (A). Let $\mathcal{M}_{ij}^{(r)}; r = 1, 2, \ldots, R$ be predictive samples of malaria incidence for the j-th child at location x_i. The resulting likelihood function is now obtained by averaging over the samples $\mathcal{M}_{ij}^{(r)}$, i.e.

$$\frac{1}{R} \sum_{r=1}^{R} [Y_{ij} | \mathcal{M}_{ij}^{(r)}] = \frac{1}{R} \sum_{r=1}^{R} \int \int [S(x_i)][U_i][Y_{ij}|S(x_i), U_i, \mathcal{M}_{ij}^{(r)}] \, dU_i \, dS(x_i), \qquad (3.15)$$

where $[\cdot]$ is a shorthand notation for "the distribution of $\cdot$" and $[Y_{ij}|S(x_i), U_i, \mathcal{M}_{ij}^{(r)}]$ is a Gaussian distribution with mean as in equation (1) of the main manuscript and variance ω^2. The expression for $[Y_{ij}|\mathcal{M}_{ij}^{(r)}]$ is given in "Additional file 1". The resulting estimate of δ and its standard error based on (3.15) incorporate the uncertainty in $\mathcal{M}_{ij}$.

However, due to limited computer memory, the predictive samples $\mathcal{M}_{ij}^{(r)}$ may not have been stored but summaries of the overall dispersion, such as standard errors, might instead be available. Let v_{ij}^2 denote the variance of the predictive distribution for malaria incidence. By approximating this with a log-Gaussian distribution, we then generate samples $\mathcal{M}_{ij}^{(r)}$ on the logarithmic scale by simulating from a Gaussian distribution with mean

$$\log \left\{ \frac{\mathcal{M}_{ij}}{\sqrt{1 + \frac{v_{ij}^2}{\mathcal{M}_{ij}}}} \right\}$$

and variance

$$\log \left\{ 1 + \frac{v_{ij}^2}{\mathcal{M}_{ij}^2} \right\}.$$

Finally, we can use the resulting samples $\mathcal{M}_{ij}^{(r)}$ as in (3.15). However, the validity of this approach largely depends on the accuracy of the log-Gaussian approximation which is not feasible in cases where $\mathcal{M}_{ij} = 0$.

Scenario (B). In this scenario, the availability of data on malaria would allow us to develop a bivariate model for HAZ and the number of malaria episodes, denoted by W_{ij}, experienced by a child during his first year of life. This is preferred to *Scenario A* because we can then model the underlying process of malaria incidence within a geostatistical framework that is consistent with our approach used for HAZ. More specifically, we would assume that W_{ij}, conditionally on a spatial Gaussian process $\tilde{S}(x_i)$, are mutually independent Poisson variables, such that

$$\log \{\mathcal{M}_{ij}\} = \tilde{e}_{ij}^\top \tilde{\gamma} + \tilde{d}(x_i)^\top \tilde{\beta} + \tilde{S}(x_i), \tag{3.16}$$

where $\tilde{e}_{ij}$ and $\tilde{d}(x_i)$ are child-specific and spatially referenced explanatory variables with associated regression coefficients $\tilde{\gamma}$ and $\tilde{\beta}$, respectively. The joint likelihood for Y_{ij} and W_{ij} is then given by

$$[Y_{ij}, W_{ij}] = \int \int \int [S(x_i)][U_i][\mathcal{M}_{ij}][Y_{ij}|S(x_i), U_i, \mathcal{M}_{ij}][W_{ij}|\mathcal{M}_{ij}] \, dU_i \, dS(x_i) \, d\mathcal{M}_{ij}. \tag{3.17}$$

Monte Carlo methods could then be used to approximate the above integral which is not available in closed form; see, for example, Christensen (2004).

References

Abuya, B. A., Ciera, J., and Kimani-Murage, E. (2012). "Effect of mother's education on child's nutritional status in the slums of Nairobi." *BMC Pediatrics*, 12(1):80.

Agresti, A. (1996). *Categorical data analysis*, volume 990. John Wiley & Sons, New York.

Allen, L. H. (1994). "Nutritional influences on linear growth: a general review." *European journal of clinical nutrition*, 48:S75–89.

Arinaitwe, E., Gasasira, A., Verret, W., Homsy, J., Wanzira, H., Kakuru, A., Sandison, T. G., Young, S., Tappero, J. W., Kamya, M. R., et al. (2012). "The association between malnutrition and the incidence of malaria among young HIV-infected and-uninfected Ugandan children: a prospective study." *Malaria journal*, 11(1):90.

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(7572):207–211.

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., De Onis, M., Ezzati, M., Grantham-McGregor, S., Katz, J., Martorell, R., et al. (2013). "Maternal and child undernutrition and overweight in low-income and middle-income countries." *The lancet*, 382(9890):427–451.

Center for International Earth Science Information Network – CIESIN – Columbia University, International Food Policy Research Institute – IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical – CIAT (2011). Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Urban Extents Grid. Palisades, NY. NASA Socioeconomic Data and Applications Center (SEDAC). Data available at http://sedac.ciesin.columbia.edu/data/collection/grump-v1. Accessed January 2018.

Christensen, O. F. (2004). "Monte Carlo Maximum Likelihood in Model-Based Geostatistics." *Journal of Computational and Graphical Statistics*, 13(3):702–718.

Crookston, B. T., Penny, M. E., Alder, S. C., Dickerson, T. T., Merrill, R. M., Stanford, J. B., Porucznik, C. A., and Dearden, K. A. (2010). "Children Who Recover from Early Stunting and Children Who Are Not Stunted Demonstrate Similar Levels of Cognition, 2." *The Journal of nutrition*, 140(11):1996–2001.

Cunha, F. and Heckman, J. (2007). "The technology of skill formation." Technical report, National Bureau of Economic Research.

Currie, J. (2000). "Child health in developed countries." *Handbook of health economics*, 1:1053–1090.

Currie, J. (2008). "Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development." Technical report, National Bureau of Economic Research.

Custodio, E., Descalzo, M. Á., Villamor, E., Molina, L., Sánchez, I., Lwanga, M., Bernis, C., Benito, A., and Roche, J. (2009). "Nutritional and socio-economic factors associated with Plasmodium falciparum infection in children from Equatorial Guinea: results from a nationally representative survey." *Malaria journal*, 8(1):1.

Daniels, M. C. and Adair, L. S. (2004). "Growth in young Filipino children predicts schooling trajectories through high school." *The Journal of nutrition*, 134(6):1439–1446.

De Beaudrap, P., Turyakira, E., Nabasumba, C., Tumwebaze, B., Piola, P., Boum II, Y., and McGready, R. (2016). "Timing of malaria in pregnancy and impact on infant growth and morbidity: a cohort study in Uganda." *Malaria journal*, 15(1):1.

Deen, J., Walraven, G., and Von Seidlein, L. (2002). "Increased risk for malaria in chronically malnourished children under 5 years of age in rural Gambia." *Journal of tropical pediatrics*, 48(2):78–83.

DHS Surveys (2003-2014). Demographic and health survey data.

Dike, N., Onwujekwe, O., Ojukwu, J., Ikeme, A., Uzochukwu, B., and Shu, E. (2006). "Influence of education and knowledge on perceptions and practices to control malaria in Southeast Nigeria." *Social science & medicine*, 63(1):103–106.

Ehrhardt, S., Burchard, G. D., Mantel, C., Cramer, J. P., Kaiser, S., Kubo, M., Otchwemah, R. N., Bienzle, U., and Mockenhaupt, F. P. (2006). "Malaria, anemia, and malnutrition in African children– defining intervention priorities." *Journal of Infectious Diseases*, 194(1):108–114.

Fink, G., Olgiati, A., Hawela, M., Miller, J. M., and Matafwali, B. (2013). "Association between early childhood exposure to malaria and children's pre-school development: evidence from the Zambia early childhood development project." *Malaria journal*, 12(1):1–9.

Gallup, J. L. and Sachs, J. D. (2001). "The economic burden of malaria." *The American journal of tropical medicine and hygiene*, 64(1 suppl):85–96.

Genton, B., Al-Yaman, F., Ginny, M., Taraika, J., and Alpers, M. P. (1998). "Relation of anthropometry to malaria morbidity and immunity in Papua New Guinean children." *The American journal of clinical nutrition*, 68(3):734–741.

Giorgi, E. and Diggle, P. J. (2017). "PrevMap: An R Package for Prevalence Mapping." *Journal of Statistical Software*, 78(8):1–29.

Guyatt, H. L. and Snow, R. W. (2004). "Impact of malaria during pregnancy on low birth weight in sub-Saharan Africa." *Clinical microbiology reviews*, 17(4):760–769.

Headey, D. D. (2013). "Developmental drivers of nutritional change: a cross-country analysis." *World Development*, 42:76–88.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." Technical report, National Bureau of Economic Research.

Holding, P. A. and Kitsao-Wekulo, P. K. (2004). "Describing the burden of Malaria on child development: What should we be measuring and how should we be measuring it?" *The American journal of tropical medicine and hygiene*, 71(2 suppl):71–79.

Kalanda, B., van, Buuren, S., Verhoeff, F., and Brabin, B. (2005). "Catch-up growth in Malawian babies, a longitudinal study of normal and low birthweight babies born in a malarious endemic area." *Early Human Development Early Hum.Dev.*, 81(10):841–850.

Kang, H., Kreuels, B., Adjei, O., Krumkamp, R., May, J., and Small, D. S. (2013). "The causal effect of malaria on stunting: a Mendelian randomization and matching approach." *International journal of epidemiology*, 42(5):1390–1398.

Kere, N., Keni, J., Kere, J., Bobogare, A., Webber, R., and Southgate, B. (1993). "The economic impact of Plasmodium falciparum malaria on education investment: a Pacific Island case study." *The Southeast Asian journal of tropical medicine and public health*, 24(4):659–663.

Lopriore, C., Guidoum, Y., Briend, A., and Branca, F. (2004). "Spread fortified with vitamins and minerals induces catch-up growth and eradicates severe anemia in stunted refugee children aged 3–6 y." *The American journal of clinical nutrition*, 80(4):973–981.

Marriott, B. P., White, A., Hadden, L., Davies, J. C., and Wallingford, J. C. (2012). "World Health Organization (WHO) infant and young child feeding indicators: associations with growth measures in 14 low income countries." *Maternal & child nutrition*, 8(3):354–370.

Matariya, Z. R., Lodhiya, K. K., and Mahajan, R. G. (2016). "Environmental correlates of undernutrition among children of 3–6 years of age, Rajkot, Gujarat, India." *Journal of family medicine and primary care*, 5(4):834.

McGregor, I. A., Wilson, M., and Billewicz, W. (1983). "Malaria infection of the placenta in The Gambia, West Africa; its incidence and relationship to stillbirth, birthweight and placental weight." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 77(2):232–244.

Müller, O., Garenne, M., Kouyaté, B., and Becher, H. (2003). "The association between protein–energy malnutrition, malaria morbidity and all-cause mortality in West African children." *Tropical Medicine & International Health*, 8(6):507–511.

Murray, M., Murray, A., Murray, N., and Murray, M. (1978). "Diet and cerebral malaria: the effect of famine and refeeding." *The American journal of clinical nutrition*, 31(1):57–61.

Nyakeriga, A., Troye-Blomberg, M., Chemtai, A., Marsh, K., and Williams, T. (2004). "Malaria and nutritional status in children living on the coast of Kenya." *Am J Clin Nutr*, 80(6):1604–1610.

Olney, D. K., Kariger, P. K., Stoltzfus, R. J., Khalfan, S. S., Ali, N. S., Tielsch, J. M., Sazawal, S., Black, R., Allen, L. H., and Pollitt, E. (2009). "Development of nutritionally at-risk young children is predicted by malaria, anemia, and stunting in Pemba, Zanzibar." *The Journal of nutrition*, 139(4):763–772.

Onis, M. (2006). "WHO Child Growth Standards based on length/height, weight and age." *Acta paediatrica*, 95(S450):76–85.

Rieger, M. and Trommlerová, S. K. (2016). "Age-Specific Correlates of Child Growth." *Demography*, pages 1–27.

Ruel, M. T., Alderman, H., Maternal, Group, C. N. S., et al. (2013). "Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition?" *The Lancet*, 382(9891):536–551.

Rutstein, S. O., Johnson, K., MEASURE, O. M., et al. (2004). "The DHS wealth index."

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

Shanks, G. D., Hay, S. I., and Bradley, D. J. (2008). "Malaria's indirect contribution to all-cause mortality in the Andaman Islands during the colonial era." *The Lancet infectious diseases*, 8(9):564–570.

Smith, L. C. and Haddad, L. (2015). "Reducing child undernutrition: past drivers and priorities for the post-MDG era." *World Development*, 68:180–204.

Snow, R., Byass, P., Shenton, F., and Greenwood, B. (1991). "The relationship between anthropometric measurements and measurements of iron status and susceptibility to

malaria in Gambian children." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(5):584–589.

Somi, M. F., Butler, J. R., Vahid, F., Njau, J., Kachur, S. P., and Abdulla, S. (2007). "Is there evidence for dual causation between malaria and socioeconomic status? Findings from rural Tanzania." *The American journal of tropical medicine and hygiene*, 77(6):1020–1027.

Sovi, A., Govoétchan, R., Tokponnon, F., Hounkonnou, H., Aïkpon, R., Agossa, F., Gnanguenon, V., Salako, A. S., Agossou, C., Ossè, R., Okè, M., Gbénou, D., Massougbodji, A., and Akogbéto, M. (2013). "Impact of land-use on malaria transmission in the Plateau region, southeastern Benin." *Parasites & Vectors*, 6(1):352.

Stevens, G. A., Finucane, M. M., Paciorek, C. J., Flaxman, S. R., White, R. A., Donner, A. J., Ezzati, M., Group, N. I. M. S., et al. (2012). "Trends in mild, moderate, and severe stunting and underweight, and progress towards MDG 1 in 141 developing countries: a systematic analysis of population representative data." *The Lancet*, 380(9844):824–834.

Teklehaimanot and Paola Mejia, A. (2008). "Malaria and poverty." *Annals of the New York Academy of Sciences*, 1136(1):32–37.

Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., and Greenwood, B. (1999). "Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results." *The American journal of tropical medicine and hygiene*, 61(1):2–8.

Thuilliez, J., Sissoko, M. S., Toure, O. B., Kamate, P., Berthelemy, J.-C., and Doumbo, O. K. (2010). "Malaria and primary education in Mali: a longitudinal study in the village of Doneguebougou." *Social science & medicine*, 71(2):324–334.

Uddenfeldt Wort, U., Hastings, I. M., Carlstedt, A., Mutabingwa, T., and Brabin, B. J. (2004). "Impact of El Nino and malaria on birthweight in two areas of Tanzania with different malaria transmission patterns." *International journal of epidemiology*, 33(6):1311–1319.

UNICEF et al. (2013). "Improving child nutrition: The achievable imperatives for global progress." *New York: UNICEF.* ISBN: 978-92-806-4686-3 https://www.unicef.org/publications/index_68661.html. Accessed December 2015.

UNICEF et al. (2015). "WHO, World Bank Group joint child malnutrition estimates." *Levels and trends in child malnutrition: Key findings of the 2015 edition. Global Database on Child Growth and Malnutrition.*

UNICEF et al. (2016). "WHO, World Bank Group joint child malnutrition estimates." *Levels and trends in child malnutrition: Key findings of the 2016 edition. Global Database on Child Growth and Malnutrition.*

Vella, V., Tomkins, A., BORGESI, A., Migliori, G. B., and Oryem, V. Y. (1994). "Determinants of stunting and recovery from stunting in northwest Uganda." *International journal of epidemiology*, 23(4):782–786.

Verhoef, H., West, C., Veenemans, J., and Begui, Y. (2002). "Stunting may determine the severity of malaria-associated anemia in African children." *Pediatrics.*

Victora, C. G., de Onis, M., Hallal, P. C., Blössner, M., and Shrimpton, R. (2010). "Worldwide timing of growth faltering: revisiting implications for interventions." *Pediatrics*, pages peds. 2009–1519.

Walker, S. P., Chang, S. M., Powell, C. A., and Grantham-McGregor, S. M. (2005). "Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted Jamaican children: prospective cohort study." *The lancet*, 366(9499):1804–1807.

Webb, P. and Block, S. (2012). "Support for agriculture during economic transformation: Impacts on poverty and undernutrition." *Proceedings of the National Academy of Sciences*, 109(31):12309–12314.

Wood, S. and Wood, M. S. (2015). "Package mgcv." *R package version*, 1:29.

World Bank Indicators (2003-2014). World bank development indicators data.

World Health Organization (2016). "World malaria report 2016."

World Health Organization (2017). "World malaria report 2017."

World Health Organization et al. (2006). Who child growth standards: length/height for age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age, methods and development. Data available at http://www.who.int/childgrowth/en/. Accessed 1 December 2015.

Chapter 4

Paper 3. A Geostatistical Framework for Combining Spatially Referenced Disease Prevalence Data from Multiple Diagnostics

Benjamin Amoah*, Emanuele Giorgi, Peter J. Diggle

*Correspondence: Benjamin Amoah, CHICAS Research Group, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, LA1 4YB, UK.

E-mail: b.amoah@lancaster.ac.uk

4.1 Summary

Multiple diagnostic tests are often used due to limited resources or because they provide complementary information on the epidemiology of a disease under investigation. Existing statistical methods to combine prevalence data from multiple diagnostics ignore the potential over-dispersion induced by the spatial correlations in the data. To address this issue, we develop a geostatistical framework that allows for joint modelling of data from multiple diagnostics by considering two main classes of inferential problems: (1) to predict prevalence for a gold-standard diagnostic using low-cost and potentially biased alternative tests; (2) to carry out joint prediction of prevalence from multiple tests. We apply the proposed framework to two case studies: mapping *Loa loa* prevalence in Central and West Africa, using miscroscopy and a questionnaire-based test called RAPLOA; mapping *Plasmodium falciparum* malaria prevalence in the highlands of Western Kenya using polymerase chain reaction and a rapid diagnostic test. We also develop a Monte Carlo procedure based on the variogram in order to identify parsimonious geostatistical models that are compatible with the data. Our study highlights (*i*) the importance of accounting for diagnostic-specific residual spatial variation and (*ii*) the benefits accrued from joint geostatistical modelling so as to deliver more reliable and precise inferences on disease prevalence.

Keywords

Author summary

Why was this study done?

> The use of different diagnostics test in disease prevalence surveys has necessitated the development of statistical methods that allow for the joint analysis of data from several diagnostics.

> Geostatistical methods for disease mapping are reliant on the spatial correlation structures in data, which might differ from one diagnostic to another.

> This study was conducted to develop geostatistical methods that allow for the joint analysis of data from several diagnostics by exploring the different (spatial) correlations in the data from several diagnostics.

What did the researchers do and find?

> We developed two geostatistical models. The first model allows for the prediction of disease prevalence as defined by a gold standard diagnostic, by exploring the data from a more economical and potentially biased alternative. The second model allows for the prediction for prevalence as defined by each of several complementary diagnostics by exploring the spatial correlations structures that are common to all diagnostics and those that are unique to each diagnostic.

> We developed Monte Carlo procedures for validating such models.

> We applied the novel methodology to two challenging disease mapping problems and found that different diagnostics can have different spatial correlation structures, which need to be accounted for in order to carry out reliable predictive inferences. We also found that the joint modelling of prevalence data from different diagnostics can lead to more precise predictive inferences.

What do these findings mean?

> The study shows two ways of combining data from multiple diagnostics depending on whether or not one of the diagnostics is to be considered a gold standard. The methods are important for both researchers in model-based geostatistics to build on, but also for spatial epidemiologist, who may use the methods in disease mapping applications.

> Our results highlight the need to check and account for different spatial correlation in data in order to make reliable predictions of disease prevalence.

4.2   Introduction

Disease mapping denotes a class of problems in public health where the scientific goal is to predict the spatial variation in disease risk on a scale that can range from sub-national to global (Liu et al., 2012, Murray et al., 2014). Understanding the geographical distribution of a disease is particularly important in the decision-making process for the planning, implementation, monitoring and evaluation of control programmes (Bhatt et al., 2015, World Health Organization, 2017). In this context, model-based geostatistical methods (Diggle et al., 1998) have been especially useful in low-resource settings (Diggle and Giorgi, 2016, Gething et al., 2012, Zouré et al., 2014) where disease registries are non-existent or geographically incomplete, and monitoring of the disease burden is carried out through cross-sectional surveys and passive surveillance systems. However, there are a lot of cohort studies in Africa are ongoing (Mudie et al., 2019) or have been carried out (Dalal et al., 2015, Haas et al., 2015).

It is often the case that prevalence data from a geographical region of interest are obtained using different diagnostic tests for the same disease under investigation. The reasons for this are manifold. For example, when the goal of geostatistical analysis is to map disease risk on a continental or global scale by combining data from multiple surveys, dealing with the use of different diagnostic tests may be unavoidable. In other cases, gold-standard diagnostic tests are often expensive and require advanced laboratory expertise and technology which may not always be available in constrained resource settings. This requires the use of more cost-effective alternatives for disease testing in order to attain a required sample size. Different diagnostics might also provide complementary information of intrinsic scientific interests into the spatial variation of disease risk and the distribution of hotspots.

In the absence of statistical methods that allow for the joint analysis of multiple diagnostics, most studies have reported separate analyses. A shortcoming of this approach is that it ignores, and therefore fails to explain, possible correlations between prevalence of different diagnostics. Statistical inference might benefit from a joint analysis, which

can yield more efficient estimation of regression parameters (Song et al., 2009) and more precise predictions of prevalence.

However, different diagnostic tests can exhibit considerable disparities in the estimates of disease prevalence for the same population, or even the same individuals. Obvious sources of such variation include differences in sensitivity and specificity. Furthermore, different diagnostics may exhibit differences in their association with the same risk factors. In a geostatistical context, there may also be differences between the spatial covariance structures of different diagnostics.

These aspects highlight the potential challenges that joint modelling of multiple diagnostics needs to take into account. In this paper, we address such issues in order to develop a general framework for geostatistical analysis and describe the application of this framework to *Loa loa* and malaria mapping in Africa. The methodology presented in this paper is also an extension of previous work published by authors; see, for example, Crainiceanu et al. (2008) and Diggle and Giorgi (2016).

The structure of the paper is as follows. In Section 4.3, we describe the two motivating applications. In Section 4.4, we review existing methods for combining prevalence data from different diagnostics. In Section 4.5, we introduce a geostatistical framework for combining data from two diagnostics and distinguish two main classes of problems that arise in this context. In Sections 4.6 and 4.7, we apply this framework to the two case studies introduced in Section 4.3. In Section 4.8, we discuss methodological extensions to more than two diagnostics.

The R code developed for the applications of Section 4.6 and Section 4.7 is available on request from the authors.

4.3 Motivating applications

4.3.1 Loa loa mapping in Central and West Africa

Loiasis is a neglected tropical disease that has received an increased attention due to its impact on the control of a more serious infectious disease, onchocerciasis, that is endemic

in large swathes of sub-Saharan Africa. Mass administration of the drug ivermectin confers protection against onchocerciasis, but individuals who are highly co-infected with *Loa loa* - the Loiasis parasite - can develop severe and occasionally fatal adverse reaction to the drug (Boussinesq et al., 1998).

Boussinesq et al. (2001) have shown that high levels in *Loa loa* prevalence within a community are strongly associated with a high parasite density. For this reason, Zouré et al. (2011) have suggested that precautionary measures should be put in place before the roll-out of mass drug administration with ivermectin in areas where prevalence of infection with *Loa loa* is greater than 20%.

In order to carry out a rapid assessment of the *Loa loa* burden in endemic areas a questionnaire instrument, named RAPLOA, was developed as a more economically feasible alternative to the standard microscopy-based microfilariae (MF) detection in blood smears (Takougang et al., 2002). To validate the RAPLOA methodology against microscopy, cross-sectional surveys using both diagnostics were carried out in four study sites in Cameroon, Nigeria, Republic of Congo and the Democratic Republic of Congo (see Wanji et al. (2012) and Figure 4.4 in Web Appendix B).

In this study, the objective of statistical inference is to develop a calibration relationship between the two diagnostic procedures. This could then be applied to map microscopy-based MF prevalence in areas where the more economical RAPLOA questionnaire is the only feasible option.

4.3.2 Malaria mapping in the highlands of Western Kenya

Malaria continues to be a global public health challenge, especially in sub-Sharan Africa which, in 2016, accounted for about 90% of all the 445,000 estimated malaria deaths worldwide (World Health Organization, 2017). Polymerase chain reaction (PCR) and a rapid diagnostic test (RDT) are two of the most commonly used procedures for detecting *Plasmodium falciparum*, the deadliest species of the malaria parasites. PCR is highly sensitive and specific, but its use is constrained by high costs and the need for highly trained technicians. RDT is simpler to use, cost-effective and requires minimal training,

but is less sensitive than PCR (Tangpukdee et al., 2009). Recent studies have reported that PCR and RDT can lead to the identification of different malaria hotspots, i.e. areas where disease risk is estimated to be unexpectedly high (Mogeni et al., 2017). In this context, mapping of both diagnostics is of epidemiological interest since their effective use is dependent on the level of malaria transmission, with PCR being the preferred testing option in low-transmission settings (Mogeni et al., 2017).

In order to investigate this issue, a malariometric survey was conducted using both RDT and PCR in two highland districts of Western Kenya (see Figure 4.5 in Web Appendix B); see Stevenson et al. (2015) for a descriptive analysis of this study. In this scenario, a joint model for the reported malaria counts from the two diagnostics could allow to exploit their cross-correlation and identify malaria hotspots more accurately.

4.4 Literature review

We formally express the format of geostatistical data from multiple diagnostics as

$$\mathcal{D} = \{(x_{ik}, n_{ik}, y_{ijk}) : j = 1, \ldots, n_{ik}; i = 1, \ldots, N; k = 1, \ldots, K\} \tag{4.1}$$

where y_{ijk} is a binary outcome taking value 1 if the j-th individual at (i, k)-th location x_{ik} tests positive for the disease under investigation using the k-th diagnostic procedure, and 0 otherwise. We use p_{ijk} to denote the probability that an individual has a positive test outcome from the k-th diagnostic. When data are only available as aggregated counts, we replace y_{ijk} in (4.1) with $y_{ik} = \sum_{j=1}^{n_{ik}} y_{ijk}$ and p_{ijk} with p_{ik}. When all diagnostic tools are used at each location, we replace x_{ik} with x_i (in which case i becomes a location index), although this is not a requirement in the development of our methodology.

In the remainder of this section, we review non-spatial methods for joint modelling of the p_{ijk} across multiple diagnostics and a geostatistical modelling approach proposed by Crainiceanu et al. (2008).

4.4.1 Non-spatial approaches

Existing non-spatial methods for the analysis of data from multiple diagnostics fall within two main classes of statistical models: generalised linear models (GLMs) and their random-effects counterpart, generalised linear mixed models (GLMMs).

Mappin et al. (2015) analysed data on *P. falciparum* prevalence from RDT and microscopy outcomes from sub-Saharan Africa, using a standard probit model

$$\Phi^{-1}(p_{i1}) = \beta_0 + \beta_1 \Phi^{-1}(p_{i2}), \tag{4.2}$$

thus assuming a linear relationship between the prevalences p_{ik} on the probit scale. Wanji et al. (2012) used a similar approach for *Loa loa* in order to study the relationship between microscopy and RAPLOA prevalence by replacing the probit link in (4.2) with the logit. This model was also used by Wu et al. (2015) to estimate the relationship between RDT, microscopy and PCR, for each pair of diagnostics. A major limitation of these approaches based on standard GLMs is that they do not account for any overdispersion that might be induced by unmeasured (spatial) risk factors. Ignoring overdispersion can result in invalid inferences on the regression relationships as a result of exceedingly narrow confidence intervals for the regression coefficients (Thomson et al., 1999).

Coffeng et al. (2013) proposes a bivariate GLMM for joint modelling of data on onchocerciasis nodule prevalence and skin MF prevalence in adult males sampled across 148 villages in 16 African countries. More specifically, the linear predictor of such model can be expressed as

$$\log\left\{\frac{p_{ijk}}{1 - p_{ijk}}\right\} = \boldsymbol{d}_{ij}^{\top}\boldsymbol{\beta}_k + Z_i + V_{ij}, \tag{4.3}$$

where the random effects terms Z_i and V_{ij} are independent zero-mean Gaussian variables accounting for unexplained variation between-villages and between-individuals within villages, respectively. Using this approach, Coffeng et al. (2013) estimated a strong

positive correlation between nodule and MF prevalence but also reported a variation in the strength of this relationship across study sites.

4.4.2 The Crainiceanu, Diggle and Rowlingson model

A standard geostatistical model for prevalence data from a single diagnostics, where y_i out of n_i individuals at each of N locations x_i have tested positive for the disease of interest is specified as follows. First, $S(x)$ is a zero-mean stationary and isotropic Gaussian process with variance σ^2, and Z_i are zero-mean independent and identically distributed Gaussian random variables with variance τ^2. Second, conditionally on $S(\cdot)$ and Z, the y_i are independent, binomial random variables with probabilities $p(x_i)$, where

$$\mathrm{logit}\{p(x_i)\} = \boldsymbol{d}^\top(x_i)\boldsymbol{\beta} + S(x_i) + Z_i, \tag{4.4}$$

where $\mathrm{logit}\{u\} = \log\{u/[1-u]\}$ and $\boldsymbol{d}(x_i)$ is a vector of spatially varying covariates corresponding regression parameters $\boldsymbol{\beta}$.

Diggle et al. (1998) give a detailed description of this framework, within which Crainiceanu et al. (2008) proposed a bivariate geostatistical model (henceforth CDRM) to analyse data on microscopy and RAPLOA *Loa loa* prevalence (see Section 4.3.1). To the best of our knowledge, CDRM is the only existing approach that attempts to model the spatial correlation between two diagnostics.

Let $k = 1$ correspond to the RAPLOA questionnaire, and $k = 2$ to microscopy. To emphasize the spatial context, we now write $p_{ik} = p_k(x_i)$; CDRM can then be expressed as

$$\begin{cases} \mathrm{logit}\{p_1(x_i)\} = \boldsymbol{d}^\top(x_i)\boldsymbol{\beta} + S(x_i) \\[2mm] \mathrm{logit}\{p_2(x_i)\} = \alpha_0 + \alpha_1\mathrm{logit}\{p_1(x_i)\} + Z_i \end{cases} \tag{4.5}$$

Crainiceanu et al. (2008) also provide empirical evidence to justify the assumption of a logit-linear relationship between the two diagnostics.

A limitation of the CDRM is that it assumes proportionality on the logit scale between the residual spatial fields associated with RAPLOA and microscopy. In our re-analysis in Section 4.6, we use a Monte Carlo procedure to test this hypothesis.

4.5 Two classes of bivariate geostatistical models

We now develop two modelling strategies that address the specific objectives of the two case studies introduced in Section 4.3. Our focus in this section will be restricted to the case of two diagnostics (hence $K = 2$). We discuss the extension to more than two in Section 4.8.

4.5.1 Case I: Predicting prevalence for a gold-standard diagnostic

Let $S_1(x)$ and $S_2(x)$ be two independent stationary and isotropic Gaussian processes; also, let $f_1\{\cdot\}$ and $f_2\{\cdot\}$ be two functions with domain on the unit interval $[0, 1]$ and image on the real line. We propose to model data from two diagnostics, with $k = 2$ denoting the gold-standard, as

$$
\begin{cases}
f_1\{p_1(x_i)\} = \boldsymbol{d}^\top(x_i)\boldsymbol{\beta_1} + S_1(x_i) + Z_{i1} \\
f_2\{p_2(x_i)\} = \boldsymbol{d}^\top(x_i)\boldsymbol{\beta_2} + S_2(x_i) + Z_{i2} + \alpha f_1\{p_1(x_i)\}.
\end{cases}
\tag{4.6}
$$

In our applications, we specify exponential correlation functions for $S_k(x)$, $k = 1, 2$, hence

$$
\text{cov}\{S_k(x), S_k(x')\} = \sigma_k^2 \exp\{\|x - x'\|/\phi_k\},
$$

where σ_k^2 is the variance of $S_k(x)$ and ϕ_k is a scale parameter regulating how fast the spatial correlation decays to zero for increasing distance. Finally, we use τ_k^2 to denote the variance of the Gaussian noise Z_{ik} which is independent of $S_1(x)$ and $S_2(x)$.

Selection of suitable functions f_1 and f_2 can be carried out, for example, by exploring the association between the empirical prevalences of the two diagnostics in order to identify transformations that render their relationship approximately linear. Alternatively,

subject matter knowledge could be used to constrain the admissible forms for f_1 and f_2; see, for example, Irvine et al. (2016) who derive a functional relationship between MF and an immuno-chromatographic test for prevalence of lymphatic filariasis by making explicit assumptions on the distribution of worms and their reproductive rate in the general population.

The proposed model in (4.6) is more flexible than the CDRM because (*i*) it allows for diagnostic-specific unstructured variation Z_{ik} and, more importantly, (*ii*) relaxes the assumption of proportionality between the residual spatial fields of the two diagnostics through the introduction of $S_2(x)$.

4.5.2 Case II: Joint prediction of prevalence from two complementary diagnostics

Let $S_1(x)$ and $S_2(x)$ be two independent Gaussian processes, and Z_{ik} Gaussian noise, each having the same properties as defined in the previous section. We now introduce a third stationary and isotropic Gaussian process $T(x)$ having unit variance and exponential correlation function with scale parameter ϕ_T.

Our proposed approach for joint prediction of prevalence from two diagnostics, when both are of interest, is expressed by the following equation for the linear predictor

$$f_k\{p_{jk}(x_i)\} = \boldsymbol{d}_{ij}^\top \boldsymbol{\beta}_k + \nu_k \big[S_k(x_i) + T(x_i) \big] + Z_{ik}. \tag{4.7}$$

The spatial processes $S_k(x)$ and $T(x)$ accounts for unmeasured risk factors that are specific to each and common to both diagnostics, respectively. The resulting variogram for the linear predictor is

$$
\begin{aligned}
\gamma_k(u) &= \mathrm{E}\Big[\big\{\big(\nu_k\big(S_k(x) + T(x)\big) + Z_k(x)\big) - \big(\nu_k\big(S_k(x') + T(x')\big) + Z_k(x')\big)\big\}^2\Big] \\
&= \tau_k^2 + \nu_k^2\big[1 - \exp(-u/\phi_T) + \sigma_k^2\{1 - \exp\big(-u/\phi_{S_k}\big)\}\big],
\end{aligned}
\tag{4.8}
$$

and the cross-variogram between the linear predictors of the two diagnostics is

$$
\begin{aligned}
\gamma_{1,2}(u) &= \mathrm{E}\Big[\Big\{\Big(\nu_1\big(S_1(x)+T(x)\big)+Z_k(x)\Big)-\Big(\nu_2\big(S_2(x')+T(x')\big)+Z_k(x')\Big)\Big\}^2\Big] \\
&= 0.5\{\tau_1^2+\tau_2^2+\nu_1^2(1+\sigma_1^2)+\nu_2^2(1+\sigma_2^2)\}-\sigma_1\sigma_2\exp(-u/\phi_T). \quad (4.9)
\end{aligned}
$$

Given the relatively large number of parameters, fitting the model may require a pragmatic approach. In order to identify a parsimonious model for the data, we recommend an incremental modelling strategy, whereby a simpler model is used in a first analysis (e.g. by setting $S_k(x)=0$ for all x) and more complexity is then added in response to an unsatisfactory validation check, as described in Section 4.5.3.

Figure 4.1 gives two directed acyclic graph representations of the models in (4.6) (left panel) and (4.7) (right panel), showing their distinctive asymmetric and symmetric structures. In the first model, stochastic independence between the two diagnostics is simply recovered by setting the parameter $\alpha=0$. If this is a scientifically relevant hypothesis, we can test it through the likelihood ratio. In the second model, independence can only be achieved if $T(x)=0$ for all x. We do not consider this to be a credible assumption in the application of Section 4.3.1.

4.5.3 Inference and model validation

We estimate the unknown parameters of the asymmetric and symmetric models using the Monte Carlo Maximum Likelihood (MCML) method (Christensen, 2004, Geyer, 1991). To carry out spatial predictions at a set of unobserved locations, we plug the MCML estimates into a Markov Chain Monte Carlo algorithm for simulation from the distribution of the random effects conditional on the data. We summarise our predictive inferences on prevalence using the mean, standard deviation and exceedance probabilities, the latter defined as the probability that the predictive distribution of prevalence exceeds a predefined threshold. Details on the derivation and approximation of the log-likelihood function are given in Web Appendix A.
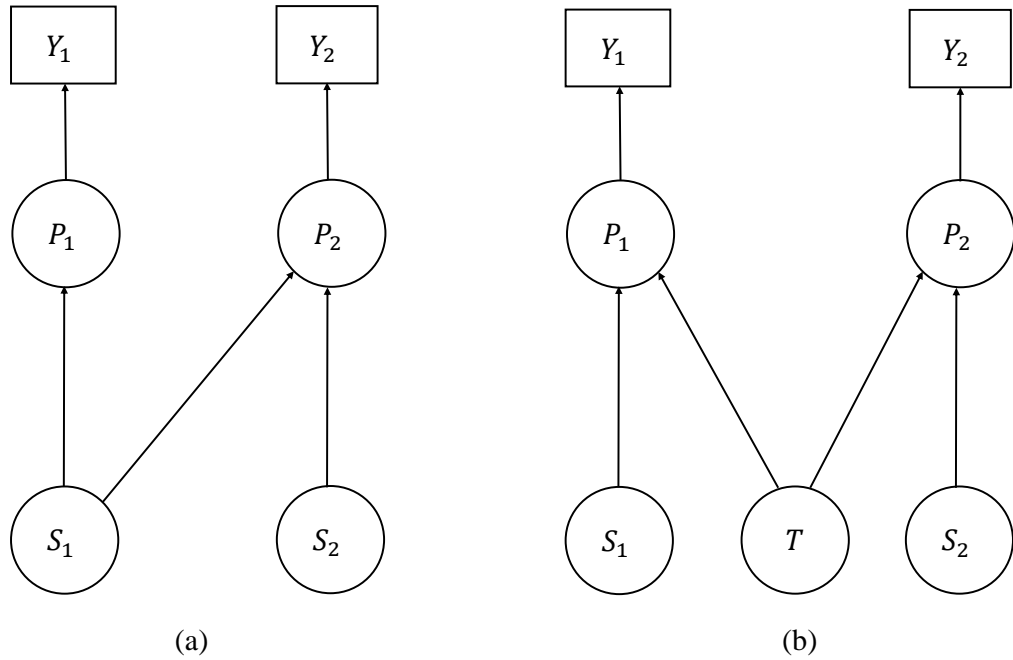
FIGURE 4.1: Directed acyclic graphs for the bivariate geostatistical models in (4.6) (left panel) and (4.7) (right panel). Circles and squares identify latent variables and the outcome random variables, respectively.

To validate the spatial covariance structure of the proposed models, while still assuming stationarity and isotropy of the Gaussian process, we proceed as follows. We first re-write both models in the general form

$$f_k\{p_{jk}(x_i)\} = \mu_{ijk} + W_k(x_i), \tag{4.10}$$

where μ_{ijk} is the mean component expressed as a regression on the available covariates. In (4.10), if we set $W_1(x_i) = S_1(x) + Z_{i1}$ and $W_2(x_i) = S_2(x_i) + Z_{i2} + \alpha\{f_1(x_i)\}$, then (4.10) reduces to the asymmetric model (4.6); if, instead, $W_k(x_i) = \nu_k\big(S_k(x_i) + T(x_i)\big) + Z_{ik}$, we recover the symmetric model (4.7).

We define the empirical variogram of $W_k(x)$ to be

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \big\{\hat{W}_k(x_i) - \hat{W}_k(x_j)\big\}^2, \tag{4.11}$$

and the empirical cross-variogram between $W_k(x)$ and $W_{k'}(x)$ to be

$$\hat{\gamma}_{k,k'}(u) = \frac{1}{2|N(u)|} \sum_{(i,j) \in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_{k'}(x_j) \right\}^2, \tag{4.12}$$

where $N(u) = \{(i,j) : ||x_i - x_j|| = u, i \neq j\}$ and $\hat{W}_k(x_i)$ is the mean of the distribution of $W_k(x_i)$ conditioned to the data. To test whether the adopted spatial structure for $W_k(x)$ is compatible with the data, we then proceed through the following steps.

Step 0. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models (i.e. $W_k(x_i) = S_k(x_i) + Z_{ik}$, where $S_k(x)$, $k = 1, 2$, are independent processes) and compute the empirical variogram, $\hat{\gamma}_k$, for $k = 1, 2$, and cross-covariogram, $\hat{\gamma}_{1,2}$.

Step 1. Simulate prevalence data as in (4.1) from the adopted model for $W_k(x)$ by plugging-in the MCML estimates. Fit separate standard geostatistical models as in *Step 0* and compute the empirical variogram, $\hat{\gamma}_k$, for $k = 1, 2$, and cross-covariogram, $\hat{\gamma}_{1,2}$, for the simulated dataset.

Step 2. Repeat *Step 1* a large enough number of times, say M.

Step 3. Use the resulting M empirical variograms and cross-variograms to generate 95% confidence intervals at each point of a set of pre-defined distance bins.

If either the empirical variogram or cross-variogram in *Step 0* fall fully or partly outside the 95% confidence intervals, we conclude that the model is not able to capture the spatial structure of the data satisfactorily.

To validate the suitability of the adopted link functions f_1 and f_2 in (4.6) and (4.7), we plot the Pearson's residuals against the fitted prevalence, where these are obtained from the separate models in *Step 0*. In the case of the asymmetric model, we also check that the chosen f_1 and f_2 are, approximately, linearly related. To this end, we fit a geostatistical model for the diagnostic $k = 2$ by setting its linear predictor to

$$f_2\{p_2(x_i)\} = \boldsymbol{d}^\top(x_i)\boldsymbol{\beta}_2 + S_2(x_i) + Z_{i2} + s\{\hat{p}_1(x_i)\}, \tag{4.13}$$

where $\hat{p}_1(x_i)$ is the plug-in estimate of $p_1(x_i)$, as obtained from the geostatistical model in *Step 0* for $k = 1$, and $s(\cdot)$ is a smoothing function, e.g. a cubic spline. We then re-fit the model in (4.13) by replacing $s(\cdot)$ with $\alpha f_1(\cdot)$. Finally, we plot the fitted $\hat{s}$ against $\hat{\alpha} f_1$ and, if this shows linear relationship, we then conclude that f_1 and f_2 are linearly related.

4.6 Application I: Re-analysis of the *Loa loa* data

A total of 223 villages were sampled in the four study sites (see Figure 4.4 in Web Appendix B). The analysis of the Pearson's residuals based on two separate models does not show any evidence against the specified logit-link functions (see Figure 4.7 (a) and 4.7(b) in Web Appendix B). Also, the log-odds from the two diagnostics show an approximately linear relationship (see Figure 4.7(c) in Web Appendix B). Each of RAPLOA and microscopy prevalences also exhibits a highly non-linear relationship with surface elevation (see Figure 5 in Web Appendix B), which we capture using a piecewise linear spline with knots at 750 meters and 1015 meters.

We consider the two following specification of the asymmetric model (4.6).

- Model 1: a slightly modified, more flexible, version of the CDRM, given by

$$
\begin{cases}
\text{logit}\{p_1(x_i)\} = \mu_1(x_i) + S_1(x_i) + Z_{i1} \\
\text{logit}\{p_2(x_i)\} = \mu_2(x_i) + \alpha\,\text{logit}\{p_1(x_i)\} + Z_{i2}
\end{cases}, \tag{4.14}
$$

  where

$$
\begin{aligned}
\mu_k(x_i) \;=\; & \beta_{k,0} + \beta_{k,1}\min\{e(x_i), e_1\} + \beta_{k,2} I(e(x_i) > e_1)\min\{e(x_i) - e_2, e_2 - e_1\} \\
& + \beta_{k,3}\max\{e(x_i) - e_2, 0\}, \quad k = 1, 2,
\end{aligned}
$$

  where $e(x)$ denotes the elevation in meters at location x, $e_1 = 750$, $e_2 = 1015$ and $I(\mathcal{P})$ is the indicator function. In this parameterisation, $\beta_{k,1}$ is the effect of

elevation on prevalence below 750 meters, $\beta_{k,2}$ its effect between 760 and 1015 meters, and $\beta_{k,3}$ its effect above 1015 meters.

- Model 2: obtained by incorporating an additional spatial process $S_2(x)$, independent of $S_1(x)$, in Model 1 to give

$$
\begin{cases}
\text{logit}\{p_1(x_i)\} = \mu_1(x_i) + S_1(x_i) + Z_{i1} \\
\text{logit}\{p_2(x_i)\} = \mu_2(x_i) + S_2(x_i) + \alpha\text{logit}\{p_1(x_i)\} + Z_{i2}.
\end{cases}, \tag{4.15}
$$

4.6.1 Results

Table 4.1 reports the MCML estimates obtained for Models 1 and 2. As expected, both models show a significant and positive logit-linear relationship between RAPLOA and miscroscopy. However, Model 2, which includes the additional spatial process $S_2(x)$, is also able to capture spatial variation in microscopy prevalence on a scale of about 24 meters. Overall, we observe that, except for τ_1^2 and τ_2^2, all other parameters common to both Models 1 and 2 have comparable point and interval estimates. The estimated parameter τ_2^2 is about three times smaller in Model 2 than the estimate from Model 1 but we also note that confidence intervals from each model are largely overlapping. We note that the regression parameters for both Models 1 and 2 are however not significant in themselves; however, they are maintained since the improved the overall fit of the model.

We use the validation procedure of Section 4.5.3 to test which of the two models better fits the spatial structure of the data. The results (see Figure 4.9) show a satisfactory assessment of Model 2, whereas for Model 1 the empirical variogram for microscopy partly falls outside the 95% confidence band, questioning its validity. The additional checks on the cross-variogram (Figure 4.10), the suitability of the logit functions (Figure 4.7(a)-(b)) and the linear relationship between the two linear predictors (Figure 4.7(c)) also yielded satisfactory results for Model 2.

We now compare the predictive inferences on microscopy prevalence between the two models in order to assess whether the introduction of $S_2(x)$ makes a tangible difference.

Figure 4.2 shows the point estimates for microscopy prevalence and the exceedance probabilities for a 20% prevalence threshold under Model 1 (upper panels), under Model 2 (middle panels), and the difference between the two (lower panels, Model 2 - Model 1). Overall, the predicted spatial pattern in prevalence from the two models is similar, but with substantial local differences. The difference between the point estimates for prevalence ranges from -0.12 to 0.13, while the difference between the two exceedance probabilities ranges from -0.44 to 0.59.

4.6.2   Simulation Study

We carry out a simulation study in order to quantify the effects on the predictive inferences for prevalence when ignoring microscopy-specific residual spatial variation, as in the case of Model 1 when Model 2 is the true model. To this end, we simulate 10,000 Binomial data-sets under Model 2 by setting its parameters to the estimates of Table 4.1 and fit both models. We then carry out predictions for microscopy prevalence over 20 unobserved locations that we randomly select from the four study sites, five from each site. We randomly select a different set of 20 prediction locations for each of the 10,000 simulations. We summarise the results using the 95% coverage probability (CP), the root-mean-square-error (RMSE) and the 95% predictive interval length (PIL). Table 4.2 shows the three metrics averaged over all simulations and prediction locations for Model 1 and Model 2. The CP of Model 1 (81.1%) is well below its nominal level of 95%. This is also reflected by a smaller PIL for Model 1, suggesting that this provides unreliably narrow 95% predictive intervals for prevalence. Finally, we note that Model 1 also has a larger RMSE than Model 2.

Having chosen Model 2 as the best model, to asses the effects of parameter uncertainty on the predictive exceedance probabilities, we predict 10,000 prevalence surfaces over all the study sites, where each realisation of simulated surface uses parameter values drawn at random from the multivariate Normal sampling distribution of the MCML estimates of Model 2. We then take the exceedance probabilities based on all the prevalence (See Figure 4.11). The mean and maximum absolute differences are respectively 0.06 and 0.2, and the mean and maximum relative differences are 21% and 30% respectively.
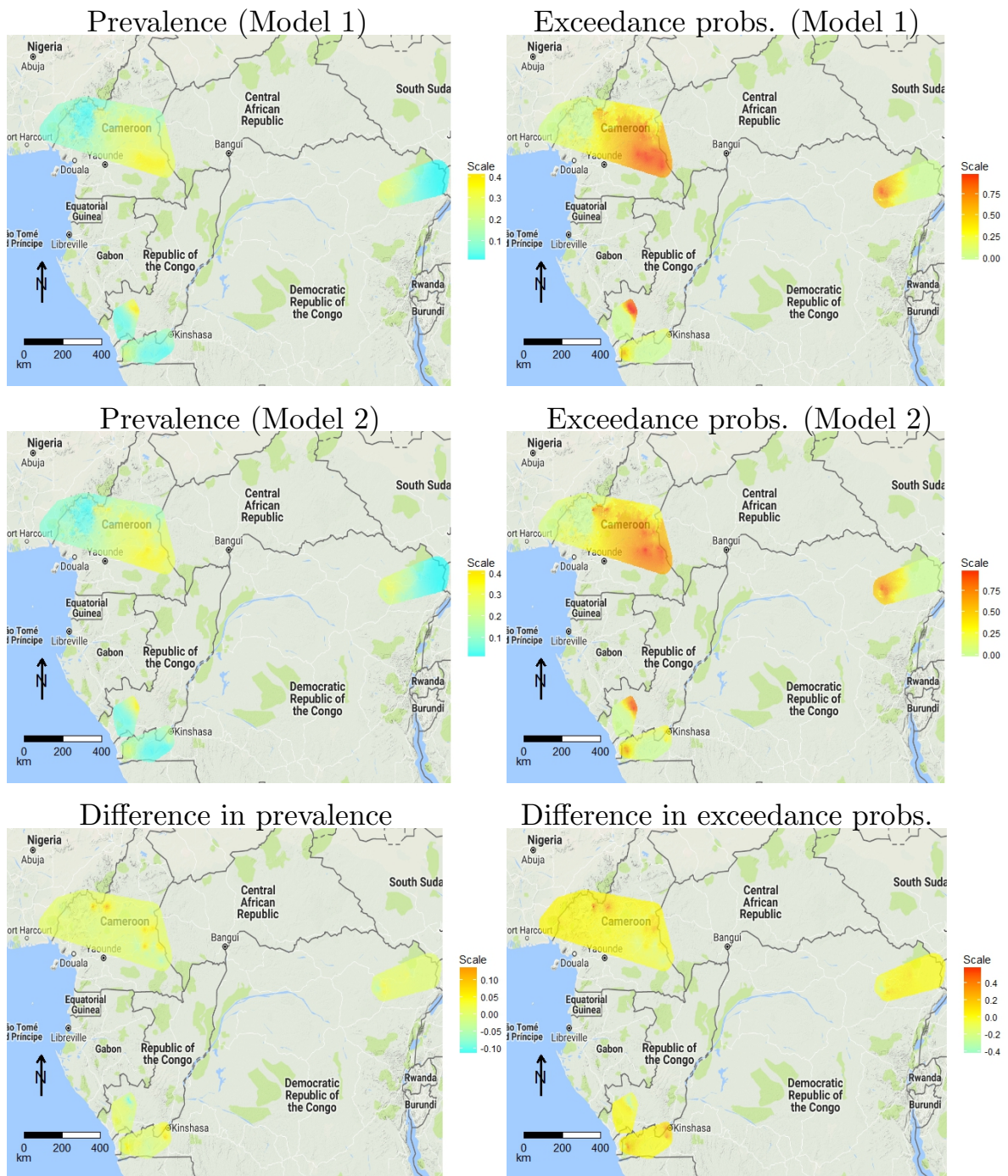
FIGURE 4.2: Predictive mean of *Loa loa* microfilariae prevalence (left panels) and probabilities of exceeding a 20% prevalence threshold (right panels), for Model 1 (top panels) and Model 2 (middle panels) of Section 4.6. The bottom panels show the differences between the predictive surfaces of Model 2 and Model 1.

Thus, the uncertainty in the MCML estimates does not have substantial influence on the exceedance probabilities.

4.7 Application II: Joint prediction of *Plasmodium falciparum* prevalence using RDT and PCR

The malaria data consist of 3,587 individuals sampled across 949 locations (see Figure 4.5). The outcomes from RDT ($k = 1$) and PCR ($k = 2$) were concordant in 92.4% of all the individuals tested for *P. falciparum.* This suggests that estimating components of residual spatial variation that are unique to each diagnostic may be difficult. For this reason our model for the data takes the following form

$$f_k(p_{jk}(x_i)) = \beta_{k,0} + \sum_{l=1}^{3} \beta_{k,l} d_{ij,l} + \nu_k T(x_i), \qquad (4.16)$$

where: $d_{ij,1}$ is a binary variable taking value 1 if the j-th individual at x_i is a male and 0 otherwise; $d_{ij,2} = \min\{a_{ij}, 5\}$ and $d_{ij,3} = \max\{a_{ij} - 5, 0\}$, i.e. the effect of age, a_{ij}, is modelled as a linear spline with a knot at 5 years.

4.7.1 Results

Table 4.3 reports point estimates and 95% confidence intervals for the model parameters. Gender has a significant effect on PCR prevalence, but its effect on RDT prevalence is not significant at the conventional 5% alpha level. The effect of age is comparable between the two diagnostics, with the probability of a positive test increasing with age up to 5 years and decreasing thereafter. The estimated variance component, $\hat{\nu}_1^2 = 0.230$, associated with RDT is about three times that for PCR, $\hat{\nu}_2^2 = 0.081$. The spatial process $T(x)$, common to both diagnostics, accounts for spatial variation in malaria prevalence up to a scale of about 11.6 kilometers, beyond which the correlation falls below 0.05. The validation procedures for the adopted stochastic spatial structures and link functions (see Section 4.5.3) do not show any strong evidence against the fitted model (see Figures 4.12, 4.13 and 4.14).

To quantify the benefit of carrying out a joint analysis for RDT and PCR, we compare the predictive inferences for prevalence that are obtained under two scenarios: (i) the fitted model in (4.16); (ii) separate fitted models that ignore the cross-correlation between the outcomes of the two diagnostic tests. Figure 4.3 shows the point predictions and standard errors for RDT and PCR prevalences for five-year-old male children under scenarios (i) (left panels) and (ii) (right panels). We observe that the point predictions for prevalence under the two models are strongly similar but the joint model in (4.16), as expected, yields smaller standard errors throughout the study area.

Having chosen (4.16) as the best model, we compare the exceedance probabilities (EPs) for a 10% threshold between RDT and PCR. Using each of the two diagnostics, we then identify malaria hotspots, as the sets of locations such that their EP is at least 90%. Figure 4.15 of Web Appendix B shows that PCR identifies a considerably larger hotspot in the north east of the study area than does RDT, and a smaller hotspot in the south west that is undetected by RDT. These results are consistent with the main findings of Mogeni et al. (2017).

4.8 Conclusions and extensions

We have developed a flexible geostatistical framework to model reported disease counts from multiple diagnostics and have distinguished two main classes of problems: (1) prediction of prevalence as defined by a gold-standard diagnostic using data obtained from a more feasible low-cost, but potentially biased, alternative; (2) joint prediction of prevalences as defined by two diagnostic tests. As the burden of disease declines in endemic regions, the use of multiple transmission metrics and diagnostics becomes necessary in order to better inform and adapt control strategies. It is thus important to develop suitable methods of inference that allow the borrowing of strength of information across multiple diagnostics. As our study has shown, the main benefit of this approach is a reduction in the uncertainty associated with the predictive inferences on disease risk.

The proposed model validation procedures can be used to identify parsimonious geostatistical models that can reliably capture the residual spatial variation of the data. Our
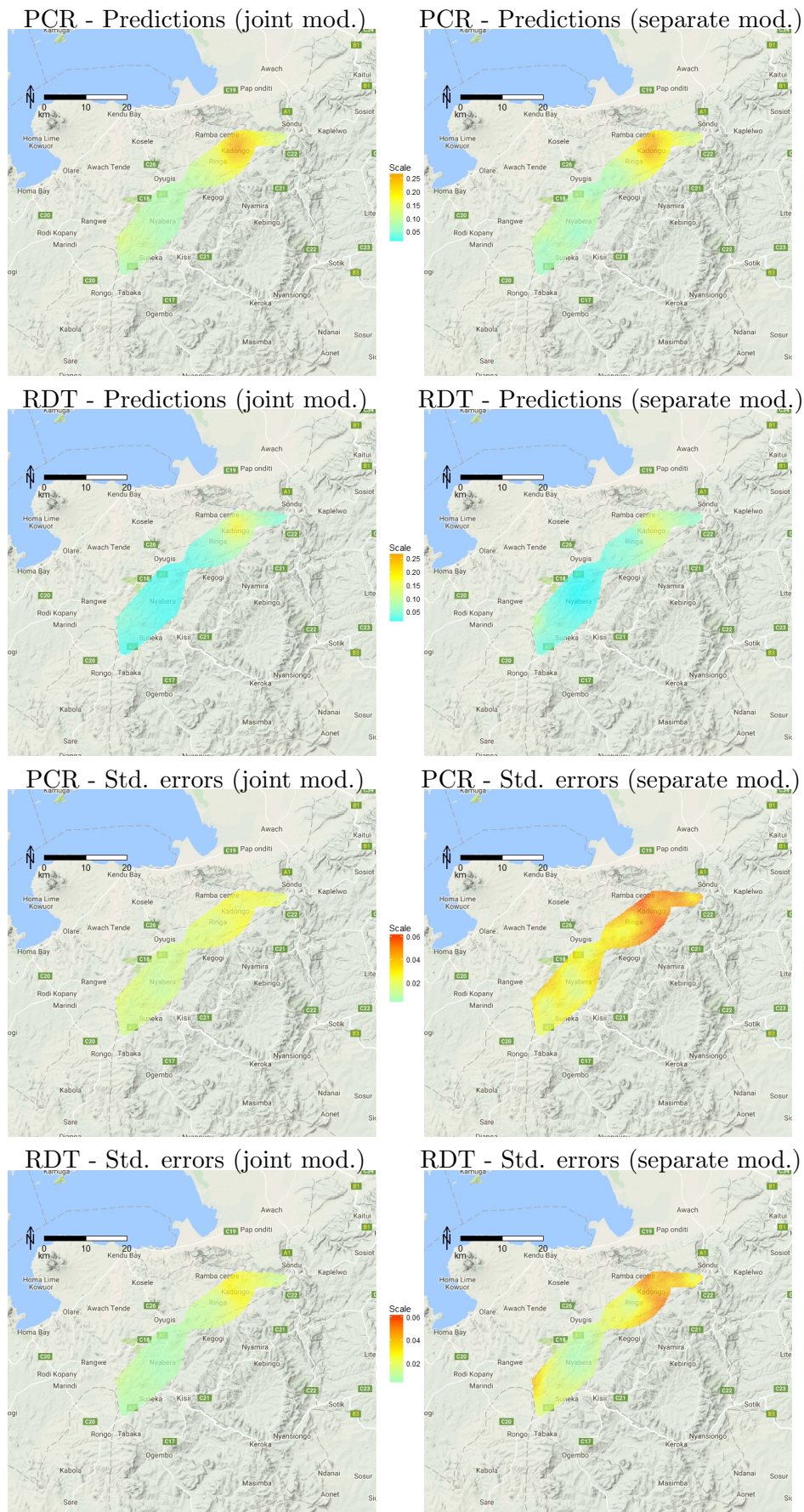
FIGURE 4.3: Point predictions (first and second rows) and standard errors (third and fourth rows) of *P. falciparum* prevalence for five-year-old children under the joint geostatistical model in (4.16) (left panels) and two separate geostatistical models (right panels) for RDT and PCR prevalence.

emphasis was on checking the validity of the specific modelling assumptions made on the spatial structure of the random effects rather than be focused exclusively on predictive performance. This is especially important when predictive probabilities are used to convey uncertainty about the exceedance, or not, of policy-relevant thresholds.

Our application to Loiasis mapping has shown the importance of acknowledging the existence of residual spatial variation specific to each diagnostic test. Through a simulation study, we have also shown that ignoring this source of extra-Binomial variation can lead to unreliably narrow prediction intervals for prevalence, with actual coverages falling well below their nominal level.

The second application on malaria mapping has highlighted the benefits of a joint analysis of data from two diagnostic tests when both are of scientific interest. A joint model can yield estimates of prevalence with smaller standard errors than estimates obtained from two separate geostatistical models.

Isotropic stationary exponential correlation functions have been used throughout the paper to model spatial dependence. An alternative and more flexible option would be to use the Matérn (1986) family of correlation functions which has expression

$$\frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{u}{\phi}\right)^{\kappa} \mathcal{K}_{\kappa}\left(\frac{u}{\phi}\right), \tag{4.17}$$

where ϕ is the scale of the spatial correlation, κ is the shape parameter that determines the differentiability of the Gaussian process and $\mathcal{K}_{\kappa}(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$. The exponential correlation function is recovered by setting $\kappa = 1/2$. However, estimating κ reliably requires a significantly larger amount of data than that available in both applications. As shown by **?**, ϕ and κ cannot be consistently estimated under in-fill asymptotics, and in practice this translates into κ often being poorly identified. In the context of counts data this issue is likely to be exacerbated and, for these reasons, in our applications, we have made the pragmatic choice of fixing κ to 0.5.

In the application of Section 4.6, we computed exceedance probabilities (EPs) by plugging-in the MCML estimates of the model parameters. To investigate how parameter uncertainty may affect the resulting EPs, we also applied a bootstrap approach, which is detailed in Section 3.2.1 of **?**. In brief, using a Gaussian approximation to the distribution of the maximum likelihood estimator (MLE), 10,000 samples were drawn for the vector of the unknown model parameters and predictive samples of prevalence were then obtained for each simulated MLE sample. We then compared the resulting EPs which also incorporate the uncertainty of the MLE with those of Section 4.6. We found no tangible difference (Figure 4.11) for the EPs from both Model 1 and Model 2. This can be explained by the fact that the uncertainty in the spatial predictions is largely dominated by the stochastic variation induced by the random effects model, as also evidenced in **?**.

Although we have only considered the case of two diagnostic tests throughout the paper, our methodology can be easily extended to more than two. However, the nature of the extension will be dependent on the specific context and scientific goal. For example, a natural extension of the models of Section 4.5.1 would be to use multiple biased diagnostic tools (for $k = 1, \ldots, K - 1$) to better predict a gold-standard ($k = K$). In this case, the cross-correlation between the outcomes of the biased diagnostic tests could be modelled using the symmetric structure of the model in Section 4.5.2, while preserving an asymmetric form for the linear predictor of the gold-standard. Formally, this is expressed as

$$
\begin{cases}
f_k\{p_{jk}(x_i)\} = \boldsymbol{d}_{ijk}^\top \boldsymbol{\beta}_k + \nu_k \left[S_k(x_i) + T(x_i) \right] + Z_{ik}, & k = 1, \ldots, K - 1 \\
f_K\{p_{jK}(x_i)\} = \boldsymbol{d}_{ijK}^\top \boldsymbol{\beta}_K + S_K(x_i) + Z_{iK} + \sum_{k=1}^{K-1} \alpha_k f_k\{p_j(x_i)\}
\end{cases}
. \quad (4.18)
$$

However, we would be wary of attempting to fit this, or other comparably complex models without an initial exploratory analysis that might help to understand the extent of the cross-correlations between the outcomes of different diagnostics, with a view to reducing the dimensionality of the model.

Acknowledgements

Supplementary Materials

Web Appendix A, referenced in Section 4.5.3; and Web Appendix B, referenced in Sections 4.3, 4.6, and 4.7 are available with this paper at the Biometrics website on Wiley Online Library.

Web-based Supplementary Materials for "A Geostatistical Framework for Combining Spatially Reference Prevalence Data from Multiple Diagnostics"

Web Appendix A: Parameter estimation and spatial prediction

As in Section 4.5, we use the same notation of the main manuscript and first re-write the models of Section 4.5.1 and 4.5.2, in the following general form

$$f_k\{p_{jk}(x_i)\} = \mu_{ijk} + W_k(x_i), \tag{4.19}$$

where μ_{ijk} is the mean component expressed as a regression on the available covariates. In the above equation, if we set $W_1(x_i) = S_1(x) + Z_{i1}$ and $W_2(x_i) = S_2(x_i) + Z_{i2} + \alpha\{f_1(x_i)\}$, then (4.10) reduces to the asymmetric model (5); if, instead, $W_k(x_i) = \nu_k\big(S_k(x_i) + T(x_i)\big) + Z_{ik}$, we recover the symmetric model (6).

We note that the components of $W_k(x_i)$, that is, $S_k(x_i)$, $T(x_i)$ and Z_{ik} are mutually independent.

Let $\boldsymbol{w}_k = \{W_k(x_i); i = 1, \ldots, N\}$ be the vector containing the random effects at location-level for the k-th diagnostic, and $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$.

In the case of model (5) of Section 4.1, we have that the covariance matrix for $\boldsymbol{w}$ is

$$\boldsymbol{\Sigma_w} = \begin{pmatrix} \sigma_1^2[\boldsymbol{R}_{S_1} + (\tau_1^2/\sigma_1^2)\boldsymbol{I}_N] & \alpha\sigma_1^2\boldsymbol{R}_{S_1} \\ \\ \alpha\sigma_1^2\boldsymbol{R}_{S_1} & \sigma_2^2[\boldsymbol{R}_{S_2} + (\tau_2^2/\sigma_2^2)\boldsymbol{I}_N] \end{pmatrix}, \tag{4.20}$$

where $\boldsymbol{I}_N$ is an N by N identity matrix and $[\boldsymbol{R}_{S_k}]_{ij} = \mathrm{Corr}\{S_k(x_i), S_k(x_j)\} = \exp\{-\|x_i - x_j\|/\phi_k\}$, $k = 1, 2$. In the model of Section 4.2, instead, we have

$$\boldsymbol{\Sigma_w} = \begin{pmatrix} \nu_1^2[\sigma_1^2\boldsymbol{R}_{S_1} + \boldsymbol{R}_T] + \tau_1^2\boldsymbol{I}_N & \sigma_1\sigma_2\boldsymbol{R}_T \\ \\ \sigma_1\sigma_2\boldsymbol{R}_T & \nu_2^2[\sigma_2^2\boldsymbol{R}_{S_2} + \boldsymbol{R}_T] + \tau_2^2\boldsymbol{I}_N \end{pmatrix}, \tag{4.21}$$

where $[\boldsymbol{R}_T]_{ij} = \mathrm{Corr}\{T(x_i), T(x_j)\} = \exp\{-\|x_i - x_j\|/\phi_T\}$.

The likelihood function for $\boldsymbol{\theta}$, the vector of the model parameters in either of the two models, is

$$L(\boldsymbol{\theta}) = \int h(\boldsymbol{y}|\boldsymbol{w})h(\boldsymbol{w})d\boldsymbol{w}, \tag{4.22}$$

where $h(\cdot)$ denotes "the distribution of $\cdot$". More specifically, we have

$$h(\boldsymbol{y}|\boldsymbol{w}) = \prod_{k=1}^{2}\prod_{i=1}^{N}\prod_{j=1}^{n_i} h(y_{ijk}|w_{ijk}) \tag{4.23}$$

with $h(y_{ijk}|w_{ijk}) = p_{jk}(x_i)^{y_{ijk}}(1-p_{jk}(x_i))^{1-y_{ijk}}$, and $h(\boldsymbol{w})$ is the density of a multivariate Gaussian distribution with mean 0 and covariance $\boldsymbol{\Sigma_w}$. Equation (4.22) can be written as

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \int \frac{h(\boldsymbol{y}|\boldsymbol{w})h(\boldsymbol{w})}{h_0(\boldsymbol{y}|\boldsymbol{w})h_0(\boldsymbol{w})}h_0(\boldsymbol{y},\boldsymbol{w})d\boldsymbol{w} \\
&\propto \int \frac{h(\boldsymbol{y}|\boldsymbol{w})h(\boldsymbol{w})}{h_0(\boldsymbol{y}|\boldsymbol{w})h_0(\boldsymbol{w})}h_0(\boldsymbol{w}|\boldsymbol{y})d\boldsymbol{w} \\
&= \mathrm{E}_{h_0(\boldsymbol{w}|\boldsymbol{y})}\left[\frac{h(\boldsymbol{y}|\boldsymbol{w})h(\boldsymbol{w})}{h_0(\boldsymbol{y}|\boldsymbol{w})h_0(\boldsymbol{w})}\right],
\end{aligned} \tag{4.24}
$$

where $\mathrm{E}_{h_0(\boldsymbol{w}|\boldsymbol{y})}[\cdot]$ is the expectation with respect to $h_0(\boldsymbol{w}|\boldsymbol{y})$. In (4.24), $h_0(\boldsymbol{w})$ and $h_0(\boldsymbol{y}|\boldsymbol{w})$ have the same distribution as $h(\boldsymbol{w})$ and $h(\boldsymbol{y}|\boldsymbol{w})$, respectively, but with parameter vector $\boldsymbol{\theta}_0$, our "best guess" of the true value for $\boldsymbol{\theta}$. We then approximate (4.24) as

$$L(\boldsymbol{\theta}) \approx L_B(\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\frac{h(\boldsymbol{w}_{(b)})h(\boldsymbol{y}|\boldsymbol{w}_{(b)})}{h_0(\boldsymbol{w}_{(b)})h_0(\boldsymbol{y}|\boldsymbol{w}_{(b)})}, \tag{4.25}$$

where $\boldsymbol{w}_{(b)}$ are B samples simulated from $h_0(\boldsymbol{w}|\boldsymbol{y})$. To simulate samples $\boldsymbol{w}_{(b)}$ we use a Metropolis-adjusted Langevin MCMC algorithm (Giorgi and Diggle, 2017), to update the standardized vector of random effects $\tilde{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{w} - \hat{\boldsymbol{w}})$, where $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{\Sigma}}$ are respectively the mode and the inverse of the negative Hessian of $f_0(\boldsymbol{w}|\boldsymbol{y})$ at $\hat{\boldsymbol{w}}$.

To improve our approximation of the likelihood function, we re-iterate this procedure by setting $\boldsymbol{\theta}_0$ to $\hat{\boldsymbol{\theta}}$ until convergence. This method of estimation converges under some regularity conditions (Christensen, 2004, **?, ?, ?**).

We now consider the problem of predicting prevalence for the k-th diagnostic, $p_k(x)$, at a set of q unobserved locations $x_1^*, \ldots, x_q^*$. Let $\boldsymbol{w}_k^* = (W_k(x_1^*), \ldots W(x_q^*))$ and $\boldsymbol{w}^* = (\boldsymbol{w}_1^*, \boldsymbol{w}_2^*)$. We first simulate B samples $\boldsymbol{w}_{(b)}$ from $h(\boldsymbol{w}|\boldsymbol{y})$ as previously described by plugging in the MCML estimates of $\boldsymbol{\theta}$. We then simulate B samples $\boldsymbol{w}_{(b)}^*$ from the conditional distribution of $\boldsymbol{w}^*|\boldsymbol{w} = \boldsymbol{w}_{(b)}$ for $b = 1, \ldots, B$. This is a multivariate Gaussian distribution with mean

$$\boldsymbol{C}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{w}_{(b)}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{w}^*} - \boldsymbol{C}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{C},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{w}^*}$ is the covariance matrix of $\boldsymbol{w}^*$ and $\boldsymbol{C}$ is the cross-covariance matrix between $\boldsymbol{w}$ and $\boldsymbol{w}^*$.

Using the resulting B samples $\boldsymbol{w}_{(b)}^*$, we then obtain predictive samples for $p_k^* = \{p_k(x_i^*); i = 1, \ldots, q\}$ as

$$p_{k,(b)}^* = f_k^{-1}\{\boldsymbol{D}^*\boldsymbol{\beta} + \boldsymbol{w}_{k,(b)}^*\}$$

where $\boldsymbol{D}^*$ is a matrices of covariates at the unobserved prediction locations.

We summarise our predictive inferences on p_k by averaging and computing the standard errors from samples $p_{k,(b)}^*$. We also compute exceedance probabilities for a threshold l as

$$\frac{1}{B} \sum_{b=1}^{B} I(p_{k,(b)}(x_i^*) > l), i = 1, \ldots, q$$

where $I(p_{k,(b)}(x_i^*) > l)$ is a indicator function taking value 1 if $p_{k,(b)}(x_i^*) > l$ and 0 otherwise.

Web Appendix B: Additional figures for the application problems
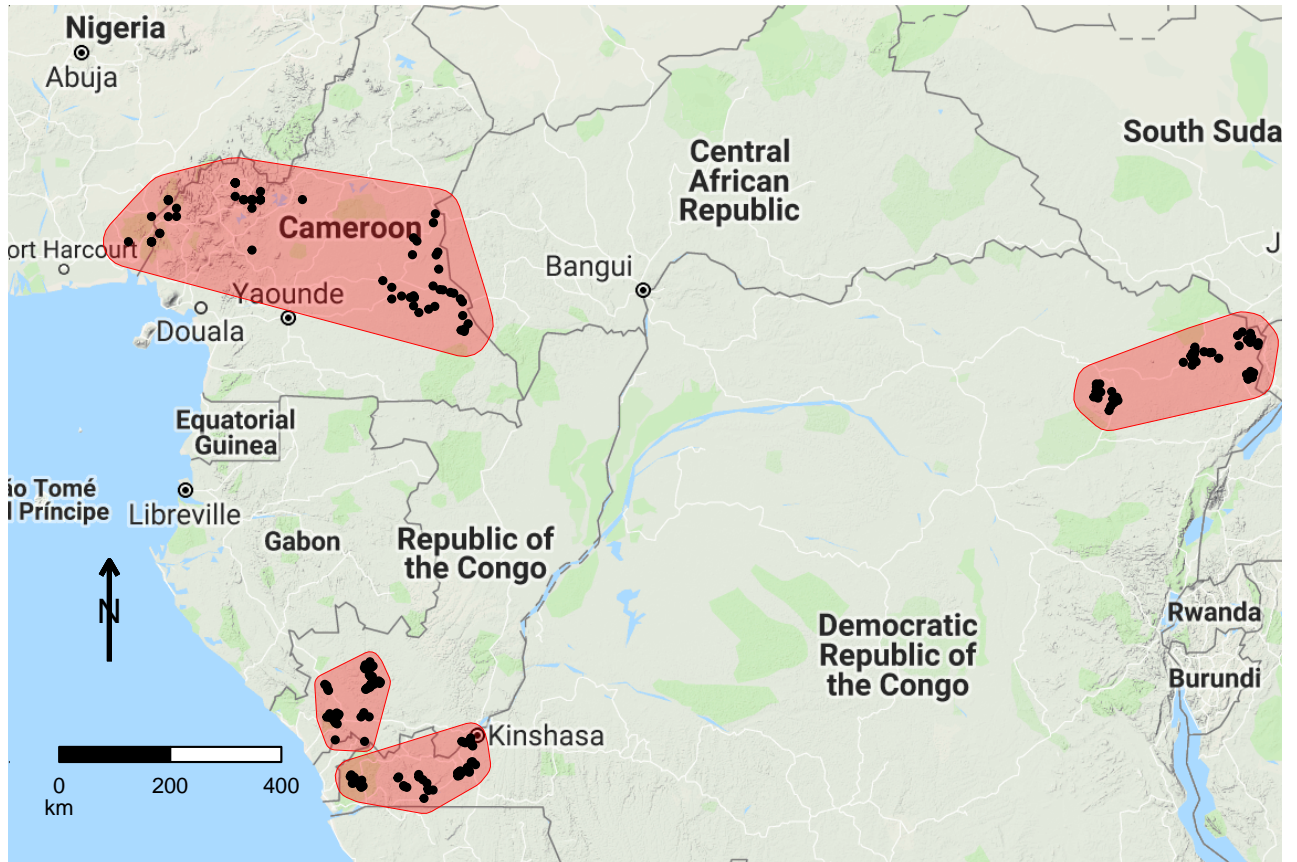
FIGURE 4.4: Map of the four sites (red polygons) of the *Loa loa* study, showing the sampled villages (black dots).
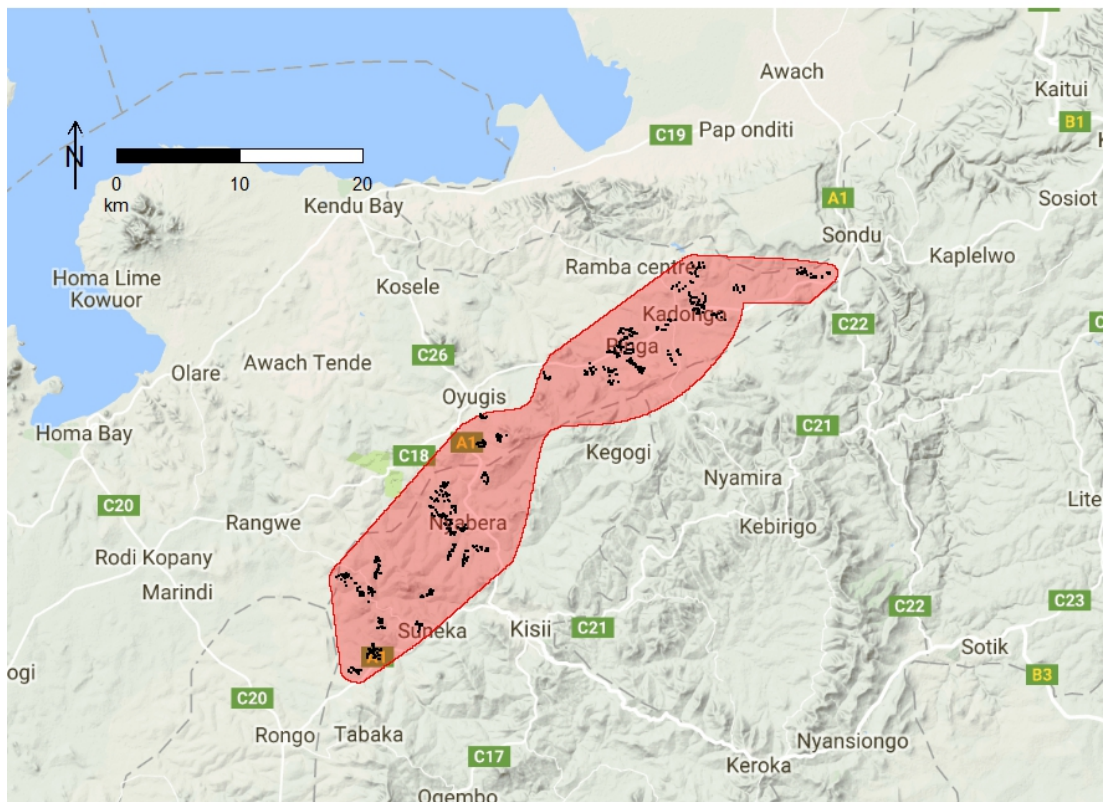
FIGURE 4.5: Map of the study site (red polygon) of the malariometric study in the highlands of Western Kenya, showing the sampled household locations (black dots).

(a)   (b)
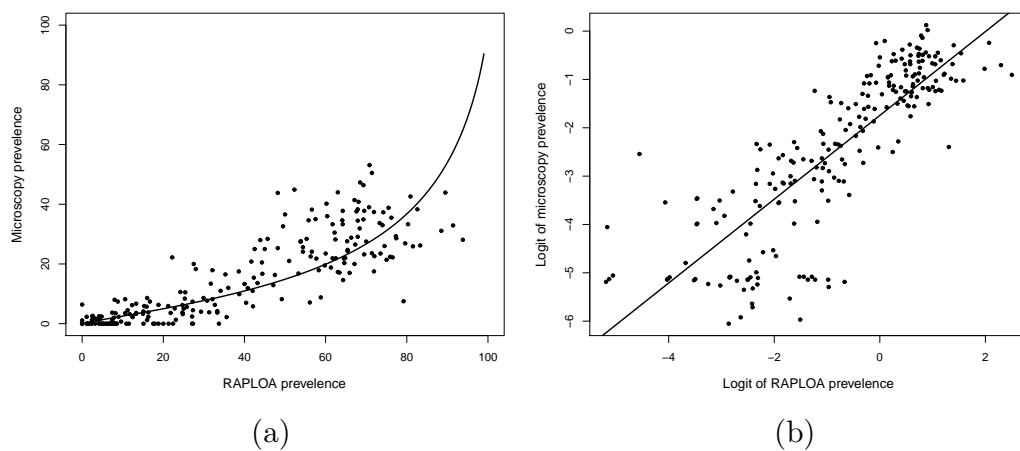
FIGURE 4.6: Scatter plots showing (a) the empirical miscroscopy prevalence against the empirical RAPLOA prevalence and (b) the same plot on the logit scale. The solid lines correspond to the ordinary least square estimate for a regression of the logit-transformed miscroscopy prevalence on the logit-transformed RAPLOA prevalence.
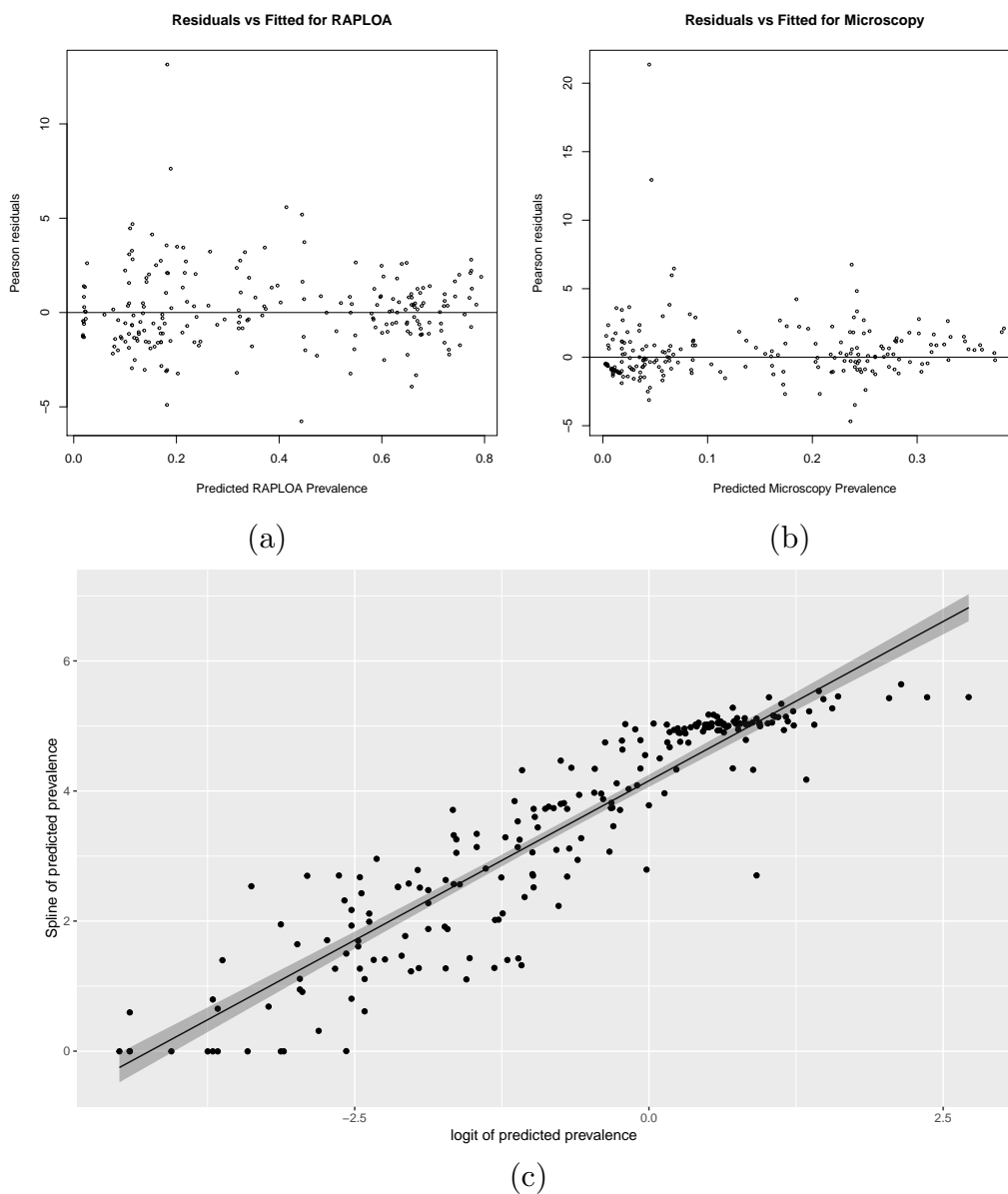
FIGURE 4.7: The top panels are scatter plots of the Prearson's residual against the fitted prevlance for RAPLOA (a) and Microscopy (b). The bottom figure (c) shows a plot of $\hat{s}\{\hat{p}_1(x_i)\}$ against $\text{logit}\{\hat{p}_1(x_i)\}$, where $s\{\cdot\}$ is a cubic spline with knots at the quantiles 0.25, 0.50 and 0.75 based on the empirical RAPLOA prevalence. The correlation coefficient between $\hat{s}\{\hat{p}_1(x_i)\}$ and $\text{logit}\{\hat{p}_1(x_i)\}$ is about 0.92.
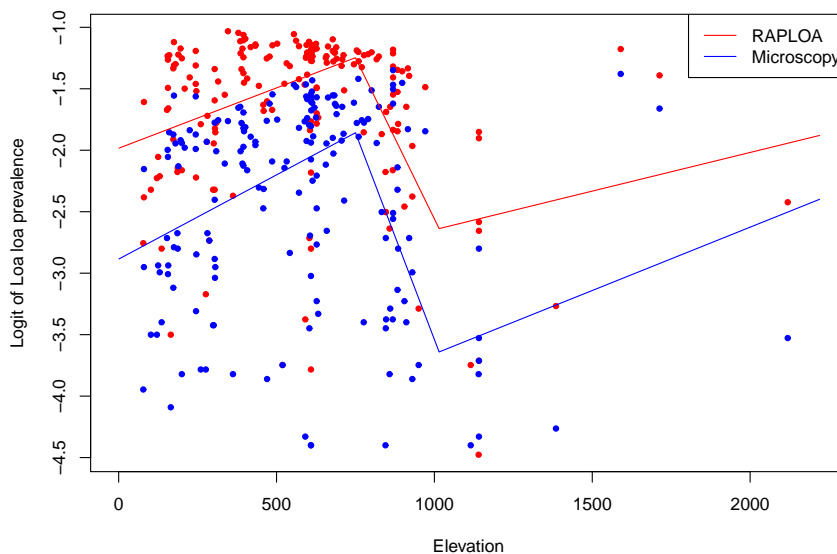
FIGURE 4.8: Scatter plots of the logit-transformed *Loa loa* prevalence against elevation in meters. The solid lines correspond to piecewise linear splines with knots at 750 and 1015 meters.
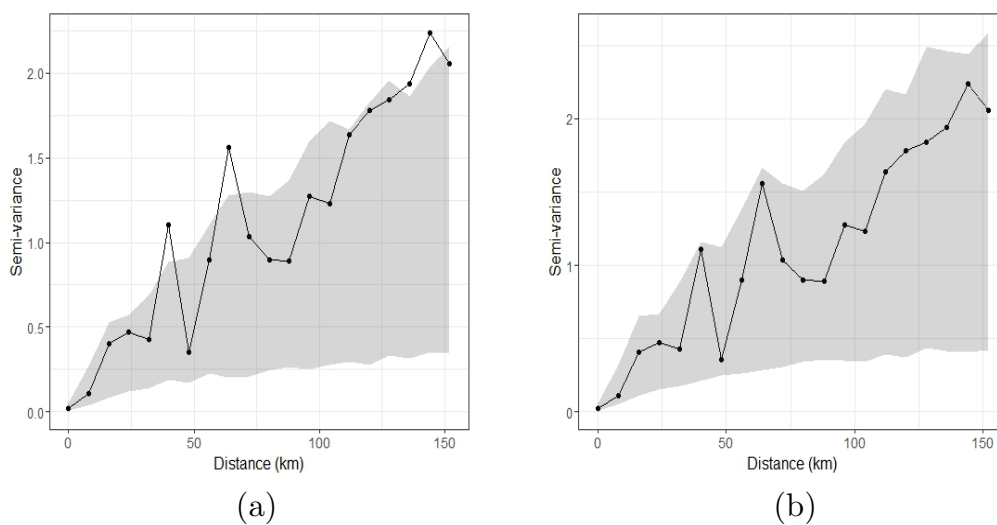
FIGURE 4.9: Empirical variogram (solid line) as defined by Equation (4.11) of the main manuscript for microscopy and RAPLOA prevalence. The shaded area is obtained using the algorithm of Section 4.5.3
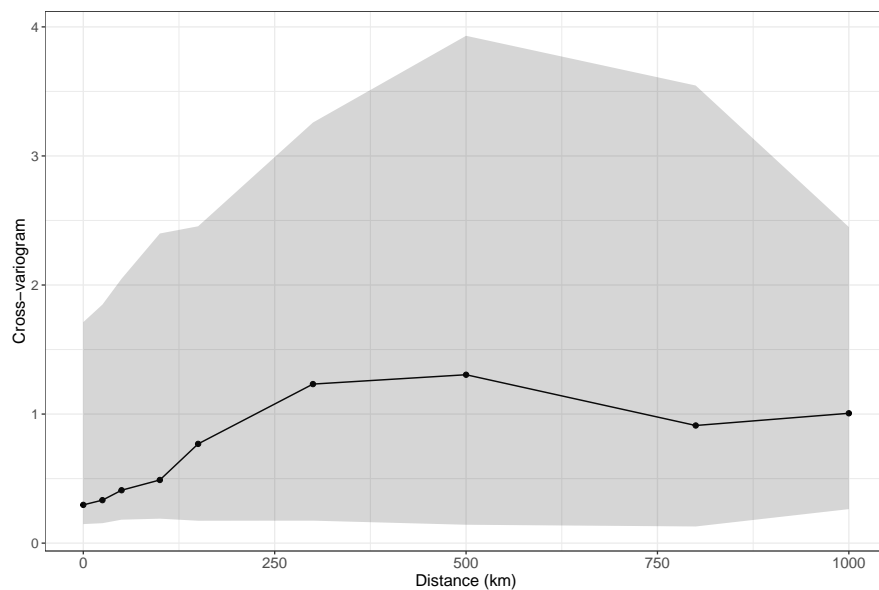
FIGURE 4.10: Empirical cross-variogram (solid line) as defined by Equation (4.12) of the main manuscript between microscopy and RAPLOA prevalence. The shaded area is obtained using the algorithm of Section 4.5.3
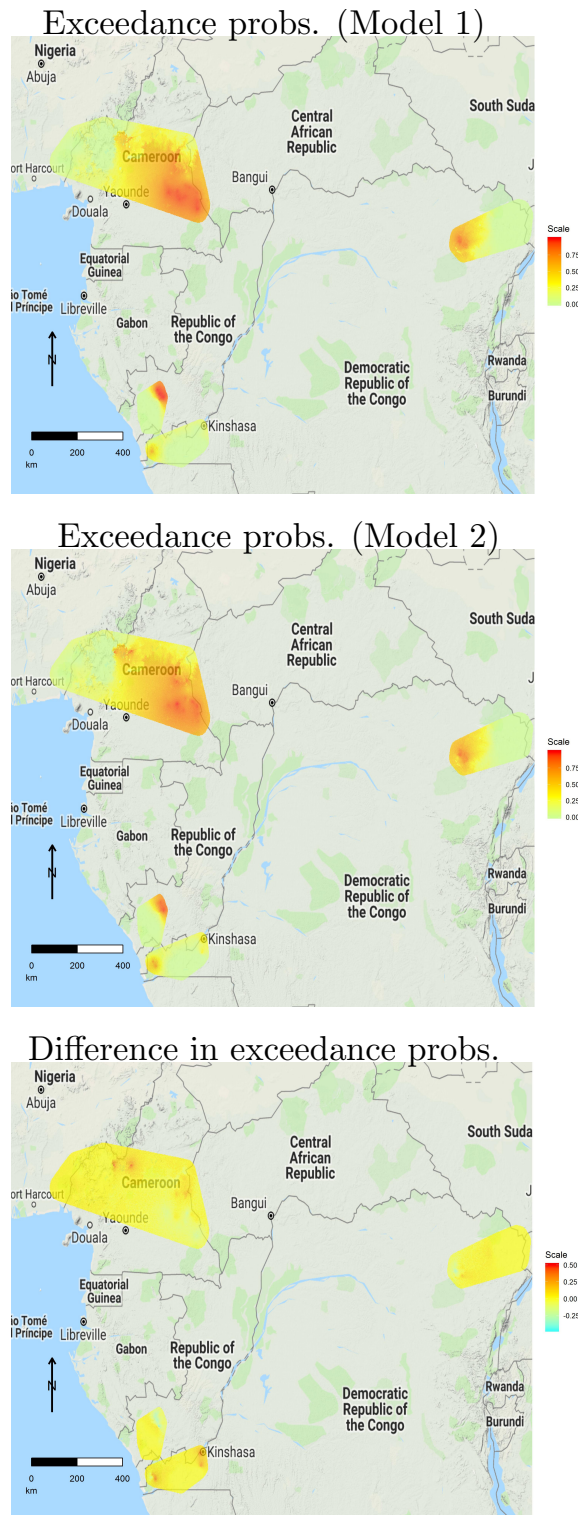
FIGURE 4.11: Predictive probabilities for the exceedance of a 20% prevalence threshold, for Model 1 (top panel) and Model 2 (middle panel) of Section 4.6. The bottom panel show the difference between the surfaces from Model 2 and Model 1. In each map, parameter uncertainty has been incorporated by averaging over the distribution of the maximum likelihood estimator, which we approximated using a multivariate Guassian distribution.
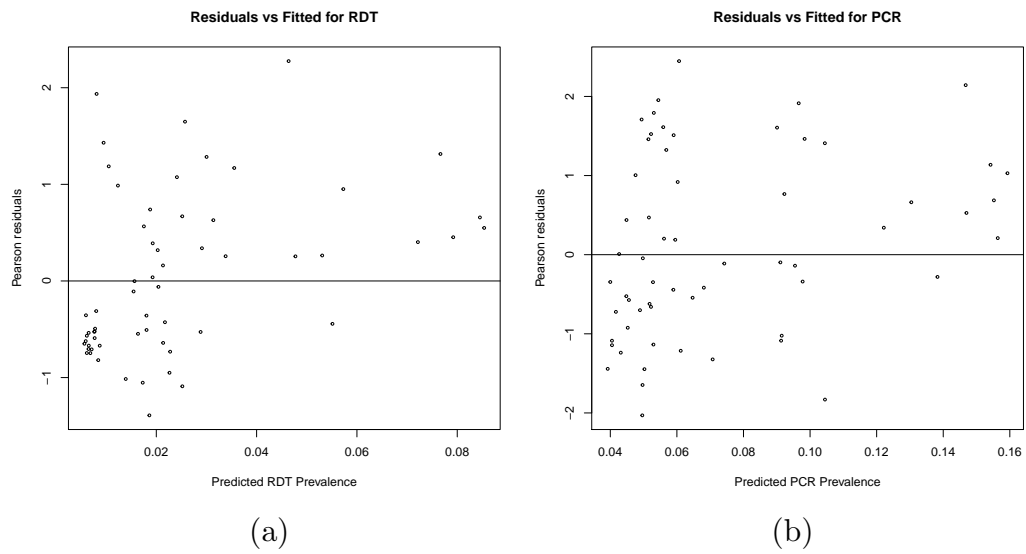
FIGURE 4.12: Scatter plots the Pearson's residuals against the fitted prevalence for RDT (a) and PCR (b).
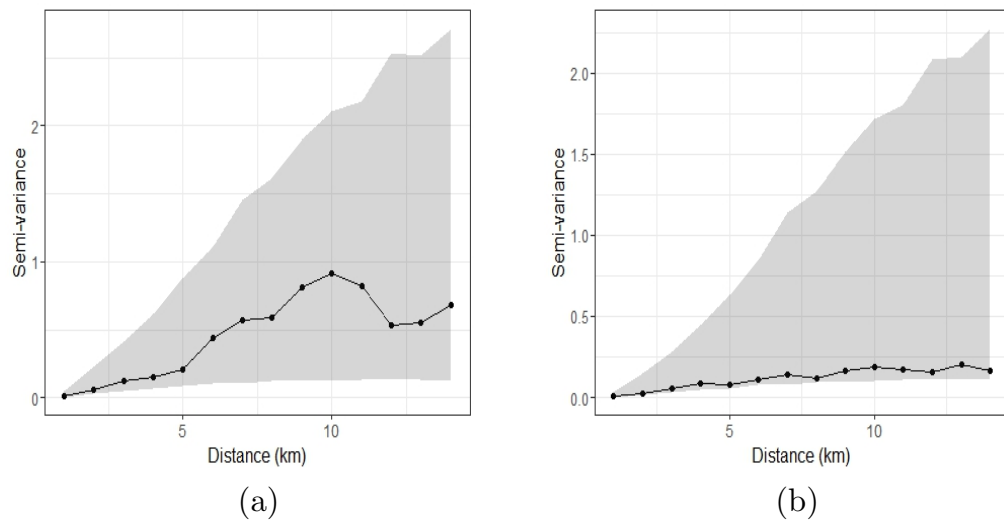
FIGURE 4.13: Empirical variogram (solid line) as defined by equation (11) of the main manuscript for RDT and PCR prevalence. The shaded area is obtained using the algorithm of Section 4.5.3
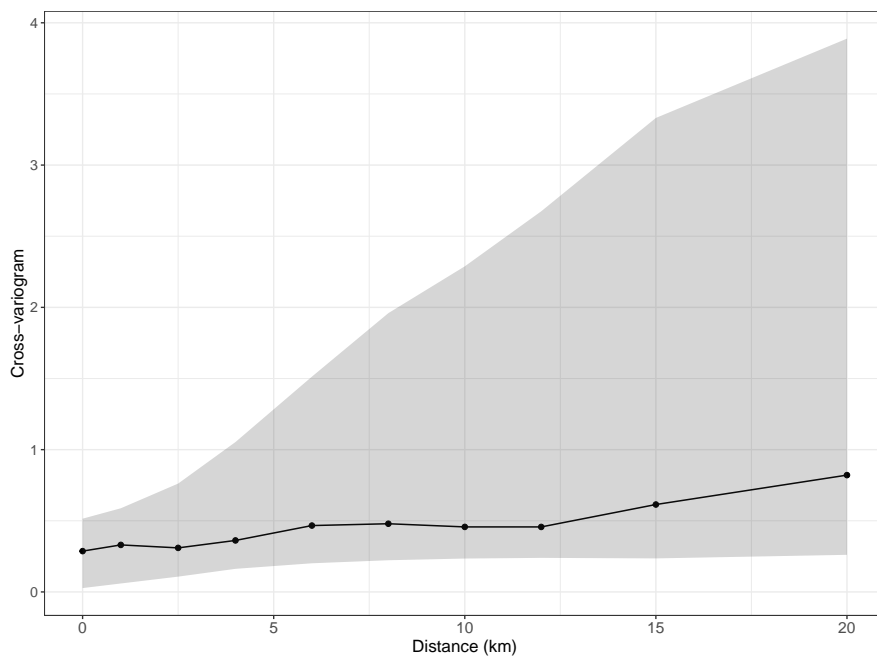
FIGURE 4.14: Empirical cross-variogram (solid line) as defined by equation (12) of the main manuscript between RDT and PCR prevalence. The shaded area is obtained using the algorithm of Section 4.5.3
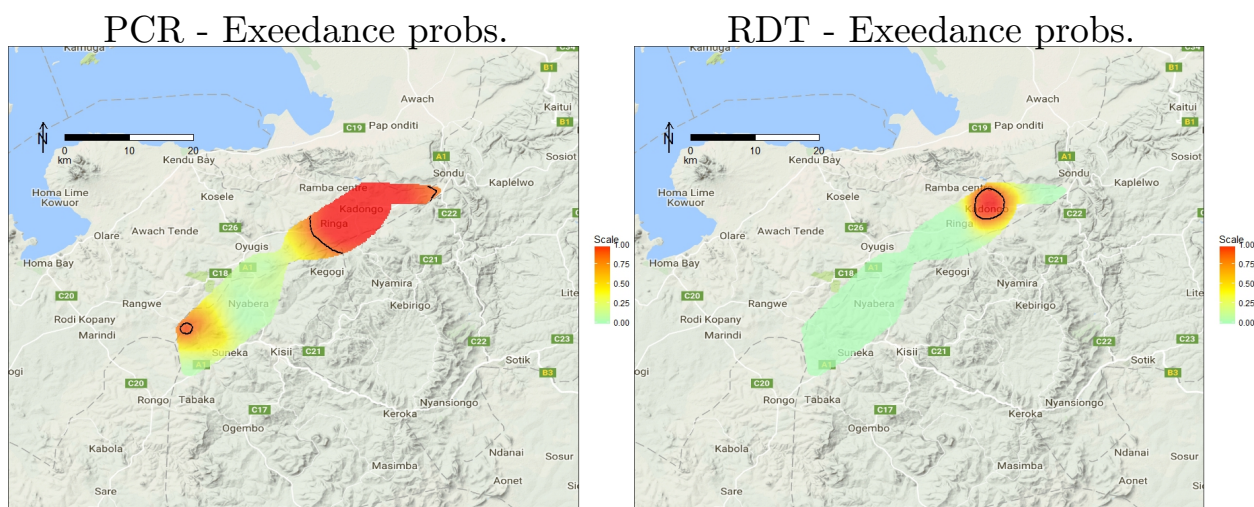
FIGURE 4.15: Exceedance probabilities for a threshold of 10% prevalence based on PCR (left panel) and RDT (right panel). The areas encompassed by solid contours include locations with an exceedance probability of no less than 90%.

TABLE 4.1: Monte Carlo maximum likelihood estimates and associated 95% confidence intervals for the fitted Model 1 and Model 2 to the *Loa loa* data; see Section 4.6 for more details.

Parameter	Model 1	Model 2
$\beta_{1,0}$	-0.791	-0.763
	(-1.984, 0.402)	(-1.963, 0.437)
$\beta_{1,1} \times 10^3$	0.515	0.588
	(-0.977, 2.008)	(-0.922, 2.098)
$\beta_{1,2} \times 10^3$	-3.529	-3.412
	(-7.314, 0.255)	(-7.155 , 0.331)
$\beta_{1,3} \times 10^3$	-0.110	-0.059
	(-1.531, 1.312)	(-1.501 , 1.382)
$\beta_{2,0}$	-1.762	-1.736
	(-2.075, -1.449)	(-2.244, -1.229)
$\beta_{2,1} \times 10^3$	0.208	0.126
	(-0.386, 0.802)	(-0.799, 1.050)
$\beta_{2,2} \times 10^3$	-0.223	-0.039
	(-2.023, 1.576)	(-2.944 , 2.865)
$\beta_{2,3} \times 10^3$	-0.591	-0.612
	(-1.666, 0.485)	(-2.429, 1.205)
σ_1^2	1.581	1.617
	(0.669, 3.738)	(0.679, 3.851)
σ_2^2	—	0.216
		(0.111, 0.419)
ϕ_1	182.037	187.388
	(64.657, 512.512)	(65.171, 538.807)
ϕ_2	—	23.686
		(6.150, 91.220)
τ_1^2	0.205	0.324
	(0.081, 0.521)	(0.052, 6.229)
τ_2^2	0.324	0.104
	(0.055, 5.873)	(0.018, 5.797)
α	1.005	1.017
	(0.902, 1.107)	(0.939, 1.095)

TABLE 4.2: Results of the simulation study including the 95% coverage probability (CP), the root-mean-square-error (RMSE), the 95% predictive interval length (PIL) averaged over the 20 unobserved locations. For more details, see the main text in Section 4.6.2.

	CP	RMSE	PIL
Model 1	0.811	0.134	0.325
Model 2	0.948	0.123	0.376

TABLE 4.3: Monte Carlo maximum likelihood estimates and associated 95% confidence intervals for the model in (4.16) fitted to the malaria data.

Parameter	RDT ($k = 1$)	PCR ($k = 2$)
$\beta_{k,0}$	-6.186	-4.373
	(-7.234, -5.138)	(-17.008, 8.261)
$\beta_{k,1}$	-0.003	0.251
	(-0.415, 0.395)	(0.009, 0.494)
$\beta_{k,2}$	0.261	0.220
	(0.070, 0.453)	(0.095, 0.344)
$\beta_{k,3}$	-0.059	-0.020
	(-0.081, -0.037)	(-0.028, -0.012)
ν_k^2	0.230	0.081
	(0.145, 0.364)	(0.052, 0.126)
ϕ_T	11.581	
	(10.618, 12.63)	

References

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(7572):207–211.

Boussinesq, M., Gardon, J., Gardon-Wendel, N., Kamgno, J., Ngoumou, P., and Chippaux, J.-P. (1998). "Three probable cases of Loa loa encephalopathy following ivermectin treatment for onchocerciasis." *The American journal of tropical medicine and hygiene*, 58(4):461–469.

Boussinesq, M., Gardon, J., Kamgno, J., Pion, S., Gardon-Wendel, N., and Chippaux, J.-P. (2001). "Relationships between the prevalence and intensity of Loa loa infection in the Central province of Cameroon." *Annals of Tropical Medicine & Parasitology*, 95(5):495–507.

Christensen, O. F. (2004). "Monte Carlo maximum likelihood in model-based geostatistics." *Journal of computational and graphical statistics*, 13(3):702–718.

Coffeng, L. E., Pion, S. D., O'Hanlon, S., Cousens, S., Abiose, A. O., Fischer, P. U., Remme, J. H., Dadzie, K. Y., Murdoch, M. E., De Vlas, S. J., et al. (2013). "Onchocerciasis: the pre-control association between prevalence of palpable nodules and skin microfilariae." *PLoS neglected tropical diseases*, 7(4):e2168.

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

Dalal, S., Holmes, M. D., Laurence, C., Bajunirwe, F., Guwatudde, D., Njelekela, M., Adebamowo, C., Nankya-Mutyoba, J., Chiwanga, F. S., Volmink, J., et al. (2015). "Feasibility of a large cohort study in sub-Saharan Africa assessed through a four-country study." *Global health action*, 8(1):27422.

Diggle, P. J. and Giorgi, E. (2016). "Model-based geostatistics for prevalence mapping in low-resource settings." *Journal of the American Statistical Association*, 111(515):1096–1120.

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

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

Geyer, C. J. (1991). "Markov chain Monte Carlo maximum likelihood."

Giorgi, E. and Diggle, P. J. (2017). "PrevMap: An R Package for Prevalence Mapping." *Journal of Statistical Software*, 78(8):1–29.

Haas, A. D., Keiser, O., Balestre, E., Brown, S., Bissagnene, E., Chimbetete, C., Dabis, F., Davies, M.-A., Hoffmann, C. J., Oyaro, P., et al. (2015). "Monitoring and switching of first-line antiretroviral therapy in adult treatment cohorts in sub-Saharan Africa: collaborative analysis." *The lancet HIV*, 2(7):e271–e278.

Irvine, M. A., Njenga, S. M., Gunawardena, S., Njeri Wamae, C., Cano, J., Brooker, S. J., and Deirdre Hollingsworth, T. (2016). "Understanding the relationship between prevalence of microfilariae and antigenaemia using a model of lymphatic filariasis infection." *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 110(2):118–124.

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., Rudan, I., Campbell, H., Cibulskis, R., Li, M., et al. (2012). "Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000." *The Lancet*, 379(9832):2151–2161.

Mappin, B., Cameron, E., Dalrymple, U., Weiss, D. J., Bisanzio, D., Bhatt, S., and Gething, P. W. (2015). "Standardizing Plasmodium falciparum infection prevalence measured via microscopy versus rapid diagnostic test." *Malaria journal*, 14(1):460.

Matérn, B. (1986). Spatial variation, volume 36 of lecture notes in statistics.

Mogeni, P., Williams, T. N., Omedo, I., Kimani, D., Ngoi, J. M., Mwacharo, J., Morter, R., Nyundo, C., Wambua, J., Nyangweso, G., et al. (2017). "Detecting Malaria Hotspots: A Comparison of Rapid Diagnostic Test, Microscopy, and Polymerase Chain Reaction." *The Journal of infectious diseases*, 216(9):1091–1098.

Mudie, K., Jin, M. M., Tan, L. K., Addo, J., dos Santos-Silva, I., Quint, J., Smeeth, L., Cook, S., Nitsch, D., Natamba, B., et al. (2019). "Non-communicable diseases in sub-Saharan Africa: a scoping review of large cohort studies." *Journal of global health*, 9(2).

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(9947):1005–1070.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). "Joint regression analysis of correlated data using Gaussian copulas." *Biometrics*, 65(1):60–68.

Stevenson, J. C., Stresman, G. H., Baidjoe, A., Okoth, A., Oriango, R., Owaga, C., Marube, E., Bousema, T., Cox, J., and Drakeley, C. (2015). "Use of different transmission metrics to describe malaria epidemiology in the highlands of western Kenya." *Malaria journal*, 14(1):418.

Takougang, I., Meremikwu, M., Wandji, S., Yenshu, E. V., Aripko, B., Lamlenn, S. B., Eka, B. L., Enyong, P., Meli, J., Kale, O., et al. (2002). "Rapid assessment method for prevalence and intensity of Loa loa infection." *Bulletin of the World Health Organization*, 80(11):852–858.

Tangpukdee, N., Duangdee, C., Wilairatana, P., and Krudsood, S. (2009). "Malaria diagnosis: a brief review." *The Korean journal of parasitology*, 47(2):93.

Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., and Greenwood, B. (1999). "Predicting malaria infection in Gambian children

from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results." *The American journal of tropical medicine and hygiene*, 61(1):2–8.

Wanji, S., Akotshi, D. O., Mutro, M. N., Tepage, F., Ukety, T. O., Diggle, P. J., and Remme, J. H. (2012). "Validation of the rapid assessment procedure for loiasis (RAPLOA) in the democratic republic of Congo." *Parasites & vectors*, 5(1):25.

World Health Organization (2017). "World malaria report 2017."

Wu, L., van den Hoogen, L. L., Slater, H., Walker, P. G., Ghani, A. C., Drakeley, C. J., and Okell, L. C. (2015). "Comparison of diagnostics for the detection of asymptomatic Plasmodium falciparum infections to inform control and elimination strategies." *Nature*, 528(7580):S86.

Zouré, H. G., Noma, M., Tekle, A. H., Amazigo, U. V., Diggle, P. J., Giorgi, E., and Remme, J. H. (2014). "The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control:(2) pre-control endemicity levels and estimated number infected." *Parasites & vectors*, 7(1):326.

Zouré, H. G. M., Wanji, S., Noma, M., Amazigo, U. V., Diggle, P. J., Tekle, A. H., and Remme, J. H. F. (2011). "The Geographic Distribution of Loa loa in Africa: Results of Large-Scale Implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA)." *PLOS Neglected Tropical Diseases*, 5(6):1–11.

Chapter 5

Conclusion and Future Research

5.1 Achievement of the objectives of the thesis

This thesis aimed to achieve three sets of objectives, each of which was investigated in a chapter of the thesis.

The first set of objectives, investigated in Paper 2 (Chapter 2), was oriented towards understanding and comparing the spatial distributions of *Plasmodium falciparum* parasite prevalence (*Pf*PR) and its entomological inoculation rates (*Pf*EIR) in the study region in Malawi, Africa. To achieve this objective, we posed the following three question: First, what are the spatio-temporal distributions of the *Pf*PR and *Pf*EIR in the study region? Second, how do the spatio-temporal patterns of *Pf*PR and *Pf*EIR compare? Third, What functional relationship exists between *Pf*PR and *Pf*EIR, and what policy relevant features of the malaria epidemiology in the study region does this relationship highlight?

To answer the first question, we developed geostatistical models for the *P. falciparum* human biting *Pf*HBR and its sporozoite rates *Pf*SR for the two species of mosquitoes known to transmit the parasite and also for *Pf*PR in the study region. Using these models, we mapped malaria transmission intensity as measured by (*Pf*EIR), and also *Pf*PR. We then developed a number of mathematical models to study the relationship

between *Pf*EIR and *Pf*PR. We concluded that the relationship between the two quantities was best described as follows: the logit of *Pf*EIR was linearly related to *Pf*PR, which translates an initial rapid rise in *Pf*PR with increasing *Pf*EIR, followed by a levelling off.

To answer the second question, we further mapped hot-spot of *Pf*EIR and P*ff*PR by mapping the predictive probabilities that these quantities exceeded predefined thresholds. We then compared the patterns of hotspots of these quantities in space and time. We found that the spatio-temporal patterns of *Pf*PR and *Pf*EIR we similar and their hot-spots mainly overlapped, but there were hot-spots that were unique to each of these two malaria risk measures.

The second set of objectives, tackled in Paper 2 (Chapter 3), was to investigate the association between malaria and child growth in Africa. We posed three research questions to explore this objective. First, what is the association between malaria and linear growth in children 0-5 years in sub-Saharan Africa? Second, what factors modulate the association between malaria and growth in children 0-5 years? Third, what is the spatial distribution of stunted growth in the study population?

We used Demographic and Health Survey (DHS) data from different countries and years to develop a linear geostatistical model for height-for-age z-scores (HAZs) as a function of malaria. We found a mixture of negative, positive and no associations, as has been reported in the literature before our study. We then investigated several factors that might modulate the association between malaria and child-growth and found that arable land (an indication of agricultural activities) significantly explained the differences in the association between malaria and child growth in different countries and years. In particular, we found malaria to be less detrimental to child growth when the quantity of arable land was high. Finally, using the predictive probabilities the HAZ is less than -2, we mapped the risk of stunted growth for each of the 20 datasets studied in this paper.

The third set of objectives aimed to develop geostatistical methods that allow for the joint analysis of disease prevalence data that have been obtained using multiple diagnostics. This objective was rephrased as two question. The first question was about how we can predict disease prevalence as defined by a gold-standard diagnostic in geographical

regions where data is available only for a low-cost, possibly biased alternative test? The second question was how we can make more precisely predictions of disease prevalence as defined by each of several complementary diagnostics by borrowing information across the others.

In Chapter 4, we developed a geostatistical framework to combine disease prevalence data obtained through multiple diagnostics. We developed an (asymmetric) model that explores the calibrations relationship between two diagnostics in order to predict prevalence for a gold-standard diagnostic using low-cost and potentially biased alternative tests. We then developed a (symmetric) model that explores the cross-correlation between two diagnostics in order to carry out joint prediction of prevalence as defined by each of multiple diagnostics. We then applied the methodology to two challenging global public health problems: the symmetric modelling approach to a malaria mapping problem in the highlands of Western Kenya, and the asymmetric modelling approach to a *Loa loa* mapping problem in Western and Eastern Africa.

5.2 Originality and contribution to knowledge

The originality and contributions of this thesis can be classified into three areas namely, the questions which we posed, the methods we used to investigate the question, and the findings we present.

The originality of Chapter 2 stems from the questions we pose, which to the best of our knowledge, until this work, had not been investigated. The possible reason why the questions had not been investigated is that the nature of data needed to answer such questions were non-existent. Our data therefore allowed us to investigate original questions such as how do spatio-temporal patterns of PfEIR compare to PfPR, do PfEIR and PfPR lead to the identification of the same hot-spots, Does PfEIR have a lagged effect on PfPR, Does the PfEIR-PfPR relationship vary across demographic groups that experience a different exposure to malaria, e.g. children between 6 and 60 months and women of reproductive age.

We showed in our results that PfEIR and PfPR show similar spatio-temporal patterns with a lagged effect of a month of PfEIR on PfPR. We further showed that hot spots of malaria risk as defined by PfEIR, and hot-spots of malaria risk as defined by PfPR coincide during the peak transmission season but not during the through of transmission. We found a logit-linear model may be the best model that describes the PfEIR-PfPR relationship. We showed that PfPR increased even when PfEIR was not detectable. Moreover, we showed that the PfEIR-PfPR relationship tends to differ between women and children. These findings had not yet been reported in the literature.

Paper 2 (Chapter 3), to the best of our knowledge, is the first paper to investigate the association between malaria and child-growth in a geostatistical framework. It also presents the first meta-analysis investigating differences in the effect of malaria on child-growth in different geographic regions and times. In this paper, we also proposed novel geostatistical models that can be used, in a spatial-longitudinal framework, to investigate the effect of malaria of growth.

In Paper 2, we also found that the association between malaria and growth could be modulated by the quantity of arable land in the country of survey in the year of survey, with malaria having a less detrimental/negative effect on growth when the quantity of arable land is high. This paper is the first study to present this finding. Furthermore, we showed that severe growth faltering during the first two years of life means less catching up of growth in later life, on average. Our study is the first paper to present this finding too, to the best of our knowledge.

The motivation of Paper 3 (Chapter 4) was the need for geostatistical methods that would allow for the joint analysis of disease prevalence data obtained from multiple diagnostics. Geostatistical data from multiple diagnostics are becoming increasingly available but methods to deal with such data were lacking. We therefore developed two novel models, one that explores and uses a calibration relationship between data from a gold standard diagnostic and a alternative one to predict prevalence of the gold standard and another that can be used to more precisely map disease prevalence as defined by each of several complementary diagnostics by borrowing information across the others. When we applied the novel models to real data, we found that combining data from multiple

diagnostics in a geostatistical analysis can lead to more precise spatial predictions than separate analyses. We also found that prevalence data from different diagnostics can have different spatial correlation structures, in which case a joint geostatistcal analysis should account for this difference in order to make reliable predictive inferences. These findings had not been reported in the literature.

5.3 The common thread of the thesis

The overall argument of this thesis is that our novel geostatistical methods can enable efficient identification of areas of high disease risk and the factors associated with the risk of diseases. One of the toughest global health challenges is undoubtedly malaria. The disease has persisted despite resilient and consistent efforts to control it. In each of the three papers that make up the core of the thesis, we developed and applied statistical methods relevant to the geospatial analysis of studies aimed at improving our understanding of the social and environmental factors affecting the spatial and/or temporal variation in malaria. Each model we developed built-up on the Model based geostatistical framework of Diggle and Ribeiro (2007). Paper 1 dwelt entirely on malaria whereas in Papers 2 and 3, we studied two other global health problems namely, child growth and *Loa loa* respectively.

5.4 Limitations of the thesis

A major improvement made by Paper 1, as compared to two previous studies (**??**) of the association between *Pf*EIR and *Pf*PR is that we used data that had been collected in the same geographic region over a period of 38 months. Previous studies had used data from different geographical locations in different parts of the African continent. However, malaria transmission in the study region during the study was generally low, which generally makes entomological indices noisy. Another limitation of Paper 1 is that we did not adjust for the sensitivity and specificity of the malaria diagnostic RDT used. This is because such data were not available. However, we show how we could have accounted for them if such data were available.

The limitation of the Paper 2 is that the malaria data used were estimates of malaria incidence from a spatio-temporal model rather than individual malaria episodes. Moreover, the uncertainty in the estimates were not available. In paper 2, we showed ways we could have accounted for the uncertainty in the malaria estimates if they were available as predictive samples of malaria incidence or summary statistics from the predictive distribution of malaria incidence. We then proposed novel models that could be used if individual malaria episodes were available, and if the data were available as spatial longitudinal data.

We developed a applied a geostatistical framework for combining data from multiple diagnostics in Paper 3. Each of our two applications problems involved data from two diagnostics. We further proposed a model for more than two diagnostics. However, data for the application of this model was not available to demonstrate the model in a real life application. Another limitation of Paper 3 is that, due to unavailability of sensitivity and specificity, we could not adjust for the sensitivity and specificity of the malaria diagnostics RDT and PCD, and of the *Loa loa* microscopy and RAPLOA used.

5.5 Recommendations and future research

In Paper 1, we found that when malaria transmission has been interrupted, there could be a residual host of individuals who may still carry the parasite. Based on this finding, we recommend that (1) monitoring malaria indices needs to continue for a while when data suggests transmission has been interrupted, (2) vector-host contact interruption should be coupled with active screening and treatment of infected individuals to be able to eliminate the disease.

In Paper 1, due to the noise in the mosquitoes data, our analysis considered PfPR and PfPR separately and modelled the relationship between the two using predictive samples from the model of each outcome. For less noisy data, we can jointly model PfPR and PfPR by embedding the logit-linear relationship in a generalized linear geostatistical

model framework. A appropriate model for this might be

$$
\begin{cases}
\log(HBR(x_i,t_i)) = d(x_i,t_i)^\top \beta_1 + f_1(t_i;\alpha_1) + S_1(x_i) + Z_{i1} = L_1(x_i,t_i) \\[2mm]
\log\left(\frac{PfSR(x_i,t_i)}{1-PfSR(x_i,t_i)}\right) = d(x_i,t_i)^\top \beta_2 + f_2(t_i;\alpha_2) + S_2(x_i) + Z_{i2} = L_2(x_i,t_i) \\[2mm]
\log\left(\frac{PfPR(x_i,t_i))}{1-PfPR(x_i,t_i)}\right) = d(x_i,t_i)^\top \beta_2 + f_3(t_i;\alpha_3) + S_3(x_i) + Z_{i3} \\[2mm]
\qquad\qquad\qquad + \delta_1 L_1(x_i,t_i) + \delta_2 \log\{1 + \exp[-L_2(x_i,t_i)]\},
\end{cases}
\tag{5.1}
$$

where all notations have the same meanings as in the original model, and δ_1 and δ_2 are parameters to be estimated.

The model given by (5.1) can have two possible advantages: (1) It takes into account the joint distribution of PfPR, PfSR and HBR, and might better explain the textitPfEIR–PfPR relationship. (2) It allows to borrow information from the entomological data in order to make better predictions of parasitaemia prevalence. However, a potential challenge in fitting this model is that it would require a lot of data to achieve efficient identification of model parameters.

An important finding in Paper 2 was that if growth faltering in the first two years of life results in severe stunting, then recovery of growth afterwards would difficult. Based on this finding, undernutrition programs should focus on the first two years of life as the best opportunity to interrupt growth faltering. We also found that malaria was less detrimental to child-growth when the quantity of arable land was high. Arable land indicates land capable of being used for crop production, which indicates that, overall, better feeding of children is key to reducing malaria's effect on child-growth. Better feeding practices that may improve nutritional outcomes are higher meal frequencies, and more diverse diet (Nti and Lartey, 2007). Consequently, a country would need to prioritise agriculture to make any meaningful impart on undernutrition nation-wide.

The effects of malaria on other undernutrition indicators such as underweight and wasting were not investigated in our study so as to allow a focused discussion on stunting. However, these would be interesting to investigate. The models developed in this paper can be used to study the association between malaria and other nutritional indicators such as weight-for-age z-scores (WAZs) and weight-for-height z-score (WHZs), which

when less than -2, respectively indicate underweight and wasting. Wasting and underweight can then also be mapped by the predictive probability that they are less than -2. In a future work to investigate these indicators however, we recommend a spatial longitudinal study as this will help to not only overcome the limitation of not being able to account for the uncertainty in malaria estimates, but it will also help to estimate the effect of malaria on the nutritional outcome during each year of life, and also establish if any effects/associations found are causal. We have proposed a novel model in Paper 2 to carry out the analysis of such spatial-longitudinal data.

Paper 3 has illustrated the gains of combining disease prevalence data from multiple diagnostics, viz, the extra gain in precision in prevalence estimates for each diagnostic and the ability to use a less accurate diagnostics to predict prevalence as defined by a more accurate one. We therefore recommend that in a situation where a survey needs to cover a large geographical region but is prevented from doing so by constraint imposed by the cost of the gold standard diagnostic of the disease in question, a less costly diagnostic may be used in combination with the more preferred, but costly diagnostic. The best approach to allocate the different diagnostics to the locations of interest to yield the most accurate and precise estimates of prevalence however needs investigation. Should the diagnostics be randomly allocated? Should the gold-standard be used at specific locations of interest and then the alternative diagnostics at other location? These questions need further investigation.

Future research following Paper 3 will include extending this framework to spatio-temporal settings and exploring the connection between different diagnostics using Bayesian shared models.

In the nutshell, this thesis has extended geostistical methods for disease mapping and has expand on our understanding of malaria, child-growth and *Loa loa*.

References

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

Nti, C. A. and Lartey, A. (2007). "Effect of caregiver feeding behaviours on child nutritional status in rural Ghana." *International Journal of Consumer Studies*, 31(3):303–309.

Appendix A

Appendix A: Published version of Paper 2

International Journal of
Health Geographics

RESEARCH   **Open Access**

CrossMark

Geostatistical modelling of the association between malaria and child growth in Africa

Benjamin Amoah[1], Emanuele Giorgi[1*] , Daniel J. Heyes[2], Stef van Burren[3,4] and Peter John Diggle[1]

Abstract

Background: Undernutrition among children under 5 years of age continues to be a public health challenge in many low- and middle-income countries and can lead to growth stunting. Infectious diseases may also affect child growth, however their actual impact on the latter can be difficult to quantify. In this paper, we analyse data from 20 Demographic and Health Surveys (DHS) conducted in 13 African countries to investigate the relationship between malaria and stunting. Our objective is to make inference on the association between malaria incidence during the first year of life and height-for-age Z-scores (HAZs).

Methods: We develop a geostatistical model for HAZs as a function of both measured and unmeasured child-specific and spatial risk factors. We visualize stunting risk in each of the 20 analysed surveys by mapping the predictive probability that HAZ is below − 2. Finally, we carry out a meta-analysis by modelling the estimated effects of malaria incidence on HAZ from each DHS as a linear regression on national development indicators from the World Bank.

Results: A non-spatial univariate linear regression of HAZ on malaria incidence showed a negative association in 18 out of 20 surveys. However, after adjusting for spatial risk factors and controlling for confounding effects, we found a weaker association between HAZ and malaria, with a mix of positive and negative estimates, of which 3 out of 20 are significantly different from zero at the conventional 5% level. The meta-analysis showed that this variation in the estimated effect of malaria incidence on HAZ is significantly associated with the amount of arable land.

Conclusion: Confounding effects on the association between malaria and stunting vary both by country and over time. Geostatistical analysis provides a useful framework that allows to account for unmeasured spatial confounders. Establishing whether the association between malaria and stunting is causal would require longitudinal follow-up data on individual children.

Keywords: Child growth, Exceedance probability, Geostatistics, Malaria, Stunting

Background

Undernutrition underlies 45% of all child deaths among children under 5 years [1]. A very low height-for-age, usually referred to as stunting, is an important indicator that reflects the cumulative effects of undernutrition and disease infections [2]. Stunted children are more prone to illness and premature death. Stunting among children is known to be associated with poor cognitive development

[3, 4]. Long-term consequences of stunting include lower adult economic productivity, higher risks of ill-health and, among women with short stature, an increased risk of death during delivery [5–8]. Globally, the rate of stunting in children under 5 years reduced from 32.7% (198 million) in year 2000 to 23.2% (156 million) in year 2015 [9]. In Africa however, the rates reduced from 38% in 2000 to 32% in 2015, representing more limited progress than in Asia, Latin America and the Caribbean where stunting rates dropped by more than one third over the same period [9]. In many low- and middle-income countries (LMICs), over 50% of 12–23 months old children

*Correspondence: e.giorgi@lancaster.ac.uk
[1] CHICAS Research Group, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK
Full list of author information is available at the end of the article

are stunted [10–12]. In 2014, less than half of all children under 5 years lived in LMICs, yet these countries accounted for two-thirds of all stunted children globally [13]. Although the main risk factor for stunting is inadequate nutrition, exposure to infectious diseases may also lead to an increase in stunting risk [14, 15]. However, there are indirect effects of malaria not fully understood [16, 17], and it is unclear if part of the stunting burden can be attributed to malaria.

Malaria is still a public health threat, although the ongoing global fight against it has resulted in 50% decrease in the infection prevalence and 40% decrease in the clinical incidence in the endemic region of Africa between 2000 and 2015 [18]. In 2015, there were an estimated 214 million malaria cases and 438 thousand deaths from malaria worldwide, of which 88% occurred in sub-Saharan Africa and 70% in children under the age of 5 years, with 10% of all deaths in children under the age of 5 years due to malaria [19]. In 2017, similar global estimates were reported: 216 million malaria cases and 445 thousand malaria deaths, of which 91% occurred in sub-Saharan Africa, with most of the deaths still occuring in children under 5 years [20]. The association between malaria and stunting is unclear and still a matter of debate, with studies showing contrasting results. For example, maternal malaria has been found to impact on child growth [21], with infants born to women who experienced malaria during pregnancy having an increased risk of impaired height and weight gain [22–25]. The risk of stunting has been found to increase for every malaria episode [26]. On the other hand, some studies suggest that stunting may modulate susceptibility to malaria, especially during the first 2 years of life [27, 28]. Whilst some studies suggest that stunted children may be at higher risk of developing malaria episodes [29], others report that stunting may have a protective effect against malaria [30, 31]. In other studies, instead, no association is found [32, 33]. More recently, Fink et al. [34] found a significant effect of malaria exposure on cognitive development and socio-emotional development, but not on height, for which they report an estimated effect of about 3.000 and associated 95% confidence interval (− 11.350, 4.606).

The height-for-age Z-score (HAZ) measures the deviation from heights based on the World Health Organization (WHO) growth standards [35, 36] and are comparable across ages and gender. Values of HAZ below − 2 are used as an indicator of stunted growth. In this paper, we analyse data from 20 Demographic and Health Surveys (DHS) conducted in Senegal, Mozambique, Ghana, Burkina Faso, Zambia, Malawi, Rwanda, Cote d'Ivoire, Burundi, Liberia, Namibia, Togo and Tanzania to pursue the following objectives: (1) to investigate the association between malaria and HAZ by developing a geostatistical framework that accounts for both measured and unmeasured risk factors for stunting; (2) to understand how such association varies across the African countries considered in this study; (3) to map the risk of stunting. We also discuss the limitations of this study and provide a detailed description on how the proposed modelling framework could be further extended to a longitudinal setting. To the best of our knowledge, this is the first study that investigates the association between the geographical distribution of malaria and HAZ using a model-based geostatistical approach.

Methods

Data

DHS are nationally representative household surveys that are generally repeated every 5 years and provide information on a range of health and population indicators, including anthropometric information. The DHS methodology is usually based on a stratified two-stage cluster design. At the first stage, enumeration areas are drawn from census files. At the second stage, for each enumeration area selected, samples of households are drawn from an updated list of households to form groups of households known as sampling clusters. The GPS location of the center of each sampling cluster is taken as the cluster location. Each child is allocated to a spatially-referenced sampling cluster. We analyse data from 20 DHS conducted between 2003 and 2014 [37]. Table 1 shows the number of clusters and individuals for each survey. The average number of children per cluster varies from one survey to another, with the highest value of about 21.7 in Burkia Faso in 2003 and the lowest of about 5.7 in Malawi in 2010.

The variables used in the analysis are the following.

Child-specific variables Data on a child's height, age and gender, family's wealth index and mother's education level were obtained from the DHS for all sampled children aged less than 5 years. Families' wealth indices are constructed using principal component analysis on household's ownership of television, radio, watch, vehicles and agricultural land, type and number of animals owned, bank account, materials used for housing construction, type of water access and sanitation facilities [38].

Urban extent indicator We use information on urban extents, available as raster data at a spatial resolution of 1 km by 1 km, from the Global Rural-Urban Mapping Project [39]. This variable is a binary indicator that classifies each spatial grid cell as urban or rural, based on a combination of population counts, settlement points, and presence of night-time lights.

Table 1 Sample size summaries for the analysed DHS data indicating the country, year of survey, number of children, number of sampled clusters, and average number of children per cluster

Country	Year	No. of children	No. of clusters	Average no. of children per cluster
Senegal	2005	2710	355	7.6
Senegal	2011	3694	384	9.6
Mozambique	2011	9595	609	15.8
Ghana	2003	3010	393	7.7
Ghana	2008	2350	393	6.0
Ghana	2014	2671	410	6.5
Burkina Faso	2003	8581	396	21.7
Burkina Faso	2010	6290	540	11.6
Zambia	2007	5243	317	16.5
Zambia	2014	4635	303	15.3
Malawi	2004	6238	386	16.2
Malawi	2010	4623	811	5.7
Rwanda	2005	3692	455	8.1
Cote d'Ivoire	2007	3305	288	11.5
Burundi	2010	3449	376	9.2
Liberia	2007	4197	270	15.5
Liberia	2013	3206	319	10.1
Namibia	2007	3669	484	7.6
Togo	2014	3209	328	9.8
Tanzania	2010	6581	453	14.5

Estimated malaria incidence rates We use raster data on estimated *Plasmodium falciparum* incidence as obtained from a Bayesian spatio-temporal model implemented by the Malaria Atlas Project [18]. The data are available at a temporal resolution of 1 year, from 2000 to 2015, and a spatial resolution of 0.05° × 0.05°. More specifically, the estimated *Plasmodium falciparum* malaria incidence at pixel-level is the predicted average clinical incidence rate per child per year in the age cohorts 0–5 years. A clinical malaria episode is an attributable febrile episode with a body temperature in excess of 37.5 °C. Multiple bouts of symptoms occurring within a 30-day period are counted as a single episode.

Model formulation and spatial prediction

Accounting for spatial effects is crucial in order to deliver valid inferences on the regression coefficients [40]. Model-based geostatistics allows us to incorporate both explained and unexplained (residual) spatial variation in HAZ and to predict the risk of stunting throughout a geographical area of interest.

Let Y_{ij} denote the HAZ for the *j th* sampled child at the cluster location x_i. We distinguish between two sources of variation in HAZ: between-cluster variation, induced by spatially varying risk factors; and within-cluster variation due to child-specific characteristics. Each of these components depends on both measured and unmeasured risk factors. In order to account for the latter, we define a hierarchical linear model as follows. Let $S(x_i)$ denote a stationary Gaussian process and U_i represent mutually independent zero-mean Gaussian variables with common variance τ^2. We assume that, conditionally on $S(x_i)$ and U_i, the Y_{ij} are Gaussian variables with means $\mu_j(x_i)$ and variance ω^2, where

$$
\begin{aligned}
\mu_j(x_i) = e_{ij}^\top \gamma + d(x_i)\beta + \delta \mathcal{M}_{ij} \\
+ f(\mathcal{A}_{ij}) + U_i + S(x_i), \text{ for } i = 1, \ldots, n \\
j = 1, \ldots, m_i.
\end{aligned} \tag{1}
$$

In (1), n is the number of cluster locations and m_i is the number of individuals at cluster location x_i. In (1) we also distinguish between three types of explanatory variables: e_{ij}, a vector of child-specific explanatory variables, including sex, family's wealth index and mother's education level; $d(x_i)$, a spatial indicator variable which takes values 1, if location x_i is classified as urban and 0 if rural; $\mathcal{M}_{ij}$, the estimated malaria incidence at location x_i during the first year of life of the *j*-th child. The parameters γ, β and δ are the regression parameters associated with each of the three types of explanatory variables, whilst $f(\mathcal{A})$ is a cubic spline function of age, $\mathcal{A}$, with knots at 12 and 24 months.

Our objective is to make inference on the parameter δ, which quantifies the effect of malaria incidence in the

first year of life on HAZ. Our assumption is that malaria has a lagged effect on height and, therefore, we use the incidence of malaria during the first year of life to determine the strength of this association. In the remainder of the paper, we shall refer to the parameter δ and the variable $\mathcal{M}_{ij}$ in (1) as the effect of malaria on HAZ and malaria incidence, respectively.

In (1), the unstructured random effect U_i conflates two sources of residual variation: spatial variation on a scale smaller than the minimum observed distance between clusters; and unexplained unstructured variation at cluster level.

The spatially structured residuals $S(x)$ are modelled as a zero-mean stationary and isotropic Gaussian process with variance σ^2 and exponential correlation function given by

$$\rho(u; \phi) = \exp(-u/\phi), \tag{2}$$

where u is the Euclidean distance between any two locations. The scale parameter ϕ regulates the rate at which the spatial correlation decays with increasing distance u.

We map the risk of stunting for male children, 24 months old, using the predictive probability that HAZ is below -2 over a $0.05° \times 0.05°$ grid. We integrate out the effect of maternal education and wealth index using the following Monte Carlo approach. We generate 10,000 samples from the joint distribution of these two variables and, conditionally on these, we then simulate values of HAZ. The stunting risk is then computed by taking the proportion of simulated HAZ samples that are below -2.

More details on the computational implementation and on the mapping of stunting risk are given in Additional file 1.

Model validation

To check the validity of the adopted spatial correlation structure for the data, we carry out the following Monte Carlo procedure. We simulate 1000 empirical variograms under the fitted model and then use these to compute 95% confidence intervals at any given spatial distance of the variogram. If the empirical variogram obtained from the data falls within the 95% tolerance bandwidth, we conclude that the adopted spatial correlation function is compatible with the data. If, instead, that falls outside the 95% tolerance bandwidth, then the data show evidence against the fitted model. More details are provided in Additional file 1.

Understanding the variation in the effect of malaria on HAZ

We carry out a meta-analysis in order to understand the variation in the estimates of the parameter of interest δ, from all the 20 DHS. Let $\hat{\delta}_k$ and s_k denote the maximum likelihood estimate of δ and its standard error,

respectively, for $k = 1, \ldots, 20$. We then model $\hat{\delta}_k$ using a weighted least squares fit to the regression model

$$\hat{\delta}_k = \alpha_0 + \alpha_1 v_k + Z_k, \tag{3}$$

where v_k is a World Bank African development indicator [41] associated with the country and year of the k-th survey, and the Z_k are independent Gaussian variables with mean zero and variance s_k^2. We select eleven development indicators belonging to the categories of "Agriculture and rural development", "Climate change", "Economy and growth", "Education" and "Environment". A full list of the indicators is given in Additional file 2.

Results

Non-spatial analysis

Figure 1 shows box-plots of HAZ by categories of family's wealth indices and mother's education level for all surveys combined. We assign integer scores 1–5 to the five levels of family wealth from very poor to very wealthy; and scores 1–6 to the six levels of mothers education, from no education to higher education. As expected, the box-plots show that the median HAZ tends to increase with increasing levels of wealth and education.

We then investigate the marginal association between malaria incidence and HAZ. Figure 2 shows the observed HAZ against malaria incidence, where the solid line is obtained from the least squares fit of a univariate linear model. The dashed horizontal lines indicate HAZ levels of 2, 0 and -2. The dashed vertical lines separate $\mathcal{M}$ into terciles. We see that Malaria incidence takes a maximum value of about 1.5 for all surveys, except Namibia in 2007, where this is about 0.7. We also note that for the surveys
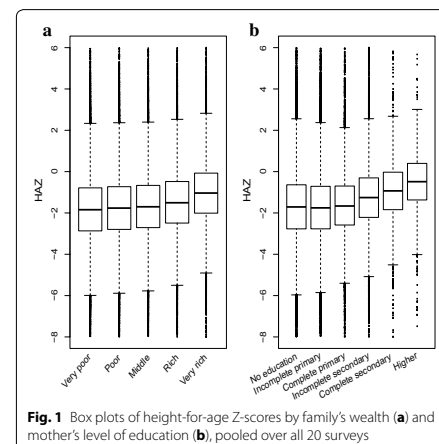
Fig. 1 Box plots of height-for-age Z-scores by family's wealth (**a**) and mother's level of education (**b**), pooled over all 20 surveys
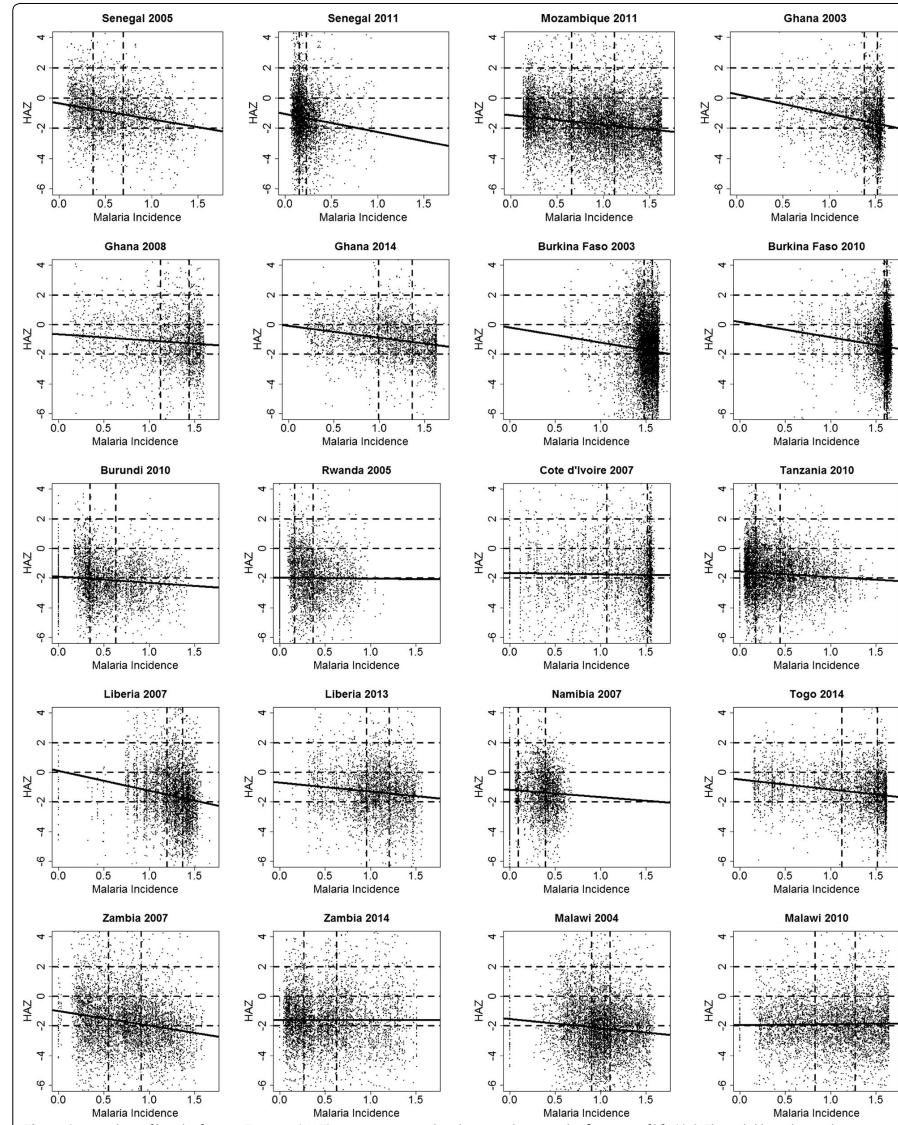
Fig. 2 Scatterplots of height-for-age Z-scores (HAZ) against expected malaria incidence in the first year of life ($\mathcal{M}$). The solid line shows the univariate linear model with malaria incidence as the predictor of HAZ. The dashed horizontal lines show HAZ levels of 2, 0 and − 2, whilst the dashed horizontal lines separates $\mathcal{M}$ into terciles

in Senegal in 2005, Mozambique in 2011, Ghana in 2003–2008–2014 and Zambia in 2007, the variation in $\mathcal{M}$ is evenly distributed, whereas it is more skewed for Senegal in 2011, Burkina Faso in 2003–2010, Malawi in 2004 and Namibia in 2007. Except for Rwanda in 2005, Zambia in 2014 and Malawi in 2010, in all the remaining 17 surveys we observe that HAZ decreases with increasing values of $\mathcal{M}$. Figure 3 shows the least squares estimates and the corresponding 95% confidence intervals. The estimated regression coefficients are negative in 18 surveys, of which 16 are significantly different from zero at 5% level.

Figure 4 shows HAZ curves as functions of age, within each of the terciles groups of $\mathcal{M}$, as indicated in Fig. 2. The fitted curves reflect the typical age-related pattern of HAZ in LMICs: after a decrease in HAZ during the first 2 years of life, child-growth slowly recovers but never reaches zero. This phenomenon, known as "growth faltering", has been widely observed; see, for example, [11, 12, 42, 43]. We also observe that in Burkina Faso in 2003, Ghana in 2008, Malawi in 2004–2010 and Rwanda in 2005, HAZ curves by terciles groups of $\mathcal{M}$ are partly overlapping, whereas in the remaining 15 surveys, children in the first tercile of $\mathcal{M}$ have the highest levels of HAZ and children in the third tercile with the lowest levels of HAZ, irrespective of age. We also notice that in Burkina Faso in 2003, Burundi in 2010, Rwanda in 2005, Cote d'Ivoire in 2007 and Malawi in 2004, where median HAZ curves fall below the -2 threshold at about 24 months of age, the curves still remain below the -2 threshold in later years.

Geostatistical analysis

Figure 5 shows estimates, with associated 95% confidence intervals, of the malaria parameter δ from the fitted geostatistical model in (1). The point estimate of δ is negative in 7 surveys with Ghana in 2014 and Liberia in 2007 being significant at the 5% level. Positive values

are estimated for the remaining 13 surveys, with only Namibia in 2007 being significant. We note that, after accounting for residual spatial variation and measured potential confounders, the magnitude of the association between malaria incidence and HAZs is smaller than for the marginal association shown in Fig. 3.

Point estimates of the covariance parameters of (1) with associated standard errors are reported in Additional file 3. We see that, for each survey, the variance corresponding to the child-specific variation is consistently larger than both the variance of the spatial process and the nugget variance.

The results from the model validation (Additional file 4) show that the fitted geostatistical models are compatible with the data for each of the 20 surveys analysed. We also point out that, although the variograms based on the residuals from the standard linear regression are relatively flat, we still find evidence of non-negligible residual spatial variation in HAZ as indicated by the interval estimates of the parameter of the scale of the spatial correlation in Additional file 3.

Mapping of stunting risk

In Fig. 6, we report the predictive maps of stunting risk for Ghana, Burkina Faso and Mozambique for boys, aged 24 months. In Ghana in 2003–2008–2014, the maps show a remarkable decrease in stunting over time, that is observed almost everywhere within the country. Similarly, in Burkina Faso, we observe a decrease in stunting risk from 2003 to 2010. Mozambique in 2011 shows high spatial heterogeneity in stunting risk, with values ranging from 0.1 to 0.9. Risk maps for the remaining surveys are shown in Additional file 5. In these maps, we observe overall higher levels of stunting risk in Burundi in 2010 and Malawi in 2004, and lower levels in Senegal in 2008 and Togo in 2014.

Variation in the effect of malaria on HAZ

The amount of arable land (defined as percentage of land under temporary crops, meadows for mowing or for pasture, market or kitchen gardens, and land temporarily fallow) in the country and year of survey is the only World Bank indicator to be significant at 5% level, with a p-value of about 0.013, explaining 26% of the total variation in the estimated effects of malaria incidence on HAZ. More specifically, we estimate that an increase of 1% in arable land leads to a 0.008 increase in the value of the estimated malaria effect, on average. See Additional file 2 for more detailed results from the meta analysis.

Discussion

The objective of our study was to model and quantify the association between malaria and HAZs in children aged less than 5 years. Using DHS data from 20 surveys

Fig. 3 Plot of estimates of the malaria effect on HAZ with associated 95% confidence intervals obtained from a univariate linear model for each survey

Fig. 4 Estimated trajectories of height-for-age Z-scores (HAZ) as a function of age, stratified by malaria incidence ($\mathcal{M}$). Each panel shows three curves. Each curve is a piecewise cubic spline with knots at 12 and 24 months and corresponds to a tercile group of $\mathcal{M}$. The solid, dotted and dashed curves respectively correspond to the first, second and third terciles of $\mathcal{M}$, as indicated in Fig. 2. The horizontal lines are the HAZ levels of 0 and -2
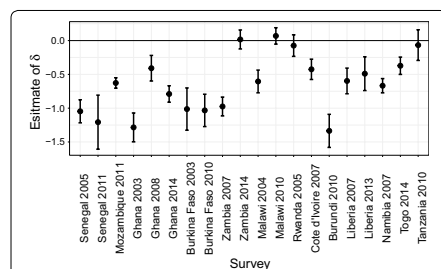
Fig. 5 Plot of estimates of the malaria effect on HAZ with associated 95% confidence intervals, obtained from the geostatistical model in (1) for each survey
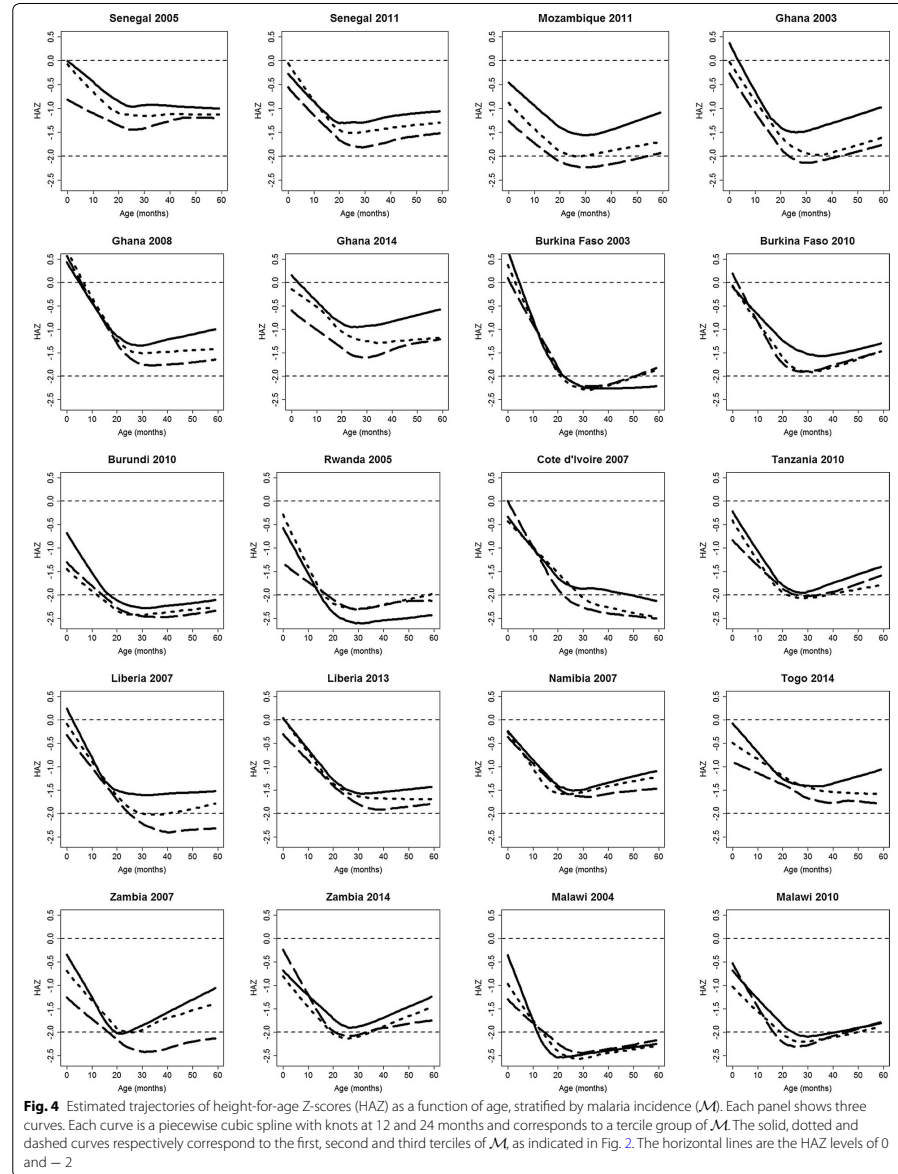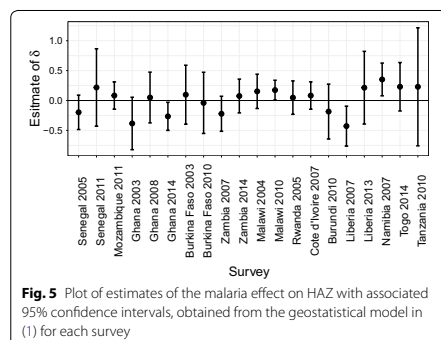
in 13 African countries between 2003 and 2014, we have developed a geostatistical framework to model HAZ as a function of both child-specific and spatial risk factors. As a proxy for malaria exposure, we used estimates of malaria incidence in the first year of life from the Malaria Atlas Project. A non-spatial univariate linear regression showed a negative effect of malaria incidence on HAZs. However, after controlling for confounding and residual spatial effects, the estimated effect of malaria on HAZ was weaker and not significant in 17 out of the 20 surveys considered.

One of the main challenges in modelling the association between malaria and HAZ is the need to take account of confounding effects. Among these, socio-economic status has been shown to be one of the most important [44–47]. Education is another important factor that affects both malaria exposure and risk of stunting [34, 48, 49]. Higher levels of education are associated with improved knowledge and practice about the appropriate strategies for the prevention and treatment of malaria [50], and about healthy practices in breastfeeding and child nutrition [51]. Our results are consistent with these findings in all of the 20 surveys here analysed.

We observed that in surveys where HAZ curves fall below the − 2 threshold in early childhood, the curves never really rise above the − 2 threshold in later years. This finding suggests that recovery to standard growth after 2 years of age may be more difficult when the decrease in HAZ in early childhood is severe. This is consistent with the findings from [52] who showed that recovery from stunting is associated with the severity of stunting in early years. Other factors that have been found to favour recovery from low HAZ are good nutrition [53] and higher levels of mother's education [54].

In our analysis, we found a mix of positive and negative point estimates of the association between malaria incidence and HAZ among the different surveys. However, findings from previous studies have shown contrasting results, with some reporting statistically significant negative associations between malaria and stunting [26, 29, 55, 56], and others reporting positive associations [30, 31]. To understand such variation in the magnitude and direction of the estimated parameters that quantify the malaria effect, we carried out a meta-analysis by considering several indicators of national development from the World Bank. Among these, the amount of arable land was the only one to show a significant association. Arable land might in fact modulate the association between malaria and HAZ, with a larger surface of arable land leading to a fall in poverty and malnutrition, especially in rural areas [57], but also to a larger number of breeding sites for mosquitoes [58]. This suggests that geo-political differences among countries should also be considered, since the implementation of policies aiming to reduce malnutrition can also impact on the epidemiology of malaria. Arable land could be indeed associated with agricultural, economic and environmental factors that are common to both malaria and stunting [59, 60].

We have quantified stunting risk by mapping the predictive probability that HAZ is below a threshold of − 2. For countries with repeated surveys, our risk maps showed reductions over time in the risk of stunting. The main factors that might be driving such reductions are improvements in health environments through increasing access to safe water and sanitation, improvements in the quality of caring practices for children through increasing women's education and promoting gender equality, including women's empowerment; and increase in food security by ensuring adequate availability of food at the national level and sufficient nutritional quality of that food [59, 61, 62]. Our risk maps showed remarkable spatial heterogeneity in the risk of stunting, identifying geographic areas with high risk that could be considered for a more targeted intervention.

It has been widely observed that HAZ undergoes a rapid decrease in the first 24 months and an increase thereafter [11, 12, 42]. For this reason we used cubic splines with knots at 12 and 24 months in order to better capture the non-linear trajectory that we observed across the 5 years of age.

Limitations of the study

The main limitation of our study is that the information available to us on malaria and HAZ is cross-sectional, rather than longitudinal, in nature. This prevents us from establishing whether our observed associations

Fig. 6 Predicted stunting risk maps for Ghana, Burkina Faso and Mozambique. The colour scale ranges from green to red with red areas being high risk areas and green areas being low risk areas

can be given a causal interpretation. A second limitation is that we have no information on the uncertainty associated with the estimates of malaria incidence. We have assumed the first year of life to be the most important in determining the strength of the association between malaria and child growth. To investigate whether exposure to malaria in other years of childhood could also have an impact on growth would require the fitting of a distributed lag-model.

In Additional file 6, we give methodological details on how to account for uncertainty in malaria incidence in a cross-sectional geostatistical setting.

To assess the cumulative effect of malaria on child-growth at different developmental stages, we would need longitudinal, individual-level data on children's actual malaria status over the first 5 years of life. We would then extend our current methodology as follows.

Novel extensions to longitudinal geostatistical data

To simplify the notation and without loss of generality, we assume that all the sampled children have identical follow up times. Then, let Y_{ijt} and W_{ijt} denote the HAZ and number of malaria episodes for the j-th child at location x_i and time t, respectively. Also, let $\tilde{S}(x, t)$ denote a latent spatio-temporal Gaussian process. Given $\tilde{S}(x, t)$, we model the W_{ijt} as a set of mutually independent Poisson variables with mean $\mathcal{M}_{ijt}$ such that

$$\log\{\mathcal{M}_{ijt}\} = \tilde{e}_{ijt}^\top \tilde{\gamma} + \tilde{d}(x_i)^\top \tilde{\beta} + \tilde{S}(x_i, t),$$

where $\tilde{e}_{ijt}$ are child-specific explanatory variables that might vary over time. We then assume that Y_{ijt}, conditionally on $\mathcal{M}_{ijt}$, a spatio-temporal Gaussian process $S(x, t)$ and random effects U_{it} and V_{ij}, are independent Gaussian variables with mean

$$\mu_j(x_i, t) = e_{ijt}^\top \gamma + d(x_i, t)\beta + \sum_{h=0}^{t-1} \delta_{t-h} \mathcal{M}_{ij(t-h)}$$
$$+ f(\mathcal{A}_{ijt}) + V_{ij} + S(x_i, t) + U_{it} \qquad (4)$$

In (4), U_{it} is unstructured unexplained variation at location x_i and time t, V_{ij} is unexplained child-specific variation and the lagged parameters δ_{t-h}, for $h = 0, \ldots, t = 1$, represents the effect of malaria incidence during the h-th year of life on HAZ. To make the model more parsimonious, the parameters δ_{t-h} can be constrained using a parametric specification, i.e. $\delta_{t-h} = g(t-h; \theta)$ where $g(\cdot; \theta)$ is a known function indexed by the vector of parameters θ.

This modelling framework would allow us to better understand the cumulative effect of malaria on HAZ at different developmental stages by overcoming the current limitation of our study where we assume that $\delta_{t-h} = 0$ for $0 \leq h \leq t - 2$.

Conclusion

Geostatistical methods provide a useful framework to account for spatially structured confounding effects that modulate the association between malaria and HAZ. This study also highlights that one of the main challenges in modelling this association is that confounding effects vary by country, as well as in time. This can change both the direction and magnitude of the effect of malaria on HAZ, making a generalization on the effect of malaria on HAZ almost impossible using only currently available data. Establishing whether the association between malaria and stunting is causal would require longitudinal follow-up data on individual children.

Additional files

Additional file 1. Computational details.

Additional file 2. Details of the World Bank development indicators.

Additional file 3. Estimates of covariance parameters.

Additional file 4. Results from the model validation.

Additional file 5. Maps of stunting risk.

Additional file 6. Accounting for the uncertainty in malaria incidence.

Author details
[1] CHICAS Research Group, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK. [2] Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK. [3] Department of Child Health, Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands. [4] Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands.

Atlas Project (http://www.map.ox.ac.uk/). The World Bank indicators data are publicly available in the World Bank database (http://data.worldbank.org/indicator).

References
1. Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, De Onis M, Ezzati M, Grantham-McGregor S, Katz J, Martorell R, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. Lancet. 2013;382(9890):427–51.
2. UNICEF et al. Improving child nutrition: the achievable imperatives for global progress. New York: UNICEF 2013. ISBN: 978-92-806-4686-3 https://www.unicef.org/publications/index_68661.html. Accessed December 2015.
3. Walker SP, Chang SM, Powell CA, Grantham-McGregor SM. Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted Jamaican children: prospective cohort study. Lancet. 2005;366(9499):1804–7.
4. Daniels MC, Adair LS. Growth in young Filipino children predicts schooling trajectories through high school. J Nutr. 2004;134(6):1439–46.
5. Cunha F, Heckman J. The technology of skill formation. Technical report, National Bureau of Economic Research 2007.
6. Currie J. Child health in developed countries. Handb Health Econ. 2000;1:1053–90.
7. Heckman JJ, Stixrud J, Urzua S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Technical report, National Bureau of Economic Research, 2006.
8. Currie J. Healthy, wealthy, and wise: socioeconomic status, poor health in childhood, and human capital development. Technical report, National Bureau of Economic Research, 2008.
9. UNICEF et al. WHO, World Bank Group joint child malnutrition estimates. Levels and trends in child malnutrition: Key findings of the 2016 edition. Global Database on Child Growth and Malnutrition. 2016.
10. Marriott BP, White A, Hadden L, Davies JC, Wallingford JC. World Health Organization (WHO) infant and young child feeding indicators: associations with growth measures in 14 low income countries. Matern Child Nutr. 2012;8(3):354–70.
11. Victora CG, de Onis M, Hallal PC, Blössner M, Shrimpton R. Worldwide timing of growth faltering: revisiting implications for interventions. Pediatrics, 2010;2009–1519.
12. Stevens GA, Finucane MM, Paciorek CJ, Flaxman SR, White RA, Donner AJ, Ezzati M, Group NIMS, et al. Trends in mild, moderate, and severe stunting and underweight, and progress towards MDG 1 in 141 developing countries: a systematic analysis of population representative data. Lancet. 2012;380(9844):824–34.
13. UNICEF et al. WHO, World Bank Group joint child malnutrition estimates. Levels and trends in child malnutrition: key findings of the 2015 edition. Global Database on Child Growth and Malnutrition. 2015.
14. Custodio E, Descalzo MÁ, Villamor E, Molina L, Sánchez I, Lwanga M, Bernis C, Benito A, Roche J. Nutritional and socio-economic factors associated with *Plasmodium falciparum* infection in children from equatorial guinea: results from a nationally representative survey. Malar J. 2009;8(1):1.
15. Verhoef H, West C, Veenemans J, Begui Y. Stunting may determine the severity of malaria-associated anemia in African children. Pediatrics. 2002;110:e48.
16. Holding PA, Kitsao-Wekulo PK. Describing the burden of malaria on child development: what should we be measuring and how should we be measuring it? Am J Trop Med Hyg. 2004;71(2 suppl):71–9.
17. Shanks GD, Hay SI, Bradley DJ. Malaria's indirect contribution to all-cause mortality in the Andaman islands during the colonial era. Lancet Infect Dis. 2008;8(9):564–70.
18. Bhatt S, Weiss D, Cameron E, Bisanzio D, Mappin B, Dalrymple U, Battle K, Moyes C, Henry A, Eckhoff P, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526(7572):207–11.
19. World Health Organization: World malaria report 2016. 2016.
20. World Health Organization: World malaria report 2017. 2017.
21. Kalanda BF, van Buuren S, Verhoeff FH, Brabin BJ. Catch-up growth in Malawian babies, a longitudinal study of normal and low birthweight babies born in a malarious endemic area. Early Human Dev. 2005;81(10):841–50.
22. De Beaudrap P, Turyakira E, Nabasumba C, Tumwebaze B, Piola P, Boum Y II, McGready R. Timing of malaria in pregnancy and impact on infant growth and morbidity: a cohort study in Uganda. Malar J. 2016;15(1):1.
23. Uddenfeldt Wort U, Hastings IM, Carlstedt A, Mutabingwa T, Brabin BJ. Impact of El Nino and malaria on birthweight in two areas of Tanzania with different malaria transmission patterns. Int J Epidemiol. 2004;33(6):1311–9.
24. McGregor IA, Wilson M, Billewicz W. Malaria infection of the placenta in The Gambia, West Africa; its incidence and relationship to stillbirth, birthweight and placental weight. Trans R Soc Trop Med Hyg. 1983;77(2):232–44.
25. Guyatt HL, Snow RW. Impact of malaria during pregnancy on low birth weight in sub-Saharan Africa. Clin Microbiol Rev. 2004;17(4):760–9.
26. Kang H, Kreuels B, Adjei O, Krumkamp R, May J, Small DS. The causal effect of malaria on stunting: a Mendelian randomization and matching approach. Int J Epidemiol. 2013;42(5):1390–8.
27. Nyakeriga A, Troye-Blomberg M, Chemtai A, Marsh K, Williams T. Malaria and nutritional status in children living on the coast of Kenya. Am J Clin Nutr. 2004;80(6):1604–10.
28. Olney DK, Kariger PK, Stoltzfus RJ, Khalfan SS, Ali NS, Tielsch JM, Sazawal S, Black R, Allen LH, Pollitt E. Development of nutritionally at-risk young children is predicted by malaria, anemia, and stunting in Pemba, Zanzibar. J Nutr. 2009;139(4):763–72.
29. Deen J, Walraven G, Von Seidlein L. Increased risk for malaria in chronically malnourished children under 5 years of age in rural Gambia. J Trop Pediatr. 2002;48(2):78–83.
30. Murray M, Murray A, Murray N, Murray M. Diet and cerebral malaria: the effect of famine and refeeding. Am J Clin Nutr. 1978;31(1):57–61.
31. Genton B, Al-Yaman F, Ginny M, Taraika J, Alpers MP. Relation of anthropometry to malaria morbidity and immunity in Papua New Guinean children. Am J Clin Nutr. 1998;68(3):734–41.
32. Snow R, Byass P, Shenton F, Greenwood B. The relationship between anthropometric measurements and measurements of iron status and susceptibility to malaria in Gambian children. Trans R Soc Trop Med Hyg. 1991;85(5):584–9.
33. Müller O, Garenne M, Kouyaté B, Becher H. The association between protein-energy malnutrition, malaria morbidity and all-cause mortality in West African children. Trop Med Int Health. 2003;8(6):507–11.
34. Fink G, Olgiati A, Hawela M, Miller JM, Matafwali B. Association between early childhood exposure to malaria and children's pre-school development: evidence from the Zambia early childhood development project. Malar J. 2013;12(1):1–9.
35. World Health Organization et al. WHO child growth standards: length/height for age, weight-for-age, weight-for-length, weight-for-height

Appendix B

Appendix B: R Code for statistical analysis using Models 4.14 and 4.15

File *README1.TXT*

```
The folder loaloa_application contains R-scripts and example
datasets used for the Loa loa application.

##########
SCRIPTS
##########

There are two R-scripts. The script loa_functions.R contains
R-function which are used in the other script,
Loaloa_Examples.R. In the script Loaloa_Examples.R,
we demonstrate the main steps of the Loa loa application.
Comments on each step of the analysis can be found
in the Loaloa_Examples.R script.

NB:  The script  loa_functions.R needs the R packages rgdal,
geoR, and PrevMap to be installed.




#########
DATASETS
#########

There are five .Rdata files whose description are as follows:
```

A) The `file` input_data.RData `is` a dataframe containing the
following variables.
    1. Cordinates of `223` geographical locations, `which` are
        displaced a few kilometres away from the
        locations of the origional data. The coordinates
        are `"web_x"` and `"web_y"`, in the web mercator
        projection (`3857`, km)
    2. The `binomial` dinominators of the origional data `"n"`
    3. The surface elevation `variable` of the locations `"elev"`.
This dataset `is` an input dataset to simulate prevalence data
under the `fitted` Model `2` specification
of the asymmetric `model`.

B) The `file` LoaloaExampleData.RData `is` an `example` data that
can be used `for` the analyses as `if` it were the original data.
In addition to the variables described in A) above,
        it has the variables,
    1. `"prev.alte"` and `"y.alte"`: respectively prevalence
        and number of positive tests as defined `by` the
        alternative (possibly baised) diagnostic
    2. `"prev.gold"` and `"y.gold"`: respectively prevalence
        and number of positive tests as defined `by` the
        gold standard diagnostic,
    3. `"elev_below750"`, `"elev_between750_1015"`,
        `"elev_above1015"` variables defining the
        elevation `spline`.

C) The `file` MCML_estimates_asymm_M1.RData and
    MCML_estimates_asymm_M.RData are `example` of MCML
    estimates of the Model `1` & `2` respective
    specification of the assymtric `model`.
    These were and can be generated using the
    script Loaloa_Examples.`R`.

D) prediction_`grid`_covariates.RData `is` an `example` `grid` `for`
    spatial predictions. It also contains the values
    of elevation at the prediction locations.

File *loa_functions.R*

```
####################################################
####################################################
# Functions for the Loaloa Analyses
####################################################
####################################################

get_elev_vars <- function(elev, k1, k2){
  elev1 <- (elev <= k1)*elev + (elev > k1)*k1
  elev2 <- ((elev > k1) & (elev <= k2))*(elev-k1)  +
```

```r
    (elev > k2)*(k2-k1)
  elev3 <- (elev > k2)*(elev-k2)

  out <- data.frame(elev)
  out[paste("elev_below", k1, sep="")] <- elev1
  out[paste("elev_between", k1, "_", k2, sep="")] <- elev2
  out[paste("elev_above", k2, sep="")] <- elev3
  out$elev <- NULL

  return(out)
}




sim_from_asymm_M2 <- function(data){

  n <- nrow(data) # Total number of locations
  # coordinates of the geolocations
  coords <- data[, c("web_x", "web_y")]
  units.m <- data$n # No. of individuals tested location

  # Parameter values to simulate the data
  # betas for the alternative diagnostic
  beta.alte <- c(-0.763, 0.588e-3, -3.412e-3, -0.059e-3)
  # betas for the gold standard
  beta.gold <- c(-1.736, 0.126e-3, -0.039e-3, -0.612e-3)
  # Parameter of the linear association between
  # f_1(P_1) and f_2(P_2)
  alpha <- 1.017
  # scale parameter of the spatial process for the
  # alternative diagnostic
  phi.alte <- 187.388
  # relative variance of the nugget effect for the
  # gold standard
  nu2.alte <- 0.20
  # scale parameter of the spatial process for the
  # alternative diagnostic
  phi.gold <- 23.686
  # relative variance of the nugget effect for the
  # gold standard
  nu2.gold <- 0.483
  # variance of the spatial process for the
  # alternative diagnostic
  sigma2.alte <- 1.617
  # variance of the spatial process for the
  # gold standard
  sigma2.gold <- 0.216

  U <- as.matrix(dist(coords))
  covariates <- get_elev_vars(data$elev, 750, 1015)

  # Get design matrices and number of parameters
  D.alte <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = covariates)
  D.gold <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = covariates)
```

```r
  # Simulated alternative diagnostic prevalence
  Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
  nugg.alte <- sigma2.alte*nu2.alte
  diag(Sigma.alte) <- diag(Sigma.alte)+ nugg.alte
  S.alte <- t(chol(Sigma.alte))%*%(rnorm(n))
  eta.alte <- as.numeric(D.alte%*%beta.alte) + S.alte
  prev.alte <- 1/(1 + exp(- eta.alte))

  # Simulated gold standard prevalence
  Sigma.gold <- sigma2.gold*exp(-U/phi.gold)
  nugg.gold <-  sigma2.gold*nu2.gold
  diag(Sigma.gold) <- diag(Sigma.gold)+nugg.gold
  S.gold <- t(chol(Sigma.gold))%*%(rnorm(n))
  eta.gold <- as.numeric(D.gold%*%beta.gold) +
    S.gold + alpha*eta.alte
  prev.gold <- 1/(1 + exp(- eta.gold))

  # Simulating the two outcome variables
  y.alte <- rbinom(n, units.m, prev.alte )
  y.gold <-  rbinom(n ,units.m, prev.gold )

  # Dataframe of the of the simulated data
  df.out <- cbind(data,  prev.alte , prev.gold, y.alte,
                  y.gold, covariates)


  return(df.out)
}



fit_asymm_M2 <- function(data){

  coords <- data[,c("web_x","web_y")]

  set.seed(1234)
  n <- nrow(data)
  U <- as.matrix(dist(coords))


  units.m <- data$n # number tested at locations
  y.alte <- data$y.alte # alternative test outcome
  y.gold <- data$y.gold # loa loa outcome

  # Design matrices
  D.alte <- model.matrix(~ elev_below750 +
        elev_between750_1015 + elev_above1015, data = data)
  D.gold <- model.matrix(~ elev_below750 +
        elev_between750_1015 + elev_above1015, data = data)

  # Number of regression parameters for the two outcomes
  p.D.alte <- ncol(D.alte)
  p.D.gold <- ncol(D.gold)


  # Initial values of the model parameters
  # betas for the alternative diagnostic
```

```r
beta.alte <- c(-0.763, 0.588e-3, -3.412e-3, -0.059e-3)
# betas for the gold standard
beta.gold <- c(-1.736, 0.126e-3, -0.039e-3, -0.612e-3)
# Parameter of the linear association
# between f_1(P_1) and f_2(P_2)
alpha <- 1.017
# scale parameter of the spatial
# process for the alternative diagnostic
phi.alte <- 187.388
# relative variance of the
# nugget effect for the gold standard
nu2.alte <- 0.20
# scale parameter of the s
# patial process for the alternative diagnostic
phi.gold <- 23.686
# relative variance of the
# nugget effect for the gold standard
nu2.gold <- 0.483 # variance of the spatial
# process for the alternative diagnostic
sigma2.alte <- 1.617
# variance of the spatial process
# for the gold standard
sigma2.gold <- 0.216



Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
# covarince matrix wrt altenative diagnostic
diag(Sigma.alte) <- diag(Sigma.alte)+sigma2.alte*nu2.alte
Sigma.alte.inv <- solve(Sigma.alte)
  Sigma.gold <- sigma2.gold*exp(-U/phi.gold)
# covariance matrix wrt gold std
diag(Sigma.gold) <- diag(Sigma.gold)+sigma2.gold*nu2.gold
Sigma.gold.inv <- solve(Sigma.gold)
n.tot <- 2*n # total number of random effects = 2* nlocations
M <- Matrix(0,n.tot,n.tot,sparse=TRUE)
M[1:n,1:n] <- -Sigma.alte.inv
M[(n+1):n.tot,(n+1):n.tot] <- -Sigma.gold.inv
mu.alte <- as.numeric(D.alte%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)

f <- function(x) log(x/(1-x))
grad.f <- function(x) -1/((x-1)*x)
hess.f <- function(x) (2*x-1)/((x-1)^2*x^2)

ind.alte <- 1:n
ind.gold <- (n+1):n.tot

##################################################
#### SIMULATING FROM THE CONDITIONAL DISTRIBUTION
#### OF THE RANDOM EFFECT GIVEN THE DATA
##################################################

integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]
  diff.S.alte <- S.alte-mu.alte
```

```r
    q.f_S.alte <- t(diff.S.alte)%*%Sigma.alte.inv%*%(diff.S.alte)
    diff.S.gold <- S.gold-mu.gold
    q.f_S.gold <- t(diff.S.gold)%*%Sigma.gold.inv%*%(diff.S.gold)
    q.f_S <- -0.5*as.numeric(q.f_S.alte+q.f_S.gold)
    p.alte <- exp(S.alte)/(1+exp(S.alte))
    eta.gold <- S.gold+alpha*f(p.alte)
    llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
      sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))
    as.numeric(q.f_S+llik)
}

grad.integrand <- function(S) {
    S.alte <- S[ind.alte]
    S.gold <- S[ind.gold]
    diff.S.alte <- S.alte-mu.alte
    diff.S.gold <- S.gold-mu.gold
    p.alte <- exp(S.alte)/(1+exp(S.alte))
    der.p.alte <- p.alte/(1+exp(S.alte))
    p.gold <-  exp(S.gold+alpha*f(p.alte))/
      (1+exp(S.gold+alpha*f(p.alte)))
    aux.gold <- y.gold-units.m*p.gold

    der.q.f_S <- -c(Sigma.alte.inv%*%diff.S.alte,
                    Sigma.gold.inv%*%diff.S.gold)
    der.llik <- c(y.alte-units.m*p.alte+
                    aux.gold*alpha*grad.f(p.alte)*der.p.alte,
                  aux.gold)
    out <- der.q.f_S+der.llik
    return(out)
}

hessian.integrand <- function(S) {
    S.alte <- S[ind.alte]
    S.gold <- S[ind.gold]

    p.alte <- exp(S.alte)/(1+exp(S.alte))
    der.p.alte <- p.alte/(1+exp(S.alte))
    der2.p.alte <- (1-exp(S.alte))*exp(S.alte)/((1+exp(S.alte))^3)
    p.gold <-  exp(S.gold+alpha*f(p.alte))/
      (1+exp(S.gold+alpha*f(p.alte)))

    aux.gold <- y.gold-units.m*p.gold
    der.aux.gold <- -units.m*exp(S.gold+alpha*f(p.alte))/
      ((1+exp(S.gold+alpha*f(p.alte)))^2)
    res <- M
    diag(res)[ind.alte] <- diag(res)[ind.alte]-
      units.m*exp(S.alte)/((1+exp(S.alte))^2)+
      aux.gold*(alpha*(hess.f(p.alte)*(der.p.alte^2)+
                       grad.f(p.alte)*der2.p.alte))+
      der.aux.gold*((alpha*grad.f(p.alte)*der.p.alte)^2)

    diag(res)[ind.gold] <- diag(res)[ind.gold]+der.aux.gold

    diag(res[ind.gold,ind.alte]) <-
      diag(res[ind.alte,ind.gold]) <-
      der.aux.gold*(alpha*grad.f(p.alte)*der.p.alte)
    return(res)
```

```r
}

S.estim <- nlminb(start=c(mu.alte, mu.gold),
                  function(x) -as.numeric(integrand(x)),
                  function(x) -as.numeric(grad.integrand(x)),
                  function(x) -as.matrix(hessian.integrand(x)),
                  control=list(trace=1)
)

S.estim$gradient <- grad.integrand(S.estim$par)
S.estim$hessian <- hessian.integrand(S.estim$par)
S.estim$Sigma.tilde <- round(solve(-S.estim$hessian), 9)
S.estim$estimate <- S.estim$par
Sigma.sroot <- t(chol(S.estim$Sigma.tilde))
A <- solve(Sigma.sroot)

library(Matrix)

Sigma.W.inv <- solve(A%*%bdiag(Sigma.alte,Sigma.gold)%*%t(A))
mu.W <- as.numeric(A%*%(c(mu.alte,mu.gold)-S.estim$estimate))

n.sim <- 12000 # number of simulations
burnin <- 2000 # burnin
thin <- 5 # thinning


n.samples <- (n.sim-burnin)/thin

h <- 1.65/((n.tot)^(1/6))
c1.h <- 0.01
c2.h <- 1e-04



# reparameterization of the model
 # rewrite likelihood based on W
cond.dens.W <- function(W,S) {
  diff.W <- W-mu.W

  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]


  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(-0.5*as.numeric(t(diff.W)%*%
                      Sigma.W.inv%*%diff.W))+llik
}


# derivetive w.r.t W for the Langevin-Hastings algorithm
lang.grad <- function(W,S) {
  diff.W <- W-mu.W
```

```r
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.llik <- c(y.alte-units.m*p.alte+
                  aux.gold*alpha*grad.f(p.alte)*der.p.alte,
              aux.gold)
  out <- as.numeric(-Sigma.W.inv%*%diff.W+
                  t(Sigma.sroot)%*%der.llik)
  return(out)
}




S.estim$mode <- S.estim$estimate
W.curr <- rep(0,n.tot)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+S.estim$mode)
mean.curr <- as.numeric(W.curr + (h^2/2)*lang.grad(W.curr,S.curr))
lp.curr <- cond.dens.W(W.curr,S.curr)

acc <- 0
sim <- matrix(NA,nrow=n.samples,ncol=n.tot)
h.vec <- rep(NA,n.sim)

for(i in 1:n.sim) {
  W.prop <- mean.curr+h*rnorm(n.tot)
  S.prop <-  as.numeric(Sigma.sroot%*%W.prop+S.estim$mode)
  mean.prop <- as.numeric(W.prop + (h^2/2)*lang.grad(W.prop,S.prop))
  lp.prop <- cond.dens.W(W.prop,S.prop)

  dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
  dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

  log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

  if(log(runif(1)) < log.prob) {
    acc <- acc+1
    W.curr <- W.prop
    S.curr <- S.prop
    lp.curr <- lp.prop
    mean.curr <- mean.prop
  }

  if( i > burnin & (i-burnin)%%thin==0) {
    sim[(i-burnin)/thin,] <- S.curr
  }

  h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
  cat("Iteration",i,"out of",n.sim,"\r")
  flush.console()
}
```

```r
if(TRUE) {
  acf.plot <- acf(sim[,1],plot=FALSE)
  plot(acf.plot$lag,acf.plot$acf,type="l",
       xlab="lag",ylab="autocorrelation",
       ylim=c(-0.1,1),
  main="Autocorrelogram of the simulated samples")
  for(i in 2:ncol(sim)) {
    acf.plot <- acf(sim[,i],plot=FALSE)
    lines(acf.plot$lag,acf.plot$acf)
  }
  abline(h=0,lty="dashed",col=2)
}




#######################################
#### APROXIMATING AND MAXIMIZING THE
#### LIKELIHOOD USING THE MCMC SAMPLES
#######################################

log.integrand <- function(S,val) { # Objective function
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-val$mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%val$R.alte.inv%*%
    (diff.S.alte)/val$sigma2.alte
  diff.S.gold <- S.gold-val$mu.gold
  q.f_S.gold <- t(diff.S.gold)%*%val$R.gold.inv%*%
    (diff.S.gold)/val$sigma2.gold
  q.f_S <- -0.5*as.numeric(val$l.det.R.alte+val$l.det.R.gold+
             n*log(val$sigma2.alte)+n*log(val$sigma2.gold)+
               q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*(S.gold+val$alpha*f(p.alte))-units.m*
          log(1+exp(S.gold+val$alpha*f(p.alte))))

  as.numeric(q.f_S+llik)
}


compute.log.f <- function(par,l.det.R.alte=NA,R.alte.inv=NA,
                           l.det.R.gold=NA,R.gold.inv=NA) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]

  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])
  val <- list()
  val$mu.alte <- as.numeric(D.alte%*%beta.alte)
```

```r
    val$mu.gold <- as.numeric(D.gold%*%beta.gold)
    val$sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
    val$sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])
    val$alpha <- par[p.D.alte+p.D.gold+1]
    if(is.na(l.det.R.alte) & is.na(as.numeric(R.alte.inv)[1])) {
      R.alte <- exp(-U/phi.alte)
      diag(R.alte) <- diag(R.alte) + nu2.alte
      R.gold <- exp(-U/phi.gold)
      diag(R.gold) <- diag(R.gold) + nu2.gold
      val$l.det.R.alte <- determinant(R.alte)$modulus
      val$R.alte.inv <- solve(R.alte)
      val$l.det.R.gold <- determinant(R.gold)$modulus
      val$R.gold.inv <- solve(R.gold)
    } else {
      val$l.det.R.alte <- l.det.R.alte
      val$R.alte.inv <- R.alte.inv
      val$l.det.R.gold <- l.det.R.gold
      val$R.gold.inv <- R.gold.inv
    }
    sapply(1:n.samples,function(i) log.integrand(sim[i,],val))
}

# theta0
par0 <- c(beta.alte, beta.gold, alpha,
          log(c(sigma2.alte,phi.alte,nu2.alte,
                sigma2.gold,phi.gold,nu2.gold)))

rm(beta.alte,beta.gold,alpha,sigma2.alte,phi.alte,
   nu2.alte,sigma2.gold,phi.gold,nu2.gold)



log.f.tilde <- compute.log.f(par0)

MC.log.lik <- function(par) {
  log(mean(exp(compute.log.f(par)-log.f.tilde)))
}



grad.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
  sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])


  R.alte <- exp(-U/phi.alte)
```

```r
diag(R.alte) <- diag(R.alte)+nu2.alte
R.alte.inv <- solve(R.alte)
l.det.R.alte <- determinant(R.alte)$modulus

R.gold <- exp(-U/phi.gold)
diag(R.gold) <- diag(R.gold)+nu2.gold
R.gold.inv <- solve(R.gold)
l.det.R.gold <- determinant(R.gold)$modulus

exp.fact <- exp(compute.log.f(par,l.det.R.alte,
      R.alte.inv,l.det.R.gold,R.gold.inv)-log.f.tilde)
L.m <- sum(exp.fact)
exp.fact <- exp.fact/L.m


R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
m2.phi.alte <- m1.phi.alte%*%R.alte.inv

t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

R.alte.phi.gold <- (U*exp(-U/phi.gold))/phi.gold
m1.phi.gold <- R.gold.inv%*%R.alte.phi.gold
t1.phi.gold <- -0.5*sum(diag(m1.phi.gold))
m2.phi.gold <- m1.phi.gold%*%R.gold.inv

t1.nu2.gold <- -0.5*sum(diag(R.gold.inv))*nu2.gold
m2.nu2.gold <- R.gold.inv%*%R.gold.inv*nu2.gold

gradient.S <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  f.p.alte <- f(p.alte)
  p.gold <- exp(S.gold+alpha*f.p.alte)/
    (1+exp(S.gold+alpha*f.p.alte))

  q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
  q.f.gold <- t(diff.S.gold)%*%R.gold.inv%*%diff.S.gold
  grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
    (diff.S.alte)/sigma2.alte
  grad.beta.gold <-  t(D.gold)%*%R.gold.inv%*%
    (diff.S.gold)/sigma2.gold

  grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

  grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+0.5*
                  q.f.alte/(sigma2.alte^2))*sigma2.alte
  grad.log.sigma2.gold <- (-n/(2*sigma2.gold)+0.5*q.f.gold/
                  (sigma2.gold^2))*sigma2.gold


  grad.log.phi.alte <- (t1.phi.alte+0.5*
```

```r
                       as.numeric(t(diff.S.alte)%*%
                     m2.phi.alte%*%(diff.S.alte))/sigma2.alte)
    grad.log.phi.gold <- (t1.phi.gold+0.5*
                      as.numeric(t(diff.S.gold)%*%
                  m2.phi.gold%*%(diff.S.gold))/sigma2.gold)

    grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
                        as.numeric(t(diff.S.alte)%*%
                 m2.nu2.alte%*%(diff.S.alte))/sigma2.alte)
    grad.log.nu2.gold <-  (t1.nu2.gold+0.5*
                       as.numeric(t(diff.S.gold)%*%
                  m2.nu2.gold%*%(diff.S.gold))/sigma2.gold)

    out <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
        grad.log.sigma2.alte,
        grad.log.phi.alte,grad.log.nu2.alte,
        grad.log.sigma2.gold,
        grad.log.phi.gold,grad.log.nu2.gold)

    return(out)
  }
  out <- rep(0,length(par))
  for(i in 1:n.samples) {
    out <- out + exp.fact[i]*gradient.S(sim[i,])
  }
  out
}




hess.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
  sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])


  R.alte <- exp(-U/phi.alte)
  diag(R.alte) <- diag(R.alte)+nu2.alte
  R.alte.inv <- solve(R.alte)
  l.det.R.alte <- determinant(R.alte)$modulus

  R.gold <- exp(-U/phi.gold)
  diag(R.gold) <- diag(R.gold)+nu2.gold
  R.gold.inv <- solve(R.gold)
  l.det.R.gold <- determinant(R.gold)$modulus
```

```r
exp.fact <- exp(compute.log.f(par,l.det.R.alte,
          R.alte.inv,l.det.R.gold,R.gold.inv)-log.f.tilde)
L.m <- sum(exp.fact)
exp.fact <- exp.fact/L.m


R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
m2.phi.alte <- m1.phi.alte%*%R.alte.inv

t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

R.alte.phi.gold <- (U*exp(-U/phi.gold))/phi.gold
m1.phi.gold <- R.gold.inv%*%R.alte.phi.gold
t1.phi.gold <- -0.5*sum(diag(m1.phi.gold))
m2.phi.gold <- m1.phi.gold%*%R.gold.inv

t1.nu2.gold <- -0.5*sum(diag(R.gold.inv))*nu2.gold
m2.nu2.gold <- R.gold.inv%*%R.gold.inv*nu2.gold

R.gold.phi.alte <- R.alte.phi.alte+(U*(U-2*phi.alte)*
                        exp(-U/phi.alte))/phi.alte^2
t2.phi.alte <- -0.5*(sum(R.alte.inv*R.gold.phi.alte)-
              sum(m1.phi.alte*t(m1.phi.alte)))
n2.phi.alte <- R.alte.inv%*%(2*R.alte.phi.alte%*%
          m1.phi.alte-R.gold.phi.alte)%*%R.alte.inv

R.gold.phi.gold <- R.alte.phi.gold+(U*(U-2*phi.gold)*
                        exp(-U/phi.gold))/phi.gold^2
t2.phi.gold <- -0.5*(sum(R.gold.inv*R.gold.phi.gold)-
                sum(m1.phi.gold*t(m1.phi.gold)))
n2.phi.gold <- R.gold.inv%*%(2*R.alte.phi.gold%*%
          m1.phi.gold-R.gold.phi.gold)%*%R.gold.inv

t2.nu2.alte <- -0.5*(sum(diag(R.alte.inv)*nu2.alte)-
          sum(diag(m2.nu2.alte)*nu2.alte))
m1.nu2.alte <- R.alte.inv*nu2.alte
nu2.alte.aux <- 2*m1.nu2.alte*nu2.alte
diag(nu2.alte.aux) <- diag(nu2.alte.aux)-nu2.alte
n2.nu2.alte <- R.alte.inv%*%nu2.alte.aux%*%R.alte.inv

t2.phi.alte.nu2.alte <- 0.5*sum(m1.phi.alte*
                        t(R.alte.inv))*nu2.alte
n2.phi.alte.nu2.alte <- R.alte.inv%*%(nu2.alte*
              m1.phi.alte+nu2.alte*R.alte.phi.alte%*%
                    R.alte.inv)%*%R.alte.inv

t2.nu2.gold <- -0.5*(sum(diag(R.gold.inv)*nu2.gold)-
            sum(diag(m2.nu2.gold)*nu2.gold))
m1.nu2.gold <- R.gold.inv*nu2.gold
nu2.gold.aux <- 2*m1.nu2.gold*nu2.gold
diag(nu2.gold.aux) <- diag(nu2.gold.aux)-nu2.gold
n2.nu2.gold <- R.gold.inv%*%nu2.gold.aux%*%R.gold.inv
```

```r
t2.phi.gold.nu2.gold <- 0.5*sum(m1.phi.gold*
                         t(R.gold.inv))*nu2.gold
n2.phi.gold.nu2.gold <- R.gold.inv%*%(nu2.gold*
                     m1.phi.gold+nu2.gold*R.alte.phi.gold%*%
             R.gold.inv)%*%R.gold.inv

ind.beta.alte <- 1:p.D.alte
ind.beta.gold <- (p.D.alte+1):(p.D.alte+p.D.gold)
ind.alpha <- p.D.alte+p.D.gold+1
ind.sigma2.alte <- p.D.alte+p.D.gold+2
ind.phi.alte <- p.D.alte+p.D.gold+3
ind.nu2.alte <- p.D.alte+p.D.gold+4
ind.sigma2.gold <- p.D.alte+p.D.gold+5
ind.phi.gold <- p.D.alte+p.D.gold+6
ind.nu2.gold <- p.D.alte+p.D.gold+7

H <- matrix(0,nrow=length(par),ncol=length(par))
H[ind.beta.alte,ind.beta.alte] <-
    -t(D.alte)%*%R.alte.inv%*%D.alte/sigma2.alte
H[ind.beta.gold,ind.beta.gold] <-
  -t(D.gold)%*%R.gold.inv%*%D.gold/sigma2.gold

hessian.S <- function(S,ef) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  f.p.alte <- f(p.alte)
  p.gold <- exp(S.gold+alpha*f.p.alte)/
    (1+exp(S.gold+alpha*f.p.alte))

  q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
  q.f.gold <- t(diff.S.gold)%*%R.gold.inv%*%diff.S.gold
  grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
        (diff.S.alte)/sigma2.alte
  grad.beta.gold <-  t(D.gold)%*%R.gold.inv%*%
    (diff.S.gold)/sigma2.gold

  grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

  grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+0.5*
                    q.f.alte/(sigma2.alte^2))*sigma2.alte
  grad.log.sigma2.gold <- (-n/(2*sigma2.gold)+0.5*
                    q.f.gold/(sigma2.gold^2))*sigma2.gold

  grad.log.phi.alte <- (t1.phi.alte+0.5*
                          as.numeric(t(diff.S.alte)%*%
                    m2.phi.alte%*%(diff.S.alte))/sigma2.alte)
  grad.log.phi.gold <- (t1.phi.gold+0.5*
                          as.numeric(t(diff.S.gold)%*%
                      m2.phi.gold%*%(diff.S.gold))/sigma2.gold)

  grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
                           as.numeric(t(diff.S.alte)%*%
                      m2.nu2.alte%*%(diff.S.alte))/sigma2.alte)
```

```
grad.log.nu2.gold <-  (t1.nu2.gold+0.5*
                            as.numeric(t(diff.S.gold)%*%
              m2.nu2.gold%*%(diff.S.gold))/sigma2.gold)

g <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
       grad.log.sigma2.alte,grad.log.phi.alte,
       grad.log.nu2.alte, grad.log.sigma2.gold,
       grad.log.phi.gold,grad.log.nu2.gold)

H[ind.alpha,ind.alpha] <- -sum((f.p.alte^2)*units.m*p.gold/
              (1+exp(S.gold+alpha*f.p.alte)))

H[ind.beta.alte,ind.sigma2.alte] <-
  H[ind.sigma2.alte,ind.beta.alte] <- -t(D.alte)%*%
        R.alte.inv%*%(diff.S.alte)/sigma2.alte
H[ind.beta.gold,ind.sigma2.gold] <-
  H[ind.sigma2.gold,ind.beta.gold] <- -t(D.gold)%*%
  R.gold.inv%*%(diff.S.gold)/sigma2.gold

H[ind.beta.alte,ind.phi.alte] <-
  H[ind.phi.alte,ind.beta.alte] <- -as.numeric(t(D.alte)%*%
    m2.phi.alte%*%(diff.S.alte))/sigma2.alte
H[ind.beta.gold,ind.phi.gold] <-
  H[ind.phi.gold,ind.beta.gold] <- -as.numeric(t(D.gold)%*%
    m2.phi.gold%*%(diff.S.gold))/sigma2.gold

H[ind.beta.alte,ind.nu2.alte] <-
  H[ind.nu2.alte,ind.beta.alte] <- -as.numeric(t(D.alte)%*%
    m2.nu2.alte%*%(diff.S.alte))/sigma2.alte
H[ind.beta.gold,ind.nu2.gold] <-
  H[ind.nu2.gold,ind.beta.gold] <- -as.numeric(t(D.gold)%*%
  m2.nu2.gold%*%(diff.S.gold))/sigma2.gold

H[ind.sigma2.alte,ind.sigma2.alte] <- (n/(2*sigma2.alte^2)-
                      q.f.alte/(sigma2.alte^3))*sigma2.alte^2+
  grad.log.sigma2.alte
H[ind.sigma2.gold,ind.sigma2.gold] <- (n/(2*sigma2.gold^2)-
                  q.f.gold/(sigma2.gold^3))*sigma2.gold^2+
  grad.log.sigma2.gold

H[ind.sigma2.alte,ind.phi.alte] <-
  H[ind.phi.alte,ind.sigma2.alte] <-
  -(grad.log.phi.alte-t1.phi.alte)
H[ind.sigma2.gold,ind.phi.gold] <-
  H[ind.phi.gold,ind.sigma2.gold] <-
  -(grad.log.phi.gold-t1.phi.gold)

H[ind.sigma2.alte,ind.nu2.alte] <-
  H[ind.nu2.alte,ind.sigma2.alte] <-
  -(grad.log.nu2.alte-t1.nu2.alte)
H[ind.sigma2.gold,ind.nu2.gold] <-
  H[ind.nu2.gold,ind.sigma2.gold] <-
  -(grad.log.nu2.gold-t1.nu2.gold)

H[ind.phi.alte,ind.phi.alte] <- t2.phi.alte-0.5*
  t(diff.S.alte)%*%n2.phi.alte%*%
  (diff.S.alte)/sigma2.alte
```

```r
    H[ind.phi.gold,ind.phi.gold] <- t2.phi.gold-0.5*
      t(diff.S.gold)%*%n2.phi.gold%*%
      (diff.S.gold)/sigma2.gold

    H[ind.nu2.alte,ind.nu2.alte] <- t2.nu2.alte-0.5*
      t(diff.S.alte)%*%n2.nu2.alte%*%
      (diff.S.alte)/sigma2.alte
    H[ind.phi.alte,ind.nu2.alte] <-
      H[ind.nu2.alte,ind.phi.alte] <- (t2.phi.alte.nu2.alte-
          0.5*t(diff.S.alte)%*%n2.phi.alte.nu2.alte%*%
            (diff.S.alte)/sigma2.alte)

    H[ind.nu2.gold,ind.nu2.gold] <- t2.nu2.gold-0.5*
      t(diff.S.gold)%*%n2.nu2.gold%*%
      (diff.S.gold)/sigma2.gold
    H[ind.phi.gold,ind.nu2.gold] <-
      H[ind.nu2.gold,ind.phi.gold] <- (t2.phi.gold.nu2.gold-
              0.5*t(diff.S.gold)%*%n2.phi.gold.nu2.gold%*%
                (diff.S.gold)/sigma2.gold)


    out <- list()
    out$mat1<- ef*(g%*%t(g)+H)
    out$g <- g*ef
    out
  }

  a <- rep(0,length(par))
  A <- matrix(0,length(par),length(par))
  for(i in 1:n.samples) {
    out.i <- hessian.S(sim[i,],exp.fact[i])
    a <- a+out.i$g
    A <- A+out.i$mat1
  }
  (A-a%*%t(a))
}

estim <- nlminb(par0,
                function(x) -MC.log.lik(x),
                function(x) -grad.MC.log.lik(x),
                function(x) -hess.MC.log.lik(x),
                control=list(trace=1))

estim$gradient <- grad.MC.log.lik(estim$par)
estim$hessian <- hess.MC.log.lik(estim$par)
estim$covariance <- solve(-estim$hessian)


names(estim$par) <- c("beta0.alte", "beta1.alte",
                       "beta2.alte", "beta3.alte",
                       "beta0.gold", "beta1.gold",
                       "beta2.gold", "beta3.gold",
                       "alpha", "log.sigma.alte",
                       "log.phi.alte", "log.nu.alte",
                       "log.sigma.gold", "log.phi.gold",
                       "log.nu.gold")
```

```r
  return(estim)
}




fit_asymm_M1 <- function(data){

  coords <- data[,c("web_x","web_y")]

  set.seed(1234)
  n <- nrow(data)
  U <- as.matrix(dist(coords))


  units.m <- data$n # number tested at locations
  y.alte <- data$y.alte # alternative test outcome
  y.gold <- data$y.gold # loa loa outcome

  # Design matrices
  D.alte <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = data)
  D.gold <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = data)

  # Number of regression parameters for the two outcomes
  p.D.alte <- ncol(D.alte)
  p.D.gold <- ncol(D.gold)


  # Parameter values to simulate the data
  # betas for the alternative diagnostic
  beta.alte <-c(-0.791, 0.515e-3, -3.529e-3, -0.110e-3)
  # betas for the gold standard
  beta.gold <- c(-1.762, 0.208e-3, -0.223e-3, -0.591e-3)
  # Parameter of the linear association
  between f_1(P_1) and f_2(P_2)
  alpha <- 1.005
  # scale parameter of the spatial process
  # for the alternative diagnostic
  phi.alte <- 182.037
  # relative variance
  # of the nugget effect for the gold standard
  nu2.alte <- 0.205
  # variance of the spatial process for
  # the alternative diagnostic
  sigma2.alte <- 1.581
  # relative variance of the
  # nugget effect for the gold standard
  tau2.gold <- 0.308



  Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
  diag(Sigma.alte) <- diag(Sigma.alte)+sigma2.alte*nu2.alte
  Sigma.alte.inv <- solve(Sigma.alte)
```

```r
Sigma.gold <- diag(tau2.gold,n)
Sigma.gold.inv <- solve(Sigma.gold)

n.tot <- 2*n # total number of random effects = 2* nlocations

M <- Matrix(0,n.tot,n.tot,sparse=TRUE)
M[1:n,1:n] <- -Sigma.alte.inv
M[(n+1):n.tot,(n+1):n.tot] <- -Sigma.gold.inv

mu.alte <- as.numeric(D.alte%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)


f <- function(x) log(x/(1-x))
grad.f <- function(x) -1/((x-1)*x)
hess.f <- function(x) (2*x-1)/((x-1)^2*x^2)

ind.alte <- 1:n
ind.gold <- (n+1):n.tot

integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%Sigma.alte.inv%*%(diff.S.alte)
  diff.S.gold <- S.gold-mu.gold
  q.f_S.gold <- t(diff.S.gold)%*%Sigma.gold.inv%*%(diff.S.gold)
  q.f_S <- -0.5*as.numeric(q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(q.f_S+llik)
}

grad.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte)) # checked
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte))) # checked
  aux.gold <- y.gold-units.m*p.gold

  der.q.f_S <- -c(Sigma.alte.inv%*%diff.S.alte,
                  Sigma.gold.inv%*%diff.S.gold)
  der.llik <- c(y.alte-units.m*p.alte+
                  aux.gold*alpha*grad.f(p.alte)*der.p.alte,
                aux.gold)
```

```r
    out <- der.q.f_S+der.llik
    return(out)
}

hessian.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  der2.p.alte <- (1-exp(S.alte))*
    exp(S.alte)/((1+exp(S.alte))^3)
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))

  aux.gold <- y.gold-units.m*p.gold
  der.aux.gold <- -units.m*exp(S.gold+alpha*
                  f(p.alte))/((1+exp(S.gold+alpha*f(p.alte)))^2)
  res <- M
  diag(res)[ind.alte] <- diag(res)[ind.alte]-
    units.m*exp(S.alte)/((1+exp(S.alte))^2)+
    aux.gold*
    (alpha*(hess.f(p.alte)*(der.p.alte^2)+
            grad.f(p.alte)*der2.p.alte))+
    der.aux.gold*((alpha*grad.f(p.alte)*der.p.alte)^2)

  diag(res)[ind.gold] <- diag(res)[ind.gold]+der.aux.gold

  diag(res[ind.gold,ind.alte]) <-  diag(res[ind.alte,ind.gold]) <-
    der.aux.gold*(alpha*grad.f(p.alte)*der.p.alte)
  return(res)
}

S.estim <-nlminb(start=c(mu.alte,mu.gold),
                  function(x) -as.numeric(integrand(x)),
                  function(x) -as.numeric(grad.integrand(x)),
                  function(x) -as.matrix(hessian.integrand(x)),
                  control=list(trace=1)
)
S.estim$gradient <- grad.integrand(S.estim$par)
S.estim$hessian <- hessian.integrand(S.estim$par)
S.estim$Sigma.tilde <- round(solve(-S.estim$hessian), 9)
S.estim$estimate <- S.estim$par
Sigma.sroot <- t(chol(S.estim$Sigma.tilde))
A <- solve(Sigma.sroot)

library(Matrix)

Sigma.W.inv <- solve(A%*%bdiag(Sigma.alte,Sigma.gold)%*%t(A))
mu.W <- as.numeric(A%*%(c(mu.alte,mu.gold)-S.estim$estimate))

n.sim <- 12000 # number of simulations
burnin <- 2000 # burnin
thin <- 5 # thinning


n.samples <- (n.sim-burnin)/thin
```

```r
h <- 1.65/((n.tot)^(1/6))
c1.h <- 0.01
c2.h <- 1e-04

# reparameterization of the model
# rewrite likelihood based on W
cond.dens.W <- function(W,S) {
  diff.W <- W-mu.W

  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]


  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(-0.5*as.numeric(t(diff.W)%*%
                      Sigma.W.inv%*%diff.W))+llik
}

# derivetive w.r.t W for the Langevin-Hastings algorithm
  lang.grad <- function(W,S) {
  diff.W <- W-mu.W

  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
      (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.llik <- c(y.alte-units.m*p.alte+
        aux.gold*alpha*grad.f(p.alte)*der.p.alte,
              aux.gold)
  out <- as.numeric(-Sigma.W.inv%*%diff.W+
                    t(Sigma.sroot)%*%der.llik)
  return(out)
}



S.estim$mode <- S.estim$estimate
W.curr <- rep(0,n.tot)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+S.estim$mode)
mean.curr <- as.numeric(W.curr + (h^2/2)*lang.grad(W.curr,S.curr))
lp.curr <- cond.dens.W(W.curr,S.curr)

acc <- 0
sim <- matrix(NA,nrow=n.samples,ncol=n.tot)
h.vec <- rep(NA,n.sim)
```

```r
for(i in 1:n.sim) {
  W.prop <- mean.curr+h*rnorm(n.tot)
  S.prop <-  as.numeric(Sigma.sroot%*%W.prop+S.estim$mode)
  mean.prop <- as.numeric(W.prop + (h^2/2)*lang.grad(W.prop,S.prop))
  lp.prop <- cond.dens.W(W.prop,S.prop)

  dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
  dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

  log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

  if(log(runif(1)) < log.prob) {
    acc <- acc+1
    W.curr <- W.prop
    S.curr <- S.prop
    lp.curr <- lp.prop
    mean.curr <- mean.prop
  }

  if( i > burnin & (i-burnin)%%thin==0) {
    sim[(i-burnin)/thin,] <- S.curr
  }

  h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
  cat("Iteration",i,"out of",n.sim,"\r")
  flush.console()
}

if(TRUE) {
  acf.plot <- acf(sim[,1],plot=FALSE)
  plot(acf.plot$lag,acf.plot$acf,type="l",
       xlab="lag",ylab="autocorrelation",
       ylim=c(-0.1,1),
       main="Autocorrelogram of the simulated samples")
  for(i in 2:ncol(sim)) {
    acf.plot <- acf(sim[,i],plot=FALSE)
    lines(acf.plot$lag,acf.plot$acf)
  }
  abline(h=0,lty="dashed",col=2)
}


 # Objective function. Rewrite like
log.integrand <- function(S,val) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-val$mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%val$R.alte.inv%*%
    (diff.S.alte)/val$sigma2.alte
  diff.S.gold <- S.gold-val$mu.gold
  q.f_S.gold <- sum(diff.S.gold^2)/val$tau2.gold
  q.f_S <- -0.5*as.numeric(val$l.det.R.alte+
                            n*log(val$sigma2.alte)+
                            n*log(val$tau2.gold)+
                            q.f_S.alte+q.f_S.gold)
```

```r
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*(S.gold+val$alpha*f(p.alte))-units.m*
         log(1+exp(S.gold+val$alpha*f(p.alte))))

  as.numeric(q.f_S+llik)
}

compute.log.f <- function(par,l.det.R.alte=NA,R.alte.inv=NA) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]

  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  val <- list()
  val$mu.alte <- as.numeric(D.alte%*%beta.alte)
  val$mu.gold <- as.numeric(D.gold%*%beta.gold)
  val$sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  val$tau2.gold <- exp(par[p.D.alte+p.D.gold+5])
  val$alpha <- par[p.D.alte+p.D.gold+1]
  if(is.na(l.det.R.alte) & is.na(as.numeric(R.alte.inv)[1])) {
    R.alte <- exp(-U/phi.alte)
    diag(R.alte) <- diag(R.alte) + nu2.alte

    val$l.det.R.alte <- determinant(R.alte)$modulus
    val$R.alte.inv <- solve(R.alte)

  } else {
    val$l.det.R.alte <- l.det.R.alte
    val$R.alte.inv <- R.alte.inv
  }
  sapply(1:n.samples,function(i) log.integrand(sim[i,],val))
}


# Initial values of the parameters
par0 <- c(beta.alte,beta.gold,alpha,
          log(c(sigma2.alte,phi.alte,nu2.alte,
                            tau2.gold)))

rm(beta.alte,beta.gold,alpha,
   sigma2.alte,phi.alte,nu2.alte,tau2.gold)



log.f.tilde <- compute.log.f(par0)

MC.log.lik <- function(par) {
  log(mean(exp(compute.log.f(par)-log.f.tilde)))
}

# THE GRADIENT OF THE MONTE CARLO LOG LIKELIHOOD
grad.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
```

```r
phi.alte <- exp(par[p.D.alte+p.D.gold+3])

mu.alte <- as.numeric(D.alte%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)
sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
nu2.alte <- exp(par[p.D.alte+p.D.gold+4])
tau2.gold <- exp(par[p.D.alte+p.D.gold+5])


R.alte <- exp(-U/phi.alte)
diag(R.alte) <- diag(R.alte)+nu2.alte
R.alte.inv <- solve(R.alte)
l.det.R.alte <- determinant(R.alte)$modulus

exp.fact <- exp(compute.log.f(par,l.det.R.alte,R.alte.inv)-
                log.f.tilde)
L.m <- sum(exp.fact)
exp.fact <- exp.fact/L.m


R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
m2.phi.alte <- m1.phi.alte%*%R.alte.inv

t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

gradient.S <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  f.p.alte <- f(p.alte)
  p.gold <- exp(S.gold+alpha*f.p.alte)/
    (1+exp(S.gold+alpha*f.p.alte))

  q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
  q.f.gold <- sum(diff.S.gold^2)
  grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
    (diff.S.alte)/sigma2.alte
  grad.beta.gold <-  t(D.gold)%*%(diff.S.gold)/tau2.gold

  grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

  grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+
              0.5*q.f.alte/(sigma2.alte^2))*sigma2.alte
  grad.log.tau2.gold <- (-n/(2*tau2.gold)+
            0.5*q.f.gold/(tau2.gold^2))*tau2.gold

  grad.log.phi.alte <- (t1.phi.alte+
              0.5*as.numeric(t(diff.S.alte)%*%
            m2.phi.alte%*%(diff.S.alte))/sigma2.alte)

  grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
```

```r
                        as.numeric(t(diff.S.alte)%*%
                  m2.nu2.alte%*%(diff.S.alte))/sigma2.alte)

    out <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
              grad.log.sigma2.alte,grad.log.phi.alte,
              grad.log.nu2.alte, grad.log.tau2.gold)

    return(out)
  }
  out <- rep(0,length(par))
  for(i in 1:n.samples) {
    out <- out + exp.fact[i]*gradient.S(sim[i,])
  }
  out
}




# THE HESSIAN OF THE LOG-LIKELIHOOD
hess.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
  sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  tau2.gold <- exp(par[p.D.alte+p.D.gold+5])


  R.alte <- exp(-U/phi.alte)
  diag(R.alte) <- diag(R.alte)+nu2.alte
  R.alte.inv <- solve(R.alte)
  l.det.R.alte <- determinant(R.alte)$modulus

  exp.fact <- exp(compute.log.f(par,l.det.R.alte,
                    R.alte.inv)-log.f.tilde)
  L.m <- sum(exp.fact)
  exp.fact <- exp.fact/L.m


  R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
  m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
  t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
  m2.phi.alte <- m1.phi.alte%*%R.alte.inv

  t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
  m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

  R.gold.phi.alte <- R.alte.phi.alte+(U*(U-2*phi.alte)*
                        exp(-U/phi.alte))/phi.alte^2
  t2.phi.alte <- -0.5*(sum(R.alte.inv*R.gold.phi.alte)-
                    sum(m1.phi.alte*t(m1.phi.alte)))
  n2.phi.alte <- R.alte.inv%*%(2*R.alte.phi.alte%*%
```

```r
                    m1.phi.alte-R.gold.phi.alte)%*%R.alte.inv

  t2.nu2.alte <- -0.5*(sum(diag(R.alte.inv)*nu2.alte)-
              sum(diag(m2.nu2.alte)*nu2.alte))
  m1.nu2.alte <- R.alte.inv*nu2.alte
  nu2.alte.aux <- 2*m1.nu2.alte*nu2.alte
  diag(nu2.alte.aux) <- diag(nu2.alte.aux)-nu2.alte
  n2.nu2.alte <- R.alte.inv%*%nu2.alte.aux%*%R.alte.inv

  t2.phi.alte.nu2.alte <- 0.5*sum(m1.phi.alte*
                          t(R.alte.inv))*nu2.alte
  n2.phi.alte.nu2.alte <- R.alte.inv%*%(nu2.alte*
              m1.phi.alte+nu2.alte*
        R.alte.phi.alte%*%R.alte.inv)%*%R.alte.inv

  ind.beta.alte <- 1:p.D.alte
  ind.beta.gold <- (p.D.alte+1):(p.D.alte+p.D.gold)
  ind.alpha <- p.D.alte+p.D.gold+1
  ind.sigma2.alte <- p.D.alte+p.D.gold+2
  ind.phi.alte <- p.D.alte+p.D.gold+3
  ind.nu2.alte <- p.D.alte+p.D.gold+4
  ind.tau2.gold <- p.D.alte+p.D.gold+5

  H <- matrix(0,nrow=length(par),ncol=length(par))
  H[ind.beta.alte,ind.beta.alte] <- -t(D.alte)%*%
    R.alte.inv%*%D.alte/sigma2.alte
  H[ind.beta.gold,ind.beta.gold] <- -t(D.gold)%*%
    D.gold/tau2.gold


# THE HESSIAN FUNCTION  OF THE RANDOM EFFECTS
hessian.S <- function(S,ef) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  f.p.alte <- f(p.alte)
  p.gold <- exp(S.gold+alpha*f.p.alte)/
    (1+exp(S.gold+alpha*f.p.alte))

  q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
  q.f.gold <- sum(diff.S.gold^2)
  grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
    (diff.S.alte)/sigma2.alte
  grad.beta.gold <-  t(D.gold)%*%(diff.S.gold)/tau2.gold

  grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

  grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+
            0.5*q.f.alte/(sigma2.alte^2))*sigma2.alte
  grad.log.tau2.gold <- (-n/(2*tau2.gold)+0.5*q.f.gold/
                (tau2.gold^2))*tau2.gold

  grad.log.phi.alte <- (t1.phi.alte+0.5*
            as.numeric(t(diff.S.alte)%*%
```

```r
            m2.phi.alte%*%(diff.S.alte))/sigma2.alte)

    grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
          as.numeric(t(diff.S.alte)%*%
        m2.nu2.alte%*%(diff.S.alte))/sigma2.alte)

    g <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
          grad.log.sigma2.alte,
        grad.log.phi.alte,grad.log.nu2.alte,
          grad.log.tau2.gold)

    H[ind.alpha,ind.alpha] <- -sum((f.p.alte^2)*
        units.m*p.gold/(1+exp(S.gold+alpha*f.p.alte)))

    H[ind.beta.alte,ind.sigma2.alte] <-
      H[ind.sigma2.alte,ind.beta.alte] <- -t(D.alte)%*%
      R.alte.inv%*%(diff.S.alte)/sigma2.alte
    H[ind.beta.gold,ind.tau2.gold] <-
      H[ind.tau2.gold,ind.beta.gold] <- -t(D.gold)%*%
      (diff.S.gold)/tau2.gold

    H[ind.beta.alte,ind.phi.alte] <-
      H[ind.phi.alte,ind.beta.alte] <- -as.numeric(t(D.alte)%*%
              m2.phi.alte%*%(diff.S.alte))/sigma2.alte

    H[ind.beta.alte,ind.nu2.alte] <-
      H[ind.nu2.alte,ind.beta.alte] <- -as.numeric(t(D.alte)%*%
              m2.nu2.alte%*%(diff.S.alte))/sigma2.alte

    H[ind.sigma2.alte,ind.sigma2.alte] <- (n/(2*sigma2.alte^2)-
                    q.f.alte/(sigma2.alte^3))*sigma2.alte^2+
      grad.log.sigma2.alte
    H[ind.tau2.gold,ind.tau2.gold] <- (n/(2*tau2.gold^2)-
                  q.f.gold/(tau2.gold^3))*tau2.gold^2+
      grad.log.tau2.gold

    H[ind.sigma2.alte,ind.phi.alte] <-
 H[ind.phi.alte,ind.sigma2.alte] <- -(grad.log.phi.alte-t1.phi.alte)

    H[ind.sigma2.alte,ind.nu2.alte] <-
H[ind.nu2.alte,ind.sigma2.alte] <- -(grad.log.nu2.alte-t1.nu2.alte)

H[ind.phi.alte,ind.phi.alte] <- t2.phi.alte-0.5*t(diff.S.alte)%*%
      n2.phi.alte%*%(diff.S.alte)/sigma2.alte

H[ind.nu2.alte,ind.nu2.alte] <- t2.nu2.alte-0.5*t(diff.S.alte)%*%
        n2.nu2.alte%*%(diff.S.alte)/sigma2.alte
      H[ind.phi.alte,ind.nu2.alte] <-
        H[ind.nu2.alte,ind.phi.alte] <- (t2.phi.alte.nu2.alte-
                  0.5*t(diff.S.alte)%*%n2.phi.alte.nu2.alte%*%
                    (diff.S.alte)/sigma2.alte)


    out <- list()
    out$mat1<- ef*(g%*%t(g)+H)
    out$g <- g*ef
    out
```

```r
  }

  a <- rep(0,length(par))
  A <- matrix(0,length(par),length(par))
  for(i in 1:n.samples) {
    out.i <- hessian.S(sim[i,],exp.fact[i])
    a <- a+out.i$g
    A <- A+out.i$mat1
  }
  (A-a%*%t(a))
}


# THE ESTIMATION
estim <- nlminb(par0,
                function(x) -MC.log.lik(x),
                function(x) -grad.MC.log.lik(x),
                function(x) -hess.MC.log.lik(x),
                control=list(trace=1))

estim$gradient <- grad.MC.log.lik(estim$par)
estim$hessian <- hess.MC.log.lik(estim$par)
estim$covariance <- solve(-estim$hessian)



fit_asymm_M2 <- function(data){

  coords <- data[,c("web_x","web_y")]

  set.seed(1234)
  n <- nrow(data)
  U <- as.matrix(dist(coords))


  units.m <- data$n # number tested at locations
  y.alte <- data$y.alte # alternative test outcome
  y.gold <- data$y.gold # loa loa outcome

  # Design matrices
  D.alte <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = data)
  D.gold <- model.matrix(~ elev_below750 + elev_between750_1015 +
                           elev_above1015, data = data)

  # Number of regression parameters for the two outcomes
  p.D.alte <- ncol(D.alte)
  p.D.gold <- ncol(D.gold)


  # Initial values of the model parameters
  beta.alte <- c(-0.763, 0.588e-3, -3.412e-3, -0.059e-3)
  beta.gold <- c(-1.736, 0.126e-3, -0.039e-3, -0.612e-3)
  alpha <- 1.017
  phi.alte <- 187.388
  nu2.alte <- 0.20
  phi.gold <- 23.686
  nu2.gold <- 0.483
```

```r
sigma2.alte <- 1.617
sigma2.gold <- 0.216



Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
diag(Sigma.alte) <- diag(Sigma.alte)+sigma2.alte*nu2.alte
Sigma.alte.inv <- solve(Sigma.alte)

Sigma.gold <- sigma2.gold*exp(-U/phi.gold)
diag(Sigma.gold) <- diag(Sigma.gold)+sigma2.gold*nu2.gold
Sigma.gold.inv <- solve(Sigma.gold)

n.tot <- 2*n

M <- Matrix(0,n.tot,n.tot,sparse=TRUE)
M[1:n,1:n] <- -Sigma.alte.inv
M[(n+1):n.tot,(n+1):n.tot] <- -Sigma.gold.inv


mu.alte <- as.numeric(D.alte%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)


f <- function(x) log(x/(1-x))
grad.f <- function(x) -1/((x-1)*x)
hess.f <- function(x) (2*x-1)/((x-1)^2*x^2)


ind.alte <- 1:n
ind.gold <- (n+1):n.tot


####################################################
#### SIMULATING FROM THE CONDITIONAL
#### DISTRIBUTION OF THE RANDOM EFFECT GIVEN THE DATA
####################################################

integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%Sigma.alte.inv%*%(diff.S.alte)
  diff.S.gold <- S.gold-mu.gold
  q.f_S.gold <- t(diff.S.gold)%*%Sigma.gold.inv%*%(diff.S.gold)
  q.f_S <- -0.5*as.numeric(q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(q.f_S+llik)
}

grad.integrand <- function(S) {
```

```r
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.q.f_S <- -c(Sigma.alte.inv%*%diff.S.alte,
                  Sigma.gold.inv%*%diff.S.gold)
  der.llik <- c(y.alte-units.m*p.alte+
      aux.gold*alpha*grad.f(p.alte)*der.p.alte,
              aux.gold)
  out <- der.q.f_S+der.llik
  return(out)
}

hessian.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  der2.p.alte <- (1-exp(S.alte))*exp(S.alte)/
    ((1+exp(S.alte))^3)
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))

  aux.gold <- y.gold-units.m*p.gold
  der.aux.gold <- -units.m*exp(S.gold+alpha*f(p.alte))/
    ((1+exp(S.gold+alpha*f(p.alte)))^2)
  res <- M
  diag(res)[ind.alte] <- diag(res)[ind.alte]-
    units.m*exp(S.alte)/((1+exp(S.alte))^2)+
    aux.gold*
    (alpha*(hess.f(p.alte)*(der.p.alte^2)+
            grad.f(p.alte)*der2.p.alte))+
    der.aux.gold*((alpha*grad.f(p.alte)*der.p.alte)^2)

  diag(res)[ind.gold] <- diag(res)[ind.gold]+der.aux.gold

  diag(res[ind.gold,ind.alte]) <-
    diag(res[ind.alte,ind.gold]) <-
    der.aux.gold*(alpha*grad.f(p.alte)*der.p.alte)
  return(res)
}

S.estim <- nlminb(start=c(mu.alte, mu.gold),
                  function(x) -as.numeric(integrand(x)),
                  function(x) -as.numeric(grad.integrand(x)),
                  function(x) -as.matrix(hessian.integrand(x)),
                  control=list(trace=1)
)
```

```r
S.estim$gradient <- grad.integrand(S.estim$par)
S.estim$hessian <- hessian.integrand(S.estim$par)
S.estim$Sigma.tilde <- round(solve(-S.estim$hessian), 9)
S.estim$estimate <- S.estim$par
Sigma.sroot <- t(chol(S.estim$Sigma.tilde))
A <- solve(Sigma.sroot)

library(Matrix)

Sigma.W.inv <- solve(A%*%bdiag(Sigma.alte,Sigma.gold)%*%t(A))
mu.W <- as.numeric(A%*%(c(mu.alte,mu.gold)-S.estim$estimate))

n.sim <- 12000 # number of simulations
burnin <- 2000 # burnin
thin <- 5 # thinning
n.samples <- (n.sim-burnin)/thin
h <- 1.65/((n.tot)^(1/6))
c1.h <- 0.01
c2.h <- 1e-04

# reparameterization of the model
cond.dens.W <- function(W,S) {
  diff.W <- W-mu.W
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(-0.5*as.numeric(t(diff.W)%*%
                  Sigma.W.inv%*%diff.W))+llik
}



lang.grad <- function(W,S) {
  diff.W <- W-mu.W
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*
            f(p.alte))/(1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.llik <- c(y.alte-units.m*p.alte+
          aux.gold*alpha*grad.f(p.alte)*der.p.alte,
              aux.gold)
  out <- as.numeric(-Sigma.W.inv%*%
            diff.W+t(Sigma.sroot)%*%der.llik)
  return(out)
}
```

```r
S.estim$mode <- S.estim$estimate
W.curr <- rep(0,n.tot)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+S.estim$mode)
mean.curr <- as.numeric(W.curr +
              (h^2/2)*lang.grad(W.curr,S.curr))
lp.curr <- cond.dens.W(W.curr,S.curr)
acc <- 0
sim <- matrix(NA,nrow=n.samples,ncol=n.tot)
h.vec <- rep(NA,n.sim)

for(i in 1:n.sim) {
  W.prop <- mean.curr+h*rnorm(n.tot)
  S.prop <-  as.numeric(Sigma.sroot%*%W.prop+S.estim$mode)
  mean.prop <- as.numeric(W.prop +
              (h^2/2)*lang.grad(W.prop,S.prop))
  lp.prop <- cond.dens.W(W.prop,S.prop)
        dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
  dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

  log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

  if(log(runif(1)) < log.prob) {
    acc <- acc+1
    W.curr <- W.prop
    S.curr <- S.prop
    lp.curr <- lp.prop
    mean.curr <- mean.prop
  }

  if( i > burnin & (i-burnin)%%thin==0) {
    sim[(i-burnin)/thin,] <- S.curr
  }

  h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
  cat("Iteration",i,"out of",n.sim,"\r")
  flush.console()
}


if(TRUE) {
  acf.plot <- acf(sim[,1],plot=FALSE)
  plot(acf.plot$lag,acf.plot$acf,type="l",
    xlab="lag",ylab="autocorrelation",
        ylim=c(-0.1,1),
  main="Autocorrelogram of the simulated samples")
  for(i in 2:ncol(sim)) {
    acf.plot <- acf(sim[,i],plot=FALSE)
    lines(acf.plot$lag,acf.plot$acf)
  }
  abline(h=0,lty="dashed",col=2)
}


#####################################
```

```r
#### APROXIMATING AND MAXIMIZING THE
#### LIKELIHOOD USING THE MCMC SAMPLES
#####################################
# Objective function
log.integrand <- function(S,val) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-val$mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%val$R.alte.inv%*%
    (diff.S.alte)/val$sigma2.alte
  diff.S.gold <- S.gold-val$mu.gold
  q.f_S.gold <- t(diff.S.gold)%*%val$R.gold.inv%*%
    (diff.S.gold)/val$sigma2.gold
  q.f_S <- -0.5*as.numeric(val$l.det.R.alte+val$l.det.R.gold+
    n*log(val$sigma2.alte)+n*
      log(val$sigma2.gold)+q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*(S.gold+val$alpha*f(p.alte))-
          units.m*log(1+exp(S.gold+val$alpha*f(p.alte))))

  as.numeric(q.f_S+llik)
}


compute.log.f <- function(par,l.det.R.alte=NA,R.alte.inv=NA,
                          l.det.R.gold=NA,R.gold.inv=NA) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]

  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])
  val <- list()
  val$mu.alte <- as.numeric(D.alte%*%beta.alte)
  val$mu.gold <- as.numeric(D.gold%*%beta.gold)
  val$sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  val$sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])
  val$alpha <- par[p.D.alte+p.D.gold+1]
  if(is.na(l.det.R.alte) & is.na(as.numeric(R.alte.inv)[1])) {
    R.alte <- exp(-U/phi.alte)
    diag(R.alte) <- diag(R.alte) + nu2.alte
    R.gold <- exp(-U/phi.gold)
    diag(R.gold) <- diag(R.gold) + nu2.gold
    val$l.det.R.alte <- determinant(R.alte)$modulus
    val$R.alte.inv <- solve(R.alte)
    val$l.det.R.gold <- determinant(R.gold)$modulus
    val$R.gold.inv <- solve(R.gold)
  } else {
    val$l.det.R.alte <- l.det.R.alte
    val$R.alte.inv <- R.alte.inv
    val$l.det.R.gold <- l.det.R.gold
    val$R.gold.inv <- R.gold.inv
```

```r
  }
  sapply(1:n.samples,function(i) log.integrand(sim[i,],val))
}

# theta0
par0 <- c(beta.alte, beta.gold, alpha,
    log(c(sigma2.alte,phi.alte,nu2.alte,
          sigma2.gold,phi.gold,nu2.gold)))
rm(beta.alte,beta.gold,alpha,sigma2.alte,
   phi.alte,nu2.alte,sigma2.gold,phi.gold,nu2.gold)

log.f.tilde <- compute.log.f(par0)

MC.log.lik <- function(par) {
  log(mean(exp(compute.log.f(par)-log.f.tilde)))
}



grad.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
  sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])


  R.alte <- exp(-U/phi.alte)
  diag(R.alte) <- diag(R.alte)+nu2.alte
  R.alte.inv <- solve(R.alte)
  l.det.R.alte <- determinant(R.alte)$modulus

  R.gold <- exp(-U/phi.gold)
  diag(R.gold) <- diag(R.gold)+nu2.gold
  R.gold.inv <- solve(R.gold)
  l.det.R.gold <- determinant(R.gold)$modulus

  exp.fact <- exp(compute.log.f(par,l.det.R.alte,
    R.alte.inv,l.det.R.gold,R.gold.inv)-log.f.tilde)
  L.m <- sum(exp.fact)
  exp.fact <- exp.fact/L.m


  R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
  m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
  t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
  m2.phi.alte <- m1.phi.alte%*%R.alte.inv

  t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
```

```r
m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

R.alte.phi.gold <- (U*exp(-U/phi.gold))/phi.gold
m1.phi.gold <- R.gold.inv%*%R.alte.phi.gold
t1.phi.gold <- -0.5*sum(diag(m1.phi.gold))
m2.phi.gold <- m1.phi.gold%*%R.gold.inv

t1.nu2.gold <- -0.5*sum(diag(R.gold.inv))*nu2.gold
m2.nu2.gold <- R.gold.inv%*%R.gold.inv*nu2.gold

gradient.S <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  f.p.alte <- f(p.alte)
  p.gold <- exp(S.gold+alpha*f.p.alte)/
    (1+exp(S.gold+alpha*f.p.alte))

  q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
  q.f.gold <- t(diff.S.gold)%*%R.gold.inv%*%diff.S.gold
  grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
    (diff.S.alte)/sigma2.alte
  grad.beta.gold <-  t(D.gold)%*%R.gold.inv%*%
    (diff.S.gold)/sigma2.gold

  grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

  grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+0.5*
              q.f.alte/(sigma2.alte^2))*sigma2.alte
  grad.log.sigma2.gold <- (-n/(2*sigma2.gold)+0.5*
              q.f.gold/(sigma2.gold^2))*sigma2.gold

  grad.log.phi.alte <- (t1.phi.alte+0.5*
                    as.numeric(t(diff.S.alte)%*%
            m2.phi.alte%*%(diff.S.alte))/sigma2.alte)
  grad.log.phi.gold <- (t1.phi.gold+0.5*
                      as.numeric(t(diff.S.gold)%*%
          m2.phi.gold%*%(diff.S.gold))/sigma2.gold)

  grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
                  as.numeric(t(diff.S.alte)%*%
            m2.nu2.alte%*%(diff.S.alte))/sigma2.alte)
  grad.log.nu2.gold <-  (t1.nu2.gold+0.5*
                as.numeric(t(diff.S.gold)%*%
            m2.nu2.gold%*%(diff.S.gold))/sigma2.gold)

  out <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
  grad.log.sigma2.alte,grad.log.phi.alte,grad.log.nu2.alte,
  grad.log.sigma2.gold,grad.log.phi.gold,grad.log.nu2.gold)

  return(out)
}
out <- rep(0,length(par))
for(i in 1:n.samples) {
```

```r
      out <- out + exp.fact[i]*gradient.S(sim[i,])
  }
  out
}




hess.MC.log.lik <- function(par) {
  beta.alte <- par[1:p.D.alte]
  beta.gold <- par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(par[p.D.alte+p.D.gold+4])

  phi.gold <- exp(par[p.D.alte+p.D.gold+6])
  nu2.gold <- exp(par[p.D.alte+p.D.gold+7])

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
  sigma2.alte <- exp(par[p.D.alte+p.D.gold+2])
  sigma2.gold <- exp(par[p.D.alte+p.D.gold+5])


  R.alte <- exp(-U/phi.alte)
  diag(R.alte) <- diag(R.alte)+nu2.alte
  R.alte.inv <- solve(R.alte)
  l.det.R.alte <- determinant(R.alte)$modulus

  R.gold <- exp(-U/phi.gold)
  diag(R.gold) <- diag(R.gold)+nu2.gold
  R.gold.inv <- solve(R.gold)
  l.det.R.gold <- determinant(R.gold)$modulus

  exp.fact <- exp(compute.log.f(par,l.det.R.alte,
    R.alte.inv,l.det.R.gold,R.gold.inv)-log.f.tilde)
  L.m <- sum(exp.fact)
  exp.fact <- exp.fact/L.m


  R.alte.phi.alte <- (U*exp(-U/phi.alte))/phi.alte
  m1.phi.alte <- R.alte.inv%*%R.alte.phi.alte
  t1.phi.alte <- -0.5*sum(diag(m1.phi.alte))
  m2.phi.alte <- m1.phi.alte%*%R.alte.inv

  t1.nu2.alte <- -0.5*sum(diag(R.alte.inv))*nu2.alte
  m2.nu2.alte <- R.alte.inv%*%R.alte.inv*nu2.alte

  R.alte.phi.gold <- (U*exp(-U/phi.gold))/phi.gold
  m1.phi.gold <- R.gold.inv%*%R.alte.phi.gold
  t1.phi.gold <- -0.5*sum(diag(m1.phi.gold))
  m2.phi.gold <- m1.phi.gold%*%R.gold.inv

  t1.nu2.gold <- -0.5*sum(diag(R.gold.inv))*nu2.gold
  m2.nu2.gold <- R.gold.inv%*%R.gold.inv*nu2.gold
```

```r
R.gold.phi.alte <- R.alte.phi.alte+(U*(U-2*phi.alte)*
                     exp(-U/phi.alte))/phi.alte^2
t2.phi.alte <- -0.5*(sum(R.alte.inv*R.gold.phi.alte)-
         sum(m1.phi.alte*t(m1.phi.alte)))
n2.phi.alte <- R.alte.inv%*%(2*R.alte.phi.alte%*%
         m1.phi.alte-R.gold.phi.alte)%*%R.alte.inv

R.gold.phi.gold <- R.alte.phi.gold+(U*(U-2*phi.gold)*
                     exp(-U/phi.gold))/phi.gold^2
t2.phi.gold <- -0.5*(sum(R.gold.inv*R.gold.phi.gold)-
        sum(m1.phi.gold*t(m1.phi.gold)))
n2.phi.gold <- R.gold.inv%*%(2*R.alte.phi.gold%*%
         m1.phi.gold-R.gold.phi.gold)%*%R.gold.inv

t2.nu2.alte <- -0.5*(sum(diag(R.alte.inv)*nu2.alte)-
         sum(diag(m2.nu2.alte)*nu2.alte))
m1.nu2.alte <- R.alte.inv*nu2.alte
nu2.alte.aux <- 2*m1.nu2.alte*nu2.alte
diag(nu2.alte.aux) <- diag(nu2.alte.aux)-nu2.alte
n2.nu2.alte <- R.alte.inv%*%nu2.alte.aux%*%R.alte.inv

t2.phi.alte.nu2.alte <- 0.5*sum(m1.phi.alte*
                     t(R.alte.inv))*nu2.alte
n2.phi.alte.nu2.alte <- R.alte.inv%*%(nu2.alte*
          m1.phi.alte+nu2.alte*R.alte.phi.alte%*%
          R.alte.inv)%*%R.alte.inv

t2.nu2.gold <- -0.5*(sum(diag(R.gold.inv)*
         nu2.gold)-sum(diag(m2.nu2.gold)*nu2.gold))
m1.nu2.gold <- R.gold.inv*nu2.gold
nu2.gold.aux <- 2*m1.nu2.gold*nu2.gold
diag(nu2.gold.aux) <- diag(nu2.gold.aux)-nu2.gold
n2.nu2.gold <- R.gold.inv%*%nu2.gold.aux%*%R.gold.inv

t2.phi.gold.nu2.gold <- 0.5*sum(m1.phi.gold*
                     t(R.gold.inv))*nu2.gold
n2.phi.gold.nu2.gold <- R.gold.inv%*%(nu2.gold*
    m1.phi.gold+nu2.gold*R.alte.phi.gold%*%
    R.gold.inv)%*%R.gold.inv

ind.beta.alte <- 1:p.D.alte
ind.beta.gold <- (p.D.alte+1):(p.D.alte+p.D.gold)
ind.alpha <- p.D.alte+p.D.gold+1
ind.sigma2.alte <- p.D.alte+p.D.gold+2
ind.phi.alte <- p.D.alte+p.D.gold+3
ind.nu2.alte <- p.D.alte+p.D.gold+4
ind.sigma2.gold <- p.D.alte+p.D.gold+5
ind.phi.gold <- p.D.alte+p.D.gold+6
ind.nu2.gold <- p.D.alte+p.D.gold+7

H <- matrix(0,nrow=length(par),ncol=length(par))
H[ind.beta.alte,ind.beta.alte] <- -t(D.alte)%*%
   R.alte.inv%*%D.alte/sigma2.alte
H[ind.beta.gold,ind.beta.gold] <- -t(D.gold)%*%
   R.gold.inv%*%D.gold/sigma2.gold

hessian.S <- function(S,ef) {
```

```r
S.alte <- S[ind.alte]
S.gold <- S[ind.gold]

diff.S.alte <- S.alte-mu.alte
diff.S.gold <- S.gold-mu.gold
p.alte <- exp(S.alte)/(1+exp(S.alte))
f.p.alte <- f(p.alte)
p.gold <- exp(S.gold+alpha*f.p.alte)/
  (1+exp(S.gold+alpha*f.p.alte))

q.f.alte <- t(diff.S.alte)%*%R.alte.inv%*%diff.S.alte
q.f.gold <- t(diff.S.gold)%*%R.gold.inv%*%diff.S.gold
grad.beta.alte <-  t(D.alte)%*%R.alte.inv%*%
  (diff.S.alte)/sigma2.alte
grad.beta.gold <-  t(D.gold)%*%R.gold.inv%*%
  (diff.S.gold)/sigma2.gold

grad.alpha <- sum(f.p.alte*(y.gold-units.m*p.gold))

grad.log.sigma2.alte <- (-n/(2*sigma2.alte)+0.5*
            q.f.alte/(sigma2.alte^2))*sigma2.alte
grad.log.sigma2.gold <- (-n/(2*sigma2.gold)+0.5*
      q.f.gold/(sigma2.gold^2))*sigma2.gold

grad.log.phi.alte <- (t1.phi.alte+0.5*
      as.numeric(t(diff.S.alte)%*%
    m2.phi.alte%*%(diff.S.alte))/sigma2.alte)
grad.log.phi.gold <- (t1.phi.gold+0.5*
          as.numeric(t(diff.S.gold)%*%
    m2.phi.gold%*%(diff.S.gold))/sigma2.gold)

grad.log.nu2.alte <-  (t1.nu2.alte+0.5*
    as.numeric(t(diff.S.alte)%*%m2.nu2.alte%*%
    (diff.S.alte))/sigma2.alte)
grad.log.nu2.gold <-  (t1.nu2.gold+0.5*
  as.numeric(t(diff.S.gold)%*%m2.nu2.gold%*%
  (diff.S.gold))/sigma2.gold)

g <- c(grad.beta.alte,grad.beta.gold,grad.alpha,
    grad.log.sigma2.alte,grad.log.phi.alte,
    grad.log.nu2.alte, grad.log.sigma2.gold,
    grad.log.phi.gold,grad.log.nu2.gold)

H[ind.alpha,ind.alpha] <- -sum((f.p.alte^2)*
    units.m*p.gold/(1+exp(S.gold+alpha*f.p.alte)))

H[ind.beta.alte,ind.sigma2.alte] <-
  H[ind.sigma2.alte,ind.beta.alte] <- -t(D.alte)%*%
  R.alte.inv%*%(diff.S.alte)/sigma2.alte
H[ind.beta.gold,ind.sigma2.gold] <-
  H[ind.sigma2.gold,ind.beta.gold] <- -t(D.gold)%*%
  R.gold.inv%*%(diff.S.gold)/sigma2.gold

H[ind.beta.alte,ind.phi.alte] <-
H[ind.phi.alte,ind.beta.alte] <- -as.numeric(t(D.alte)%*%
    m2.phi.alte%*%(diff.S.alte))/sigma2.alte
H[ind.beta.gold,ind.phi.gold] <-
```

```r
H[ind.phi.gold,ind.beta.gold] <- -as.numeric(t(D.gold)%*%
    m2.phi.gold%*%(diff.S.gold))/sigma2.gold

 H[ind.beta.alte,ind.nu2.alte] <-
   H[ind.nu2.alte,ind.beta.alte] <-
   -as.numeric(t(D.alte)%*%m2.nu2.alte%*%
                 (diff.S.alte))/sigma2.alte
 H[ind.beta.gold,ind.nu2.gold] <-
   H[ind.nu2.gold,ind.beta.gold] <-
   -as.numeric(t(D.gold)%*%m2.nu2.gold%*%
                 (diff.S.gold))/sigma2.gold

 H[ind.sigma2.alte,ind.sigma2.alte] <-
   (n/(2*sigma2.alte^2)-q.f.alte/
      (sigma2.alte^3))*sigma2.alte^2+
   grad.log.sigma2.alte
 H[ind.sigma2.gold,ind.sigma2.gold] <-
   (n/(2*sigma2.gold^2)-q.f.gold/
      (sigma2.gold^3))*sigma2.gold^2+
   grad.log.sigma2.gold

 H[ind.sigma2.alte,ind.phi.alte] <-
   H[ind.phi.alte,ind.sigma2.alte] <-
   -(grad.log.phi.alte-t1.phi.alte)
 H[ind.sigma2.gold,ind.phi.gold] <-
   H[ind.phi.gold,ind.sigma2.gold] <-
   -(grad.log.phi.gold-t1.phi.gold)

 H[ind.sigma2.alte,ind.nu2.alte] <-
   H[ind.nu2.alte,ind.sigma2.alte] <-
   -(grad.log.nu2.alte-t1.nu2.alte)
 H[ind.sigma2.gold,ind.nu2.gold] <-
   H[ind.nu2.gold,ind.sigma2.gold] <-
   -(grad.log.nu2.gold-t1.nu2.gold)

 H[ind.phi.alte,ind.phi.alte] <- t2.phi.alte-
   0.5*t(diff.S.alte)%*%n2.phi.alte%*%
   (diff.S.alte)/sigma2.alte
 H[ind.phi.gold,ind.phi.gold] <- t2.phi.gold-
   0.5*t(diff.S.gold)%*%n2.phi.gold%*%
   (diff.S.gold)/sigma2.gold

 H[ind.nu2.alte,ind.nu2.alte] <- t2.nu2.alte-
   0.5*t(diff.S.alte)%*%n2.nu2.alte%*%
   (diff.S.alte)/sigma2.alte
 H[ind.phi.alte,ind.nu2.alte] <-
   H[ind.nu2.alte,ind.phi.alte] <-
   (t2.phi.alte.nu2.alte-0.5*t(diff.S.alte)%*%
 n2.phi.alte.nu2.alte%*%(diff.S.alte)/sigma2.alte)

 H[ind.nu2.gold,ind.nu2.gold] <- t2.nu2.gold-
   0.5*t(diff.S.gold)%*%n2.nu2.gold%*%
   (diff.S.gold)/sigma2.gold
 H[ind.phi.gold,ind.nu2.gold] <-
   H[ind.nu2.gold,ind.phi.gold] <-
   (t2.phi.gold.nu2.gold-0.5*t(diff.S.gold)%*%
      n2.phi.gold.nu2.gold%*%(diff.S.gold)/sigma2.gold)
```

```r
      out <- list()
      out$mat1<- ef*(g%*%t(g)+H)
      out$g <- g*ef
      out
    }

    a <- rep(0,length(par))
    A <- matrix(0,length(par),length(par))
    for(i in 1:n.samples) {
      out.i <- hessian.S(sim[i,],exp.fact[i])
      a <- a+out.i$g
      A <- A+out.i$mat1
    }
    (A-a%*%t(a))
  }

  estim <- nlminb(par0,
                  function(x) -MC.log.lik(x),
                  function(x) -grad.MC.log.lik(x),
                  function(x) -hess.MC.log.lik(x),
                  control=list(trace=1))

  estim$gradient <- grad.MC.log.lik(estim$par)
  estim$hessian <- hess.MC.log.lik(estim$par)
  estim$covariance <- solve(-estim$hessian)
  names(estim$par) <- c("beta0.alte", "beta1.alte",
                        "beta2.alte", "beta3.alte",
                        "beta0.gold", "beta1.gold",
                        "beta2.gold", "beta3.gold",
                        "alpha", "log.sigma.alte",
                        "log.phi.alte", "log.nu.alte",
                        "log.nu.gold")
      return(estim)
  }

  return(estim)
}



predictions.M2 <- function(data, grid.pred, covariate, estim){

  n <- nrow(data)
  coords <- data[,c("web_x","web_y")]
  U <- as.matrix(dist(coords))
  units.m <- data$n
  y.alte <- data$y.alte
  y.gold <- data$y.gold
    D.alte <- model.matrix(~ elev_below750 +
        elev_between750_1015 + elev_above1015, data = data)
  D.gold <- model.matrix(~ elev_below750 +
        elev_between750_1015 + elev_above1015, data = data)

    p.D.alte <- ncol(D.alte)
```

```r
p.D.gold <- ncol(D.gold)

# I'm defining theta0
beta.alte    <- estim$par[1:p.D.alte]
beta.gold     <- estim$par[(p.D.alte+1):(p.D.alte+p.D.gold)]
alpha        <- estim$par[p.D.alte+p.D.gold+1]
sigma2.alte <- exp(estim$par[p.D.alte+p.D.gold+2])
phi.alte     <- exp(estim$par[p.D.alte+p.D.gold+3])
nu2.alte     <- exp(estim$par[p.D.alte+p.D.gold+4])
sigma2.gold  <- exp(estim$par[p.D.alte+p.D.gold+5])
phi.gold      <- exp(estim$par[p.D.alte+p.D.gold+6])
nu2.gold       <- exp(estim$par[p.D.alte+p.D.gold+7])


Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
diag(Sigma.alte) <- diag(Sigma.alte)+sigma2.alte*nu2.alte
Sigma.alte.inv <- solve(Sigma.alte)

Sigma.gold <- sigma2.gold*exp(-U/phi.gold)
diag(Sigma.gold) <- diag(Sigma.gold)+sigma2.gold*nu2.gold
Sigma.gold.inv <- solve(Sigma.gold)

n.tot <- 2*n # total number of random effects = 2* nlocations

M <- Matrix(0,n.tot,n.tot,sparse=TRUE)
M[1:n,1:n] <- -Sigma.alte.inv
M[(n+1):n.tot,(n+1):n.tot] <- -Sigma.gold.inv

mu.alte <- as.numeric(D.alte%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)


f <- function(x) log(x/(1-x))
grad.f <- function(x) -1/((x-1)*x)
hess.f <- function(x) (2*x-1)/((x-1)^2*x^2)

ind.alte <- 1:n
ind.gold <- (n+1):n.tot

integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  q.f_S.alte <- t(diff.S.alte)%*%Sigma.alte.inv%*%(diff.S.alte)
  diff.S.gold <- S.gold-mu.gold
  q.f_S.gold <- t(diff.S.gold)%*%Sigma.gold.inv%*%(diff.S.gold)
  q.f_S <- -0.5*as.numeric(q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(q.f_S+llik)
}
```

```r
grad.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.q.f_S <- -c(Sigma.alte.inv%*%diff.S.alte,
                  Sigma.gold.inv%*%diff.S.gold)
  der.llik <- c(y.alte-units.m*p.alte+
   aux.gold*alpha*grad.f(p.alte)*der.p.alte,
             aux.gold)
  out <- der.q.f_S+der.llik
  return(out)
}

hessian.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  der2.p.alte <- (1-exp(S.alte))*
    exp(S.alte)/((1+exp(S.alte))^3)
  p.gold <-  exp(S.gold+alpha*
    f(p.alte))/(1+exp(S.gold+alpha*f(p.alte)))

  aux.gold <- y.gold-units.m*p.gold
  der.aux.gold <- -units.m*exp(S.gold+
alpha*f(p.alte))/((1+exp(S.gold+alpha*f(p.alte)))^2)
  res <- M
  diag(res)[ind.alte] <- diag(res)[ind.alte]-
  units.m*exp(S.alte)/((1+exp(S.alte))^2)+
    aux.gold*
    (alpha*(hess.f(p.alte)*(der.p.alte^2)+
             grad.f(p.alte)*der2.p.alte))+
    der.aux.gold*((alpha*grad.f(p.alte)*der.p.alte)^2)
diag(res)[ind.gold] <- diag(res)[ind.gold]+der.aux.gold

  diag(res[ind.gold,ind.alte]) <-
    diag(res[ind.alte,ind.gold]) <-
    der.aux.gold*(alpha*grad.f(p.alte)*der.p.alte)
  return(res)
}

S.estim <-nlminb(start=c(mu.alte, mu.gold),
       function(x) -as.numeric(integrand(x)),
       function(x) -as.numeric(grad.integrand(x)),
       function(x) -as.matrix(hessian.integrand(x)),
     control=list(trace=0)
)
```

```r
S.estim$gradient <- grad.integrand(S.estim$par)
S.estim$hessian <- hessian.integrand(S.estim$par)
S.estim$Sigma.tilde <- round(solve(-S.estim$hessian) , 9)
S.estim$estimate <- S.estim$par
Sigma.sroot <- t(chol(S.estim$Sigma.tilde))
A <- solve(Sigma.sroot)


Sigma.W.inv <- solve(A%*%bdiag(Sigma.alte,Sigma.gold)%*%t(A))
mu.W <- as.numeric(A%*%(c(mu.alte,mu.gold)-S.estim$estimate))

n.sim <- 1400 # number of simulations
burnin <- 400 # burnin
thin <- 10 # thinning
n.samples <- (n.sim-burnin)/thin
h <- 1.65/((n.tot)^(1/6))
c1.h <- 0.01
c2.h <- 1e-04

# reparameterization of the model
cond.dens.W <- function(W,S) {
  diff.W <- W-mu.W
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))
  as.numeric(-0.5*as.numeric(t(diff.W)%*%
                      Sigma.W.inv%*%diff.W))+llik
}

lang.grad <- function(W,S) {
  diff.W <- W-mu.W

  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.llik <- c(y.alte-units.m*p.alte+
          aux.gold*alpha*grad.f(p.alte)*der.p.alte,
              aux.gold)
  out <- as.numeric(-Sigma.W.inv%*%
                      diff.W+t(Sigma.sroot)%*%der.llik)
  return(out)
}


S.estim$mode <- S.estim$estimate
W.curr <- rep(0,n.tot)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+S.estim$mode)
mean.curr <- as.numeric(W.curr +
```

```r
                       (h^2/2)*lang.grad(W.curr,S.curr))
  lp.curr <- cond.dens.W(W.curr,S.curr)

  acc <- 0
  sim <- matrix(NA,nrow=n.samples,ncol=n.tot)
  h.vec <- rep(NA,n.sim)

  for(i in 1:n.sim) {
    W.prop <- mean.curr+h*rnorm(n.tot)
    S.prop <-  as.numeric(Sigma.sroot%*%W.prop+S.estim$mode)
    mean.prop <- as.numeric(W.prop +
                    (h^2/2)*lang.grad(W.prop,S.prop))
    lp.prop <- cond.dens.W(W.prop,S.prop)

    dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
    dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

    log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

    if(log(runif(1)) < log.prob) {
      acc <- acc+1
      W.curr <- W.prop
      S.curr <- S.prop
      lp.curr <- lp.prop
      mean.curr <- mean.prop
    }

    if( i > burnin & (i-burnin)%%thin==0) {
      sim[(i-burnin)/thin,] <- S.curr
    }

    h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
cat("Simulating from cond distn of the random effects: Iteration",
        i," out of ",n.sim, "\r")
    flush.console()
  }


  ##################
  #################

  # Design matrices at at prediction locations
  D.pred.alte <- model.matrix(~ elev_below750 +
      elev_between750_1015 + elev_above1015, data = covariate)
  D.pred.gold <- model.matrix(~ elev_below750 +
        elev_between750_1015 + elev_above1015, data = covariate)
    # covariates effects at prediction locations
  mu.pred.alte <- as.numeric(D.pred.alte%*%beta.alte)
  mu.pred.gold <- as.numeric(D.pred.gold%*%beta.gold)
    # covariates effects at observed locations
  mu.alte <- as.numeric(D.gold%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)
    n.samples <- nrow(sim)
  n.pred <- nrow(grid.pred)
  # Alternative diagnostic predictions
  U.pred <- as.matrix(pdist(grid.pred,coords))
  C.alte <- sigma2.alte*exp(-U.pred/phi.alte)
```

```r
A.alte <- C.alte%*%Sigma.alte.inv

# Simulations for linear predictor of alternative diagnostic
mu.cond.alte <- mu.pred.alte+sapply(1:n.samples,
          function(i) A.alte%*%(sim[i,1:n]-mu.alte))
sd.cond.alte <- sqrt(sigma2.alte-apply(A.alte*C.alte,1,sum))
  eta.alte.samples <- sapply(1:n.samples,
  function(i) mu.cond.alte[,i]+sd.cond.alte*rnorm(n.pred))
prev.alte.samples <- exp(eta.alte.samples)/
  (1+exp(eta.alte.samples))
prev.alte.mean <- apply(prev.alte.samples,1,mean)

prev.mean.alte <- apply(prev.alte.samples,1,mean) #
prev.sd.alte <- apply(prev.alte.samples,1,sd)
CI.lower.alte <- apply(prev.alte.samples,1,
        function(x) quantile(x,0.025))
CI.upper.alte <- apply(prev.alte.samples,1,
        function(x) quantile(x,0.975))
exceed.20p.prob.alte <- sapply(1:dim(prev.alte.samples)
    [1],function(i) mean(prev.alte.samples[i,] > 0.2) )


# Gold standard predictions
C.gold <- sigma2.gold*exp(-U.pred/phi.gold)
A.gold <- C.gold%*%Sigma.gold.inv

# Simulations for linear predictor of the gold standard
mu.cond.gold <- mu.pred.gold+sapply(1:n.samples,
      function(i) A.gold%*%(sim[i,(n+1):(2*n)]-mu.gold))
sd.cond.gold <- sqrt(sigma2.gold-apply(A.gold*C.gold,1,sum))

eta.gold.samples <- sapply(1:n.samples,
function(i) mu.cond.gold[,i]+sd.cond.gold*rnorm(n.pred)+
              alpha*eta.alte.samples[,i])
prev.gold.samples <- exp(eta.gold.samples)/
  (1+exp(eta.gold.samples))
  prev.mean.gold <- apply(prev.gold.samples,1,mean) #
prev.sd.gold <- apply(prev.gold.samples,1,sd)
CI.lower.gold <- apply(prev.gold.samples,1,
            function(x) quantile(x,0.025))
CI.upper.gold <- apply(prev.gold.samples,1,
            function(x) quantile(x,0.975))
exceed.20p.prob.gold <- sapply(1:dim(prev.gold.samples)[1],
      function(i) mean(prev.gold.samples[i,] > 0.2) )



df <- data.frame(grid.pred, prev.mean.gold, prev.sd.gold,
        CI.lower.gold, CI.upper.gold, exceed.20p.prob.gold,
        prev.mean.alte, prev.sd.alte, CI.lower.alte,
        CI.upper.alte, exceed.20p.prob.alte)

return(df)
}
```

```r
pred.asymm.M1 <- function(data, grid.pred, covariate, estim){
  coords <- data[,c("web_x","web_y")]
  U <- as.matrix(dist(coords))
  n <- nrow(data)
  units.m <- data$n
  y.alte <- data$y.alte
  y.gold <- data$y.gold
    D.alte <- model.matrix(~ elev_below750 +
      elev_between750_1015 + elev_above1015, data = data)
  D.gold <- model.matrix(~ elev_below750 +
    elev_between750_1015 + elev_above1015, data = data)


  p.D.alte <- ncol(D.alte)
  p.D.gold <- ncol(D.gold)
    # I'm defining theta0
  beta.alte <- estim$par[1:p.D.alte]
  beta.gold <- estim$par[(p.D.alte+1):(p.D.alte+p.D.gold)]
  alpha <- estim$par[p.D.alte+p.D.gold+1]
  phi.alte <- exp(estim$par[p.D.alte+p.D.gold+3])
  nu2.alte <- exp(estim$par[p.D.alte+p.D.gold+4])

  sigma2.alte <- exp(estim$par[p.D.alte+p.D.gold+2])
  sigma2.gold <- exp(estim$par[p.D.alte+p.D.gold+5])

  Sigma.alte <- sigma2.alte*exp(-U/phi.alte)
  diag(Sigma.alte) <- diag(Sigma.alte)+sigma2.alte*nu2.alte
  Sigma.alte.inv <- solve(Sigma.alte)

  Sigma.gold <- diag(sigma2.gold,n)
  Sigma.gold.inv <- solve(Sigma.gold)

  n.tot <- 2*n

  M <- Matrix(0,n.tot,n.tot,sparse=TRUE)
  M[1:n,1:n] <- -Sigma.alte.inv
  M[(n+1):n.tot,(n+1):n.tot] <- -Sigma.gold.inv

  mu.alte <- as.numeric(D.alte%*%beta.alte)
  mu.gold <- as.numeric(D.gold%*%beta.gold)


  f <- function(x) log(x/(1-x))
  grad.f <- function(x) -1/((x-1)*x)
  hess.f <- function(x) (2*x-1)/((x-1)^2*x^2)

  ind.alte <- 1:n
  ind.gold <- (n+1):n.tot

  integrand <- function(S) {
    S.alte <- S[ind.alte]
    S.gold <- S[ind.gold]

    diff.S.alte <- S.alte-mu.alte
    q.f_S.alte <- t(diff.S.alte)%*%Sigma.alte.inv%*%(diff.S.alte)
    diff.S.gold <- S.gold-mu.gold
```

```r
  q.f_S.gold <- t(diff.S.gold)%*%Sigma.gold.inv%*%(diff.S.gold)
  q.f_S <- -0.5*as.numeric(q.f_S.alte+q.f_S.gold)

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
  llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))

  as.numeric(q.f_S+llik)
}

grad.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  diff.S.alte <- S.alte-mu.alte
  diff.S.gold <- S.gold-mu.gold

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.q.f_S <- -c(Sigma.alte.inv%*%diff.S.alte,
                  Sigma.gold.inv%*%diff.S.gold)
  der.llik <- c(y.alte-units.m*p.alte+
      aux.gold*alpha*grad.f(p.alte)*der.p.alte,
               aux.gold)
  out <- der.q.f_S+der.llik
  return(out)
}

hessian.integrand <- function(S) {
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  der2.p.alte <- (1-exp(S.alte))*
     exp(S.alte)/((1+exp(S.alte))^3)
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))

  aux.gold <- y.gold-units.m*p.gold
  der.aux.gold <- -units.m*exp(S.gold+
                    alpha*f(p.alte))/
        ((1+exp(S.gold+alpha*f(p.alte)))^2)
  res <- M
  diag(res)[ind.alte] <- diag(res)[ind.alte]-
    units.m*exp(S.alte)/((1+exp(S.alte))^2)+
    aux.gold*
    (alpha*(hess.f(p.alte)*(der.p.alte^2)+
             grad.f(p.alte)*der2.p.alte))+
    der.aux.gold*((alpha*grad.f(p.alte)*der.p.alte)^2)

  diag(res)[ind.gold] <- diag(res)[ind.gold]+der.aux.gold
```

```r
diag(res[ind.gold,ind.alte]) <- diag(res[ind.alte,ind.gold]) <-
    der.aux.gold*(alpha*grad.f(p.alte)*der.p.alte)
  return(res)
}

S.estim <-nlminb(start=c(mu.alte,mu.gold),
    function(x) -as.numeric(integrand(x)),
    function(x) -as.numeric(grad.integrand(x)),
    function(x) -as.matrix(hessian.integrand(x)),
                control=list(trace=0)
)
S.estim$gradient <- grad.integrand(S.estim$par)
S.estim$hessian <- hessian.integrand(S.estim$par)
S.estim$Sigma.tilde <- round(solve(-S.estim$hessian), 9)
S.estim$estimate <- S.estim$par
Sigma.sroot <- t(chol(S.estim$Sigma.tilde))
A <- solve(Sigma.sroot)

Sigma.W.inv <- solve(A%*%bdiag(Sigma.alte,Sigma.gold)%*%t(A))
mu.W <- as.numeric(A%*%(c(mu.alte,mu.gold)-S.estim$estimate))
n.sim <- 1400 # number of simulations
burnin <- 400 # burnin
thin <- 10 # thinning
n.samples <- (n.sim-burnin)/thin
h <- 1.65/((n.tot)^(1/6))
c1.h <- 0.01
c2.h <- 1e-04

# reparameterization of the model
cond.dens.W <- function(W,S) {
  diff.W <- W-mu.W
  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]
  p.alte <- exp(S.alte)/(1+exp(S.alte))
  eta.gold <- S.gold+alpha*f(p.alte)
llik <- sum(y.alte*S.alte-units.m*log(1+exp(S.alte)))+
    sum(y.gold*eta.gold-units.m*log(1+exp(eta.gold)))
      as.numeric(-0.5*as.numeric(t(diff.W)%*%
                        Sigma.W.inv%*%diff.W))+llik
}

lang.grad <- function(W,S) {
  diff.W <- W-mu.W

  S.alte <- S[ind.alte]
  S.gold <- S[ind.gold]

  p.alte <- exp(S.alte)/(1+exp(S.alte))
  der.p.alte <- p.alte/(1+exp(S.alte))
  p.gold <-  exp(S.gold+alpha*f(p.alte))/
    (1+exp(S.gold+alpha*f(p.alte)))
  aux.gold <- y.gold-units.m*p.gold

  der.llik <- c(y.alte-units.m*p.alte+
          aux.gold*alpha*grad.f(p.alte)*der.p.alte,
        aux.gold)
```

```r
  out <- as.numeric(-Sigma.W.inv%*%diff.W+
                     t(Sigma.sroot)%*%der.llik)
  return(out)
}


S.estim$mode <- S.estim$estimate
W.curr <- rep(0,n.tot)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+S.estim$mode)
mean.curr <- as.numeric(W.curr +
                (h^2/2)*lang.grad(W.curr,S.curr))
lp.curr <- cond.dens.W(W.curr,S.curr)
acc <- 0
sim <- matrix(NA,nrow=n.samples,ncol=n.tot)
h.vec <- rep(NA,n.sim)


for(i in 1:n.sim) {
  W.prop <- mean.curr+h*rnorm(n.tot)
  S.prop <-  as.numeric(Sigma.sroot%*%W.prop+S.estim$mode)
  mean.prop <- as.numeric(W.prop +
                (h^2/2)*lang.grad(W.prop,S.prop))
  lp.prop <- cond.dens.W(W.prop,S.prop)
      dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
  dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))
  log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

  if(log(runif(1)) < log.prob) {
    acc <- acc+1
    W.curr <- W.prop
    S.curr <- S.prop
    lp.curr <- lp.prop
    mean.curr <- mean.prop
  }

  if( i > burnin & (i-burnin)%%thin==0) {
    sim[(i-burnin)/thin,] <- S.curr
  }

  h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
  cat("Simulating from cond distn of random effects: Iteration",
      i," out of ",n.sim, "\r")
  flush.console()
}


###########################
###########################


# Design matrices at at prediction locations
D.pred.alte <- model.matrix(~ elev_below750 +
    elev_between750_1015 + elev_above1015, data = covariate)
D.pred.gold <- model.matrix(~ elev_below750 +
   elev_between750_1015 + elev_above1015, data = covariate)
```

```r
# covariates effects at prediction locations
mu.pred.alte <- as.numeric(D.pred.alte%*%beta.alte)
mu.pred.gold <- as.numeric(D.pred.gold%*%beta.gold)


# covariates effects at observed locations
mu.alte <- as.numeric(D.gold%*%beta.alte)
mu.gold <- as.numeric(D.gold%*%beta.gold)


n.samples <- nrow(sim)
n.pred <- nrow(grid.pred)

# RAPLOA predictions
U.pred <- as.matrix(pdist(grid.pred,coords))
C.alte <- sigma2.alte*exp(-U.pred/phi.alte)
A.alte <- C.alte%*%Sigma.alte.inv

# Simulations for linear predictor for RAPLOA
mu.cond.alte <- mu.pred.alte+sapply(1:n.samples,
      function(i) A.alte%*%(sim[i,1:n]-mu.alte))
sd.cond.alte <- sqrt(sigma2.alte-apply(A.alte*C.alte,1,sum))

eta.alte.samples <- sapply(1:n.samples,
function(i) mu.cond.alte[,i]+sd.cond.alte*rnorm(n.pred))
prev.alte.samples <- exp(eta.alte.samples)/
   (1+exp(eta.alte.samples))


prev.mean.alte <- apply(prev.alte.samples,1,mean) #
prev.sd.alte <- apply(prev.alte.samples,1,sd)
CI.lower.alte <- apply(prev.alte.samples,1,
         function(x) quantile(x,0.025))
CI.upper.alte <- apply(prev.alte.samples,1,
         function(x) quantile(x,0.975))
exceed.20p.prob.alte <- sapply(1:dim(prev.alte.samples)[1],
     function(i) mean(prev.alte.samples[i,] > 0.2) )

eta.gold.samples <- sapply(1:n.samples,
               function(i) mu.pred.gold+
           alpha*eta.alte.samples[,i])
prev.gold.samples <- exp(eta.gold.samples)/
   (1+exp(eta.gold.samples))

prev.mean.gold <- apply(prev.gold.samples,1,mean) #
prev.sd.gold <- apply(prev.gold.samples,1,sd)
CI.lower.gold <- apply(prev.gold.samples,1,
             function(x) quantile(x,0.025))
CI.upper.gold <- apply(prev.gold.samples,1,
            function(x) quantile(x,0.975))
exceed.20p.prob.gold <- sapply(1:dim(prev.gold.samples)[1],
function(i) mean(prev.gold.samples[i,] > 0.2) )

df <- data.frame(grid.pred, prev.mean.gold, prev.sd.gold,
               CI.lower.gold, CI.upper.gold,
               exceed.20p.prob.gold,
               prev.mean.alte, prev.sd.alte,
```

```
                         CI.lower.alte, CI.upper.alte,
                         exceed.20p.prob.alte)



  return(df)
}
```

File *Loaloa_Examples.R*

```r
#  rm(list=ls())
# It is important to keep all the files in a single folder and set
# this folder as the working directory. Edit the next line to set
# your working directory.

workingdir <- "~/../Examples_Code_and_Data/loaloa_application"
setwd(workingdir)
library(rgdal)
library(geoR)
library(PrevMap)
source("loa_functions.R")

#####################################################################
######################################################################
# Simulating data under the fitted asymmetric Model 2 in the paper,
# using jittered locations of the original Loa loa study
#####################################################################
###################################################################

# An example data that mimics the origional loa loa data
# has been provided as LoaloaExampleData.RData. However,
# you can simulate your own data under the fitted asymmetric
# Model 2 in the paper. The next two lines simulate such data.
input.data <- get(load("input_data.RData"))
data <- sim_from_asymm_M2(input.data)
head(data)
plot(data$prev.alte, data$prev.gold)
# save(data, file="LoaloaExampleData.RData")



###################################################################
################################################################
# ANALYSES USING THE MODEL 1 SPECIFICATION OF ASYMMETRIC MODEL
###################################################################
################################################################

# Loading the example loa loa prevalence data
data <- get(load("LoaloaExampleData.RData"))

# Estimating the Model 1 specification of the asymmetric model
# You could estimate the parameters of the model or
# used alread estimated
# and saved parameters for your predictions.
```

```r
estim <- fit_asymm_M1(data)
# save(estim, file="MCML_estimates_asymm_M1.RData")
# You can change the number of MCMC samples, thining,
# and burnin in the loa_functions.R file.


# Getting the predictions grid
grid.pred <-
 get(load("prediction_grid_covariates.RData"))[, c("web_x", "web_y")]

# Getting the values of elevation at the prediction locations
elev <- prediction_grid_covariates$elev

# Comuping the spline variables for elevation at the
# prediction location.
covariate <- get_elev_vars(elev, 750, 1015)

# Loading the already estimated parameters the asymetric model
# You can skip this line if you wish to use parameters you have
# estimated using the function fit_asymm_M1()
estim <- get(load("MCML_estimates_asymm_M1.RData"))
# You can change the number of MCMC samples, thining,
# and burnin in the loa_functions.R file.

# Predicting with the M1 specification of the assymetric model
# You can change the number of MCMC samples, thining,
# and burnin in the loa_functions.R file.
predictions.asymm.M1 <- pred.asymm.M1(data, grid.pred,
                                   covariate, estim)

# web_x , web_y   : Cordinates of the prediction locations
# that form the grid
# prev.mean.gold  : Mean prevalence as defined by the gold
# standard diagnostic
# prev.sd.gold    : Standard deviation based on the gold
# standard diagnostic
# CI.lower.gold   : lower end of the 95% predictive interval
# based on the gold standard diagnostic
# CI.upper.gold   : upper end of the 95% predictive interval
# based on the gold standard diagnostic
# exceed.20p.prob.gold  : predictive probability of prevalence
# based on the gold standard diagnostic exceeding 20%
# prev.mean.alte  : Mean prevalence as defined by the
# alternate standard diagnostic
# prev.sd.alte    : Standard deviation based on the
# alternate diagnostic
# CI.lower.alte   : lower end of the 95% predictive interval
# based on the alternate diagnostic
# CI.upper.alte   : upper end of the 95% predictive interval
# based on the alternate diagnostic
# exceed.20p.prob.alte : predictive probability of prevalence
# based on the alternate diagnostic exceeding 20%

# ploting the mean prevalence as defined by the
# gold standard diagnostic
plot(rasterFromXYZ(predictions.asymm.M1
            [, c("web_x", "web_y", "prev.mean.gold")]))
```

```r
####################################################################
###################################################################
# ANALYSES USING THE MODEL 2 SPECIFICATION OF ASYMMETRIC MODEL
###################################################################
###################################################################

# Loading the example loa loa prevalence data
data <- get(load("LoaloaExampleData.RData"))

# Fitting the M2 specification of the assymetric model
# You could estimate the parameters of the model or
# used alread estimated
# and saved parameters for your predictions.
# You can change the number of MCMC samples, thining,
# and burnin in the loa_functions.R file.
estim <- fit_asymm_M2(data)
#save(estim, file="MCML_estimates_asymm_M2.RData")


# Getting the predictions grid
grid.pred <- get(
  load("prediction_grid_covariates.RData"))[, c("web_x", "web_y")]

# Getting the values of elevation at the prediction locations
elev <- prediction_grid_covariates$elev

# Comuping the spline variables for elevation at the
# prediction location.
covariate <- get_elev_vars(elev, 750, 1015)

# Geting the estimates of the model parameters
# for the plugged-in predictions
# Loading the already estimated parameters the asymetric model
# You can skip this line if you wish to use parameters you have
# estimated using the function fit_asymm_M1()
estim <- get(load("MCML_estimates_asymm_M2.RData"))

# Predicting with the M2 specification of the assymetric model
# You can change the number of MCMC samples, thining,
# and burnin in the loa_functions.R file.
predictions.asymm.M2 <-
  predictions.M2(data, grid.pred, covariate, estim)

head(predictions.asymm.M2)
# web_x , web_y   : Cordinates of the prediction locations
# that form the grid
# prev.mean.gold  : Mean prevalence as defined by the
# gold standard diagnostic
# prev.sd.gold    : Standard deviation based on the
# gold standard diagnostic
# CI.lower.gold   : lower end of the 95% predictive interval
# based on the gold standard diagnostic
# CI.upper.gold   : upper end of the 95% predictive interval
```

```r
# based on the gold standard diagnostic
# exceed.20p.prob.gold  : predictive probability of prevalence
# based on the gold standard diagnostic exceeding 20%
# prev.mean.alte  : Mean prevalence as defined by the
# alternate standard diagnostic
# prev.sd.alte    : Standard deviation based on the
# alternate diagnostic
# CI.lower.alte   : lower end of the 95% predictive interval
# based on the alternate diagnostic
# CI.upper.alte   : upper end of the 95% predictive interval
# based on the alternate diagnostic
# exceed.20p.prob.alte : predictive probability of prevalence
# based on the alternate diagnostic exceeding 20%

# ploting the mean prevalence as defined by the gold standard diagnostic
plot(rasterFromXYZ(predictions.asymm.M2
                   [, c("web_x", "web_y", "prev.mean.gold")])))
```

Appendix C

Appendix C: R Code for statistical analysis using Model 4.16

File *README2.TXT*

```
The folder malaria_application contains two R scripts and four
.RData files.  The R script malaria_Example.R demonstrates the
malaria analysis with an example data. It calls functions in the
other R script malaria_functions.R. Detailed description of the
malaria analysis are given as comments in the malaria_Example.R.


The .RData files are as follows:

malaria_input4sim.RData  is a dataframe used as input data to
simulated the RDT and PCR prevalence datasets. It contains the
following variables.
ageless5, agegreater5: variables defining the age spline
sex: gender of the individual
web_x,   web_y: gitter geolocations of the original data
diagnostic: the diagnostic used



MalariaExampleData.RData is an example dataset for the malaria
analysis. It contains, in addition to the variables of
malaria_input4sim.RData  described above, the variable malaria,
which is a binary outcome of weather the individual has malaria
(1) or not (0). malaria_site_grid.RData is the prediction grid of
the malaria analyses. MCML_estimates_malaria.RData is a list
containing the MCML estimates for the malaria analyses.
```

File *malaria_functions*

```r
###########################################################
###########################################################
# SIMULATING FROM THE MODEL OF THE MALARIA APPLICATION
###########################################################
###########################################################

simulate_malaria_data <- function(){

data <- get(load("malaria_input4sim.RData"))
estim <- get(load("MCML_estimates_malaria.RData"))


D <- model.matrix( ~ diagnostic+diagnostic:(sex +
                                 ageless5 + agegreater5),
                    data=data,family = binomial, x =TRUE)
p <- ncol(D)

beta0 <- estim$par[1:p]
sigma2.pcr.0 <- exp(estim$par[p+1])
sigma2.rdt.0 <- exp(estim$par[p+2])
phi0 <-  exp(estim$par[p+3])

n.tot <- nrow(data)
n.indiv <- n.tot/2
mu0.pcr <- as.numeric(D%*%beta0)[1:n.indiv]
mu0.rdt <- as.numeric(D%*%beta0)[(n.indiv+1):n.tot]
units.m <- rep(1,2*n.indiv)

indiv.coords <- data[1:n.indiv,c("web_x","web_y")]
ID.indiv.coords <- create.ID.coords(indiv.coords,~web_x+web_y)
locs.coords <- unique(indiv.coords)
n.locs <- nrow(locs.coords)

y.pcr = data$malaria[1:n.indiv]
y.rdt = data$malaria[(n.indiv+1):n.tot]
U <- as.matrix(dist(locs.coords))
R <- ((exp(-U/phi0)))
R.sroot <- t(chol(R))
C.S <- sapply(1:n.locs,function(i) ID.indiv.coords==i)*1

T.process <- R.sroot%*%rnorm(n.locs)
eta.rdt <- mu0.rdt + C.S %*% (sqrt(sigma2.rdt.0)*T.process)
eta.pcr <- mu0.pcr + C.S %*% (sqrt(sigma2.pcr.0)*T.process)
prob.pos.pcr <- exp(eta.pcr)/(1 + exp(eta.pcr))
prob.pos.rdt <- exp(eta.rdt)/(1 + exp(eta.rdt))
malaria <- rbinom(n.indiv, 1, c(prob.pos.pcr,prob.pos.rdt) )

outdata <- cbind(data, malaria)

return(outdata)
```

```r
}




###########################################################
###########################################################
# ESTIMATING THE MODEL OF THE MALARIA APPLICATION
###########################################################
##########################################################

estim_mal <- function(data.mod){

D <- model.matrix(~ diagnostic+diagnostic:(sex +
                    ageless5 + agegreater5), data=data.mod)


# Initializing the model parameters
estim <- get(load("MCML_estimates_malaria.RData"))
p <- ncol(D)
beta0 <- estim$par[1:p]
sigma2.pcr.0 <- exp(estim$par[p+1])
sigma2.rdt.0 <- exp(estim$par[p+2])
phi0 <-  exp(estim$par[p+3])-2



coords <- unique(data.mod[,c("web_x","web_y")])
U <- as.matrix(dist(coords))
Sigma <- exp(-U/phi0)
Sigma.inv <- solve(Sigma)
n.x <- dim(Sigma.inv)[1]
n <- nrow(data.mod)/2
scale0 <- sqrt(c(rep(sigma2.pcr.0,n),rep(sigma2.rdt.0,n)))
mu0 <- as.numeric(D%*%beta0)
ID.coords <- create.ID.coords(data.mod,~web_x+web_y)
C.S <- t(sapply(1:n.x,function(i) ID.coords==i))
y <- data.mod$malaria
units.m <- rep(1,2*n)




# Implementation of the MCMC

integrand <- function(S) {
  eta <- mu0+as.numeric(scale0*S[ID.coords])

  llik <- sum(y*eta-units.m*log(1+exp(eta)))

  q.f_S <- t(S)%*%Sigma.inv%*%(S)
```

```
    -0.5*q.f_S+llik
}

grad.integrand <- function(S) {
  eta <- as.numeric(mu0+scale0*as.numeric(S[ID.coords]))
  h <- units.m*exp(eta)/(1+exp(eta))

  as.numeric(-Sigma.inv%*%S+
                sapply(1:n.x,
              function(i) sum((scale0*(y-h))[C.S[i,]])))
}

hessian.integrand <- function(S) {
  eta <- as.numeric(mu0+scale0*as.numeric(S[ID.coords]))
  h <- units.m*exp(eta)/(1+exp(eta))
  h1 <- h/(1+exp(eta))


  grad.S.S <-  -Sigma.inv
  diag(grad.S.S) <- diag(grad.S.S)-sapply(1:n.x,
             function(i) sum(((scale0^2)*h1)[C.S[i,]]))
  as.matrix(grad.S.S)
}

estim <- maxBFGS(integrand,grad.integrand,
                   hessian.integrand,rep(0,n.x))
mode <- estim$estimate
Sigma.tilde <- solve(-estim$hessian)

Sigma.sroot <- t(chol(Sigma.tilde))
A <- solve(Sigma.sroot)
Sigma.w.inv <- solve(A%*%Sigma%*%t(A))
mu.w <- -as.numeric(A%*%mode)

cond.dens.W <- function(W,S) {
  eta <- mu0+scale0*as.numeric(S[ID.coords])

  llik <- sum(y*eta-units.m*log(1+exp(eta)))

  diff.w <- W-mu.w
  -0.5*as.numeric(t(diff.w)%*%Sigma.w.inv%*%diff.w)+
    llik
}

lang.grad <- function(W,S) {
  diff.w <- W-mu.w
  eta <- mu0+scale0*as.numeric(S[ID.coords])

  der <- units.m*exp(eta)/(1+exp(eta))

  grad.S <- sapply(1:n.x,
      function(i) sum(((scale0)*(y-der))[C.S[i,]]))
  as.numeric(-Sigma.w.inv%*%(W-mu.w)+ t(Sigma.sroot)%*%grad.S)
}

h <- 1.65/(n.x^(1/6))
```

```r
n.sim <- 9000
burnin <- 1000
thin <- 2
c1.h <- 0.001
c2.h <- 0.0001
W.curr <- rep(0,n.x)
S.curr <- as.numeric(Sigma.sroot%*%W.curr+mode)
C.S <- t(sapply(1:n.x,function(i) ID.coords==i))
mean.curr <- as.numeric(W.curr + (h^2/2)*lang.grad(W.curr,S.curr))
lp.curr <- cond.dens.W(W.curr,S.curr)
acc <- 0
S.sim <- matrix(NA,nrow=(n.sim-burnin)/thin,ncol=n.x)

h.vec <- rep(NA,n.sim)

for(i in 1:n.sim) {
  W.prop <- mean.curr+h*rnorm(n.x)
  S.prop <-  as.numeric(Sigma.sroot%*%W.prop+mode)
  mean.prop <- as.numeric(W.prop +
                      (h^2/2)*lang.grad(W.prop,S.prop))
  lp.prop <- cond.dens.W(W.prop,S.prop)

  dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
  dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

  log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

  if(log(runif(1)) < log.prob) {
    acc <- acc+1
    W.curr <- W.prop
    S.curr <- S.prop
    lp.curr <- lp.prop
    mean.curr <- mean.prop
  }

  if( i > burnin & (i-burnin)%%thin==0) {
    S.sim[(i-burnin)/thin,] <- S.curr
  }

  h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
  cat("Iteration",i,"out of",n.sim,"\r")
  flush.console()
}




####################################################
log.integrand <- function(S,val) {
  n.x <- length(S)
  scale <- sqrt(c(rep(val$sigma2.pcr,n),rep(val$sigma2.rdt,n)))
  eta <- scale*S[ID.coords]+val$mu
  llik <- sum(y*eta-units.m*log(1+exp(eta)))

  q.f_S <- t(S)%*%val$R.inv%*%S
  as.numeric(-0.5*(val$ldetR+q.f_S)+
              llik)
}
```

```r
# Maximization of the likelihood
compute.log.f <- function(par,ldetR=NA,R.inv=NA) {
  beta <- par[1:p]
  sigma2.pcr <- exp(par[p+1])
  sigma2.rdt <- exp(par[p+2])

  phi <- exp(par[p+3])
  val <- list()
  val$sigma2.pcr <- sigma2.pcr
  val$sigma2.rdt <- sigma2.rdt

  val$mu <- as.numeric(D%*%beta)
  if(is.na(ldetR) & is.na(as.numeric(R.inv)[1])) {
    R <- varcov.spatial(dists.lowertri=U,cov.model="matern",
                        cov.pars=c(1,phi),
                        nugget=0,kappa=0.5)$varcov
    val$ldetR <- determinant(R)$modulus
    val$R.inv <- solve(R)
  } else {
    val$ldetR <- ldetR
    val$R.inv <- R.inv
  }
  sapply(1:(dim(S.sim)[1]),
    function(i) log.integrand(S.sim[i,],val))
}
U <- dist(coords)
par0 <- c(beta0,log(c(sigma2.pcr.0,sigma2.rdt.0,phi0)))
log.f.tilde <- compute.log.f(par0)

MC.log.lik <- function(par) {
  log(mean(exp(compute.log.f(par)-log.f.tilde)))
}

der.phi <- function(u,phi,kappa) {
  u <- u+10e-16
  if(kappa==0.5) {
    out <- (u*exp(-u/phi))/phi^2
  } else {
    out <- ((besselK(u/phi,kappa+1)+besselK(u/phi,kappa-1))*
        phi^(-kappa-2)*u^(kappa+1))/(2^kappa*gamma(kappa))-
      (kappa*2^(1-kappa)*besselK(u/phi,kappa)*phi^(-kappa-1)*
        u^kappa)/gamma(kappa)
  }
  out
}

der2.phi <- function(u,phi,kappa) {
  u <- u+10e-16
  if(kappa==0.5) {
    out <- (u*(u-2*phi)*exp(-u/phi))/phi^4
  } else {
    bk <- besselK(u/phi,kappa)
    bk.p1 <- besselK(u/phi,kappa+1)
    bk.p2 <- besselK(u/phi,kappa+2)
    bk.m1 <- besselK(u/phi,kappa-1)
    bk.m2 <- besselK(u/phi,kappa-2)
```

```
    out <- (2^(-kappa -1)*phi^(-kappa -4)*
      u^kappa*(bk.p2*u^2+2*bk*u^2+
      bk.m2*u^2-4*kappa*bk.p1*phi*u-4*
      bk.p1*phi*u-4*kappa*bk.m1*phi*u-4*bk.m1*phi*u+
      4*kappa^2*bk*phi^2+4*kappa*bk*phi^2))/(gamma(kappa))
  }
  out
}

matern.grad.phi <- function(U,phi,kappa) {
  n <- attr(U,"Size")
  grad.phi.mat <- matrix(NA,nrow=n,ncol=n)
  ind <- lower.tri(grad.phi.mat)
  grad.phi <- der.phi(as.numeric(U),phi,kappa)
  grad.phi.mat[ind] <-  grad.phi
  grad.phi.mat <- t(grad.phi.mat)
  grad.phi.mat[ind] <-  grad.phi
  diag(grad.phi.mat) <- rep(der.phi(0,phi,kappa),n)
  grad.phi.mat
}

matern.hessian.phi <- function(U,phi,kappa) {
  n <- attr(U,"Size")
  hess.phi.mat <- matrix(NA,nrow=n,ncol=n)
  ind <- lower.tri(hess.phi.mat)
  hess.phi <- der2.phi(as.numeric(U),phi,kappa)
  hess.phi.mat[ind] <-  hess.phi
  hess.phi.mat <- t(hess.phi.mat)
  hess.phi.mat[ind] <-  hess.phi
  diag(hess.phi.mat) <- rep(der2.phi(0,phi,kappa),n)
  hess.phi.mat
}

grad.MC.log.lik <- function(par) {
  beta <- par[1:p]; mu <- D%*%beta
  sigma2.pcr <- exp(par[p+1])
  sigma2.rdt <- exp(par[p+2])
  phi <- exp(par[p+3])

  R <- varcov.spatial(dists.lowertri=U,cov.model="matern",
                      cov.pars=c(1,phi),
                      nugget=0,kappa=0.5)$varcov
  R.inv <- solve(R)
  ldetR <- determinant(R)$modulus

  exp.fact <- exp(compute.log.f(par,ldetR,R.inv)-log.f.tilde)
  L.m <- sum(exp.fact)
  exp.fact <- exp.fact/L.m

  R1.phi <- matern.grad.phi(U,phi,0.5)
  m1.phi <- R.inv%*%R1.phi
  t1.phi <- -0.5*sum(diag(m1.phi))
  m2.phi <- m1.phi%*%R.inv; rm(m1.phi)
  scale <- sqrt(c(rep(sigma2.pcr,n),rep(sigma2.rdt,n)))


  gradient.S <- function(S) {
```

```r
    eta <- mu+scale*S[ID.coords]
    h <- units.m*exp(eta)/(1+exp(eta))


    q.f <- t(S)%*%R.inv%*%S

    grad.beta <-  t(D)%*%(y-h)

    grad.log.sigma2.pcr <- sum((y[1:n]-h[1:n])*
                S[ID.coords[1:n]])*sqrt(sigma2.pcr)/2

    grad.log.sigma2.rdt <- sum((y[-(1:n)]-h[-(1:n)])*
              S[ID.coords[-(1:n)]])*sqrt(sigma2.rdt)/2

    grad.log.phi <- (t1.phi+
        0.5*as.numeric(t(S)%*%m2.phi%*%(S)))*phi


    out <- c(grad.beta,grad.log.sigma2.pcr,
            grad.log.sigma2.rdt,grad.log.phi)

    out
  }
  out <- rep(0,length(par))
  for(i in 1:(dim(S.sim)[1])) {
    out <- out + exp.fact[i]*gradient.S(S.sim[i,])
  }
  out
}

hess.MC.log.lik <- function(par) {
  beta <- par[1:p]; mu <- D%*%beta
  sigma2.pcr <- exp(par[p+1])
  sigma2.rdt <- exp(par[p+2])
  phi <- exp(par[p+3])

  R <- varcov.spatial(dists.lowertri=U,cov.model="matern",
                      cov.pars=c(1,phi),
                      nugget=0,kappa=0.5)$varcov
  R.inv <- solve(R)
  ldetR <- determinant(R)$modulus

  exp.fact <- exp(compute.log.f(par,ldetR,R.inv)-log.f.tilde)
  L.m <- sum(exp.fact)
  exp.fact <- exp.fact/L.m

  R1.phi <- matern.grad.phi(U,phi,0.5)
  m1.phi <- R.inv%*%R1.phi
  t1.phi <- -0.5*sum(diag(m1.phi))
  m2.phi <- m1.phi%*%R.inv; rm(m1.phi)

  R2.phi <- matern.hessian.phi(U,phi,0.5)
  t2.phi <- -0.5*sum(diag(R.inv%*%
            R2.phi-R.inv%*%R1.phi%*%R.inv%*%R1.phi))
  n2.phi <- R.inv%*%(2*R1.phi%*%R.inv%*%R1.phi-R2.phi)%*%R.inv

  H <- matrix(0,nrow=length(par),ncol=length(par))
```

```r
  scale <- sqrt(c(rep(sigma2.pcr,n),rep(sigma2.rdt,n)))

  hessian.S <- function(S,ef) {
    eta <- mu+scale*S[ID.coords]
    h <- units.m*exp(eta)/(1+exp(eta))
    h1 <- as.numeric(h/(1+exp(eta)))
    q.f <- t(S)%*%R.inv%*%S
    grad.beta <-  t(D)%*%(y-h)
    grad.log.sigma2.pcr <- sum((y[1:n]-h[1:n])*
                 S[ID.coords[1:n]])*sqrt(sigma2.pcr)/2
    grad.log.sigma2.rdt <- sum((y[-(1:n)]-h[-(1:n)])*
              S[ID.coords[-(1:n)]])*sqrt(sigma2.rdt)/2
    grad.log.phi <- (t1.phi+0.5*
      as.numeric(t(S)%*%m2.phi%*%(S)))*phi
    g <- c(grad.beta,grad.log.sigma2.pcr,
          grad.log.sigma2.rdt,grad.log.phi)
    grad2.log.lsigma2.lsigma2.pcr <- grad.log.sigma2.pcr/2+
      -sum(h1[1:n]*((S[ID.coords[1:n]])^2))*sigma2.pcr/4
    grad2.log.lsigma2.beta.pcr <- apply(-D[1:n,]*
      h1[1:n]*S[ID.coords[1:n]]*sqrt(sigma2.pcr)/2,2,sum)
    grad2.log.lsigma2.lsigma2.rdt <- grad.log.sigma2.rdt/2+
      -sum(h1[-(1:n)]*((S[ID.coords[-(1:n)]])^2))*sigma2.rdt/4

    grad2.log.lsigma2.beta.rdt <- apply(-D[-(1:n),]*
              h1[-(1:n)]*S[ID.coords[-(1:n)]]*
                sqrt(sigma2.rdt)/2, 2,sum)
    grad2.log.lphi.lphi <-(t2.phi-0.5*t(S)%*%
              n2.phi%*%(S))*phi^2+grad.log.phi

    H[1:p,1:p]  <- -t(D)%*%(D*h1)
    H[p+1,p+1]  <-  grad2.log.lsigma2.lsigma2.pcr
    H[1:p,p+1]  <- H[p+1,1:p] <- grad2.log.lsigma2.beta.pcr
    H[p+2,p+2]  <-  grad2.log.lsigma2.lsigma2.rdt
    H[1:p,p+2]  <- H[p+2,1:p] <- grad2.log.lsigma2.beta.rdt
    H[p+3,p+3]  <- grad2.log.lphi.lphi
    out <- list()
    out$mat1<- ef*(g%*%t(g)+H)
    out$g <- g*ef

    out
  }

  a <- rep(0,length(par))
  A <- matrix(0,length(par),length(par))
  for(i in 1:(dim(S.sim)[1])) {
    out.i <- hessian.S(S.sim[i,],exp.fact[i])
    a <- a+out.i$g
    A <- A+out.i$mat1
  }
  (A-a%*%t(a))

}

estim <- nlminb(start = par0,
                function(x) -MC.log.lik(x),
                function(x) -grad.MC.log.lik(x),
                function(x) -hess.MC.log.lik(x),
```

```r
                        control=list(trace=1))

estim$gradient <- grad.MC.log.lik(estim$par)
estim$hessian <- hess.MC.log.lik(estim$par)
estim$covariance <- solve(-estim$hessian)

return(estim)
}




##########################################################
##########################################################
# PREDICTION WITH THE MODEL OF THE MALARIA APPLICATION
##########################################################
#########################################################

predict_mal <- function(data.mod, estim, grid.pred){

mcml.estimates <- estim

D <- model.matrix(~ diagnostic+diagnostic:(sex +
                ageless5 + agegreater5), data=data.mod)

p <- ncol(D)
coords <- unique(data.mod[,c("web_x","web_y")])
U <- as.matrix(dist(coords))
n <- nrow(data.mod)/2
y <- data.mod$malaria
units.m <- rep(1,2*n)




beta0 <- estim$par[1:p]
sigma2.pcr.0 <- exp(estim$par[p+1])
sigma2.rdt.0 <- exp(estim$par[p+2])
phi0 <-  exp(estim$par[p+3])




Sigma <- exp(-U/phi0)
Sigma.inv <- solve(Sigma)
n.x <- dim(Sigma.inv)[1]

scale0 <- sqrt(c(rep(sigma2.pcr.0,n),rep(sigma2.rdt.0,n)))
mu0 <- as.numeric(D%*%beta0)
ID.coords <- create.ID.coords(data.mod,~web_x+web_y)
C.S <- t(sapply(1:n.x,function(i) ID.coords==i))
```

```r
###########################################
#### Some important functions
###########################################
integrand <- function(S) {
  eta <- mu0+as.numeric(scale0*S[ID.coords])

  llik <- sum(y*eta-units.m*log(1+exp(eta)))

  q.f_S <- t(S)%*%Sigma.inv%*%(S)

  -0.5*q.f_S+llik
}

grad.integrand <- function(S) {
  eta <- as.numeric(mu0+scale0*as.numeric(S[ID.coords]))
  h <- units.m*exp(eta)/(1+exp(eta))

  as.numeric(-Sigma.inv%*%S+
 sapply(1:n.x,function(i) sum((scale0*(y-h))[C.S[i,]])))
}

hessian.integrand <- function(S) {
  eta <- as.numeric(mu0+scale0*as.numeric(S[ID.coords]))
  h <- units.m*exp(eta)/(1+exp(eta))
  h1 <- h/(1+exp(eta))


  grad.S.S <-  -Sigma.inv
  diag(grad.S.S) <- diag(grad.S.S)-sapply(1:n.x,
        function(i) sum(((scale0^2)*h1)[C.S[i,]]))
  as.matrix(grad.S.S)
}

cond.dens.W <- function(W,S) {
  eta <- mu0+scale0*as.numeric(S[ID.coords])

  llik <- sum(y*eta-units.m*log(1+exp(eta)))

  diff.w <- W-mu.w
  -0.5*as.numeric(t(diff.w)%*%Sigma.w.inv%*%diff.w)+
    llik
}

lang.grad <- function(W,S) {
  diff.w <- W-mu.w
  eta <- mu0+scale0*as.numeric(S[ID.coords])

  der <- units.m*exp(eta)/(1+exp(eta))

  grad.S <- sapply(1:n.x,
    function(i) sum(((scale0)*(y-der))[C.S[i,]]))
  as.numeric(-Sigma.w.inv%*%(W-mu.w)+
              t(Sigma.sroot)%*%grad.S)
}


# simulating from conditional distribution of random effects
```

```r
estim <- maxBFGS(integrand,grad.integrand,
                 hessian.integrand,rep(0,n.x))
mode <- estim$estimate
Sigma.tilde <- solve(-estim$hessian)

Sigma.sroot <- t(chol(Sigma.tilde))
A <- solve(Sigma.sroot)
Sigma.w.inv <- solve(A%*%Sigma%*%t(A))
mu.w <- -as.numeric(A%*%mode)


  h <- 1.65/(n.x^(1/6))
  n.sim <- 1150
  burnin <- 150
  thin <- 5
  c1.h <- 0.001
  c2.h <- 0.0001
  W.curr <- rep(0,n.x)
  S.curr <- as.numeric(Sigma.sroot%*%W.curr+mode)
  C.S <- t(sapply(1:n.x,function(i) ID.coords==i))
  mean.curr <- as.numeric(W.curr + (h^2/2)*
                             lang.grad(W.curr,S.curr))
  lp.curr <- cond.dens.W(W.curr,S.curr)
  acc <- 0
  S.sim <- matrix(NA,nrow=(n.sim-burnin)/thin,ncol=n.x)

  h.vec <- rep(NA,n.sim)

    for(i in 1:n.sim) {
    W.prop <- mean.curr+h*rnorm(n.x)
    S.prop <-  as.numeric(Sigma.sroot%*%W.prop+mode)
    mean.prop <- as.numeric(W.prop +
                (h^2/2)*lang.grad(W.prop,S.prop))
    lp.prop <- cond.dens.W(W.prop,S.prop)

    dprop.curr <- -sum((W.prop-mean.curr)^2)/(2*(h^2))
    dprop.prop <- -sum((W.curr-mean.prop)^2)/(2*(h^2))

    log.prob <- lp.prop+dprop.prop-lp.curr-dprop.curr

    if(log(runif(1)) < log.prob) {
      acc <- acc+1
      W.curr <- W.prop
      S.curr <- S.prop
      lp.curr <- lp.prop
      mean.curr <- mean.prop
    }

    if( i > burnin & (i-burnin)%%thin==0) {
      S.sim[(i-burnin)/thin,] <- S.curr
    }

    h.vec[i] <- h <- max(0,h + c1.h*i^(-c2.h)*(acc/i-0.57))
    cat("Iteration",i,"out of",n.sim,"\r")
    flush.console()
  }
```

```r
# Doing spatial predictions
  estim <- mcml.estimates
  beta <- estim$par[1:p]
  mu <- as.numeric(D%*%beta)

  sigma2.pcr <- exp(estim$par[p+1])
  sigma2.rdt <- exp(estim$par[p+2])
  phi <- exp(estim$par[p+3])

  mu <- as.numeric(D%*%beta)
  sigma2.pcr <- exp(estim$par[p+1])
  sigma2.rdt <- exp(estim$par[p+2])
  phi <- exp(estim$par[p+3])

  U <- as.matrix(dist(coords))
  Sigma <- exp(-U/phi)
  Sigma.inv <- solve(Sigma)
  n.x <- dim(Sigma.inv)[1]



  U.pred <- as.matrix(pdist(grid.pred,coords))
  C <- exp(-U.pred/phi)
  A <- C%*%Sigma.inv

  n.samples <- nrow(S.sim)

  mu.cond <- t(sapply(1:n.samples,
                  function(i) A%*%S.sim[i,]))
  sd.cond <- sqrt(1-apply(A*C,1,sum))

  T.samples <- sapply(1:n.samples,
              function(i) mu.cond[i,]+sd.cond*
                rnorm(nrow(grid.pred)))

  # Predict RDT
  D.pred.rdt <- data.frame(sex=
                  factor(rep("male",nrow(grid.pred)),
                  levels=levels(data.mod$sex)),
                  diagnostic=factor(rep("rdt",
                  nrow(grid.pred)),
                levels=levels(data.mod$diagnostic)),
                     ageless5=5, agegreater5=0)

  D.pred.rdt <- model.matrix(~ diagnostic+
                  diagnostic:(sex + ageless5 +
                      agegreater5),D.pred.rdt)
  mu.rdt <- as.numeric(D.pred.rdt%*%beta)

  prev.rdt.samples <- sapply(1:n.samples,
            function(i) 1/(1+exp(-mu.rdt-
            sqrt(sigma2.rdt)*T.samples[,i])))
```

```
  # Predict PCR
  D.pred.pcr <- data.frame(sex=factor(rep("male",
                 nrow(grid.pred)),levels=levels(data.mod$sex)),
                 diagnostic=factor(rep("pcr",nrow(grid.pred)),
                 levels=levels(data.mod$diagnostic)),
                          ageless5=5, agegreater5=0)

  D.pred.pcr <- model.matrix(~ diagnostic+
        diagnostic:(sex + ageless5 + agegreater5),D.pred.pcr)

  mu.pcr <- as.numeric(D.pred.pcr%*%beta)

  prev.pcr.samples <- sapply(1:n.samples,
function(i) 1/(1+exp(-mu.pcr-sqrt(sigma2.pcr)*T.samples[,i])))

  mean.prev.rdt <- apply(prev.rdt.samples, 1, mean)
  mean.prev.pcr <- apply(prev.pcr.samples, 1, mean)
  mean.sd.rdt <- apply(prev.rdt.samples, 1, sd)
  mean.sd.pcr <- apply(prev.pcr.samples, 1, sd)
  ex10_rdt <- sapply(1:dim(prev.rdt.samples)[1],
  function(i) sum(prev.rdt.samples[i,] > 0.1 ))/n.samples
  ex10_pcr <- sapply(1:dim(prev.pcr.samples)[1],
    function(i) sum(prev.pcr.samples[i,] > 0.1 ))/n.samples

  out.df <- data.frame(grid.pred, mean.prev.rdt, mean.prev.pcr,
         mean.sd.rdt, mean.sd.pcr, ex10_rdt, ex10_pcr)
 return(out.df)
}
```

File *malaria_Examples.R*

```
###############################################################
###############################################################
# ANALYSES OF THE MALARIA DATA USING THE MODEL 1
# SPECIFICATION OF ASYMMETRIC MODEL
###############################################################
###############################################################

# It is important to keep all the files in a single folder and set
# this folder as the working directory. Edit the next line to set
# your working directory.

workingdir <- "~/Examples_Code_and_Data/malaria_application"
setwd(workingdir)


library(rgdal)
library(geoR)
library(PrevMap)
```

```r
# Loading the script that has the functions
# for the malaria analysis
source("malaria_functions.R")


###################################################################
# Simulating data under the fitted symmetric model for the
# malaria application # in the paper, using jittered locations
# of the original malaria study
###################################################################
# An example data that mimics the origional loa loa data
# has been provided as MalariaExampleData.RData. However,
# you can simulate your own data under the fitted symmetric
# model in the paper. The next line simulates such data.
malaria_simulated_data <- simulate_malaria_data()
head(malaria_simulated_data)
# save(MalariaExampleData, file="MalariaExampleData.RData")
# The variables in the data are as follows
# ageless5, agegreater5    : Variables defining the age spline
# sex                      : gender of the individual
# web_x    web_y  : The (gittered) geolocations of the individuals
# diagnostic     : the diagnostic used pcr/rdt
# malaria        : 0 for negative malaria test and 1 for positive.




ppp <- get(load("MalariaExampleData.RData"))
head(ppp)



#############################################################
# Estimating the specified symmetric model
# for the malaria analysis
#############################################################
# Loading the example malaria prevalence data
data.mod <- get(load("MalariaExampleData.RData"))
# Estimating the model specified symmetric model for
# the malaria analysis
# You could estimate the parameters of the model or used
# already estimated
# and saved parameters for your predictions.
estim <- estim_mal(data.mod)
# estim <- estim_mal(data.mod=malaria_simulated_data)
# use this line instead of the preceeding if
# you intend not to use the example data provided,
# but data you have simulated above.
# You can change the number of MCMC samples, thining,
# and burnin in the malaria_functions.R file.
# save(estim,file="MCML_estimates_malaria2.RData")
```

```r
############################################################
# Predicting with the specified symetric
# model for the malaria analysis
############################################################
# Loading the example malaria prevalence data
# If you simlated
data.mod <- get(load("MalariaExampleData.RData"))
# a different dataset, set data.mod to that data using
# data.mod = malaria_simulated_data

# Loading the prediction grid
grid.pred <- get(load("malaria_site_grid.RData"))

# Loading the saved model estimates
# You can also use the estimates you obtained above
estim <- get(load("MCML_estimates_malaria.RData"))
# You can change the number of MCMC samples, thining,
# and burnin in the malaria_functions.R file.

# Spatial predictions using the malaria model
predictions.mal <- predict_mal(data.mod, estim, grid.pred)
# web_x , web_y  : Cordinates of the prediction locations
# that form the grid
# mean.prev.rdt  : Mean prevalence as defined by the rdt
# mean.prev.pcr : Mean prevalence as defined by the pcr
# mean.sd.rdt : Standard deviation of rdt prevalence
# mean.sd.pcr : Standard deviation of pcr prevalence
# ex10_rdt : Probability of exceeding 10%
# threshold of rdt prevalence
# ex10_pcr : Probability of exceeding 10% threshold of
# rdt prevalence
plot(rasterFromXYZ(predictions.mal[, c("web_x",
                        "web_y", "mean.prev.rdt")]))
plot(rasterFromXYZ(predictions.mal[, c("web_x",
                        "web_y", "mean.prev.pcr")]))
```