LANCASTER UNIVERSITY

Geostatistical methods for disease prevalence mapping

by

Emanuele Giorgi

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Health and Medicine
Lancaster Medical School

September 2015

Declaration of Authorship

I, EMANUELE GIORGI, declare that this thesis titled, 'GEOSTATISTICAL METHODS FOR DISEASE PREVALENCE MAPPING' 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: ______________________________

"A good theory is better then a lot of data without a theory."

Denis Lindley

LANCASTER UNIVERSITY

Abstract

Faculty of Health and Medicine
Lancaster Medical School

Doctor of Philosophy

by Emanuele Giorgi

Geostatistical methods are increasingly used in low-resource settings where disease registries are either non-existent or geographically incomplete. In this thesis, which is comprised of four papers, we address some of the common issues that arise from analysing disease prevalence data. In the first paper we consider the problem of combining data from multiple spatially referenced surveys so as to account for two main sources of variation: temporal variation, when surveys are repeated over time; data-quality variation, e.g. between randomised and non-randomised surveys. We then propose a multivariate binomial geostatistical model for the combined analysis of data from multiple surveys. We also show an application to malaria prevalence data from three surveys conducted in two consecutive years in Chikwawa District, Malawi, one of which used a more economical convenience sampling strategy. In the second paper, we analyse river-blindness prevalence data from a survey conducted in 20 African countries enrolled in the African Programme of Onchocerciasis Control (APOC). The main challenge of this analysis is computational, as a binomial geostatistical model has to be fitted to more than 14,000 village locations and predictions carried out on about 10 millions locations across Africa. To make the computation feasible and efficient, we then develop a low rank approximation based on a convolution-kernel representation which avoids matrix inversion. The third paper is a tutorial on the use of a new R package, namely "PrevMap", which provides functions for both likelihood-based and Bayesian analysis of spatially referenced prevalence data. In the fourth paper, we present some extensions of the standard geostatistical model for spatio-temporal analysis of prevalence data and modelling of spatially structured zero-inflation. We then describe three applications that have arisen through our collaborations with researchers and public health programmers in African countries.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Peter Diggle for the continuous support of my Ph.D. study and research, for his guidance, motivation, enthusiasm, and knowledge. I am also grateful for all his Thursday lifts to Leahurst and Liverpool that have provided fruitful scientific discussions.

I would like to thank my second supervisor Dr. Anja Terlouw for her support of my research and for giving me the opportunity to contribute to the improvement of the effectiveness of the malaria control programme in Chikwawa District, Malawi.

My sincere thanks also goes to all the members of CHICAS for their professional help and for making me feel part of the group since the first day of my Ph.D.

Contents

List of Figures

List of Tables

List of Papers

Paper 1. *Combining data from multiple spatially referenced surveys using generalized linear geostatistical models.*

Giorgi, E., Sesay, S. S. S., Terlouw, D. J., Diggle, P. J.

Published in *Journal of the Royal Statistical Society, Series A (2014), 178:445-464.*

Contribution: lead author, implementation of the simulation studies, statistical analysis and writing of the paper.

Paper 2. *The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control: pre-control endemicity levels and estimated number infected.*

Zouré, H. G. M., Noma, M., Tekle, A. H., Amazigo, U. V., Diggle, P. J., Giorgi, E. and Remme, J. H. F.

Published in *Parasites and Vectors (2014), 7:326, doi:10.1186/1756-3305-7-326.*

Contribution: development of a computationally efficient algorithm for parameter estimation and spatial prediction, statistical analysis and writing of the paper.

Paper 3. *PrevMap: an R package for prevalence mapping.*

Giorgi, E., Diggle, P. J.

Under review in *Journal of Statistical Software.*

Contribution: lead author, development of statistical software, statistical analysis and writing of the paper.

Paper 4. *Model-based geostatistics for prevalence mapping in low-resource settings.*

Diggle, P. J., Giorgi, E.

Under review in *Journal of the American Statistical Association.*

Contribution: statistical analysis and writing of the paper.

To Iulia

Chapter 1

Introduction

In this thesis we address some of the issues related to disease prevalence mapping with a particular focus on its application in low-resource settings. In the first paper we extend the standard geostatistical model (Diggle, Tawn, and Moyeed, 1998) for prevalence data to allow for temporal and data quality variation across surveys. In the second, we describe an application to a large spatial data-set on river blindness prevalence data. In the third paper we illustrate how to use newly developed statistical software for prevalence mapping. In the fourth paper, we present additional extensions of the standard geostatistical model to model spatio-temporal variation in disease prevalence and spatially structured zero-inflation.

1.1 The standard geostatistical model for prevalence data

Let $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^2$ denote a set of n distinct spatial locations that represent the geographical coordinates of sampled households or villages in a prevalence survey. At each of the locations x_i, we then sample m_i individuals and perform a test on each of them for the disease of interest. Let Y_i denote the resulting number of positive tests at location x_i. Conditionally on a spatial stochastic process $S(x_i)$ and mutually independent zero-mean Gaussian latent variables Z_i, we assume that Y_i are mutually independent binomial variables with probability of having a positive test p_i. A logit-link function is then used for p_i, assuming the form

$$\log\{p_i/(1-p_i)\} = d(x_i)^\top \beta + S(x_i) + Z_i, \tag{1.1}$$

where $d(x_i)$ is a vector of explanatory variables which are often obtained by remotely-sensed images (e.g. temperature, rainfall, NDVI, population density) or represent household specific information (e.g. material of the house wall, type of roof, ownership of bed-nets, Socio-Economic-Status). Each of these are associated with regression coefficients given by the elements of the vector β. The spatial random effect $S(x_i)$ is used to account for unmeasured spatially structured risk factors which induce residual spatial correlation amongst the observations. The unstructured residuals Z_i, often referred to as "nugget effect", can be dually interpreted either as small-range spatial variation (on a range smaller than the observed minimum distance between locations) or extra-binomial variation within households (e.g. genetic variation). It is generally difficult to disentangle the two effects without multiple observations at each location x_i, which are often not available. However, the main concern of almost any geostatistical analysis is in the prediction of S at an unobserved location x while Z is usually a nuisance.

The model in (1.1) can also be modified to allow for individual-specific information. We then introduce an additional subscript j to denote the j-th individual within the i-th household (or village). The response variable Y_{ij} is now a binary indicator of the test outcome for each individual and takes value 1 if the test is positive and 0 if negative. Conditionally on $S(x_i)$ and Z_i, Y_{ij} now follows a Bernoulli distribution with probability of a positive test p_{ij}. The logit-link function then assumes the form

$$\log\{p_{ij}/(1-p_{ij})\} = d_{ij}^\top \gamma + d(x_i)^\top \beta + S(x_i) + Z_i, \tag{1.2}$$

where d_{ij} are individual-specific covariates, e.g. age and gender, with associated vector of regression coefficients γ.

We also assume that $S(x)$ is a zero-mean, stationary and isotropic Gaussian process, i.e. with invariant distribution under translation and rotation. In this thesis, we restrict our attention to the class of Matérn (1986) covariance functions, given by the following expression

$$\text{cov}(S(x_i), S(x_j)) = \sigma^2\{2^{k-1}\Gamma(\kappa)\}^{-1}(u_{ij}/\phi)^\kappa K_\kappa(u_{ij}/\phi), \tag{1.3}$$

where u_{ij} is the Euclidean distance between x_i and x_j, σ^2 is the variance of $S(x)$, ϕ is a scale parameter and $K_\kappa(\cdot)$ is the Bessel function of the second kind of order $\kappa > 0$. The parameter κ determines the smoothness of the process $S(x)$ which is then $\lceil \kappa \rceil - 1$ times differentiable, with $\lceil a \rceil$ indicating the smallest integer not less than a. Since estimating κ is generally difficult, a practical approach is to fix κ at a plausible value. In the case $\kappa = 1/2$, (1.3) simplifies to the exponential covariance function, i.e.

$$\text{cov}(S(x_i), S(x_j)) = \sigma^2 \exp\{-u_{ij}/\phi\}, \tag{1.4}$$

which corresponds to a mean-square continuous process.

1.2 Structure of the thesis

In Chapter 2 (Paper 1), the specific research questions that we address are the following.

- How to account for spatially structured bias inherent to data from non-randomised surveys?
- How to account for temporal variation across surveys that are repeated over time?
- What is the gain in accuracy of prevalence estimates when combining the data in a joint model with respect to a marginal analysis of unbiased prevalence data only?

To answer these questions, we develop a multivariate binomial geostatistical model for the joint analysis of data from multiple spatially referenced surveys. An important assumption of our approach is that there is at least one "gold-standard" survey able to deliver unbiased estimates of prevalence. In our application, we analyse data from three malaria prevalence surveys two of which are "gold-standard" but conducted in two consecutive years, whilst the third, at the time of the second "gold-standard", uses a more economic, but potentially biased, convenience sampling approach. The resulting model for the data is then characterised by a tri-variate process $(S_1(x), S_2(x), B(x))$ where: $S_1(x)$ and $S_2(x)$ correspond to temporally correlated spatial random effects for the first and second "gold-standard" surveys, respectively, and $B(x)$ is spatially structured bias inherent to the convenience survey. We then compare the accuracy in the estimates of $S_2(x)$ obtained by using the proposed joint approach with two simpler approaches that only make use of "gold-standard" data.

In Chapter 3 (Paper 2) we present an application to river-blindness prevalence data from 20 countries enrolled in the African Programme of Onchocerciasis Control (APOC). A standard geostatistical model is used to analyse the spatial variation in river-blindness prevalence in 14,309 villages across Africa. Given the high dimensionality of the resulting random effect structure, standard fitting algorithms are very inefficient. Indeed, evaluation of the likelihood function would require the inversion of a 14,309 by 14,309 covariance matrix. To circumvent the problem of matrix inversion, we then propose to approximate the spatial process $S(x)$ using a low-rank approximation based on convolution kernel representations (Higdon, 1998; Higdon, 2002), i.e.

$$S(x) \approx \sum_{j=1}^{k} K(x - \tilde{x}_j) Z_j, \tag{1.5}$$

where $\tilde{x}_j$ for $j = 1 \ldots, k$ is a set of pre-defined spatial knots, $K(\cdot)$ is a suitable kernel function and Z_j are independent identically distributed zero-mean Gaussian variables with variance σ^2. An additional advantage of (1.5) is that prediction of $S(x)$ at unobserved locations can be carried out without computing the cross-covariance matrix. In our case, this would be computationally infeasible, as we have more than 10 million prediction locations across Africa.

Chapter 4 (Paper 3) is a tutorial paper on a newly developed R package, “PrevMap”, that fits the models given by (1.1) and (1.2) using both likelihood-based and Bayesian methods of inference. More details on the fitting algorithms used in Paper 1 and Paper 2 are given. We also illustrate the use of the package through the analysis of Loa-loa prevalence data from Nigeria and Cameroon and show an example of the low-rank approximation in (1.5) for a large simulated spatial data-set. Other functionalities of the package include fitting of linear geostatistical models that can be used as a faster but approximate procedure based on logit-transformed empirical prevalence data.

In Chapter 5 (Paper 4), we first review the standard geostatistical model and available open-source statistical software. We then describe three extensions of the standard geostatistical model to address four specific methodological questions: combining data from multiple spatially referenced surveys (see also Chapter 2); spatio-temporal modelling of disease prevalence; estimating the impact of control interventions, such as distribution of insecticide treated nets and indoor residual spraying in the case of malaria control; geostatistical modelling of zero-inflated prevalence data. For each of these we show applications that have arisen from our collaborations with researchers and public health programmers in African countries. More specifically we analyse: malaria prevalence data from a community survey and a school-based, hence potentially biased, survey, both conducted in Nyanza Province, Kenya; malaria prevalence data from a continuous Malaria Indicator Survey conducted in Chikwawa District, Malawi, from May 2010 to June 2013; a subset of the river-blindness prevalence data of Chapter 3 corresponding to Mozambique, Malawi and Tanzania.

Chapter 6 is a concluding discussion where we briefly explore possible extensions of the developed methodology in the previous chapters.

References

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). “Model-based geostatistics (with discussion)”. In: *Applied Statistics* 47, pp. 299–350.

Higdon, D. (1998). “A process-convolution approach to modeling temperatures in the North Atlantic Ocean”. In: *Environmental and Ecological Statistics* 5, pp. 173–190.

Higdon, D. (2002). “Space and space-time modeling using process convolutions”. In: *Quantitative methods for current environmental issues*. Ed. by C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi. Springer-Verlag, New York, pp. 37–56.

Matérn, B. (1986). *Spatial Variation*. Second. Springer, Berlin.

Chapter 2

Paper 1. Combining data from multiple spatially referenced surveys using generalized linear geostatistical models

E. Giorgi[1], S. S. S. Sanie[2], D. J. Terlouw[2] and P. J. Diggle[1]

[1] Lancaster Medical School, Lancaster University, Lancaster, UK

[2] Liverpool School of Tropical Medicine, Liverpool, UK

Summary

Data from multiple prevalence surveys can provide information on common parameters of interest, which can therefore be estimated more precisely in a joint analysis than by separate analyses of the data from each survey. However, fitting a single model to the combined data from multiple surveys is inadvisable without testing the implicit assumption that all of the surveys are directed at the same inferential target. In this paper we propose a multivariate generalized linear geostatistical model that accommodates two sources of heterogeneity across surveys so as to correct for spatially structured bias in non-randomised surveys and to allow for temporal variation in the underlying prevalence surface between consecutive survey-periods. We describe a Monte Carlo maximum likelihood procedure for parameter estimation, and show through simulation experiments how accounting for the different sources of heterogeneity among surveys in a joint model leads to more precise inferences. We describe an application to multiple surveys of malaria prevalence conducted in Chikhwawa District, Southern Malawi, and discuss how this approach could inform hybrid sampling strategies that combine data from randomised and non-randomised surveys so as to make the most efficient use of all available data.

Keywords: convenience sampling; generalized linear geostatistical models; malaria mapping; Monte Carlo maximum likelihood; multiple surveys; spatio-temporal models.

2.1 Introduction

In studies of spatial variation in disease prevalence, it is often necessary to combine information from multiple prevalence surveys. This is particularly the case in low-resource settings, where disease registries typically do not exist. A methodological challenge in these circumstances is that survey designs are severely constrained by cost constraints. The available surveys may therefore be of variable quality and/or conducted at different times. In this paper, we propose a class of generalized linear geostatistical models (GLGMs) to address two specific issues. The first is variation in quality, for example between randomised and non-randomised surveys, in which case our proposed methodology assumes that at least one of the surveys provides an unbiased "gold-standard". The second is variation in the underlying prevalence when surveys are conducted at different times. In this case, by modelling the underlying prevalence over time we are able to use data collected at all times to estimate the underlying prevalence surface at the specific time of interest, typically the time of the most recent survey.

Methods for the combined analysis of data from multiple surveys have previously used meta-analysis and small area statistics approaches; see Moriarity and Scheuren (2001), Elliot and Davis (2005), Lohr and Rao (2006) and Turner et al. (2009). More recently, Manzi et al. (2011) used Bayesian hierarchical models to combine smoking prevalence estimates from multiple surveys. They noted that commercial surveys are often ignored in constructing official estimates because of poor information about the sampling designs used, but argued that these surveys can nevertheless provide useful additional information because they are more frequently updated than official surveys.

Raghunathan et al. (2007) noted the potential benefits that might accrue from spatial modelling of multiple survey data, but to the best of our knowledge, explicit spatial modelling of biases and/or temporal variation in the outcome of interest has not previously been addressed, except in a few specific applications. For example, Wanji et al. (2012) established a logit-linear calibration relationship between estimates of *Loa loa* prevalence in part of equatorial Africa based on two different methods, finger-prick blood sampling and a short questionnaire instrument. Crainiceanu, Diggle, and Rowlingson (2008) incorporated this calibration relationship into a bivariate geostatistical model for the two corresponding prevalence maps.

As discussed in Turner et al. (2009), if information from multiple surveys is to be combined, it is important to understand the limitations of their designs in order to take account of potential biases in the associated estimates of prevalence. As a minimal condition, the study subjects in each survey should be drawn from the same target population. One potential source of bias is that some members of the target population may be less likely than others to be included. Convenience samples provide an example of this. In resource-poor settings, the relatively low cost of convenience sampling is tempting, but its potential to produce biased estimates is clear. In a non-spatial context, Hedt and Pagano (2011) propose a hybrid prevalence estimator that combines information from randomised and convenience surveys. They demonstrate that, with suitable adjustment for the bias, their hybrid estimator can give better prevalence estimates than would be obtained by using only the data from the randomised surveys.

A second source of heterogeneity amongst multiple prevalence surveys is temporal variation in prevalence. When spatially referenced prevalence surveys are repeated over time it is usually of interest to estimate changes in prevalence over time. When the outcomes from consecutive surveys are correlated, there is also a potential gain in efficiency if comparisons are made through the use of a joint model. This is especially advantageous when the surveys do not use the same set of sampling locations, because a joint analysis can then exploit both the temporal and spatial correlation structure of the combined data.

In Section 2.2 of the paper we propose a class of generalised linear geostatistical models (GLGMs) for the combined analysis of data from multiple prevalence surveys. The model allows both for biased sampling and temporal variation in prevalence provided that one of the surveys delivers unbiased "gold-standard" estimates of prevalence. In Section 2.3 we describe the methods that we use to fit the model. In Section 2.4 we report the results of simulation experiments that illustrate how a joint model leads to gains in efficiency of estimation and spatial prediction. In Section 2.5 we describe an application to malaria prevalence data from three surveys conducted in Chikhwawa District, Southern Malawi. Section 2.6 is a concluding discussion. All computations for the paper were run on the High End Computing Cluster at Lancaster University, using the R software environment (R Core Team, 2012).

2.2 A multivariate generalized linear geostatistical model

The ingredients of a univariate GLGM are the following. Random variables Y_j and explanatory variables d_j are associated with sampling locations x_j in a region of interest $A \subseteq \mathbb{R}^2$. Each d_j is a vector of length $p \geq 1$. Conditional on the realisation of a zero-mean latent Gaussian process $S(x)$ and a set of mutually independent zero-mean latent Gaussian variables Z_j, the Y_j follow a classical generalized linear model (McCullagh and Nelder, 1989), hence:

(i) the Y_j are mutually independent conditional on the $S(x_j)$ and Z_j, with conditional expectations $\mu_j = m_j g^{-1}(\eta_j)$, where m_j is a known scalar and $g(\cdot)$ a known *link function*;

(ii) $\eta_j = d_j^\top \beta + S(x_j) + Z_j$;

(iii) the conditional distribution of the Y_j falls within the exponential family.

In the remainder of the paper, we assume that the conditional distributions in (iii) are binomial, with the y_j representing the number of positives amongst m_j individuals sampled at location x_j. We also adopt the standard logistic link function, $g(\mu/m) = \log\{\mu/(m-\mu)\}$, but other link functions could also be used. We specify the Gaussian process $S(x)$ to have covariance function $\text{Cov}\{S(x), S(x')\} = \sigma^2\rho(x, x'; \phi)$, and the mutually independent Z_j to have variance τ^2. The Z_j have a dual interpretation as either non-spatial extra-binomial variation or spatial variation at scales smaller than the smallest distance between sampling locations; the two interpretations can only be disentangled unambiguously if repeated measurements are taken at coincident locations. Finally, we write $d_j = d(x_j)$ to emphasise its spatial context.

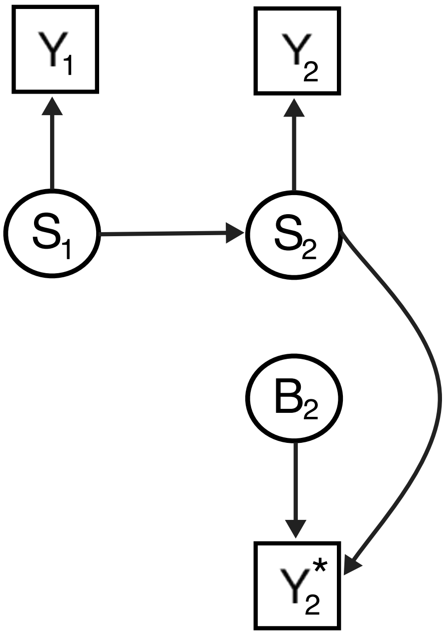

FIGURE 2.1: Representation of the multivariate generalized linear geostatistical model (2.1) as a directed acyclic graph; S_1 and S_2 represent prevalences at times t_1 and $t_2 > t_1$; B_2 represents bias; Y_1, Y_2 and Y_2^* are observed prevalences from unbiased surveys at times t_1 and t_2, and a biased survey at time t_2, respectively. The target for predictive inference is S_2.

We now extend the model to accommodate multiple surveys taken at possibly different times, some of which may be biased. To motivate the extension, consider a prevalence survey that includes a community at a particular location x in which the odds for disease is $r_d(x)$ but a member of the community that participates in the survey has a relative risk of the disease $r_p(x)$ with respect to those who do not participate. Then, the odds for disease within the survey is $r_c(x) = r_d(x) r_p(x)$, hence $\log r_c(x) = \log r_d(x) + \log r_p(x) = S(x) + B(x)$, say. Under randomised sampling $r_p(x) = 1$ for all x. Otherwise, provided that $r_p(x)$ is either known or can be modelled as a function of the design and/or observed covariates, standard methods can be used to obtain unbiased estimates of $r_d(x)$. In our context, we want to allow both $S(x)$ and $B(x)$ to vary spatially, and potentially to depend on both observed and unobserved covariates.

Now, let $i = 1, \ldots, r$ denote the index of the survey and $x_{ij} : j = 1, \ldots, n_i$ the corresponding set of sample locations. We replace the single process $S(x)$ by a set of r processes $S_i(x)$ which relate to the true prevalence at different times. We assume that at least the first of the surveys ($i = 1$) is known to be unbiased, define $\mathcal{B}$ to be the index set of the potentially biased surveys and introduce an additional set of latent Gaussian processes $B_i(x) : i \in \mathcal{B}$ to represent the spatially varying biases. Finally, we assume that data from different surveys are generated by conditionally independent univariate GLGMs, with link functions

$$\begin{aligned} g_i(\mu_{ij}/m_{ij}) \quad = \quad & \eta_{ij} = d(x_{ij})^\top \beta_1 + S_i(x_{ij}) + I(i \in \mathcal{B})[B_i(x_{ij}) + d(x_{ij})^\top \beta_i] + Z_{ij}, \\ & j = 1, \ldots, n_i; i = 1, \ldots, r. \qquad (2.1) \end{aligned}$$

On the right-hand-side of (2.1), we assume that the marginal properties of each $S_i(x)$ are the same as previously specified for $S(x)$, and add a set of cross-covariance functions, $\text{Cov}\{S_i(x), S_{i'}(x')\} = \sigma^2 \alpha_{ii'} \rho(x, x'; \phi)$, where $-1 < \alpha_{ii'} < 1$. The parameters $\alpha_{ii'}$ capture the temporal correlation between the true prevalence surfaces at different times, hence if surveys i and i' are taken at the same time, $S_i(x) = S_{i'}(x)$ for all x and $\alpha_{ii'} = 1$. Note that if $r > 2$, some combinations of $\alpha_{ii'}$ result in a non-positive-definite variance matrix. If r is small, this can be handled by setting the likelihood to zero for all such combinations. When r is large the issue can be avoided by imposing a spatio-temporally continuous parametric structure. This has the incidental benefit of making the model more parsimonious. One such example would be an exponentially decaying cross-covariance structure with $\alpha_{ii'} = \exp\{-|t_i - t_{i'}|/\psi\}$, where t_i is the time at which the ith survey is taken. The processes $B_i(x)$ in (2.1) are assumed to be independent, with zero mean and covariance functions $\text{Cov}\{B_i(x), B_i(x')\} = \nu_i^2 \rho(x, x'; \delta_i)$. Finally, the random variables Z_{ij} are again assumed to be mutually independent and Normally distributed with common mean 0 and variances τ_i^2.

As already noted, when all surveys are taken at the same time $S_i(x) = S_1(x)$ for all i, which formally corresponds to $\alpha_{ii'} = 1$ for all (i, i'). When all surveys are unbiased but are taken at different times, $\mathcal{B}$ is the empty set and the terms $[B_i(x_{ij}) + d(x_{ij})^\top \beta_i]$ in (2.1) are omitted; formally, this corresponds to $\nu_i^2 = 0 : i = 2, \ldots, r$. If it is appropriate to use different explanatory variables to model the true prevalence and the bias, this is accommodated by setting some elements of the β_i to zero. The dependence structure of the model is illustrated by the directed acyclic graph in Figure 2.1 for the special case of two gold-standard surveys conducted at two different times and a biased survey at the second time period. This scenario corresponds to the case study analysed in Section 2.5, where the aim is predictive inference for $S_2(x)$. In this case, the potential gains in efficiency by jointly modelling the data from all three surveys stem from the direct links between S_2 and both Y_2 and Y_2^* and the indirect link between S_2 and Y_1 via S_1.

2.3 Inference

In this section, we focus on the case $r = 2$. The generalization to more than two surveys is straightforward. We set $B_1(x) = 0$, write $B(x)$ in place of $B_2(x)$ and write the parameters of this bivariate version of (2.1) as $\beta^\top = (\beta_1^\top, \beta_2^\top)$ and $\theta^\top = (\sigma^2, \nu^2, \tau_1^2, \tau_2^2, \phi, \delta, \alpha)$.

2.3.1 Likelihood

Let $y_i^\top = (y_{i1}, \ldots, y_{in_i})$ denote the outcome data from surveys $i = 1, 2$ and let D_i be the n_i by p matrix whose jth row contains the values $d(x_{ij})^\top = (d_1(x_{ij}), \ldots, d_p(x_{ij}))$. Similarly, let T_i denote the vector of the n_i values of the linear predictor for survey i, hence $T_i = D_i\{\beta_1 + I(i = 2)\beta_2\} + W_i$, where $W_i^\top = (W_{i1}, \ldots, W_{in_i})$ and

$$W_{ij} = S_i(x_{ij}) + I(i = 2)B(x_{ij}) + Z_{ij}. \tag{2.2}$$

Now, let T denote the $(n_1 + n_2)$-element vector $T^\top = (T_1^\top, T_2^\top)$ and D the $(n_1 + n_2)$ by $2p$ matrix,

$$D = \begin{bmatrix} D_1 & 0 \\ D_2 & D_2 \end{bmatrix}. \tag{2.3}$$

Also, write $R_{ii'}(\phi)$ for the n_i by $n_{i'}$ matrix with (h, k)th element $\rho(x_{ih}, x_{i'k}; \phi)$ and $R_b(\delta)$ for the n_2 by n_2 matrix with (h, k)th element $\rho(x_{2h}, x_{2k}; \delta)$. Then,

$$T \sim \mathrm{MVN}(D\beta, V(\theta)) \tag{2.4}$$

where

$$V(\theta) = \begin{bmatrix} \sigma^2 R_{11}(\phi) + \tau_1^2 I & \sigma^2 \alpha R_{12}(\phi) \\ \sigma^2 \alpha R_{21}(\phi) & \sigma^2 R_{22}(\phi) + \nu^2 R_b(\delta) + \tau_2^2 I \end{bmatrix}. \tag{2.5}$$

The conditional distribution of Y given $T = t$ is a product of independent binomial probability mass functions. We write this as

$$f(y|t) = \prod_{i=1}^{2} \prod_{j=1}^{n_i} f(y_{ij}|t_{ij}). \tag{2.6}$$

Combining (2.3), (2.4), (2.5) and (2.6) then gives the likelihood function as the high-dimensional integral

$$L(\beta, \theta) = \int h(t; D\beta, V(\theta)) f(y|t)\, dt, \tag{2.7}$$

where $h(\cdot|\mu, V)$ is the density function of a multivariate Normal distribution with mean μ and covariance matrix V.

2.3.2 Conditional simulation

We propose to use Monte Carlo methods to evaluate the high-dimensional integral in (2.7). These methods require us to simulate from the conditional distribution of the spatial random effect T given the data $Y = y$. Using Bayes' formula, this conditional density is

$$\pi(t|y) \propto h(t|D\beta, V(\theta)) f(y|t). \tag{2.8}$$

To simulate from (2.8), Christensen, Roberts, and Sköld (2006) propose a Langevin-Hastings (LH) Markov chain Monte Carlo (MCMC) algorithm. This operates by updating a linear transformation of T, chosen to make the components of $T|y$ approximately independent. Christensen, Roberts, and Sköld (2006) use a Gaussian approximation to the distribution of $T|y$, with mean $D\beta$ and covariance matrix

$$\tilde{V} = \{V(\theta) + \Lambda(\hat{t})\}^{-1}. \tag{2.9}$$

In (2.9), $\Lambda(t)$ is a diagonal matrix with entries $-\partial^2/\partial t_i^2 \log f(y|t)$ and $\hat{t}$ is a typical value of T such as the mode of $f(y|t)$. For the binomial model with logistic link, this gives $\Lambda(\hat{t}) = \mathrm{diag}\{y_i(1 - y_i/m_i)\}$. Christensen, Roberts, and Sköld (2006) demonstrate that updating the centred random variable $\tilde{T} = \tilde{V}^{-1/2}(T - D\beta)$ gives better mixing and convergence properties than the analogous MCMC algorithms based on either T or on $\bar{T} = V^{-1/2}(T - D\beta)$, as suggested by Christensen and Waagepetersen (2002).

2.3.3 Monte Carlo Maximum Likelihood: estimation and spatial prediction

The Monte Carlo Maximum Likelihood (MCML) method (Geyer and Thompson, 1992; Geyer, 1994; Geyer, 1996; Geyer, 1999) uses conditional simulations of T given Y to obtain a computationally efficient approximation to the intractable likelihood function. From (2.7), the likelihood function can be written as

$$\begin{aligned} L(\beta,\theta) &= \int h(t|D\beta,V(\theta))f(y|t)\,dt = \int \frac{h(t|D\beta,V(\theta))f(y|t)}{\tilde{f}(y,t)}\tilde{f}(y,t)\,dt \\ &\propto \int \frac{h(t|D\beta,V(\theta))f(y|t)}{\tilde{h}(t)f(y|t)}\tilde{f}(t|y)\,dt = E_{\tilde{f}}\left[\frac{h(t|D\beta,V(\theta))}{\tilde{h}(t)}\right]. \end{aligned} \tag{2.10}$$

In (2.10), $\tilde{f}(t,y) = f(y|t)\tilde{h}(t)$, where $\tilde{h}(t)$ is any fixed density function with support in $\mathbb{R}^n$, and $E_{\tilde{f}}$ denotes expectation with respect to $\tilde{f}(\cdot|y)$. MCML estimates are then obtained by maximizing

$$L_m(\beta,\theta) = \frac{1}{m}\sum_{h=1}^{m}\frac{h(t_h|D\beta,V(\theta))}{\tilde{h}(t_h)}, \tag{2.11}$$

where $t_1,\ldots,t_m$ are samples from $\tilde{f}(\cdot|y)$.

The accuracy of the approximation for a given value of m depends critically on the choice of $\tilde{h}(\cdot)$. A suitable choice is $h(t_h|D\beta_0,V(\theta_0))$, where β_0 and θ_0 are as close as possible to the maximum likelihood estimates, $\hat{\beta}$ and $\hat{\theta}$. In practice, we embed the maximisation of $L_m(\beta,\theta)$ within the following iterative procedure as suggested in Geyer and Thompson (1992) and Geyer (1994): let $(\hat{\beta}_1,\hat{\theta}_1)$ denote the values that maximise $L_m(\beta,\theta)$ using an initial guess at suitable values (β_0,θ_0); repeat the maximisation with $(\hat{\beta}_1,\hat{\theta}_1)$ replacing (β_0,θ_0); continue until convergence.

For the numerical maximization of (2.11) we use a similar procedure to the one presented in Christensen (2004). Write $V(\theta) = \sigma^2 V(\psi)$ where $\psi = (\nu^2/\sigma^2, \tau_1^2/\sigma^2, \tau_2^2/\sigma^2, \phi, \delta, \alpha)^\top$ For a given value of ψ, the first and second derivatives of (2.11) with respect to β and σ^2 are analytically tractable and we use an iterative Newton-Raphson algorithm. We then plug into (2.11) the values $\hat{\beta}(\psi)$ and $\hat{\sigma}(\psi)^2$ and maximize with respect to ψ using direct numerical optimization with a further re-parameterisation to remove any restrictions on the permissible ranges of the parameters; we use a log-transformation for all elements of ψ except α, for which we use $\log\{(1+\alpha)/(1-\alpha)\}$ to correspond to the range $-1 < \alpha < 1$. We also consider a variety of starting values to guard against false convergence to either a local maximum or an arbitrary point on a plateau of the likelihood surface.

We now consider the prediction of $T^{*\top} = (T(x_{n+1}), \ldots, T(x_{n+q}))$ at q additional prediction locations that are not included in any of the prevalence surveys. This requires all relevant explanatory variables to be available at the prediction locations. We include the mutually independent random variables Z_{ij} in (2.2) as part of our target for prediction. Note that in a linear Gaussian geostatistical model, the Z_{ij} would be conflated with Normally distributed measurement errors, whereas in a GLGM for prevalence survey data the analogue of measurement error is binomial sampling variation and is formally distinguishable from the extra-binomial variation induced by the Z_{ij}.

Zhang (2002) gives approximate expressions for the minimum mean square predictor $\mathrm{E}[T^*|y]$ and its variance using samples from the conditional distribution of $T|y$ generated by conditional simulation. For prediction of non-linear functionals of the prevalence surface, we first use our MCMC algorithm to generate samples $t_h : h = 1, ..., m$ from the conditional distribution of $T|y$, then simulate samples $t_h^* : h = 1, ..., m$ directly from the multivariate Normal conditional distribution of $T^*|T = t_h$. This has expectation

$$D^*\beta + C^\top V^{-1}(t_h - D\beta), \tag{2.12}$$

where D^* is the matrix of covariates at the prediction locations, and covariance matrix

$$V^* - C^\top V^{-1} C, \tag{2.13}$$

where V^* is the covariance matrix of T^* and C is the cross-covariance matrix between T and T^*. Finally, we transform the sampled values t_h^* to predicted prevalences,

$$p_h^* = g^{-1}(t_h^*)^\top = (g^{-1}(t_{n+1,h}^*), \ldots, g^{-1}(t_{n+q,h}^*)),$$

where $g^{-1}(\cdot)$ is the inverse link function. Typically, the prediction locations will form a fine grid to cover the area of interest, A, so as to approximate a set of predicted surfaces, $\mathcal{P}^* = \{p_h^*(x) : x \in A\}$ which can then be summarised according to the needs of each application. For example, we might want to map pointwise means, or selected quantiles, or predictive probabilities of the exceedance of policy-relevant thresholds.

2.4 Simulation study

We have conducted a simulation study of our proposed methodology with three aims: to show that the parameters in (2.1) are identifable; to illustrate the finite sample properties of the MCML estimators; and to demonstrate the potential gains in predictive performance that can be obtained by combining data from unbiased and biased surveys.

Throughout, we consider a generalized linear mixed model with a binomial response and logistic link.

2.4.1 Identifiability and finite sample properties

For this part of the simulation study we simulated data from two surveys, the first of which was unbiased, the second biased. We specified the covariance structure of the model to correspond to the MCML estimates that were obtained in the analysis of malaria prevalence data to be reported in Section 2.5. We also used the same sample sizes as in the malaria application, hence $n_1 = 425$ (to correspond to the second of the two randomised surveys) and $n_2 = 249$ (to correspond to the convenience survey), and the same binomial denominators m_{ij}. We specified constant means β_1 for the first survey and $\beta_1 + \beta_2$ for the second survey. We generated the sampling locations for the unbiased survey as an independent random sample from the uniform distribution in the rectangle $[34.700, 34.900] \times [-16.170, -15.880]$. The usefulness of the data from the biased survey may depend on the degree of overlap between the two sets of sampling locations. For this reason we generated the sampling locations for the biased survey from each of three inhomogeneous Poisson processes, with intensity $\lambda(x) = \exp\{-\|x - x_0\|/0.02\}$ and x_0 set to each of the three following locations: $(34.800, -16.025)$, the center of the the rectangle; $(34.700, -16.170)$, the lower left corner of the rectangle; $(34.600, -16.315)$, a point outside the rectangle. Figure 2.2 shows an example of simulated locations for the biased survey under each of these three scenarios. For each simulation we computed the mean and relative bias of the MCML estimates of the covariance parameters and the eigenvalues of their correlation matrix, based on 1000 replications of each of the three scenarios. The results are shown in Table 2.1. The estimates of β_1, β_2, σ^2, τ^2 and ϕ are approximately unbiased under all three scenarios whereas the estimates of ν^2 and δ, which relate to the process $B(x)$, become increasingly biased as the overlap between

Table 2.1: Estimated means and relative biases (RB) of the MCML estimators for the covariance parameters, and ordered eigenvalues (EV) of their correlation matrix under three scenarios.

		(1)			(2)			(3)		
	True value	Mean	RB	EV	Mean	RB	EV	Mean	RB	EV
β_1	1.000	0.997	-0.003	1.677	1.011	0.011	1.677	0.998	-0.002	1.811
β_2	-1.000	-1.011	0.011	1.287	-1.013	0.013	1.425	-0.980	-0.020	1.472
σ^2	2.186	2.132	-0.025	1.173	2.093	-0.042	1.298	2.005	-0.083	1.141
τ^2	0.558	0.465	-0.166	0.903	0.476	-0.148	0.840	0.486	-0.130	0.835
ν^2	0.672	0.900	0.339	0.772	1.011	0.504	0.715	1.193	0.776	0.806
ϕ	0.017	0.016	-0.045	0.695	0.016	-0.033	0.577	0.016	-0.085	0.503
δ	0.004	0.005	0.249	0.492	0.006	0.496	0.468	0.008	1.037	0.433

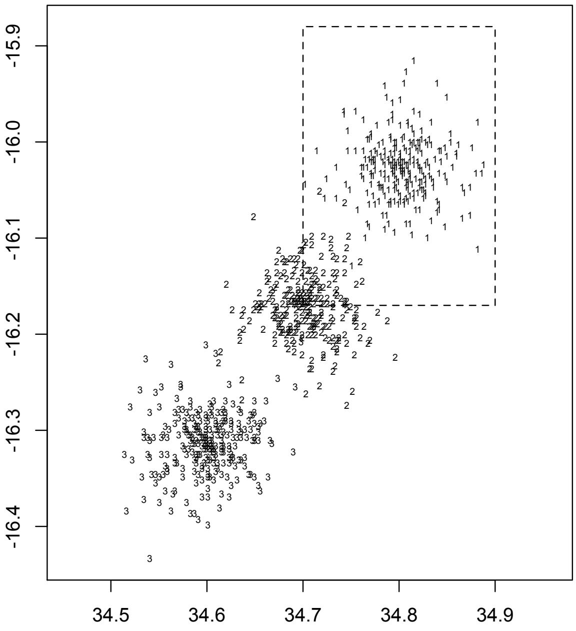

FIGURE 2.2: Example of simulated locations from a biased survey under three different scenarios as defined in Section 2.4.1; the dashed lines encompass the region within which locations of an unbiased survey are uniformly generated.

the two sampled areas decreases. Under all three scenarios, the smallest eigenvalue of the correlation matrix corresponds to about 6% of its total variation as measured by the sum of the eigenvalues. Also, the off-diagonal elements of the correlation matrix are never greater than 0.47 in absolute value, this largest value representing the correlation between the estimates of τ^2 and ϕ in the third scenario.

The overall conclusion from this part of the simulation study is that all of the model parameters are identifiable, and that the parameter estimates are approximately unbiased provided that there is a substantial overlap in the spatial coverage of the unbiased and biased surveys. This is as expected, because without such overlap the two surveys can only estimate the properties of the sum, $S(x) + B(x)$, in the area covered by the biased survey.

2.4.2 Quality variation and temporal variation

In this part of the simulation study we focus on predictive performance. Our main objective is to indicate to what extent the inclusion of data from a biased survey can improve predictive inference, under circumstances similar to those that hold in our malaria application. A secondary objective, as suggested by a reviewer, is to demonstrate the unreliability of a naive analysis that ignores bias and temporal variation. We therefore conducted three analyses of each simulated data-set as follows.

- Joint (J). The combined data are analysed using the bivariate GLGM as specified in Section 2.2.
- First-survey-only (FSO). Only the data from the first, unbiased survey are used.
- "Naive" (N). The data from the two surveys are analysed using a GLGM that does not account for bias or temporal variation.

We consider a *quality variation* (QV) scenario, in which one survey is unbiased and the other biased, and a *temporal variation* (TV) scenario, in which both surveys are unbiased but at different times, with predictions required for the first time period.

The following features are common to both scenarios. The processes $S_i(x) : i = 1, 2$ have mean $\beta_1 = 1$, variance $\sigma^2 = 1$ and correlation function $\rho(u) = \exp(-u/\phi)$ with $\phi = 0.15$. Locations of unbiased surveys are uniformly generated in the unit square centred on $x_0 = (1/2, 1/2)$. Both surveys have the same number of sampling locations, $n_1 = n_2 = 300$. The binomial denominators at each sampling location are all set equal to 1. Our primary focus is on prediction of prevalence at x_0 but we also consider estimation of the parameters β_1, $\log \sigma^2$ and $\log \phi$ that define the model for the underlying prevalence process $S_1(x)$.

In the QV scenario, $S_1(x) = S_2(x)$ for all x and the process $B_2(x)$ has mean $\beta_2 = -1$ and correlation function $\rho(u) = \exp(-u/\delta)$ with $\delta = 0.15$. Locations from the biased survey are generated from a Poisson process with intensity $\lambda(x) = \exp\{-\|x - x_0\|/0.15\}$ so that points closer to x_0 are more likely to be sampled, as might occur when using a convenience sampling strategy and x_0 is the location of a health-care facility. Finally, we consider four values, $\nu^2 = 0.5, 1, 2, 4$, for the variance of the process $B_2(x)$, corresponding to increasingly severe spatial variation in the bias.

In the TV scenario, the cross-correlation function between $S_1(x)$ and $S_2(x)$ is $\alpha \exp(-u/\phi)$. We consider three values, $\alpha = 0.2, 0.5, 0.8$, to correspond to weak, moderate and strong correlation between the two prevalence surfaces.

The results are summarised in Tables 2.2 and 2.3. These show estimates of the root-mean-square-error (RMSE) and coverage of nominal 95% confidence intervals (CIC) for MCML estimates of the parameters β_1, $\log \sigma^2$ and $\log \phi$, and for the minimum mean square error predictors of $S_1(x_0)$ and $\beta_1 + S_1(x_0)$. Each entry is calculated from 1000 independent replicates of the simulation model.

Table 2.2: Estimated RMSE, bias, SD and 95% CIC for the MCML estimates of β_1, $\log \sigma^2$, $\log \phi$, for the minimum mean square error predictor of $S_1(x_0)$ at location x_0 and $\beta_1 + S(x_0)$, under QV scenarios.

Model	Parameter	RMSE				CIC			
		ν^2				ν^2			
		0.5	1	2	4	0.5	1	2	4
J	β_1	0.37	0.36	0.36	0.36	0.95	0.95	0.95	0.94
FSO		0.37	0.36	0.35	0.35	0.95	0.95	0.95	0.95
N		0.50	0.52	0.51	0.48	0.82	0.79	0.83	0.77
J	$\log \sigma^2$	0.94	0.60	0.63	1.33	0.99	0.94	0.95	0.99
FSO		1.08	1.14	1.04	0.99	0.97	0.97	0.97	0.97
N		0.78	0.49	0.52	0.54	0.99	0.96	0.95	0.91
J	$\log \phi$	0.84	0.81	0.98	0.95	0.92	0.90	0.92	0.92
FSO		1.45	1.42	1.32	1.44	0.94	0.94	0.92	0.94
N		0.69	0.68	0.66	0.70	0.92	0.90	0.88	0.83
J	$S_1(x_0)$	0.80	0.79	0.88	0.86	0.95	0.95	0.95	0.95
FSO		0.91	0.86	0.92	0.85	0.95	0.95	0.95	0.94
N		0.93	0.96	1.13	1.21	0.80	0.80	0.77	0.76
J	$\beta_1 + S_1(x_0)$	0.72	0.73	0.82	0.83	0.95	0.95	0.95	0.95
FSO		0.83	0.80	0.84	0.81	0.95	0.94	0.95	0.94
N		1.10	1.15	1.33	1.35	0.78	0.78	0.75	0.74

Table 2.3: Estimated RMSE, bias, SD and 95% CIC for the MCML estimates of β_1, $\log \sigma^2$, $\log \phi$, for the minimum mean square error predictor of $S_1(x_0)$ at location x_0 and $\beta_1 + S(x_0)$, under TV scenarios.

Model	Parameter	RMSE			CIC		
		α			α		
		0.2	0.5	0.8	0.2	0.5	0.8
J	β_1	0.35	0.35	0.35	0.95	0.93	0.94
FSO		0.36	0.36	0.35	0.94	0.93	0.93
N		0.63	0.63	0.64	0.29	0.38	0.42
J	$\log \sigma^2$	0.60	0.68	0.69	0.94	0.94	0.93
FSO		1.04	1.46	1.47	0.97	0.96	0.96
N		1.18	0.95	0.87	0.91	0.92	0.95
J	$\log \phi$	0.93	0.92	0.91	0.91	0.93	0.93
FSO		1.47	1.56	1.75	0.92	0.93	0.94
N		1.32	1.09	1.02	0.92	0.94	0.94
J	$S_1(x_0)$	1.37	1.30	1.28	0.95	0.95	0.94
FSO		1.37	1.30	1.31	0.94	0.93	0.92
N		1.39	1.32	1.28	0.87	0.91	0.92
J	$\beta_1 + S_1(x_0)$	1.34	1.28	1.26	0.96	0.96	0.94
FSO		1.35	1.27	1.27	0.95	0.95	0.94
N		1.50	1.39	1.36	0.86	0.90	0.91

Overall, J outperforms FSO, which in turn outperforms N. Under the QV scenario, the main benefits of J are in the prediction of $S_1(x_0)$ for values of ν^2 smaller than 4. The N approach yields much higher values of RMSE for the estimates of β_1, $S_1(x_0)$ and $\beta_1 + S_1(x_0)$ and very poor CIC. Under the TV scenario, the biggest gains achieved by J over FSO are in estimating the parameters $\log \sigma^2$ and $\log \phi$. Both J and FSO perform similarly with respect to prediction of $S_1(x_0)$ and $\beta_1 + S_1(x_0)$. The N approach, which in this scenario consists of combining the data under the assumption that $S_1(x) = S_2(x)$ for all x, i.e. $\alpha = 1$, has the worst performance.

2.5 Application: malaria prevalence mapping

In this Section, we use our proposed methodology to construct malaria prevalence maps for an area of Malawi by combining information from three surveys. All three surveys were directed at the same target population, covering a 400 square km area within Chikhwawa District, Southern Malawi. Two of the surveys were "rolling" Malaria Indicator Surveys (MIS) (Roca-Feltrer et al., 2012), that used two different practical strategies to obtain random, and therefore unbiased, samples from the population at risk. The third was a facility-based survey that used a convenience sampling strategy, in which recruitment took place at a central child-vaccination clinic at the main hospital in the centre of the study area. We refer to this as the Easy Access Group (EAG) study. All three surveys recorded the numbers of participating children from each community and the number of those who tested positive using a rapid diagnostic test (RDT) for malaria parasites.

2.5.1 Data

Two population-level continuous malaria indicator surveys were conducted over the period May 2010 to April 2012. Both surveys recruited children aged less than five years in a sample of 50 village communities in order to monitor the malaria intervention coverage and childhood burden of malaria in a designated area containing the sampled villages, which was chosen to represent the catchment area of the Chikhwawa District Hospital (CDH).

The two surveys differed in the sampling strategy used, as described below. We refer to these two surveys as the rMIS, covering the period May 2010 to April 2011 and the eMIS, covering the period May 2011 to April 2012. Throughout the two-year period seven or eight villages were randomly selected per month so as to sample all 50 villages twice yearly, once during the high-transmission season and once during the low-transmission

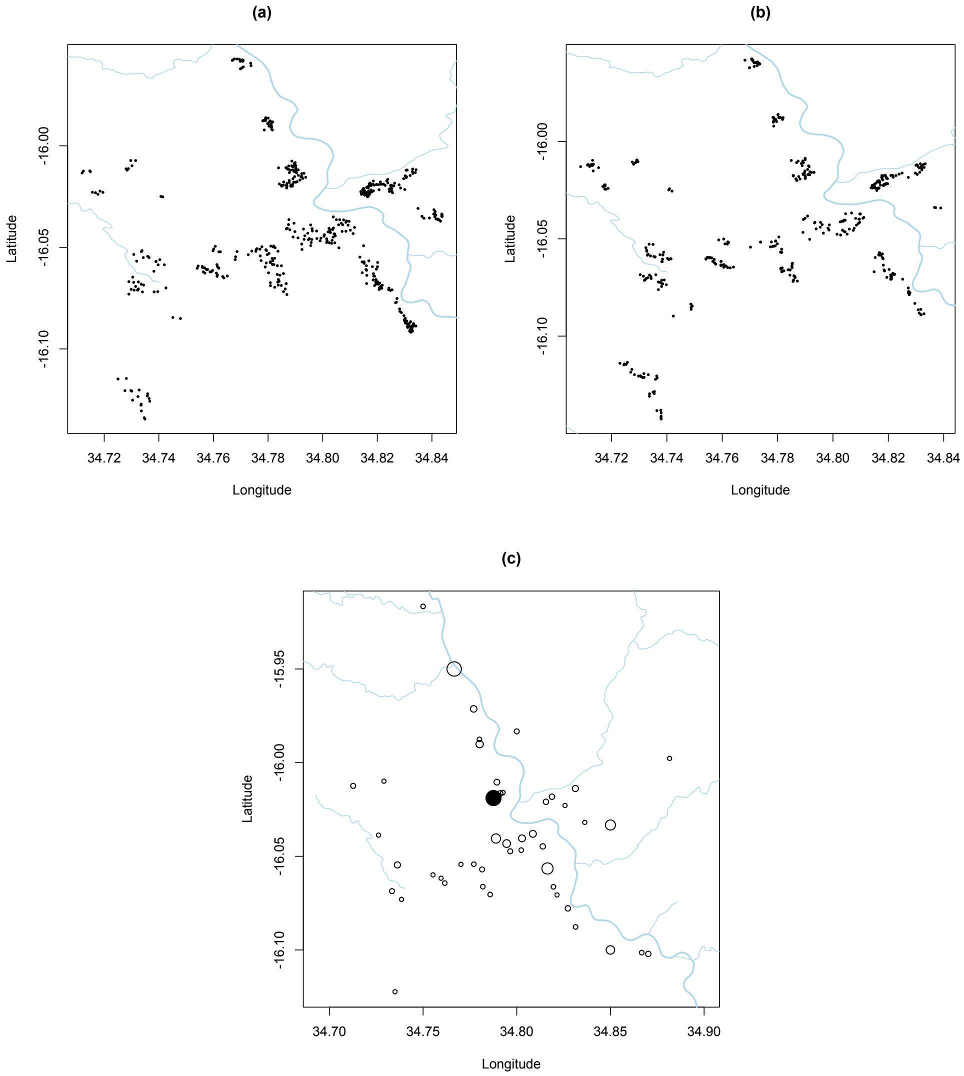

FIGURE 2.3: Sampled locations for (a) rMIS, (b) eMIS and (c) EAG. Coordinates are of individual houses for rMIS and eMIS, and of the villages for EAG; in (c), the radius of each circle is proportional to the number of the sampled households from the respective village and the black solid point corresponds to CDH village. The light blue lines represent waterways, with the thicker line corresponding to Shire river.

season. Within sampled villages, selection of households was as follows. In the rMIS, households were randomly selected within each village from a list of households, with sampling probability proportional to village population size, based on a population enumeration exercise. In the eMIS, a more economical "spin-the-bottle" method was used to identify a random set of households within villages. A bottle was placed in the center of a village and used to select random directions. A virtual line was drawn in each chosen direction to the border of the village, the households that intersected this line were counted, and from these a random household number was chosen as the starting point. The number of houses selected within each village was proportional to the estimated village population size. Figures 2.3 (a)-(b) show the sampled locations for the rMIS and the eMIS.

The third survey is a continuous facility-based MIS in children attending the immunization clinic at the CDH, conducted from May 2011 to April 2012. The objective of this study was to determine if estimates of uptake of control interventions and the burden of malaria from convenience sampling were comparable to those from a randomised MIS conducted within the same catchment area of CDH. Children from 3 months of age who attended the vaccination clinic, and any accompanying sibling below 5 years, were recruited. Between 30 and 50 children were recruited per month. Village of origin was extracted by direct questioning. If the village was not one of the 50 eMIS/rMIS villages for which the location was already known, its coordinates were determined retrospectively. The results for villages within 15km of CDH were extracted to make the catchment area of the EAG comparable to that of the rMIS and eMIS. Malaria control efforts by the national control program during the first period included a district-wide household indoor residual spraying campaign between February and April 2011. Practical difficulties resulted in this campaign being conducted at the end of the rainy season rather than, as would have been ideal, before the start of the rainy season. This will have reduced its potential impact. Insecticide-treated net control efforts were stable over the three months of the campaign, with distribution to women attending antenatal clinics and mother and child clinics.

2.5.2 Results

The response from each child was a binary indicator of the outcome of the RDT used to test for the presence of malaria from a finger-prick blood sample. Six explanatory variables were considered, as defined in Table 2.4. Socio-Economic-Status (SES), an indicator of household wealth taking discrete values from 1 (poor) to 5 (wealthy), was derived by an application of principal component analysis as discussed in Vyas and Kumuranayake (2006).

Table 2.4: Explanatory variables used in the analysis of the Chikhwawa malaria prevalence surveys

1	intercept
2	at least one treated bed-net in the household (yes/no)
3	indoor residual spraying in the past two months (yes/no)
4	high-transmission season (January-June/July-December)
5	distance from the closest waterway (km)
6	Socio-Economic-Status (SES, 1 to 5)

It was thought that health facility utilization might be associated with SES as previously observed in Gahutu et al. (2011), where children with relatively high SES were more likely to attend a CDH. Table 2.5 shows the average SES observed in each of our three surveys. Enrolled children in the EAG study show a higher average SES than those in the two other surveys. Additionally, Table 2.6 shows that the relationship between SES and the distribution of the number of RDT positive results per household differs between the two gold-standard surveys and the convenience survey. We therefore allowed SES to have a direct effect on the spatially structured bias of the EAG survey in addition to its possible association with prevalence.

Table 2.5: Mean and standard deviation (SD) of SES in the three surveys.

	SES		
	rMIS	eMIS	EAG
Mean	2.76	2.50	3.45
SD	1.45	1.37	1.39

Table 2.6: Distribution (percentage) of the number of positive RDTs per household for each value of SES, in the convenience survey (EAG, left-columns) and in the gold-standard surveys (rMIS and eMIS, right columns)

		SES (EAG)					SES (rMIS and eMIS)				
		1	2	3	4	5	1	2	3	4	5
	0	75.76	80.56	77.50	79.10	89.04	54.58	63.09	71.72	73.29	83.74
RDT	1	21.21	19.44	20.00	20.90	10.96	40.49	33.56	25.25	22.60	15.45
positives	2	3.03	0.00	2.50	0.00	0.00	4.58	3.35	3.03	3.42	0.81
	3	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.69	0.00

The resulting model for the combined data therefore included seven regression parameters, $\beta_1, \beta_2, ..., \beta_7$. Let $\beta^\top = (\beta_1, \ldots, \beta_6)$ and denote by $d(x_{ij})$ the vector of covariates associated with location x_{ij}. Use $i = 1, 2, 3$ to denote rMIS, eMIS and EAG, respectively. Then, the linear predictor is

$$\begin{aligned}\eta_{ij} &= d(x_{ij})^\top \beta + S_i(x_{ij}) + I(i = 3)[B(x_{ij}) + \beta_7 \mathrm{SES}_{ij}] + Z_{ij},\\ & \quad i = 1, \ldots, 3; j = 1, \ldots, n_i,\end{aligned}$$

where $n_1 = 475$, $n_2 = 425$ and $n_3 = 249$. Note that in the joint model for $S_1(x)$, $S_2(x)$ and $S_3(x)$, $\alpha_{23} = 1$ because the EAG study took place over the same period as the eMIS. We therefore use α to denote α_{12} and set $S_3(x) = S_2(x)$. We also assume equal variances τ^2 for the nugget term Z across all three surveys. Finally, we define $\mathrm{Cov}\{S_1(x), S_2(x')\} = \sigma^2 \alpha \exp\{-\|x - x'\|/\phi)$ where $\sigma^2 > 0$, $\phi > 0$ and $-1 < \alpha < 1$.

Table 2.7: Monte Carlo maximum likelihood estimates and 95% confidence intervals.

Term	Estimate	95% confidence interval
β_1	-0.272	(-1.382, 0.862)
β_2	-0.439	(-0.623, -0.277)
β_3	-0.399	(-0.621, -0.189)
β_4	0.415	(0.206, 0.598)
β_5	-0.373	(-0.970, 0.116)
β_6	-0.151	(-0.233, -0.072)
β_7	-0.096	(-0.222, 0.021)
σ^2	2.186	(0.955, 3.155)
τ^2	0.558	(0.089, 1.231)
ν^2	0.672	(0.525, 0.802)
α	0.859	(0.483, 0.924)
ϕ	0.017	(0.006, 0.032)
δ	0.004	(0.001, 0.025)

Table 2.7 shows the Monte Carlo maximum likelihood estimates of the model parameters together with 95% confidence intervals. Each evaluation of the log-likelihood used 5000 simulated values, obtained by conditional simulation of 110000 values and sampling every 20th realization after discarding a burn-in of 10000 values. Figure 2.4 shows two diagnostic plots for the average random effect: convergence of the MCMC algorithm appears to be satisfactory.

The confidence intervals in Table 2.7 were calculated using the following parametric bootstrap procedure. Using the parameter estimates in Table 2.7 we simulated 1000 data-sets from the model, applied to each simulated data set the Monte Carlo maximum likelihood method with 5000 conditional simulations, and computed the empirical quantiles of the 1000 resulting estimates of each parameter. Although this procedure introduces additional Monte Carlo error, it allows us to compute confidence intervals without relying on questionable Normal approximations for the distribution of the Monte Carlo maximum likelihood estimates.

From Table 2.7, we see that the ownership of at least one treated bed net, the presence of residual indoor spraying and an increase in SES are all associated with a reduction in the prevalence of a positive RDT. The distance from the closest waterway is not significant, although the sign of the regression coefficient suggests that prevalence decreases with increasing distance. The period January to June, which is known to be a period of high malaria transmission, is associated with a significant increase in prevalence, by an estimated factor of $\exp(0.415) \approx 1.51$.

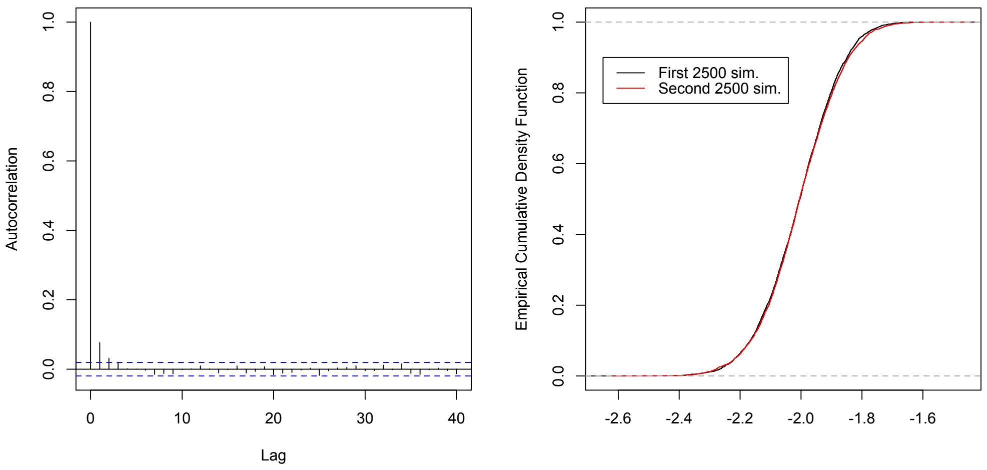

Figure 2.4: Diagnostic plots for the convergence of the posterior average of the random effect. Left panel: correlogram of the 5000 simulated values. Right panel: empirical cumulative density function for the first (black line) and second (red line) 2500 simulated values.

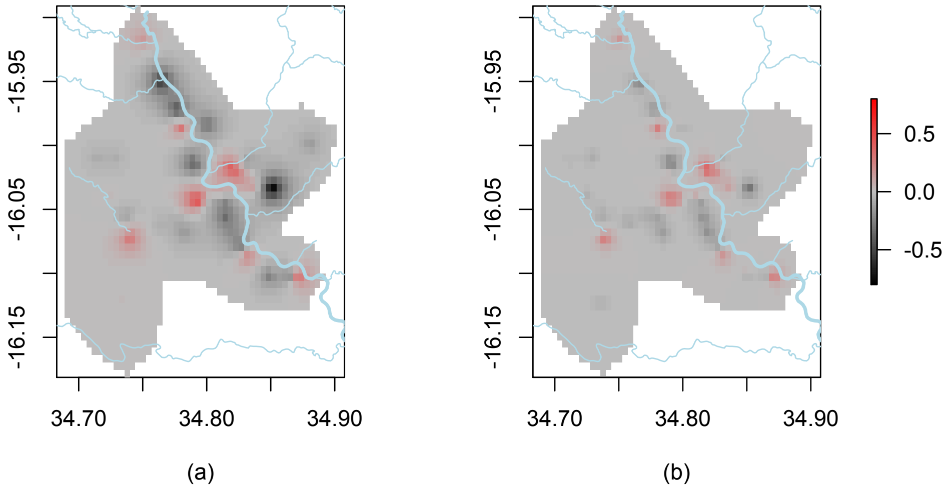

Figure 2.5: Predicted bias surface $B(x)$ (a) without the interaction term of SES and (b) including the effect of SES on the spatial bias.

The regression coefficient β_7, which represents the additional effect of SES on the bias of the EAG data, is not significant, but its inclusion nevertheless makes a noticeable difference to the predicted bias surface. Figures 2.5(a) and 2.5(b) show the minimum mean square error predictions of the bias with and without including the regression on SES.

The estimate $\hat{\alpha} = 0.859$, albeit with a wide confidence interval, indicates a strong correlation between prevalences in the two sampling periods, 2010-2011 and 2011-2012.

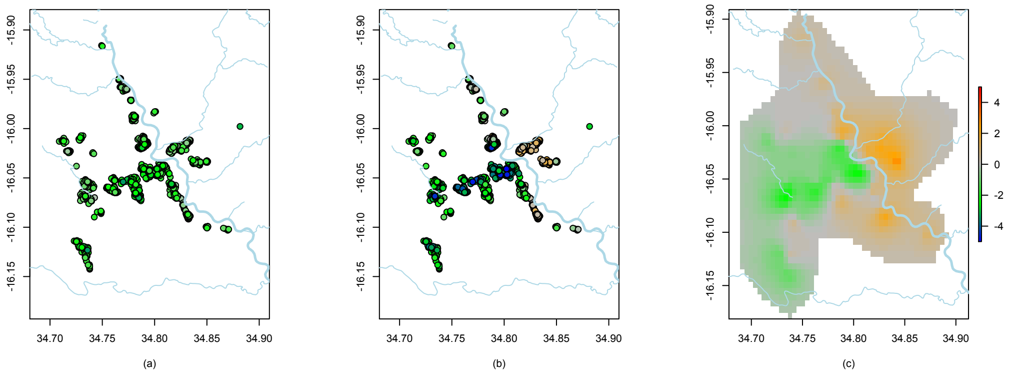

FIGURE 2.6: Predictions of (a) $d(x_{2j})^\top\beta$ and (b) $d(x_{2j})^\top\beta + S_2(x_{2j})$ at observed locations; (c) predicted surface of the unexplained spatial variation $S_2(x)$. The same colour scale has been used for the point predictions.

Figures 2.6(a) and (b) show the contributions of the linear regression and of the unexplained spatial variation to the predicted log-odds of prevalence at each of the observed locations. Figure 2.6(c) shows the unexplained component, $\hat{S}(x)$, of the predicted prevalence as a spatially continuous surface. The clear and substantial difference between adjacent areas to the east and west of the river Shire strongly suggests the existence of one, or more, social or environmental risk-factors that are not captured by the available explanatory variables.

Figure 2.7 shows pairwise scatter plots to compare the prediction standard deviations for $S_2(x)$ at the sampling locations. Figure 2.7 (a) shows that analysing rMIS and eMIS data in the joint model for temporal variation results in substantially better precision than using only the eMIS; Figures 2.7 (c) and (d) show the further, but more modest, gains resulting from addition of the data from the EAG; in contrast, Figure 2.7 (b) suggests little or no benefit from adding the EAG data to the eMIS data, with predictive standard deviations decreasing at some locations but increasing at others.

2.6 Discussion

We have developed a class of multivariate GLGMs for the combined data from multiple spatially referenced surveys, and associated Monte Carlo methods for maximum likelihood estimation and spatial prediction within the proposed class of models.

The model as defined by (2.1) is the minimally parameterised model that captures the essential features of our motivating application: variation in data-quality arising from non-randomised sampling; variation in prevalence over time; binomial and extra-binomial sampling variation. We have shown that all of the model parameters are identifiable from surveys of comparable size to the ones available to us for the application. If substantially

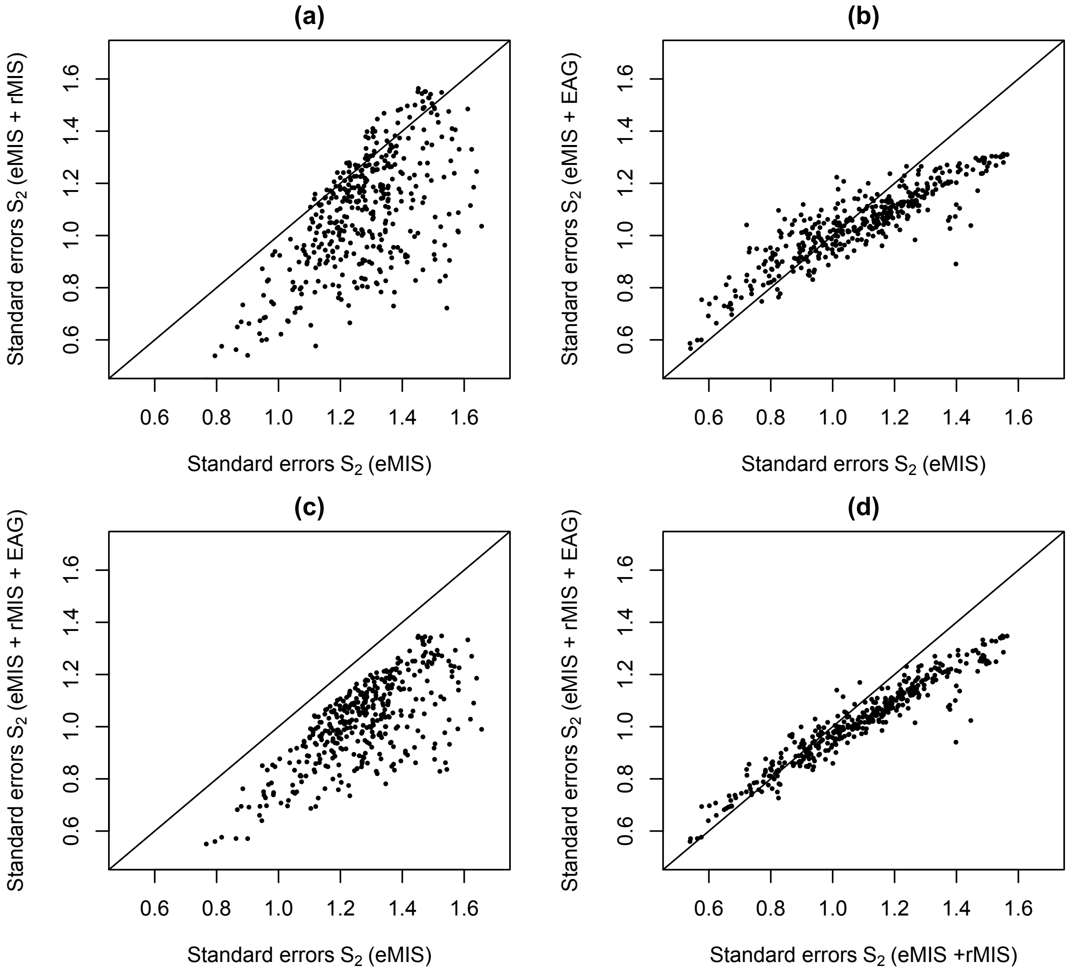

FIGURE 2.7: Scatter plots of the prediction standard errors for $S_2(x)$ at sampled locations x, using models fitted to the data from: (a) eMIS against eMIS and rMIS; (b) eMIS against eMIS and EAG; (c) eMIS, rMIS and EAG against eMIS; (d) eMIS, rMIS and EAG against eMIS and rMIS.. The solid line represents the identity line.

larger data-sets were available, it would be of interest to extend the model in various ways, for example by relaxing the assumption of common parameters for the prevalence surfaces $S_i(x)$ at different times or by allowing cross-correlation between the $S_i(x)$ and their paired bias surfaces $B_i(x)$. Additionally, if a large number of surveys were conducted at irregularly spaced time-points within partly overlapping time periods, the use of a structured spatio-temporally continuous process $S(x,t)$, as mentioned in Section 2.2, would be more appealing than a discrete set of processes $S_i(x)$ at specific times t_i.

The Monte Carlo maximum likelihood estimation procedure is computationally intensive, primarily because of the need to use parametric bootstrapping to compute standard errors reliably. For this reason, we are currently developing a much faster Monte Carlo method for approximate evaluation of the likelihood function.

In our application to malaria prevalence surveys, we combined data from three surveys, two of which were unbiased and conducted in two consecutive years,whilst the third was a potentially biased convenience survey conducted over the same time-period as the second unbiased survey. We obtained substantial gains in the precision of spatial predictions by combining the data from the two unbiased surveys and further, but smaller, gains from combining the data from all three surveys.

One of the limitations of our approach is that it assumes that at least one of the available surveys represents an unbiased gold-standard. This is a reasonable assumption when, as in our application, at least one of the surveys uses a properly randomised sampling scheme. When we cannot assume that one of the surveys is unbiased by design, it is difficult to see how any method could deliver reliable predictions without additional assumptions that would be difficult or impossible to validate empirically.

The problem that we have addressed in this paper is related to, but distinct from, the problem of preferential sampling as formulated in Diggle, Menezes, and Su (2010). In both settings, the goal is to predict the realisation of a latent spatial process $S(x)$ using data obtained by a potentially biased sampling scheme. In preferential sampling, the bias arises from a direct relationship between the value of $S(x)$ and the probability that the location x will be sampled. In the present paper, the bias is a function of the location x itself, rather than of the value of $S(x)$. In the context of disease prevalence mapping, properties of a location could be intrinsic to that location (e.g. height above sea-level), or to a person who happens to live at that location (e.g. age). In our application a relationship between a child's location x and the probability, say $p(X)$, that they present at the CDH would not result in bias unless at least one of the factors that affect $p(X)$ is both unmeasured and related to malaria risk. The bias surface $B(x)$ allows for the possibility that the sub-population of children who present at the CDH differs from the general population with respect to their exposure to unmeasured risk-factors for malaria.

Our approach is of potentially wide application to disease monitoring and control in low-resource settings, where registry data are typically not available. The ability to combine data from surveys that vary in their level of bias and timing can inform more accurate, local-area burden maps, allowing for improved risk stratification of high burden areas and identification of transmission hot-spots. For example, although substantial progress has been made over the past decade with malaria control by homogeneous scaling up of interventions at national level, it is increasingly recognized by funders and policy makers that a more targeted approach focused on high-burden areas or hot-spots may be more cost-effective. Furthermore, apart from its potential to optimize the use of available data, our approach can also inform improved prospective data collection, by using the fitted model in simulation studies to identify efficient prospective hybrid sampling approaches that combine convenience and random sampling strategies in ways that acknowledge and exploit spatial and/or temporal heterogeneity as revealed by analyses of the kind described in Section 2.5.

In conclusion, our proposed approach provides a way of making use of mixed source prevalence data to improve estimates of spatial predictions. These are urgently needed to support control programmes and develop more accurate local spatio-temporal risk stratification maps that can inform more targeted control efforts. Malaria is one of a number of diseases that bring a high public health burden in low-resource settings, whilst exhibiting highly heterogeneous distributions across space and time. Control of such diseases needs methods of continuous monitoring of prevalence and evaluation of control measures that make the best possible use of limited resources, and will therefore benefit greatly from the ability to combine national household surveys with more local convenience sampling strategies without compromising the validity of the resulting prevalence estimates.

Acknowledgements

We thank the participants and staff involved in the data collection of the presented household prevalence surveys, especially Dr Roca-Feltrer who oversaw the survey field teams between 2010 and 2011

Funding

Emanuele Giorgi holds an ESRC-NWDTC funded Ph.D. studentship. Dr Sanie Sesay holds an MCDC funded Ph.D. studentship. Dr Dianne Terlouw acknowledges support from the ACT consortium for the presented household prevalence surveys from Malawi. Prof Peter Diggle is supported by the UK Medical Research Council (grant number G0902153).

References

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

Christensen, O. F., G. O. Roberts, and M. Sköld (2006). “Robust Markov Chain Monte Carlo methods for Spatial Generalized Linear Mixed Models”. In: *Journal of Computational and Graphical Statistics* 15, pp. 1–17.

Christensen, O. F. and R. P. Waagepetersen (2002). “Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models”. In: *Biometrics* 58, pp. 280–286.

Crainiceanu, C., P.J. Diggle, and B.S. Rowlingson (2008). “Bivariate modelling and prediction of spatial variation in *Loa loa* prevalence in tropical Africa (with Discussion)”. In: *Journal of the American Statistical Association* 103, pp. 21–43.

Diggle, P. J., R. Menezes, and T. Su (2010). “Geostatistical inference under preferential sampling”. In: *Journal of the Royal Statistical Society, Series C* 59, pp. 191–232.

Elliot, M. R. and W. W. Davis (2005). “Obtaning Risk Factor Prevalence Estimates in Small Areas: combining Data From Two Surveys”. In: *Journal of the Royal Statistical Society, Series C* 54, pp. 595–609.

Gahutu, J.-B., C. Steininger, C. Shyirambere, I. Zeile, N. Cwinya-Ay, I. Danquah, C. Larsen, T. Eggelte, A. Uwimana, C. Karema, A. Musemakweri, G. Harms, and F. Mockenhaupt (2011). “Prevalence and risk factors of malaria among children in southern highland Rwanda”. In: *Malaria Journal* 10.1, p. 134. DOI: 10.1186/1475-2875-10-134.

Geyer, C. J. (1994). “On the Convergence of Monte Carlo Maximum Likelihood Calculations”. In: *Journal of the Royal Statistical Society, Series B* 56, pp. 261–274.

Geyer, C. J. (1996). “Estimation and Optimization of Functions”. In: *Markov Chain Monte Carlo in Practice*. Ed. by W. Gilks, S. Richardson, and D. Spiegelhalter. London: Chapman and Hall, 241–âĂŞ258.

Geyer, C. J. (1999). “Likelihood Inference for Spatial Point Processes”. In: *Stochastic Geometry, Likelihood and Computation*. Ed. by O. E. Barndorff-Nielsen, W. S.Kendall, and M. N. M. van Lieshout. Boca Raton, FL: Chapman and Hall/CRC, 79–âĂŞ140.

Geyer, C. J. and E. A. Thompson (1992). “Constrained Monte Carlo Maximum Likelihood for Dependent Data”. In: *Journal of the Royal Statistical Society, Series B* 54, pp. 657–699.

Hedt, B. L. and M. Pagano (2011). “Health indicators: Eliminating bias from convenience sampling estimator”. In: *Statistics in Medicine* 30, pp. 560–568.

Lohr, S. L. and J. N. K. Rao (2006). “Estimation in Multiple-Frame Surveys”. In: *Journal of the American Statistical Association* 101, pp. 1019–1030.

Manzi, G., D. J. Spiegelhalter, R. M. Turner, J. Flowers, and S. G. Thompson (2011). “Modelling bias in combining small area prevalence estimates from multiple sruveys”. In: *Journal of the Royal Statistical Society, Series A* 174, pp. 31–50.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models*. Second. Chapman and Hall, London.

Moriarity, C. and F. Scheuren (2001). “Statistical Matching: A Paradigm for Assesing the Uncertainty in the Procedure”. In: *Journal of Official Statistics* 17, pp. 407–422.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raghunathan, T. E., D. Xie, N. Schenker, and V. L. Parsons (2007). “Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening”. In: *Journal of the American Statistical Association* 102, pp. 474–486.

Roca-Feltrer, A., D. Lalloo, K. Phiri, and D. J. Terlouw (2012). “Rolling Malaria Indicator Surveys (rMIS): a potential district-level malaria monitoring and evaluation (M&E) tool for programme managers”. In: *American Journal of Tropical Medicine and Hygiene* 86, pp. 96–98.

Turner, R. M., D. J. Spiegelhalter, G. C. S. Smith, and S. G. Thompson (2009). “Bias modelling in evidence synthesis”. In: *Journal of the Royal Statistical Society, Series A* 172, pp. 21–47.

Vyas, S. and L. Kumuranayake (2006). “Constructing socio-economic status indices: how to use principal component analysis”. In: *Health Policy Plan* 21, pp. 459–468.

Wanji, S., D. Akotshi, M. Mutro, F. Tepage, T. Ukety, P. Diggle, and J. Remme (2012). “Validation of the rapid assessment procedure for loiasis (RAPLOA) in the democratic republic of Congo”. In: *Parasites & Vectors* 5.1, p. 25. DOI: 10.1186/1756-3305-5-25.

Zhang, H. (2002). “On Estimation and Prediction for Spatial Generalized Linear Mixed Models”. In: *Biometrics* 58, pp. 129–136.

Chapter 3

Paper 2. The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control

H. G. M. Zouré[1], M. Noma[1], A. H. Tekle, U[1]. V. Amazigo[2], P. J. Diggle[3], E. Giorgi[3] and J. H. F. Remme[4]

[1] African Programme for Onchocerciasis Control, Ouagadougou BP 549, Burkina Faso

[2] Consultant, Box 3397, Main Post office, Enugu, Nigeria

[3] Lancaster Medical School, Lancaster University, Lancaster, UK

[4] Consultant, 120 Rue des Campanules, Ornex 01210, France

Summary

The original aim of the African Programme for Onchocerciasis Control (APOC) was to control onchocerciasis as a public health problem in 20 African countries. In order to identify all high risk areas where ivermectin treatment was needed to achieve control, APOC used Rapid Epidemiological Mapping of Onchocerciasis (REMO). REMO involved spatial sampling of villages to be surveyed, and examination of 30 to 50 adults per village for palpable onchocercal nodules. REMO has now been virtually completed and we report the results in two articles. A companion article reports the delineation of high risk areas based on expert analysis. The present article reports the results of a geostatistical analysis of the REMO data to map endemicity levels and estimate the number infected.

A model-based geostatistical analysis of the REMO data was undertaken to generate high-resolution maps of the predicted prevalence of nodules and of the probability that the true nodule prevalence exceeds the high risk threshold of 20%. The number infected was estimated by converting nodule prevalence to microfilaria prevalence, and multiplying the predicted prevalence for each location with local data on population density. The geostatistical analysis included the nodule palpation data for 14,473 surveyed villages.

The generated map of onchocerciasis endemicity levels, as reflected in the prevalence of nodules, is a significant advance with many new endemic areas identified. The prevalence of nodules wasâĂĹgreater than 20% over an area of 2.5 million km^2 with an estimated population of 62 million people. The results were consistent with the delineation of high risk areas of the expert analysis except for borderline areas where the prevalence fluctuated around 20%. It is estimated that 36 million people would have been infected in the APOC countries by 2011 if there had been no ivermectin treatment.

The map of onchocerciasis endemicity levels has proven very valuable for onchocerciasis control in the APOC countries. Following the recent shift to onchocerciasis elimination, the map continues to play an important role in planning treatment, evaluating impact and predicting treatment end dates in relation to local endemicity levels.

Keywords: Onchocerciasis; APOC; Onchocercal nodule; Mapping; REMO; Geostatistics; Endemicity level.

3.1 Background

Onchocerciasis, or river blindness, used to be endemic in some 30 countries in Africa where over 99% of all cases in the world were found (World Health Organization, 1995). The Onchocerciasis Control Programme in West Africa (OCP) has successfully controlled onchocerciasis by large scale vector control in the savanna belt of nine West African countries (Boatin, 2008). In the remaining endemic African countries, where some 85% of onchocerciasis cases lived, onchocerciasis control became feasible with the registration of ivermectin for the treatment of human onchocerciasis in 1987 and its donation free of charge for as long as needed (Tielsch and Beeche, 2004; Gustavsen, Hopkins, and Sauerbrey, 2011). Clinical and community trials demonstrated that annual ivermectin treatment could effectively control the disease (Tielsch and Beeche, 2004), and Non-Governmental Development Organizations initiated the first ivermectin distribution efforts (Bush and Hopkins, 2011). In 1995 the African Programme for Onchocerciasis Control (APOC) was created to support the establishment of community directed treatment with ivermectin (CDTI) in all remaining areas in Africa where onchocerciasis was a public health problem (Remme, 1995).

One of the first challenges for APOC was to determine where exactly onchocerciasis was a public health problem. The existing information on the geographic distribution of onchocerciasis in the 20 APOC countries (World Health Organization, 1995; Baker and Abdelnur, 1986; Zein, 1990) was incomplete and not reliable enough for targeting ivermectin treatment programmes, and there was an urgent need for comprehensive mapping of the geographic distribution of onchocerciasis in all potentially endemic countries in Africa outside the OCP (Remme, 1995; Noma et al., 2002). This was a vast area of some 14 million km^2 and the survey methods available were difficult to implement at such a large scale. In anticipation of this problem, the WHO Special Programme for Research and Training in Tropical Diseases (TDR) developed a rapid assessment method in 1993, Rapid Epidemiological Mapping of Onchocerciasis (REMO) (Ngoumou and Walsh, 1993). In REMO, sample villages are selected using a sampling methodology that takes the spatial epidemiology of onchocerciasis into account. Rapid assessment surveys are then carried out in the selected villages to estimate the prevalence of palpable onchocercal nodules as a proxy for the prevalence of onchocerciasis infection. Following its successful field testing in Cameroon and Nigeria (Ngoumou, Walsh, and Mace, 1994), APOC adopted REMO for large-scale mapping of onchocerciasis in all APOC countries in order to identify priority areas for CDTI. Large scale application of REMO started in 1996, and has since been applied in phase with the expansion of CDTI to cover all potentially endemic areas in APOC countries (Noma et al., 2002).

To date, virtually all potentially endemic areas in the 20 APOC countries have been mapped for onchocerciasis through REMO. In a companion paper we summarize the REMO surveys and show the results of an expert analysis that was undertaken to delineate high-risk areas where onchocerciasis was a major public health problem and where ivermectin treatment was a priority (World Health Organization, 2012). Based on these maps of high risk areas, CDTI projects were created that by 2012 were treating over 80 million people in the APOC countries (World Health Organization, 2012).

In the present article we report the results of a more advanced analysis of the REMO data using a model-based geostatistical methodology that has allowed a more effective utilization of the extensive REMO data. One important application was the mapping of the geographic distribution of onchocerciasis endemicity levels as reflected in the prevalence of palpable onchocercal nodules. Endemicity is a key concept in onchocerciasis epidemiology. The severity of the disease and the public health importance of onchocerciasis in a given area are directly related to the local level of endemicity (Murdoch et al., 2002; Remme et al., 1989). The endemicity level is also an important indicator of the local potential for transmission, as well as a predictor of the intensity and duration of interventions needed to control or eliminate onchocerciasis in an onchocerciasis focus (World Health Organization, 2010). It is therefore important for onchocerciasis control programmes to have a detailed map of onchocerciasis endemicity levels throughout their operational area.

In 1979, Prost, Hervouet, and Thylefors (1979) defined three levels of onchocerciasis endemicity in terms of the community prevalence of *Onchocerca volvulus* microfilaria in the skin: hyperendemic onchocerciasis (prevalence of microfilaria greater than 60%),where the disease is very severe and associated with onchocercal blindness rates in excess of 4 to 5% in the West African savanna; hypoendemic onchocerciasis (prevalence of microfilaria smaller than 35%) where ocular complications are rare and the disease is socially not apparent, and mesoendemic onchocerciasis (prevalence of microfilaria between 35% and 60%) where the disease pattern varied between these two extremes. The prevalence of nodules is related to the prevalence of skin microfilaria. Using the quantification of this relationship by Coffeng et al. (2013), the above endemicity classes translate into hyperendemic onchocerciasis for a prevalence of palpable nodules in adults greater than 45%, mesoendemic onchocerciasis for a nodule prevalence between 20% and 45%, and hypoendemicity for a prevalence of nodules smaller than 20%.

When ivermectin became available for onchocerciasis control, a WHO expert meeting recommended that in order to control onchocerciasis as a public health problem, ivermectin treatment was urgent in communities with a prevalence of nodules in adult males greater than 40% and highly desirable for a nodule prevalence greater than 20%, i.e. in all meso and hyper endemic communities (World Health Organization, 1991). Based on this recommendation, APOC aim was to establish CDTI in all high risk areas where the prevalence of palpable nodules in adults was greater than 20% (Noma et al., 2002). A first application of the geostatistical analysis was to delineate all areas where the estimated prevalence was greater than 20% and to compare the results with the classification of high risk areas from the expert analysis as reported in the companion paper. We also used the geostatistical analysis to provide population estimates by endemicity level, and to predict how many people would have been infected with *O. volvulus* in the APOC countries if there had been no onchocerciasis control.

3.2 Methods

3.2.1 REMO methodology

The geographic distribution of onchocerciasis is determined by the availability of breeding sites for the *Simulium* vectors in fast flowing rivers and streams, and the limited flight range of the vector when seeking a blood meal. The spatial epidemiology of onchocerciasis is therefore closely related to the distribution and suitability of local river systems. REMO is based on this knowledge and consists of three stages (Ngoumou and Walsh, 1993):

1) The division of the area to be mapped into biogeographic zones that are reasonably uniform with regard to their potential for onchocerciasis and that cover the watersheds of the main local drainage systems. Areas that are known to be unsuitable for the vector for ecological reasons (absence of fast flowing water, high altitude, etc.) and uninhabited areas (e.g. national parks) are excluded at this stage.

2) The selection of a sample of villages to be surveyed in order to determine whether onchocerciasis is present or not and, if present, to give a rough indication of the distribution and severity of onchocerciasis in the zone. This sampling uses the available information on the local river system

3) Rapid epidemiological assessment (REA) surveys in the selected villages. A sample of 30 to 50 adult males who have resided in the village for more than 10 years are examined for the presence of nodules, and the percentage of males with palpable nodules is calculated. The geographic coordinates of each village are collected by applying a Global Positioning System (GPS) in a central location in the village.

More details of the REMO methodology are provided in the companion paper (Noma et al., 2014) and the WHO Manual for Rapid Epidemiological Mapping of Onchocerciasis

(Ngoumou and Walsh, 1993). The companion paper also describes the implementation of REMO in APOC countries and ethical considerations in undertaking the REMO surveys.

3.2.2 Analysis of REMO data

The analysis of the REMO data was undertaken using two analytical approaches: an expert analysis using the original REMO analytical methodology for which the results are reported in the companion article (Noma et al., 2014), and a geostatistical analysis which is described here.

3.2.3 Geographic information system (GIS)

All relevant geographic information was processed using ArcGIS 10 (ESRI Inc., Redlands, USA).

The geographic information used for the analysis included:

- National and administrative boundaries, rivers and lakes, national parks, main roads, villages and urban settlements (source WHO HealthMapper http://health-mapper-release-5.software.informer.com.
- Topography and relief (source ESRI http://services.arcgisonline.com/ArcGIS/rest/services/World_Shaded_Relief/MapServer.
- Population density at 30 arc seconds resolution (source LandScan http://www.ornl.gov/sci/landscan/index.shtml.
- Total surface area per country, including water bodies (http://wdi.worldbank.org/table/1.1.
- Areas that are unsuitable for onchocerciasis as defined during the first REMO phase (see above).
- Geographic coordinates of all surveyed villages and for each surveyed village the percentage of examined adults who had palpable nodules, referred to as the "prevalence of nodules" or nodule prevalence.
- Surveyed area: this is the total area within 50 km from the nearest surveyed village. The threshold of 50 km corresponds to the maximum acceptable distance between sample villages as defined in the REMO manual (Ngoumou and Walsh, 1993). Areas beyond 50 km from the nearest surveyed village are classified as non-surveyed. Excluded from both the surveyed and non-surveyed areas are unsuitable areas, national parks and water bodies.

3.2.4 Geostatistical analysis

For probabilistic prediction of the true prevalence at both sampled and unsampled locations, a geostatistical model (Diggle and Ribeiro, 2007) was fitted in which conditional on the true prevalence $P(x)$ at location x, the number of positives, Y, amongst a sample of N individuals follows a binomial distribution with N trials and "success" probability $P(x)$. We used a standard logistic link function $\log\{P(x)/(1+P(x))\} = \mu + S(x)$, where $S(x)$ is a low-rank approximation to a zero-mean isotropic Gaussian process (Higdon, 1998). For the main analysis, which excluded the spatially separate areas of Liberia and the island of Bioko, this process is defined as follows: (1) choose a discrete set of M points, say X_j, over the region of interest; (2) represent $S(x)$ as a weighted average of M independent, identically distributed zero-mean Gaussian variables Z_j with variance σ^2, i.e. $\sum_{j=1}^{M} w(X_j - x)Z_j$, where the weights $w(X_j - x)$ are chosen as functions of the great-circle distance, say u_j, between x and each of the X_j, so as to approximate the required correlation function of S(x). Note that, in this case, the variance σ^2 does not represent variability on the logit scale since the range of variation of the Z_j variables is scaled by the kernel weights $w(X_j - x)$. Following the procedure suggested by Rodrigues and Diggle (2010), we used $M = 10734$ points X_j in a regular lattice at spacing 0.1 by 0.1 degrees and weights $w(X_j - x) = \exp(-2\sqrt{2}u_j/\phi)/\phi$ to approximate a Matérn correlation function (Diggle and Ribeiro, 2007, p.51-52) with scale parameter ϕ and smoothness parameter $\kappa = 2$.

In the separate analyses for Liberia and Bioko the dimensionality was much lower and there was no need of a low rank approximation of $S(x)$. In these analyses, the zero-mean isotropic Gaussian process $S(x)$ has Matérn correlation function, as previously defined, and variance τ^2, which represents, unlike σ^2, variation on the logit scale.

In each of the three analyses, model parameters were then estimated using the method of maximum likelihood based on the Laplace approximation method (Pinheiro and Chao, 2006). Maximum likelihood estimates, with associated 95% confidence intervals, of the geostatistical model parameters were for the main analysis (all REMO data excluding Liberia and Bioko) $\hat{\mu} = -2.451$ $(-2.469, -2.432)$, $\hat{\sigma}^2 = 31.570$ $(31.038, 32.112)$ and $\hat{\phi} = 65.208$ $(64.993, 66.301)$. For Liberia the parameter estimates were $\hat{\mu} = -1.759$ $(-1.779, -1.739)$, $\hat{\tau}^2 = 0.486$ $(0.432, 0.547)$ and $\hat{\phi} = 57.945$ $(52.151, 64.381)$. Finally for Bioko the estimates were $\hat{\mu} = -0.079$ $(-0.283, 0.125)$, $\hat{\tau}^2 = 0.133$ $(0.057, 0.310)$ and $\hat{\phi} = 1.950$ $(0.535, 7.112)$. From the estimates of the scale parameters ϕ we determined that the range of the spatial correlation, defined as the distance at which the spatial correlation is 0.05 (Diggle and Ribeiro, 2007), is about 350 km for the main area, 311 km for Liberia and 10 km for Bioko. Hence pairs of observations within these distances in each of the three areas will show non-negligible spatial correlation.

The output from the fitted geostatistical model is a sample, of whatever desired size, from the joint predictive distribution of $P(x)$, i.e. the conditional distribution of $P(x)$ given all of the data, at locations x forming a regular grid at spacing 1 km over the entire surveyed area. A Monte Carlo Markov Chain method for conditional simulation of $P(x)$ is used, based on the approach proposed by Giorgi et al. (2015). Any desired summaries of the predictive distributions can then be calculated and mapped. The two most relevant summaries for the current population are the mean of the predictive distribution of $P(x)$ and the probability that $P(x)$ exceeds 0.2 (20%), which corresponds to the operational criterion for delineating high-risk areas.

In order to deal with the high number of zero reported disease cases, we added zero prevalence data-points in areas free from the disease (ocean, deserts) when simulating from the predictive distribution of $P(x)$. The fraction of added zeros corresponds to 5% of the total sample size beyond which very little impact was observed on the predicted prevalence surface. This approach decreases prevalence estimates in proximity of boundaries with areas free from the disease and avoids unrealistic high estimates of prevalence in such boundary areas. All computations were run on the High End Computer Cluster at Lancaster University, using the R statistical software environment (R Development Core Team, 2011).

3.2.5 Estimation of population by endemicity level and number infected

The "at risk population" of the surveyed areas in each APOC country was estimated by multiplying the surface of the surveyed area in the country with the country-specific average population density for CDTI projects. The latter was obtained for each APOC country by dividing the total population of the CDTI projects in the country in 2011 by the total surface area of these projects.

The nodule prevalence map was used to divide the surveyed area in each country into three endemicity classes with nodule prevalence of $0 - 4.5\%$, $5 - 19.9\%$ and greater than 20% respectively. The population in each class was estimated by multiplying the surface area with the average population density for CDTI projects in the country. For all surface calculations, the geographic coordinates were first projected using the ARCGIS (World) Cylindrical Equal Area projection.

In order to estimate the number of persons that would have been infected with *O. volvulus* by the year 2011 if there had been no onchocerciasis control, we used the recently published results of a study on the relationship between the prevalence of skin microfilaria in a village (all age groups combined) and the prevalence of palpable nodules in adult males in the same villages (Coffeng et al., 2013). From this publication we used the main relationship for all study areas except one (Mbam), for which the pattern was different. This relationship was used to convert the 1 km resolution predicted nodule prevalence in adults, as generated during the geostatistical analysis, into the corresponding predicted prevalence of microfilaria for all ages combined. For each country, the predicted prevalence of microfilaria was then averaged over the total surveyed area and multiplied with the estimated at risk population of the surveyed areas in the country to obtain an estimate of the total number, T, infected with *O. volvulus* if there had been no onchocerciasis control. To obtain a confidence interval for this estimate, we sampled repeatedly from the joint predictive distribution of prevalence surface $P(x)$, and from each sample calculated the corresponding estimate of T. Then, a 95% confidence interval for T is the range from the 2.5-th to the 97.5-th percentile of the empirical distribution of these estimates. For the APOC-wide total we used a similar procedure. Since nodule prevalence was modelled using three independent spatial processes with different means for the main area, Liberia and Bioko, we obtained a simulated sample for each from the joint predictive distribution of $P(x)$, the estimated number of infected for the three areas separately and added these up. The 95% confidence intervals were then calculated from the resulting APOC-wide total distribution of T.

3.3 Results

3.3.1 Surveyed and excluded areas

The first step in the implementation of REMO was the exclusion of areas that were considered unsuitable for onchocerciasis transmission, and where, therefore, no nodule surveys were carried out. Also excluded at this stage were large water bodies and national parks that were considered uninhabited. The extent of the excluded areas in the different countries is summarised in Table 3.1. Large excluded areas covering more than 50% of the country surface were identified in Chad, Ethiopia, Kenya and Sudan. A description of the main unsuitable areas is provided in the companion paper (Noma et al., 2014).

The remaining areas after the above exclusions were considered potentially endemic areas that needed to be surveyed for onchocerciasis. Table 3.1 shows for each country the extent of the areas that were surveyed and of the remaining non-surveyed area. In 8 countries all of the potentially endemic areas were surveyed. In 6 other countries, all (Central African Republic and Gabon) or nearly all (Angola, Cameroon, Congo and South Sudan) of the non-surveyed areas were uninhabited or had a very low population density of less than 1 person per km^2. In only 2 of the remaining countries was the non-surveyed area

Table 3.1: Extent of excluded and surveyed areas in the 20 APOC countries.

	Country surface	Excluded area		Surveyed area		Non-surveyed area	
Country	1000 km^2	1000 km^2	%	1000 km^2	%	1000 km^2	%
Angola	1,247	84	6.7	1,015	81.4	148	11.9
Burundi	28	4	13.0	24	87.0	0	0.0
Cameroon	475	25	5.2	430	90.5	21	4.3
CAR	623	129	20.6	448	71.9	46	7.4
Chad	1284	1027	80.0	257	20.0	0	0.0
Congo	342	59	17.1	271	79.3	12	3.5
DRC	2,345	183	7.8	2,053	87.6	109	4.6
Eq. Guinea	28	4	15.2	23	80.4	0	0.0
Ethiopia	1,104	583	52.8	446	40.4	75	6.8
Gabon	268	19	7.3	191	71.5	57	21.2
Kenya	584	505	86.3	57	9.8	23	3.9
Liberia	96	1	0.8	96	99.2	0	0.0
Malawi	118	39	33.3	77	64.9	2	1.9
Mozambique	799	60	7.5	549	68.7	190	23.8
Nigeria	924	42	4.6	858	92.9	0	0.0
Rwanda	25	5	21.3	20	78.0	0	0.0
South Sudan	644	68	10.6	535	83.0	41	6.4
Sudan	1,861	1,516	81.4	346	18.6	0	0.0
Tanzania	947	380	40.1	393	41.5	174	18.4
Uganda	242	62	25.5	180	74.4	0	0.0
Total	13,986	4,793	34.3	8,270	59.1	898	6.4

more than 10% of the country surface: Mozambique (17%) and Tanzania (16%). REMO surveys were carried out in a total of 14,473 sample villages in the surveyed areas in the 20 APOC countries. Figure 3.1 provides a map showing the location and observed prevalence in the sample villages and the extent of the surveyed area.

3.3.2 Map of the estimated prevalence of palpable nodules

The model-based geostatistical analysis generated a map of the predicted prevalence of palpable nodules at 1 km resolution throughout the surveyed area in the 20 countries (see Figure 3.2). This map provides the best estimate of the geographic distribution of onchocerciasis endemicity levels based on the model based analysis of the REMO data. It shows substantial spatial variation in onchocerciasis endemicity levels. There are some vast areas where the endemicity levels are very high with the estimated prevalence of nodules exceeding 40%. A vast belt of hyperendemic onchocerciasis extends from the Democratic Republic of Congo through the west of South Sudan and the Central African Republic to Cameroon and south east Nigeria. There are also large hyperendemic foci in south Tanzania and west Ethiopia. On the other end of the endemicity scale there are several large areas where the prevalence of nodules is close to 0. This includes

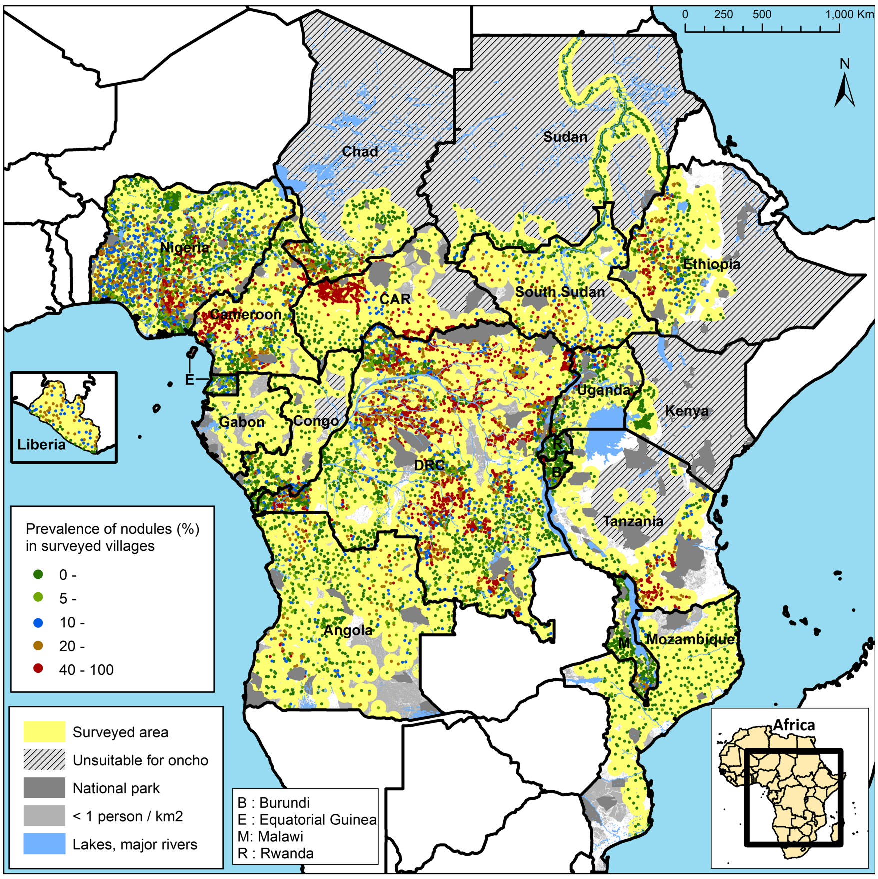

Figure 3.1: Map of the observed prevalence of palpable nodules in the 14,473 surveyed villages.

an area of some 500,000 km^2 in North and Central Congo, South West of the Central African Republic and border areas of Cameroon, Gabon and the Democratic Republic of Congo where the results suggest that onchocerciasis is not endemic. Similar results were obtained for most of Mozambique, Malawi and Uganda, and large sections of Tanzania, Ethiopia, Sudan and Chad.

3.3.3 Estimated population by endemicity level

Table 3.2 shows for each APOC country the classification of the surveyed areas into three endemicity classes with nodule prevalences of $0 - 4.9\%$, $5 - 19.9\%$ and greater than 20% respectively. The table also gives the estimated population for these three categories for the year 2011. Overall, the predicted prevalence of nodules is greater than 20% over an area of 2.5 million km^2 where an estimated 62 million people live. Another 77 million people are estimated to live in an area of 2.8 million km^2 where the predicted nodule prevalence is between 5% and 20%. There are four countries, namely Cameroon, Central African Republic, Democratic Republic of Congo and Liberia, where more than 50% of the surveyed population live in areas where the predicted nodule prevalence is greater

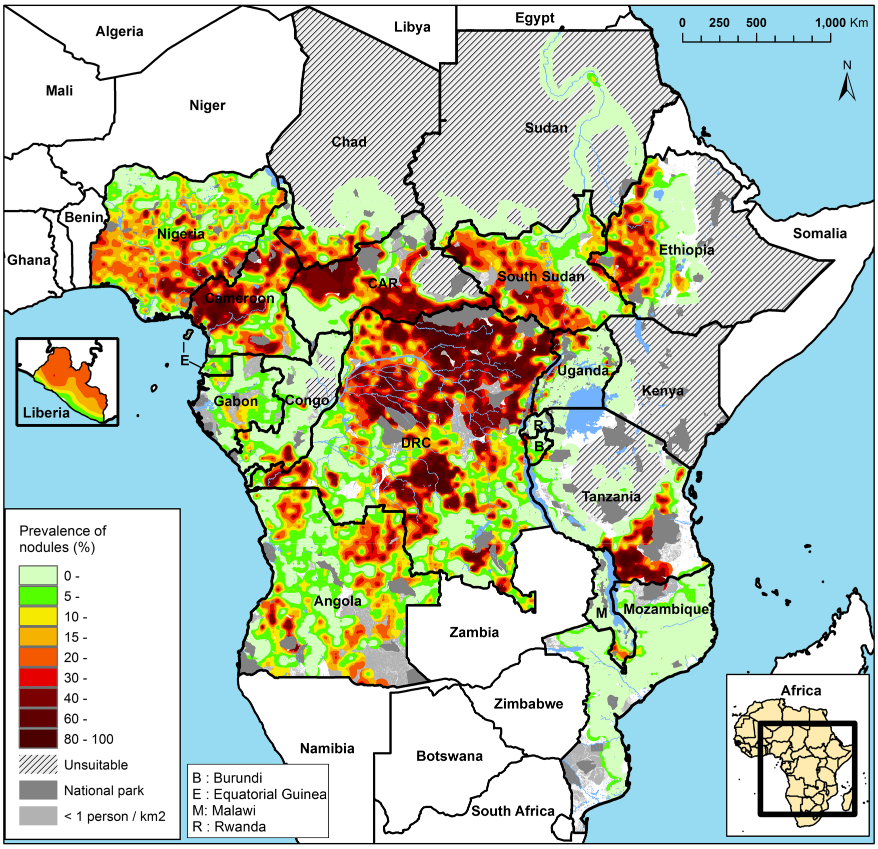

FIGURE 3.2: Map of the estimated prevalence of palpable nodules in the 20 APOC countries.

than 20%. In absolute numbers, the main countries are the Democratic Republic of Congo with 23.3 million people living in areas with more than 20% prevalence, Nigeria (14.3 million), Ethiopia (5.9 million), and Cameroon (5.2 million).

Table 3.2: Surveyed area and population by estimated nodule prevalence in the 20 APOC countries.

	Total surveyed area			Surface km^2			Estimated population (1000)			Estimated population as % of total population of surveyed area		
Country	Population per km^2 in CTDi area	Surface (1000 km^2)	Estimated population (1000)	Prevalence of nodules 0-4.9%	Prevalence of nodules 5-19.9%	Prevalence of nodules $\geq$ 20%	Prevalence of nodules 0-4.9%	Prevalence of nodules 5-19.9%	Prevalence of nodules $\geq$ 20%	Prevalence of nodules 0-4.9%	Prevalence of nodules 5-19.9%	Prevalence of nodules $\geq$ 20%
Angola	3.8	1,015	3,812	278	570	166	1,046	2,142	625	27.4	56.2	16.4
Burundi	341.0	24	8,252	11	11	2	3,821	3,760	671	46.3	45.6	8.1
Cameroon	24.2	430	10,389	64	149	217	1,555	3,599	5,234	15.0	34.6	50.4
CAR	5.2	448	2,312	135	81	232	697	418	1,198	30.1	18.1	51.8
Chad	21.6	257	5,557	152	51	54	3,286	1,099	1,171	59.1	19.8	21.1
Congo	34.8	271	9,432	194	62	15	6,750	2,159	523	71.6	22.9	5.5
DRC	22.1	2,053	45,391	318	683	1,052	7,040	15,089	23,262	15.5	33.2	51.2
Eq. Guinea	22.0	23	433	3	18	1	54	298	81	12.4	68.8	18.7
Ethiopia	46.7	446	20,842	168	152	126	7,844	7,094	5,904	37.6	34.0	28.3
Gabon	NA	191	722	88	101	2	333	381	8	46.1	52.8	1.2
Kenya	NA	57	3,035	57	0	0	3,035	0	0	100.0	0.0	0.0
Liberia	30.2	96	2,884	0	42	53	10	1,269	1,604	0.3	44.0	55.6
Malawi	237.4	77	18,245	61	11	4	14,529	2,708	1,008	79.6	14.8	5.5
Mozambique	NA	549	9,889	483	65	1	8,694	1,170	25	87.9	11.8	0.2
Nigeria	65.3	858	56,016	175	463	220	11,440	30,239	14,336	20.4	54.0	25.6
Rwanda	NA	20	9,550	19	1	0	9,550	0	0	100.0	0.0	0.0
South Sudan	13.8	535	7,380	115	206	214	1,591	2,842	2,947	21.6	38.5	39.9
Sudan	14.6	346	5,053	331	13	2	4,841	190	23	95.8	3.8	0.4
Tanzania	19.4	393	7,631	221	66	106	4,289	1,276	2,065	56.2	16.7	27.1
Uganda	56.4	180	10,135	128	28	24	7,208	1,556	1,371	71.1	15.4	13.5
Total	27.1	8,270	236,959	3,004	2,774	2,493	97,611	77,291	62,056	41.2	32.6	26.2

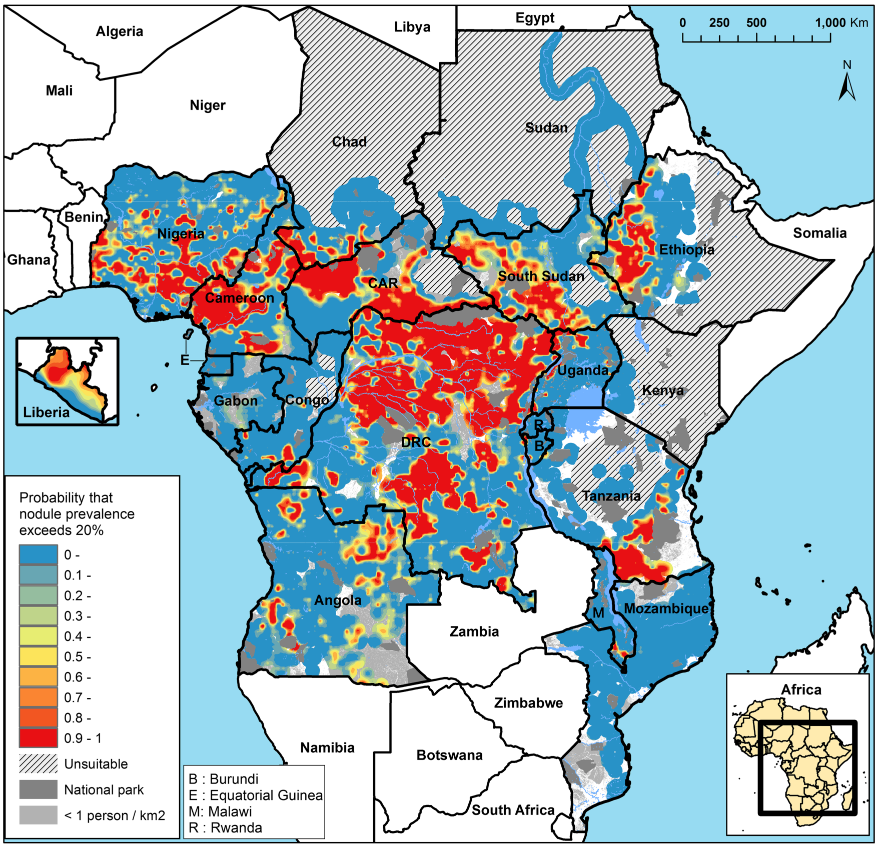

FIGURE 3.3: Map of the predictive probability that the local prevalence of nodules exceeds 20%.

3.3.4 Priority areas for large scale treatment

The main objective of the REMO surveys was to identify priority areas for large-scale ivermectin treatment, i.e. areas where the prevalence of nodules is greater than 20%. The geostatistical analysis provides an objective method for defining such areas while taking the statistical uncertainty of the estimates into account. Figure 3.3 provides a map of the predicted probability that the local prevalence of palpable nodules exceeds the threshold of 20%. The map shows that for most of the surveyed area there is little uncertainty whether the prevalence of nodules exceeds the threshold or not: the probability is in most areas less than 0.1 (highly unlikely that the prevalence exceeds 20%) or greater than 0.9 (very likely that the prevalence is greater than 20%). Only for a few areas is the exceedance probability around 0.5, indicating that it is uncertain whether the prevalence exceeds the threshold. Most of these concern transition areas between high and low endemicity zones.

Table 3.3 provides a summary of the classification of the surveyed area according to the probability that the nodule prevalence exceeds 20%, and compares the results with those

Table 3.3: Comparison of priority areas for treatment identified by the two analytical approaches: expert analysis and geostatistical analysis.

Exceedance	High risk (expert analysis)								
probability	Yes			No			Total		
of 20%	Surface	Population		Surface	Population		Surface	Population	
	(1000 km^2)	(1000)	% of total	(1000 km^2)	(1000)	% of total	(1000 km^2)	(1000)	% of total
> 0.9	1,440	35,447	98.0%	25	714	2.0%	1,465	36,161	100.0%
> 0.5	2,193	55,066	93.4%	152	3,877	6.6%	2,345	58,943	100.0%
> 0.1	2, 794	72, 479	80.4%	837	17,682	19.6%	3,631	90,162	100.0%
Total surveyed area	3,180	83,972	35.4%	5,089	152,986	64.6%	8,270	236,959	100.0%

of the classification of high risk areas in the expert analysis described in the companion paper (Noma et al., 2014). Using an exceedance probability of 0.5, it is estimated that the nodule prevalence exceeds 20% over a total surface of 2.3 million km^2 with a population of 59 million people. However, these estimates are subject to considerable statistical uncertainty. Using an exceedance probability of 0.9 (highly likely that the nodule prevalence exceeds 20%), the corresponding population is only 36 million. Using an exceedance probability of 0.1, the population increases to as much as 90 million. The expert analysis reported in the companion article identified high risk areas with a total surface of 3.2 million km^2 and a population of 84 million (this figure refers to high risk areas within the surveyed area; the experts also classified an additional 0.11 million km^2 of unsurveyed area as “assumed” high risk based on circumstantial evidence, giving a total of 3.3 million km^2 of high risk areas and a population of 86 million reported in the companion paper). Table 3.3 shows the overlap between the two approaches. 98% of the priority areas for treatment that were identified with an exceedance probability of 0.9 in the geostatistical analysis were classified as high risk areas by the experts. The few differences between the two classification methods concerned minor differences in the delineation of boundaries of priority areas for treatment, with the expert analysis drawing boundaries according to river basins and the model based analysis, which currently does not include spatial information on rivers, drawing the boundaries often slightly wider. For the priority areas identified with the low exceedance probability of 0.1, the agreement with the experts was, unsurprisingly, poorer; only 80% of this area was classified as high risk by the experts.

3.3.5 Estimated number infected

The map of the predicted prevalence of nodules in adults in the 20 APOC countries, together with the recently published relationship between the prevalence of skin microfilaria and the prevalence of nodules, allowed the estimation of the total number of people that would have been infected with *O. volvulus* in the APOC countries if there had been

Table 3.4: Estimated number of people that would have been infected with *Onchocerca volvulus* in the 20 APOC countries in 2011 if there had been no ivermectin treatment.

	Population per km^2	Surveyed area	Rural population	Number infected with *O. volvulus* (1000)		
	in CTDi area	area (1000 km^2)	in surveyed area (1000)	Estimate	Quantile 0.025	Quantile 0.975
Angola	3.8	1,015	3,812	440	410	475
Burundi	341.0	24	8,252	658	603	717
Cameroon	24.2	430	10,389	2,810	2,674	2,956
CAR	5.2	448	2,312	592	562	624
Chad	21.6	257	5,557	551	516	591
Congo	34.8	271	9,432	512	442	605
DRC	22.1	2,053	45,391	13,155	12,869	13,462
Equatorial Guinea	22.0	23	433	58	55	62
Ethiopia	46.7	446	20,842	2,882	2,677	3,117
Gabon	NA	191	722	49	35	66
Kenya	NA	57	3,035	68	37	123
Liberia	30.2	96	2,884	554	515	596
Malawi	237.4	77	18,245	817	727	968
Mozambique	NA	549	9,889	330	275	398
Nigeria	65.3	858	56,016	8,510	8,292	8,750
Rwanda	NA	20	9,550	228	179	283
South Sudan	13.8	535	7,380	1,361	1,269	1,464
Sudan	14.6	346	5,053	58	50	68
Tanzania	19.4	393	7,631	1,061	975	1,152
Uganda	56.4	180	10,135	865	814	925
Total	27.1	8,270	236,959	35,559	35,085	36,116

no onchocerciasis control. The results of this analysis are shown in Table 3.4. It is estimated that overall 35.6 million people (95% confidence interval 35.1 to 36.1 million) would have been infected by 2011 if there had been no CDTI. Of those, 13.2 million are from the Democratic Republic of Congo and 8.5 million from Nigeria.

As reported in the companion article, the prevalence of nodules was virtually zero in Kenya and Rwanda, suggesting that these two countries are non-endemic for onchocerciasis (Noma et al., 2014). However, the necessarily imperfect calibration relationship between the prevalence of nodules and skin mf prevalence of Coffeng et al. (2013) shows that a zero nodule prevalence is compatible with skin mf prevalence between zero and about four percent. This explains why, in each of these two presumed non-endemic countries, our point estimate of the number of infected is approximately 2% of the population of the surveyed area.

3.4 Discussion

The geostatistical analysis of the extensive REMO data for 14,473 surveyed villages has produced a detailed map of the pre-control geographic distribution of onchocerciasis endemicity levels in the 20 APOC countries. This map has been proven very valuable for onchocerciasis control and elimination.

Nearly all potentially endemic areas in the 20 APOC countries have been mapped for onchocerciasis. Of the total surface area of the 20 countries, 94% has been surveyed for onchocerciasis or classified as unsuitable for onchocerciasis transmission. By design, no surveys were done in the unsuitable areas. Although we have no reason to doubt the classification of unsuitability, we were not able to validate it with survey data. Most of the remaining 6% of unsurveyed area is either not populated or has a very low population density of less than 1 person per km^2. It also includes a few zones for which it can reasonably be assumed that onchocerciasis is not endemic: the belts between surveyed and unsuitable areas in central Ethiopia and Kenya where the prevalence of nodules was zero in all neighbouring REMO villages; the unsurveyed areas in Mozambique south of latitude 18°S given that only 1 single nodule was detected in 37 villages surveyed below this latitude; and the coastal low lands of Tanzania where onchocerciasis vectors have never been reported (Mwaiko, Mtoi, and Mkufya, 1990; Raybould and White, 1979; Wegesa, 1970). Only for less than 1% of the total surface area of the 20 APOC countries may surveys still be needed to estimate the level of onchocerciasis endemicity. Hence the mapping of onchocerciasis in all potentially endemic areas in the APOC countries can be considered more than 99% complete.

The map of the pre-control prevalence of nodules that was generated in the geostatistical analysis predicts that before the start of CDTI, onchocerciasis was endemic in 18 of the 20 APOC countries. In Rwanda and Kenya (where onchocerciasis has been eliminated through vector control in the 1960s) the prevalence of nodules was virtually zero and these countries were classified as non-endemic. In Mozambique, the predicted prevalence of nodules was around zero throughout the country except for two small border areas with Tanzania and Malawi. In these two neighbouring countries there are hyperendemic onchocerciasis foci close to the border and this resulted in a predicted nodule prevalence of 15% to 20% just across the border in Mozambique. In the remaining 17 endemic countries, the endemicity levels of onchocerciasis varied significantly. There was a vast belt of hyperendemic onchocerciasis covering most of the Democratic Republic of Congo and extending across west Uganda, South Sudan, Central African Republic, Chad and Cameroon into Nigeria. In all of these countries the estimated nodule prevalence reached levels of over 40%, corresponding to skin microfilaria prevalence levels of about 60%. There were also large hyperendemic zones in Ethiopia and Tanzania with equally high prevalence levels. On the other hand, the estimated prevalence was close or equal to zero in most of Malawi, Uganda and Sudan, and in large sections of Burundi, Congo, Gabon, Tanzania, central Ethiopia and south-west Central African Republic. An intermediate pattern with low to medium prevalence levels was seen in the mainland of Equatorial Guinea and most of Angola. Overall, the predicted prevalence of nodules was greater than 20% over a surveyed area of 2.5 million km^2 with an estimated population of 62 million, while the prevalence was between 5% and 20% over 2.8 million km^2 with an estimated population of 77 million.

Beyond the APOC countries, onchocerciasis was known to be endemic in West Africa where the disease has been mapped by the OCP (Boatin, 2008; De Sole et al., 1991). To the north of the surveyed area in the APOC countries are arid zones that are not suitable for *Simulium* vectors and which are therefore onchocerciasis free. For the same reason, Somalia is also considered onchocerciasis free even though the presence of *S. damnosum s.l.* (though not the disease) was reported from one area in the 1950s (Raybould and White, 1979). To the south of APOC, all countries except one are located below the most southern latitude at which onchocerciasis has ever been reported. The exception is Zambia. Since Zambia is not a participating country of APOC, REMO surveys have not been done in this country. In the literature there is only one report from 1983 of an infection with *O. volvulus* in a child (Beaver, Hira, and Patel, 1983), otherwise onchocerciasis has never been reported from Zambia. However, in the absence of systematic survey data, we cannot be certain that the country is onchocerciasis free, especially for some border areas.

Compared to the historical information on the geographic distribution of onchocerciasis in the APOC countries, the nodule prevalence map is a significant advance. The WHO Expert Committee on Onchocerciasis Control of 1995 produced a provisional map of endemic onchocerciasis in Africa on the basis of information available at that time (World Health Organization, 1995). Much of the area that the Committee identified as endemic for onchocerciasis has been confirmed endemic in the geostatistical analysis of the REMO data. However, there were several large areas that the Committee labeled as non-endemic but that were shown to have medium to high prevalence levels in the nodule prevalence map. These include endemic foci in North Nigeria, South Cameroon, South Sudan, much of Angola, and several large hyperendemic zones in the Democratic Republic of Congo where the prevalence of nodules exceeded 50%-80%. Conversely, several areas labeled as endemic by the Committee had an estimated nodule prevalence around zero, e.g. the zone in the south-west of the Central African Republic and the north of Congo.

A second limitation of the historical data was the lack of information on onchocerciasis endemicity levels for most areas. The REMO surveys filled this gap and generated detailed information on onchocerciasis endemicity that was critically important for APOC to identify priority areas for ivermectin treatment, i.e. areas where the prevalence of nodules exceeded 20%. Wherever REMO data became available, they were subjected to an expert analysis that delineated high risk areas where the prevalence of nodules was greater than 20% and where CDTI was subsequently implemented to control the disease as a public health problem. The results of the expert analysis are described in the companion paper. The expert analysis used a standard methodology to analyse the REMO data within the context of other relevant geographic information. The ability to take data from multiple sources into account was a strength of this methodology but a perceived weakness was its subjective component: the expertsâĂŹ interpretation of the information. The geostatistical analysis involves an objective statistical method that can take statistical uncertainty into account in the decision making process on priority areas. Given these fundamental differences between the two analytical approaches, it was of interest to compare their results.

Using the geostatistical analysis it was predicted that the local prevalence of nodules was equal to or greater than the threshold of 20% over a total surface area of 2.5 million km^2 with a population of 62 million people. This is less than the high risk area of 3.2 million km^2 with a population of 84 million identified in the expert analysis. However, in contrast to the expert analysis, the geostatistical estimate has the advantage that it is accompanied by an estimate of its statistical uncertainty. Taking into account the probability that the local prevalence exceeds the 20% threshold, the surface area ranges from 1.5 million km^2 to 3.6 million km^2 for exceedance probabilities of 0.9 and 0.1 respectively. For exceedance probabilities of 0.9, nearly all the surface area classified as having a prevalence of nodules greater than 20% was also classified as high risk in the expert analysis. For the low exceedance probability of 0.1, there was agreement with the expert analysis for only 80% of the area classified as exceeding the 20% threshold. The results indicate that the two methods gave comparable results for areas where the prevalence of nodules clearly exceeds (i.e. exceedance probability greater than âĽ0.9) the threshold of 20%, and where ivermectin treatment is therefore needed to control onchocerciasis as a public health problem, but that there is some disagreement for borderline areas where the prevalence of nodules fluctuates around or below 20%. We conclude that the expert analysis has correctly identified all areas for which there is strong evidence that ivermectin treatment is needed to control onchocerciasis as a public health problem. It also includes many borderline areas for which the evidence of high risk is less strong, but this has been considered justified for ethical reasons so as not to exclude isolated high-risk communities from treatment (Noma et al., 2014).

The geostatistical analysis has also been used to estimate the total number of people that would have been infected with *O. volvulus* in the 20 APOC countries if there had been no CDTI. Based on the nodule prevalence map and the recently published quantification of the relationship between the prevalence of skin microfilaria and the prevalence of onchocercal nodules (Coffeng et al., 2013), we estimate that some 35.6 million people (95% confidence interval 35.1 to 36.1 million) would have been infected by the year 2011 if there had been no CDTI. This estimate is significantly higher than the most commonly quoted estimate from the WHO Expert Committee on Onchocerciasis Control which estimated that in 1995 a total of 17.7 million people were infected globally, of which 15.0 million lived in APOC countries (World Health Organization, 1995). Using an annual rural population growth rate of 2.2% for the APOC countries ("Rural population growth rate"), our estimate of 36 million infected for 2011 corresponds to 25 million infected in 1995, i.e. 10 million more than the previous WHO estimate for the APOC countries. This difference is not surprising given that REMO identified many new endemic areas and generated prevalence estimates for all areas. However, compared to other, more recent estimates our figure appears low. Coffeng et al. (2013) reported an estimate of 32 million people infected in the APOC countries in 1995, and Remme et al. (2006) estimated 37 million people infected globally in 1995 and also about 32 million for the APOC countries. These estimates are also largely based on REMO data of APOC. The difference with our estimate is mainly due to two methodological factors. One concerns the formula used to quantify the relationship between the prevalence of microfilaria and the prevalence of nodules. We used a formula from a recently published analysis of data from West, Central and East Africa (Coffeng et al., 2013) which predicts a lower prevalence of microfilariae for a given prevalence of nodules than the formulas used previously. The second factor concerns the way the REMO sampling design has been taken into account. The previous estimates assume that sample villages were selected randomly from a given area. However, in the REMO sampling method villages are selected spatially at regular distances along rivers with potential breeding sites and at lower sampling density between rivers. Because of this design, the selection of villages to be surveyed is biased towards villages with a high endemicity level close to breeding sites and this bias may have resulted in an overestimate of the number infected in previous studies. The current geostatistical analysis partially corrects for this bias by taking into account the spatial distribution of the survey data. Specifically, in estimating the total number infected, one effect of the spatial correlation is that the observed prevalence from an isolated surveyed village acts as a proxy for the results that would have been obtained had surrounding villages also been surveyed, and therefore has greater influence than any one of a number of surveyed villages at mutually close locations. This results in a discrepancy between the crude average prevalence and the spatially averaged modelled prevalence.

A possible improvement of the geostatistical analysis of the REMO data would be to include relevant geographical covariates in the geostatistical model (Diggle, Menezes, and Su, 2010), such as the distance to the nearest river with breeding sites, local *Simulium* species and vegetation. This will not be easy as the distribution of the different *Simulium* species is not well known for most areas while the identification of rivers with potential for *Simulium* breeding is a challenge, especially in forest areas. However, recent progress in the development of a remote sensing model to identify *S. damnosum s.l.* breeding sites in Africa appears promising (Jacob et al., 2013). If this approach can be made to work also in forest areas, and if the cost of its large scale application can be reduced, it should be possible to improve the nodule prevalence map by including in the model the distance to the nearest potential *S. damnosum* breeding site as identified by remote sensing data. Another possible improvement of the model concerns predictions in areas where the prevalence is zero. A common feature of prevalence survey data, here and elsewhere, is an excess of zeros by comparison with the best-fitting binomial distribution. In a spatial setting, this zero-inflation can be artificial; for example, it could be the result of over-sampling in low-prevalence areas. In principle, geographical covariate information could again be used to model genuine zero-inflation (Giardina et al., 2012). In our analysis, we dealt with this by adding dummy zero prevalence data at points within areas known to be disease-free (eg deserts and large water-bodies), thereby ensuring that our estimated prevalence approaches zero at the boundaries of each of these areas. We intend to develop an extended model which treats zero-inflation as a second spatial stochastic process for applications where areas of true zero prevalence are not known beforehand and prediction of such areas is important. One such application is the use of the REMO data for helping to revise ivermectin treatment boundaries for the purpose of onchocerciasis elimination. Finally, bias would arise if implementers deliberately sampled communities whose prevalence was atypical of their general localities, a phenomenon called preferential sampling. Correcting for the effects of preferential sampling is difficult unless it can be explained by measured covariates such as distance to the nearest river in the case of onchocerciasis (Diggle, Menezes, and Su, 2010).

The original objective of REMO was to identify target areas for ivermectin treatment with the aim of controlling onchocerciasis as a public health problem. In recent years evidence has emerged that in the long term onchocerciasis infection and transmission can even be eliminated with CDTI (Higazi et al., 2013; Tekle et al., 2012; Traore et al., 2012). Based on this new evidence, APOC has adopted an additional objective to eliminate onchocerciasis where feasible (World Health Organization, 2012). Because of this paradigm shift, the target areas for CDTI are currently being revised to include all areas with local onchocerciasis transmission. The nodule prevalence map provides the starting point for determining the new treatment boundaries. Furthermore, the number of years of ivermectin treatment that is required to achieve elimination depends strongly on the local endemicity level (Winnen et al., 2002). Information on pre-control endemicity levels is therefore essential for the correct interpretation of the results of epidemiological evaluations of the impact of CDTI on onchocerciasis infection levels, and for the prediction of the remaining number of years of CDTI needed in a given area (World Health Organization, 2010). This information is now also available for all CDTI areas from the nodule prevalence map.

3.5 Conclusions

APOC is close to achieving the objective of controlling onchocerciasis as a public health problem throughout the APOC countries, and the REMO data and nodule prevalence maps have played an essential role in targeting treatment where needed to achieve this objective (World Health Organization, 2011). Following the shift from onchocerciasis control to onchocerciasis elimination, the nodule prevalence map will continue to play an important role and help with adjusting treatment boundaries, interpreting epidemiological evaluation data on progress towards elimination and predicting when elimination will be achieved in different areas. REMO was a major undertaking but it has been worthwhile and the results have been very valuable for onchocerciasis control and elimination in Africa.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

MN, AT, HZ, UVA and JHFR were involved in the design and implementation of REMO. HZ was responsible for data processing. PJD, HZ, EG and JHFR conceptualized the geostatistical analysis. PJD, EG, HZ, AT and JHFR did the analysis. JHFR drafted the manuscript and all authors contributed to and approved the final manuscript.

Acknowledgements

The work presented in this paper is the result of an effective collaboration between many people and organizations at national and international levels. A list of the many partners who facilitated and implemented the REMO surveys is provided in the acknowledgements section of the companion paper. We like to repeat our gratitude to the communities for their active participation in the surveys.

References

Baker, R. H. and O. M. Abdelnur (1986). "Onchocerciasis in Sudan: the distribution of the disease and its vectors". In: *Tropical medicine and parasitology* 37.4, pp. 341–355.

Beaver, P. C., P. R. Hira, and B. G. Patel (1983). "Onchocerciasis in Zambia: report of O. volvulus in a child and its differentiation from O. dukei in cattle". In: *Transactions of the Royal Society of Tropical Medicine and Hygiene* 77.2, pp. 162–166. DOI: 10.1016/0035-9203(83)90057-3.

Boatin, B. (2008). "The Onchocerciasis Control Programme in West Africa (OCP)". In: *Annals of tropical medicine and parasitology* 102.Suppl 1, pp. 13–17.

Bush, S. and A. D. Hopkins (2011). "Public-private partnerships in neglected tropical disease control: the role of nongovernmental organisations". In: *Acta Tropica* 120.Suppl 1, S169–S172.

Coffeng, L. E., W. A. Stolk, H. G. Zoure, J. L. Veerman, K. B. Agblewonu, M. E. Murdoch, M. Noma, G. Fobi, J. H. Richardus, D. A. Bundy, D. Habbema, S. J. de Vlas, and U. V. Amazigo (2013). "African Programme For Onchocerciasis Control 1995-2015: model-estimated health impact and cost". In: *PLoS Neglected Tropical Diseases* 7.1, e2032. DOI: 10.1371/journal.pntd.0002032.

Coffeng, L. E., S. D. Pion, S. O'Hanlon, S. Cousens, A. O. Abiose, P. U. Fischer, J. H. F. Remme, K. Y. Dadzie, M. E. Murdoch, S. J. de Vlas, M. G. Basanez, W. A. Stolk, and M. Boussinesq (2013). "Onchocerciasis: The Pre-control Association between Prevalence of Palpable Nodules and Skin Microfilariae". In: *PLoS Neglected Tropical Diseases* 7.4, e2168. DOI: 10.1371/journal.pntd.0002168.

De Sole, G., R. Baker, K. Y. Dadzie, J. Giese, P. Guillet, F. M. Keita, and J. Remme (1991). "Onchocerciasis distribution and severity in five West African countries". In: *Bulletin World Health Organization* 69.6, pp. 689–698.

Diggle, P. J., R. Menezes, and T. Su (2010). "Geostatistical inference under preferential sampling". In: *Journal of the Royal Statatistical Society, Series C* 59.2, pp. 191–232. DOI: 10.1111/j.1467-9876.2009.00701.x.

Diggle, P. J. and P. J. Ribeiro (2007). In: *Model-Based Geostatistics.*

Giardina, F., L. Gosoniu, L. Konate, M. B. Diouf, R. Perry, O. Gaye, O. Faye, and P. Vounatsou (2012). "Estimating the burden of malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 data". In: *PLoS One* 7.3, e32625. DOI: 10.1371/journal.pone.0032625.

Giorgi, E., S. S. S. Sesay, D. J. Terlouw, and P. J. Diggle (2015). "Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models". In: *Journal of the Royal Statistical Society, Series A* 178, pp. 445–464.

Gustavsen, K., A. Hopkins, and M. Sauerbrey (2011). "Onchocerciasis in the Americas: from arrival to (near) elimination". In: *Parasites & Vectors* 4, p. 205. DOI: 10.1186/1756-3305-4-205.

Higazi, T. B., I. M. Zarroug, H. A. Mohamed, W. A. Elmubark, T. C. Deran, N. Aziz, M. Katabarwa, H. K. Hassan, T. R. Unnasch, C. D. Mackenzie, F. Richards, and K. Hashim (2013). "Interruption of Onchocerca volvulus Transmission in the Abu Hamed Focus". In: *The American Journal of Tropical Medicine and Hygiene* 89.1, pp. 51–57. DOI: 10.4269/ajtmh.13-0112.

Higdon, D. (1998). "A process-convolution approach to modelling temperatures in the North Atlantic Ocean". In: *Environmental and Ecological Statistics* 5.2, pp. 173–190. DOI: 10.1023/A:1009666805688.

Jacob, B. G., R. J. Novak, L. D. Toe, M. Sanfo, D. A. Griffith, T. L. Lakwo, P. Habomugisha, M. N. Katabarwa, and T. R. Unnasch (2013). "Validation of a remote sensing model to identify simulium damnosum s.l. Breeding sites in Sub-Saharan Africa". In: *PLoS Neglected Tropical Diseases* 7.7, e2342. DOI: 10.1371/journal.pntd.0002342.

Murdoch, M. E., M. C. Asuzu, M. Hagan, W. H. Makunde, P. Ngoumou, K. F. Ogbuagu, D. Okello, G. Ozoh, and J. Remme (2002). "Onchocerciasis: the clinical and epidemiological burden of skin disease in Africa". In: *Annals of tropical medicine and parasitology* 96.3, pp. 283–296. DOI: 10.1179/000349802125000826.

Mwaiko, G. L., R. S. Mtoi, and A. R. Mkufya (1990). "Onchocerciasis prevalence in Tanzania". In: *Central African Journal of Medicine* 36.4, pp. 94–96.

Ngoumou, P. and F. Walsh (1993). In: *A manual for Rapid Epidemiological Mapping of Onchocerciasis (REMO). Document TDR/TDE/ONCHO/93.4.*

Ngoumou, P., J. F. Walsh, and J. M. Mace (1994). "A rapid mapping technique for the prevalence and distribution of onchocerciasis: a Cameroon case study". In: *Annals of tropical medicine and parasitology* 88.5, pp. 463–474.

Noma, M., B. E. Nwoke, I. Nutall, P. A. Tambala, P. Enyong, A. Namsenmo, J. Remme, U. V. Amazigo, O. O. Kale, and A. Seketeli (2002). "Rapid epidemiological mapping of onchocerciasis (REMO): its application by the African Programme for Onchocerciasis Control (APOC)". In: *Annals of tropical medicine and parasitology* 96.Suppl 1, S29–39.

Noma, M., H. Zoure, A. H. Tekle, P. Enyong, B. E. B. Nwoke, and J. H. F. Remme (2014). "The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control: (1) priority areas for ivermectin treatment". In: *Parasites & Vectors* 7, p. 325. DOI: 10.1186/1756-3305-7-325.

Pinheiro, J. C. and E. C. Chao (2006). “Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models”. In: *Journal of Computational and Graphical Statistics* 15.1, pp. 58–81. DOI: 10.1198/106186006X96962.

Prost, A., J. P. Hervouet, and B. Thylefors (1979). “The degrees of endemicity of onchocerciasis”. In: *Bulletin World Health Organization* 57, pp. 655–662.

R Development Core Team (2011). “R: A language and environment for statistical computing”. In: http://www.R-project.org/.

Raybould, J. N. and G. B. White (1979). “The distribution, bionomics and control of Onchocerciasis vectors (Diptera: Simuliidae) in Eastern Africa and the Yemen”. In: *Tropenmedizin und Parasitologie* 30, pp. 505–547.

Remme, J., K. Y. Dadzie, A. Rolland, and B. Thylefors (1989). “Ocular onchocerciasis and intensity of infection in the community. I. West African savanna”. In: *Tropical Medicine and Parasitology* 40.3. issn: 0177-2392, pp. 340–347.

Remme, J. H. F. (1995). “The African Programme for Onchocerciasis Control: preparing to launch”. In: *Parasitology Today* 11, pp. 403–406. DOI: 10.1016/0169-4758(95)80017-4.

Remme, J. H. F., P. Feenstra, P. R. Lever, A. C. Medici, C. M. Morel, M. Noma, K. D. Ramaiah, F. Richards, A. Seketeli, G. Schmunis, W. H. van Brakel, and A. Vassall (2006). “Tropical Diseases Targeted for Elimination: Chagas Disease, Lymphatic Filariasis, Onchocerciasis and Leprosy”. In: *Disease Control Priorities in Developing Countries*, pp. 433–449.

Rodrigues, A. and P. J. Diggle (2010). “A class of convolutionbased models for spatiotemporal processes with Nonseparable covariance structure”. In: *Scandinavian Journal of Statistics* 37.4, pp. 553–567. DOI: 10.1111/j.1467-9469.2009.00675.x.

Tekle, A. H., E. Elhassan, S. Isiyaku, U. V. Amazigo, S. Bush, M. Noma, S. Cousens, A. Abiose, and J. H. F. Remme (2012). “Impact of long-term treatment of onchocerciasis with ivermectin in Kaduna State, Nigeria: first evidence of the potential for elimination in the operational area of the African Programme for Onchocerciasis Control”. In: *Parasites & Vectors* 5, p. 28. DOI: 10.1186/1756-3305-5-28.

Tielsch, J. M. and A. Beeche (2004). “Impact of ivermectin on illness and disability associated with onchocerciasis”. In: *Tropical Medicine & International Health* 9.4, A45–A56. DOI: 10.1111/j.1365-3156.2004.01213.x.

Traore, M. O., M. D. Sarr, A. Badji, Y. Bissan, L. Diawara, K. Doumbia, S. F. Goita, L. Konate, K. Mounkoro, A. F. Seck, L. Toe, S. Toure, and J. H. F. Remme (2012). “Proof-of-principle of onchocerciasis elimination with ivermectin treatment in endemic foci in Africa: final results of a study in Mali and Senegal". In: *PLoS Neglected Tropical Diseases* 6.9, e1825. DOI: 10.1371/journal.pntd.0001825.

Wegesa, P. (1970). "The present status of onchocerciasis in Tanzania. A review of the distribution and prevalence of the disease". In: *Tropical and geographical medicine* 22, pp. 345–351.

Winnen, M., A. P. Plaisier, E. S. Alley, N. J. Nagelkerke, G. van Oortmarssen, B. A. Boatin, and J. D. Habbema (2002). "Can ivermectin mass treatments eliminate onchocerciasis in Africa?" In: *Bulletin World Health Organization* 80.5, pp. 384–391.

World Development Indicators. "Rural population growth rate". In: URL: http://data.worldbank.org/indicator/SP.RUR.TOTL.ZG.

World Health Organization (1991). In: *Strategies for Ivermectin Distribution Through Primary Health Care Systems. Document WHO/PBL/91.24*.

World Health Organization (1995). In: *Onchocerciasis and its Control. Report of a WHO Expert Committee on Onchocerciasis Control*, pp. 1–104.

World Health Organization (2010). In: *Conceptual and Operational Framework of Onchocerciasis Elimination With Ivermectin Treatment.*

World Health Organization (2011). In: *Report of the External mid-Term Evaluation of the African Programme for Onchocerciasis Control. Document JAF16.8*, p. 77.

World Health Organization (2012). In: *African Programme for Onchocerciasis Contro: 18th Session of the Joint Action Forum. Bujumbura, Burundi. Document JAF19.4.*

World Health Organization (2012). "African Programme for Onchocerciasis Control: meeting of national onchocerciasis task forces, September 2012". In: *Weekly Epidemiological Record* 87.49/50, pp. 494–502.

Zein, Z. A. (1990). "An appraisal of the epidemiologic situation of onchocerciasis in Ethiopia". In: *Parassitologia* 32.2, pp. 237–244.

Chapter 4

Paper 3. PrevMap: an R package for prevalence mapping

E. Giorgi and P. J. Diggle

Lancaster Medical School, Lancaster University, Lancaster, UK

Summary

In this paper we introduce a new `R` package, `PrevMap`, for the analysis of spatially referenced prevalence data, including both classical maximum likelihood and Bayesian approaches to parameter estimation and plug-in or Bayesian prediction. More specifically, the new package implements fitting of geostatistical models for binomial data, based on two distinct approaches. The first approach uses a generalized linear mixed model with logistic link function, binomial error distribution and a Gaussian spatial process as a stochastic component in the linear predictor. A simpler, but approximate, alternative approach consists of fitting a linear Gaussian model to empirical-logit-transformed data. The package also includes implementations of convolution-based low-rank approximations to the Gaussian spatial process to enable computationally efficient analysis of large spatial data-sets. We illustrate the use of the package through the analysis of *Loa loa* prevalence data from Cameroon and Nigeria. We illustrate the use of the low rank approximation using a simulated geostatistical data-set.

Keywords: Bayesian analysis; Geostatistics; Low-rank approximations; Monte Carlo maximum likelihood; Prevalence data; R software environment.

4.1 Introduction

This article introduces PrevMap, an R package for classical and Bayesian inference on spatially referenced prevalence data. The package implements fitting and spatial prediction for the standard geostatistical model used in the context of prevalence mapping (Diggle, Tawn, and Moyeed, 1998). This model falls within the generalized linear mixed model framework whereby, conditionally on a Gaussian spatial process, a binomial error distribution with logistic-link function is used to model the data. For classical analysis, we estimate parameters by Monte Carlo maximum likelihood (MCML), which uses importance sampling techniques so as to approximate the high-dimensional intractable integral that defines the likelihood function; see for example Christensen (2004). Plug-in spatial prediction is then carried out by fixing the model parameters at the corresponding MCML estimates. In order to account for uncertainty in the model parameter estimates, we also consider a Bayesian approach in which plug-in predictive distributions at different values of the model parameters are weighted according to their posterior probabilities. A simpler, but approximate, procedure consists of fitting a geostatistical linear Gaussian model to empirical-logit-transformed prevalences.

Table 4.1 summarises the common functionalities required for prevalence mapping that are available in PrevMap and the existing packages geoR (Diggle and Ribeiro, 2007; Ribeiro and Diggle, 2001), geoRglm (Christensen and Ribeiro, 2002), geostatsp (Brown, 2015), geoBayes and spBayes (Finley, Banerjee, and Carlin, 2007; Finley, Banerjee, and Gelfand, 2015). Overall, PrevMap provides the most extensive functionality. Specifically, PrevMap provides the following features: implementation of a convolution-based low-rank approximation that can be used to reduce the computational burden when analysing large spatial data-sets; accurate numerical computation of MCML standard errors for both regression and covariance parameters estimates; inclusion of both individual-level and location-level explanatory variables with random effects defined at location-level when repeated observations are made at the same locations; more flexible prior specifications for the covariance parameters; implementation of an efficient Hamiltonian Monte Carlo algorithm for Bayesian parameter estimation.

The paper is structured as follows. In Section 4.2, we briefly introduce the geostatistical binomial logistic (henceforth BL) model, describe methods for classical and Bayesian inference, and outline approximate procedures based on the empirical logit transformation and low-rank approximations. Section 4.3 is a geostatistical analysis of *Loa loa* prevalence data using the empirical logit transformation of the data; we also show how to fit a BL model using Monte Carlo methods, both for classical and Bayesian analysis. In Section 4.4, we illustrate the use of the low-rank approximation by fitting a BL model

TABLE 4.1: List of functionalities that are currently available (✓) and not available (✕) in PrevMap and other R packages used to analyse geostatistical data.

	PrevMap	geoR	geoRglm	geostatsp	geoBayes	spBayes
– Fitting of geostatistical binomial models.	✓	✕	✓	✓	✓	✓
– Likelihood-based inference (binomial model).	✓	✕	✓	✕	✕	✕
– Bayesian inference (binomial model).	✓	✕	✓	✓	✓	✓
– Option for inclusion of the nugget effect (binomial model).	✓	✕	✓	✓	✓	✕
– Low-rank approximations.	✓	✕	✕	✕	✕	✓
– Fitting of two-levels models.	✓	✕	✕	✓	✕	✕
– Spatial prediction of non-linear properties.	✓	✓*	✓	✕	✓	✓
– Spatial prediction of multivariate properties.	✓	✓*	✓	✕	✓	✓
– Option for inclusion of anisotropy.	✕	✓*	✓	✓*	✕	✕
– Specification of non-Matérn correlation functions (e.g. Gaussian, spherical).	✕	✓*	✓	✕	✓	✓

* Available only for the linear model.

to a simulated geostatistical data-set. Section 4.5 is a concluding discussion on planned extensions to the package.

4.2 Methodological framework

The ingredients of a geostatistical BL model are: random variables Y_i of positive counts, binomial denominators m_i, explanatory variables $d_i \in \mathbb{R}^p$ and associated sampling locations $x_i : i = 1, \ldots, n$ in a given region of interest $A \subseteq \mathbb{R}^2$. Conditionally on a zero-mean Gaussian process $S(x)$ and mutually independent zero-mean Gaussian variables Z_i, Y_i follows a binomial distribution with mean $E[Y_i|S(x_i), Z_i] = m_i p_i$ such that

$$\log\left\{\frac{p_i}{1-p_i}\right\} = T_i = d(x_i)^\top\beta + S(x_i) + Z_i, \tag{4.1}$$

where we set $d_i = d(x_i)$ to emphasize the spatial context and β is a vector of regression coefficients. In (4.1), we write τ^2 for the variance of Z_i and model $S(x)$ as a stationary isotropic Gaussian process with variance σ^2 and Matérn (1986) correlation function given by

$$\rho(u; \phi, \kappa) = \{2^{k-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa \mathcal{K}_\kappa(u/\phi), u > 0,$$

where $\phi > 0$ is a scale parameter, $\mathcal{K}_\kappa(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$ and u is the distance between two sampling locations. The shape parameter κ determines the smoothness of $S(x)$, in the sense that $S(x)$ is $\lceil \kappa \rceil - 1$ times mean-square differentiable, with $\lceil \kappa \rceil$ denoting the smallest integer greater than or equal to κ.

In most of the functions available in `PrevMap`, the Matérn shape parameter κ is treated as fixed. One reason for this is that, as shown by Zhang (2004), not all of the three parameters σ^2, ϕ and κ can be consistently estimated under in-fill asymptotics and in practice this translates to κ often being poorly identified. Additionally, the parameter κ is rarely of direct scientific interest. We therefore recommend either fixing κ at a plausible value, or considering a discrete set of values e.g., $\{1/2, 3/2, 5/2\}$ corresponding to different levels of smoothness, and profiling on κ.

4.2.1 Monte Carlo maximum likelihood

The likelihood function for the parameters β and $\theta^\top = (\sigma^2, \phi, \tau^2)$ is given by the marginal distribution of the random variables Y_i. This is obtained by integrating out the random effects in T_i as defined by (4.1). Let D denote the n by p matrix of explanatory variables and $y^\top = (y_1, \dots, y_n)$ the vector of binomial observations. The marginal distribution of T is multivariate Gaussian with mean vector $D\beta$ and covariance matrix $\Sigma(\theta)$ with diagonal elements $\sigma^2 + \tau^2$ and off-diagonal elements $\sigma^2 \rho(u_{ij})$, where u_{ij} is the distance between locations x_i and x_j. The conditional distribution of $Y^\top = (Y_1, \dots, Y_n)$ given $T^\top = t^\top = (t_1, \dots, t_n)$ is

$$f(y|t) = \prod_{i=1}^{n} f(y_i|t_i), \tag{4.2}$$

a product of independent binomial probability functions. The likelihood function for β and θ follows as

$$L(\beta, \theta) = f(y; \beta, \theta) = \int_{\mathbb{R}^n} N(t; D\beta, \Sigma(\theta)) f(y|t)\, dt \tag{4.3}$$

where $N(\cdot; \mu, \Sigma)$ is the density function of a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ.

The MCML method (Geyer and Thompson, 1992; Geyer, 1994; Geyer, 1996; Geyer, 1999) uses conditional simulation from the distribution of T given $Y = y$ to approximate the high-dimensional integral in (4.3). Specifically, the likelihood function can be rewritten as

$$L(\beta,\theta) = \int_{\mathbb{R}^n} \frac{N(t; D\beta, \Sigma(\theta)) f(y|t)}{N(t; D\beta_0, \Sigma(\theta_0)) f(y|t)} f(y,t)\, dt$$
$$\propto \int_{\mathbb{R}^n} \frac{N(t; D\beta, \Sigma(\theta))}{N(t; D\beta_0, \Sigma(\theta_0))} f(t|y)\, dt = E_{T|y}\left[\frac{N(t; D\beta, \Sigma(\theta))}{N(t; D\beta_0, \Sigma(\theta_0))}\right] \qquad (4.4)$$

where $f(y,t) = N(t; D\beta_0, \Sigma(\theta_0)) f(y|t)$ is the joint distribution of Y and T for pre-defined, fixed values of β_0 and θ_0. We use a Markov Chain Monte Carlo (MCMC) algorithm to obtain m samples $t_{(i)}$ from the conditional distribution of T given $Y = y$ under β_0 and θ_0 and approximate (4.4) with

$$L_m(\beta,\theta) = \frac{1}{m}\sum_{i=1}^{n} \frac{N(t_{(i)}; D\beta, \Sigma(\theta))}{N(t_{(i)}; D\beta_0, \Sigma(\theta_0))}. \qquad (4.5)$$

Note that $L_m(\beta,\theta)$ is a consistent estimator of $L(\beta,\theta)$, whether or not the samples $t_{(i)}$ are correlated. The optimal choices for β_0 and θ_0 are the maximum likelihood estimates of β and θ, for which $\max_{\beta,\theta} L_m(\beta,\theta) \to 1$ as $m \to \infty$. Since our choices for β_0 and θ_0 will necessarily differ from the actual maximum likelihood estimates, the distance of $L_m(\hat{\beta}_m, \hat{\theta}_m)$ from 1, where $\hat{\beta}_m$ and $\hat{\theta}_m$ are the MCML estimates, can be used as a measure of quality of the Monte Carlo approximation. In practice, we embed the maximisation of $L_m(\beta,\theta)$ within the following iterative procedure. Let $\hat{\beta}_m$ and $\hat{\theta}_m$ denote the values that maximise $L_m(\beta,\theta)$ using an initial guess at suitable values β_0 and θ_0, repeat the maximisation with $\beta_0 = \hat{\beta}_m$ and $\theta_0 = \hat{\theta}_m$ and continue until convergence.

For maximization of the approximation to the log-likelihood $l_m(\beta,\theta) = \log L_m(\beta,\theta)$ in `PrevMap`, the user can choose between a BFGS algorithm or unconstrained optimization with PORT routines. Let $\psi = \log\theta$; analytical expressions for the first and second derivatives of $l_m(\beta,\psi)$ with respect to β and ψ are internally passed to the optimization functions `maxBFGS` of the `maxLik` package (Henningsen and Toomet, 2011) in the former case and to the `nlminb` function in the latter. This can be very useful in order to better locate the global maximum on a generally flat likelihood surface, as it is often the case for the ψ parameter. The MCML standard errors are then estimated by taking the square-roots of the diagonal elements of the inverse of the negative Hessian of $l_m(\hat{\beta}_m, \hat{\psi}_m)$. The accuracy of such an approximation for the standard errors is context-specific and is also affected by the Monte Carlo error. As a partial check, the resulting standard errors for β are typically larger than those estimated using an ordinary logistic regression. In the examples of Section 4.3.3 and Section 4.4, the number of simulated samples is sufficiently large to make the Monte Carlo error negligible.

In the `PrevMap` package, conditional simulation of T given y with fixed parameters β and

TABLE 4.2: Some of the main functions available in the `PrevMap` package. Note that all of the listed functions include an option to use a low-rank approximation procedure.

Function	Model	Method of inference	Type of use
`binomial.logistic.MCML`	Binomial	Classical	Parameter estimation
`binomial.logistic.Bayes`	Binomial	Bayesian	Parameter estimation
`linear.model.MLE`	Linear	Classical	Parameter estimation
`linear.model.Bayes`	Linear	Bayesian	Parameter estimation
`spatial.pred.binomial.MCML`	Binomial	Classical	Spatial prediction
`spatial.pred.binomial.Bayes`	Binomial	Bayesian	Spatial prediction
`spatial.pred.linear.MLE`	Linear	Classical	Spatial prediction
`spatial.pred.linear.Bayes`	Linear	Bayesian	Spatial prediction

θ is implemented by the function `Laplace.sampling`. This function uses a Langevin-Hastings algorithm to update the random variable $\hat{T} = \hat{\Sigma}^{1/2}(T - \hat{t})$, where $\hat{t}$ and $\hat{\Sigma}$ are respectively the mode and the inverse of the negative Hessian of the density of the conditional distribution. The objective of this linear transformation is to break the dependence between the different components of T so as to allow for faster convergence of the MCMC algorithm. However, when using the function `binomial.logistic.MCML` for parameter estimation, conditional simulation is carried out internally; see Section 4.3.3.1.

4.2.2 Bayesian inference

In the Bayesian framework, a joint prior distribution for β and θ is combined with the likelihood function through Bayes' theorem so as to obtain the corresponding posterior distribution. We assume that the prior distributions for θ and β are of the form

$$\begin{aligned} \theta &\sim g(\cdot), \\ \beta|\sigma^2 &\sim N(\cdot; \xi, \sigma^2\Omega) \end{aligned}$$

where $g(\cdot)$ can be any distribution for θ, and ξ and Ω are the mean vector and a p by p covariance matrix for the Gaussian prior of β. The posterior distribution for β, θ and T is given by

$$\pi(\beta, \theta, t|y) \propto g(\theta)N(\beta; \xi, \sigma^2\Omega)N(t; D\beta, \Sigma(\theta))f(y|t). \tag{4.6}$$

The function `binomial.logistic.Bayes` can be used to obtain samples from the above posterior distribution. This uses an MCMC algorithm, where θ, β and T are updated in turn using the following procedure.

1. Initialise β, θ and T.

2. Following the procedure proposed by Christensen, Roberts, and Sköld (2006), use the following re-parametrization for the covariance parameters

$$(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3) = (\log \sigma, \log(\sigma^2/\phi^{2\kappa}), \log \tau^2)$$

and update each of them in turn using a random-walk Metropolis Hastings (RWMH). In each of the three RWMH for $\tilde{\theta}_1$, $\tilde{\theta}_2$ and $\tilde{\theta}_3$, the standard deviation, h say, of the Gaussian proposal at i-th iteration is given by

$$h_i = h_{i-1} + c_1 i^{-c_2}(\alpha_i - 0.45), \tag{4.7}$$

where $c_1 > 0$ and $c_2 \in (0, 1]$ are pre-defined constants, α_i is the acceptance probability at the i-th iteration and 0.45 is the optimal acceptance probability for a univariate Gaussian distribution.

3. Update β using a Gibbs step. The required conditional distribution of β given θ and T is Gaussian, independent of y and with mean $\tilde{\xi}$ and covariance matrix $\sigma^2\tilde{\Omega}$ given by

$$\begin{aligned} \tilde{\xi} &= \tilde{\Omega}(\Omega^{-1}\xi + D^\top R(\theta)^{-1} T) \\ \sigma^2\tilde{\Omega} &= \sigma^2(\Omega^{-1} + D^\top R(\theta)^{-1} D)^{-1}, \end{aligned}$$

where $\sigma^2 R(\theta) = \Sigma(\theta)$.

4. Update the distribution of T given β, θ and y using a Hamiltonian Monte Carlo algorithm (Neal, 2011). Specifically, let $H(t, u)$ be the Hamiltonian function

$$H(t, u) = u^\top u/2 - \log f(t|y, \beta, \theta),$$

where $u \in \mathbb{R}^n$ is the vector of the momentum variables and $f(t|y, \beta, \theta)$ is the conditional density of T given β, θ and y. The partial derivatives of $H(u, t)$ determine how u and t change over time v according to the Hamiltonian equations

$$\begin{aligned} \frac{dt_i}{dv} &= \frac{\partial H}{\partial u_i}, \\ \frac{du_i}{dv} &= -\frac{\partial H}{\partial t_i} \end{aligned}$$

for $i = 1, \ldots, n$. In order to implement the Hamiltonian dynamic, the above differential equations are discretized using the *leapfrog* method (Neal, 2011, pp. 121-122) and approximate solutions are then found.

Two auxiliary functions, `control.prior` and `control.mcmc.Bayes`, define prior distributions and tuning parameters for the above MCMC scheme.

4.2.3 Empirical logit transformation

An alternative approach to exact fitting methods is to use a trans-Gaussian approximation of the model in (4.1). This consists of fitting a linear model to the empirical logit transformation of the data,

$$\tilde{Y}_i = \log\left(\frac{Y_i + 1/2}{m_i - Y_i + 1/2}\right) : i = 1, \ldots, n. \tag{4.8}$$

The method then assumes that $\tilde{Y}_i | S(x_i) \sim N(d(x_i)^\top \beta + S(x_i), \tau^2)$ with $S(x)$ having the same properties as previously defined. Guidance on when this model can be safely used is given by Stanton and Diggle (2013).

In the `PrevMap` package the empirical logit transformation is implemented both for classical and Bayesian inference in the functions `linear.model.MLE` and `linear.model.Bayes`.

4.2.4 Low-rank approximation

The Gaussian process $S(x)$ in (4.1) can be represented as a convolution of Gaussian noise (Higdon, 1998; Higdon, 2002),

$$S(x) = \int_{\mathbb{R}^2} K(\|x - t\|; \phi, \kappa)\, dB(t) \tag{4.9}$$

where B is Brownian motion, $\|\cdot\|$ is the Euclidean distance and $K(\cdot)$ is the Matérn kernel given by the following expression

$$K(u; \phi, \kappa) = \frac{\Gamma(\kappa+1)^{1/2}\kappa^{(\kappa+1)/4}u^{(\kappa-1)/2}}{\pi^{1/2}\Gamma((\kappa+1)/2)\Gamma(\kappa)^{1/2}(2\kappa^{1/2}\phi)^{(\kappa+1)/2}}\mathcal{K}_\kappa(u/\phi), u > 0. \tag{4.10}$$

Let $(\tilde{x}_1, \ldots, \tilde{x}_r)$ be a grid of spatial knots. By discretization of equation (4.9), and for r sufficiently large, we obtain a low-rank approximation

$$S(x) \approx \sum_{i=1}^{r} K(\|x - \tilde{x}_i\|; \phi, \kappa)U_i, \tag{4.11}$$

where the U_i are independent zero-mean Gaussian variables with variance σ^2. This approximation is particularly beneficial for relatively large values of the scale parameter ϕ, when a small number of spatial knots is required to give a good approximation over the study-region. Note also that the number of spatial knots r is independent of the sample size n, making this approach computationally attractive when n is large.

Since the resulting approximation in (4.11) is no longer a stationary process, we adjust the value of σ^2 by multiplying it by the following quantity

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} K(\|\tilde{x}_j - \tilde{x}_i\|; \phi, \kappa)^2.$$

The adjusted value of σ^2 is then a closer approximation to the actual variance of the Gaussian process $S(x)$.

Different implementations of this method are possible, depending on whether we use an exact fitting method or an empirical logit approximation. In the `PrevMap` package, low-rank approximations can be used in each of the fitting functions listed in Table 4.2; we give an example in Section 4.4.

Implementations of the low-rank approximation for the BL and linear model are as follows.

- *BL model.* In this implementation the nugget effect is not included, hence $\tau^2 = 0$. For both the classical and Bayesian analysis, conditional simulation from the distribution of the random effect U given the data y (and the model parameters in the Bayesian case) is used, hence avoiding matrix inversion.

- *Linear model.* The low-rank approximation is here used for the empirical logit transformation of prevalence. In this case $\tau^2 > 0$, since the nugget effect is now a proxy for binomial sampling variation. Inversion of the covariance matrix and computation of the determinant are simplified as follows. Let $K(\theta)$ denote the n by r kernel matrix. The covariance matrix now assumes the form

  $$\Sigma(\theta) = \sigma^2 K(\theta)K(\theta)^\top + \tau^2 I_n$$

  where I_n is the n by n identity matrix. The Woodbury identity for matrix inversion gives

  $$\Sigma(\theta)^{-1} = \sigma^2\nu^{-2}(I_n - \nu^{-2}K(\theta)(\nu^{-2}K(\theta)^\top K(\theta) + I_r)^{-1}K(\theta)^\top)$$

  where $\nu^2 = \tau^2/\sigma^2$. Inversion of $\Sigma(\theta)$ now requires inversion of an r by r matrix. Computation of the determinant, denoted by $|\cdot|$, can also be simplified by using

Sylvester's determinant theorem. This gives

$$\begin{aligned} |\Sigma(\theta)| &= |\sigma^2 K(\theta)K(\theta)^\top + \tau^2 I_n| \\ &= \tau^{2n}|\nu^{-2}K(\theta)^\top K(\theta) + I_r|, \end{aligned}$$

which again reduces the dimensionality of the required matrix operations from n by n to r by r.

4.2.5 Spatial prediction

We now consider the prediction of $T^* = (T(x_{n+1}), \ldots, T(x_{n+q}))^\top$ at q additional locations not included in the data. This requires all relevant explanatory variables to be available at the prediction locations. We do not include the mutually independent random variables Z_i in (4.1) as part of our target for prediction, hence $T(x_{n+i}) = d(x_{n+i})^\top \beta + S(x_{n+i})$ for $i = 1, \ldots, q$.

Conditionally on $T^\top = (T_1, \ldots, T_n)$, β, θ and y, the target for prediction T^* follows a multivariate Gaussian distribution with mean and covariance matrix

$$\mu^*(T) = D^*\beta + C\Sigma^{-1}(T - D\beta), \tag{4.12}$$

$$\Sigma^* = V - C\Sigma^{-1}C^\top, \tag{4.13}$$

where C is the cross-covariance matrix between T and T^*, V is the covariance matrix of T^* and D^* is a q by p matrix of explanatory variables at the prediction locations. Let $T^*_{(j)}$ denote the j-th simulated sampled from the posterior distribution of T^* for $j = 1, \ldots, m$. If the sample mean is to be used as a point predictor of T, the package uses the following result to reduce the associated Monte Carlo error,

$$E_{T^*|y}E[T^*] = E_{T,\beta,\theta|y}[E_{T^*|T,\beta,\theta,y}[T^*]] = E_{T,\beta,\theta|y}[\mu^*(T)] \approx \frac{1}{m}\sum_{j=1}^{m}\mu^*(T_{(j)}).$$

Prediction of the functional $W(T^*)^\top = (W(T(x_{n+1})), \ldots, W(T(x_{n+q})))$ follows immediately by computing $W_{(j)} = W(T^*_{(j)})$ for $j = 1, \ldots, m$. The `PrevMap` package provides automatic computation of the following functionals.

- *Prevalence:* $W(T(x_{n+i})) = \exp\{T(x_{n+i})\}/(1 + \exp\{T(x_{n+i})\})$.

- *Odds:* $W(T_{n+i}) = \exp\{T(x_{n+i})\}$. Let $\sigma^{2*} = \text{diag}(\Sigma^*)$ denote the vector of conditional variances. In this case, the Monte Carlo error in the computation of the posterior mean is reduced by noticing that $E_{T^*|T,\beta,\theta,y}[\exp\{T^*\}] = \exp\{\mu^*(T) + \sigma^{2*}/2\}$, hence
$$E_{T^*|y}[\exp\{T^*\}] \approx \frac{1}{m}\sum_{j=1}^{m} \exp\{\mu^*(T_{(j)}) + \sigma^{2*}_{(j)}/2\}.$$

Another summary of the posterior distribution that is often relevant, particularly in problems of hotspot detection, is the exceedance probability $P(T(x_{n+i}) > l \mid y)$ for a given threshold l and $i = 1, \ldots, q$. We estimate this as

$$\frac{1}{m}\sum_{j=1}^{m} I\left(T_{(j)}(x_{n+i}) > l\right),$$

where $I(a > l)$ is 1 if $a > l$ and 0 otherwise, and $T_{(j)}(x_{n+i})$ is the i-th element of $T^*_{(j)}$.

The `spatial.pred.binomial.MCML` and `spatial.pred.binomial.Bayes` functions can be used for classical and Bayesian spatial prediction, respectively. As we later illustrate, one of the available options is also the computation of either joint or marginal predictions. For example, joint predictions are needed when the target for prediction is an average over a sub-region. Spatial prediction for the empirical logit transformation using classical and Bayesian approaches is implemented in the `spatial.pred.linear.MLE` and `spatial.pred.linear.Bayes` functions, respectively. Low-rank approximations for each of the above functions are also available; see Section 4.3.3.

4.3 Example: Loa loa prevalence mapping

The data that we analyse relate to a study of the prevalence of *Loa loa* (eyeworm) in a series of surveys undertaken in 197 villages in Cameroon and southern Nigeria; see Diggle et al. (2007) for more details. Figure 4.2(a) shows the locations of the sampled villages.

4.3.1 Choosing initial values

Choosing initial values for the model parameters is the first step in both classical and Bayesian analysis. Initial values for the regression coefficients can be easily obtained from an ordinary logistic regression fit. Choosing initial values for the covariance parameters is less straightforward. The shape parameter κ of the Matérn function is typically chosen from a discrete set of candidate values, which can be compared by evaluating a profile

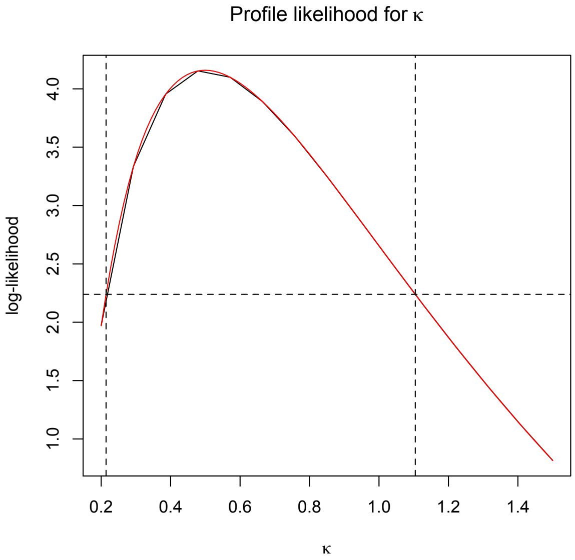

FIGURE 4.1: Profile likelihood for the shape parameter κ of the Matérn covariance function, obtained using the function `shape.matern`; the profile likelihood (black solid line) is interpolated by a spline (red solid line), which is then used to obtain a confidence interval of coverage 95% (vertical dashed lines).

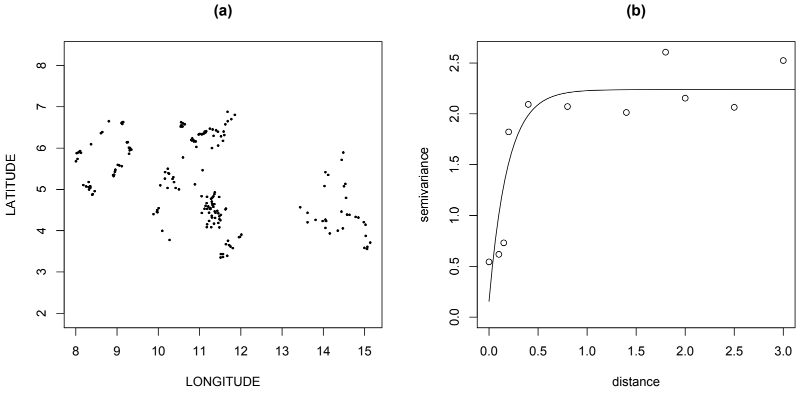

FIGURE 4.2: (a) Sampling locations for the *Loa loa* data. (b) Empirical variogram for the empirical logit transformation of the observed prevalence with theoretical variogram (solid line) obtained by least-squares estimation.

likelihood for κ based on the empirical logit transformation of the observed prevalence, as in the following example.

```
R> library("PrevMap")
R> data("loaloa")
R> loaloa$logit <- log((loaloa$NO_INF + 0.5)/
+                      (loaloa$NO_EXAM - loaloa$NO_INF + 0.5))
R> profile.kappa <- shape.matern(formula = logit ~ 1,
+               coords = ~ LONGITUDE + LATITUDE,
+               data = loaloa, set.kappa = seq(0.2,1.5, length = 15),
+               start.par = c(0.2,0.05), coverage = 0.95)
R>
R> c(profile.kappa$lower, profile.kappa$upper)
```

```
[1] 0.2140705 1.1044392
```

```
R> profile.kappa$kappa.hat
```

```
[1] 0.4991899
```

The shape.matern function evaluates the profile likelihood for κ and obtains a corresponding confidence interval with coverage specified by the argument coverage. The set of values that are used for evaluation of the profile likelihood is specified through the set.kappa argument. Computation of the confidence interval uses the interpolated profile log-likelihood as shown in Figure 4.1: the red line corresponds to an interpolating spline and the likelihood threshold, denoted by the horizontal dashed line, is obtained using the asymptotic distribution of a chi-squared with one degree of freedom. Since the maximum likelihood estimate is very close to 1/2, we then fix the shape parameter κ at this value for the subsequent analysis.

The package geoR provides several functions that are useful for an initial exploratory analysis of geostatistical data. For example, using the function variofit, a least-squares estimation of the empirical variogram can be used in order to choose initial values for the covariance parameters of the Gaussian spatial process.

```
R> library("geoR")
R> coords <- as.matrix(loaloa[, c("LONGITUDE", "LATITUDE")])
R> vari <- variog(coords = coords, data = loaloa$logit,
+               uvec = c(0, 0.1, 0.15, 0.2,
+               0.4, 0.8, 1.4, 1.8, 2, 2.5, 3))
R> vari.fit <- variofit(vari, ini.cov.pars = c(2, 0.2),
+                cov.model = "matern",
+                 fix.nugget = FALSE, nugget = 0 ,
+                 fix.kappa = TRUE, kappa = 0.5)
R> par(mfrow = c(1,2))
R> plot(coords, pch = 20, asp = 1, cex = 0.5, main = "(a)")
R> plot(vari, main = "(b)")
R> lines(vari.fit)
R> vari.fit
```

```
variofit: model parameters estimated by WLS (weighted least squares):
covariance model is: matern with fixed kappa = 0.5 (exponential)
parameter estimates:
  tausq sigmasq     phi
 0.1554  2.0827  0.1890
Practical Range with cor = 0.05 for asymptotic range: 0.5662674

variofit: minimised weighted sum of squares = 780.6663
```

The above code computes the empirical logit transformation of the observed *Loa loa* prevalence, uses this to obtain the empirical variogram with the `variog` function and fits an exponential correlation function to the empirical variogram with the `variofit` function, which uses a least squares curve-fitting criterion. The results are shown in Figure 4.2(b).

4.3.2 Linear model

In this section we show how to fit a linear model with Matérn correlation function to the empirical logit transformation of the *Loa loa* data using the maximum likelihood method. The `linear.model.MLE` function has its counterpart in the `likfit` function in `geoR` but, unlike `likfit`, uses analytic expressions for the gradient function and Hessian matrix, and delivers an estimated covariance matrix of the maximum likelihood estimates accordingly. As shown in the next section, the `binomial.logistic.MCML` function uses the same approach in fitting a BL model.

```
R> fit.MLE <- linear.model.MLE(formula = logit ~ 1,
+                  coords = ~ LONGITUDE + LATITUDE, data = loaloa,
+                  start.cov.pars = c(0.2, 0.15), kappa = 0.5)
R>
R> summary(fit.MLE, log.cov.pars = FALSE)


Geostatistical linear Gaussian model
Call:
geo.linear.MLE(formula = formula, coords = coords, data = data,
    kappa = kappa, fixed.rel.nugget = fixed.rel.nugget,
    start.cov.pars = start.cov.pars,
    method = method)

            Estimate  StdErr z.value   p.value
(Intercept)  -2.2986  0.5469  -4.203 2.634e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -94.34047

Covariance parameters Matern function (kappa = 0.5)
        Estimate StdErr
sigma^2  2.45148 0.1393
phi      0.84398 0.4933
tau^2    0.36865 1.1717

Legend:
sigma^2 = variance of the Gaussian process
phi = scale of the spatial correlation
tau^2 = variance of the nugget effect
```

The first argument of `linear.model.MLE` specifies the covariates used in the regression as a `formula` object; in this case `formula = logit ~ 1` since we only fit an intercept. The argument `start.cov.pars` provides the initial values of ϕ and $\nu^2(=\tau^2/\sigma^2)$, respectively, used in the optimization algorithm. The argument `fixed.rel.nugget` allows the relative variance of the nugget effect ν^2 to be fixed if desired. Additionally, two different maximisation algorithms are available: if `method = "BFGS"` (set by default), the `maxBFGS` function in the `maxLik` package is used, otherwise `method = "nlminb"` and the `nlminb` function is then used for unconstrained optimization using PORT routines.

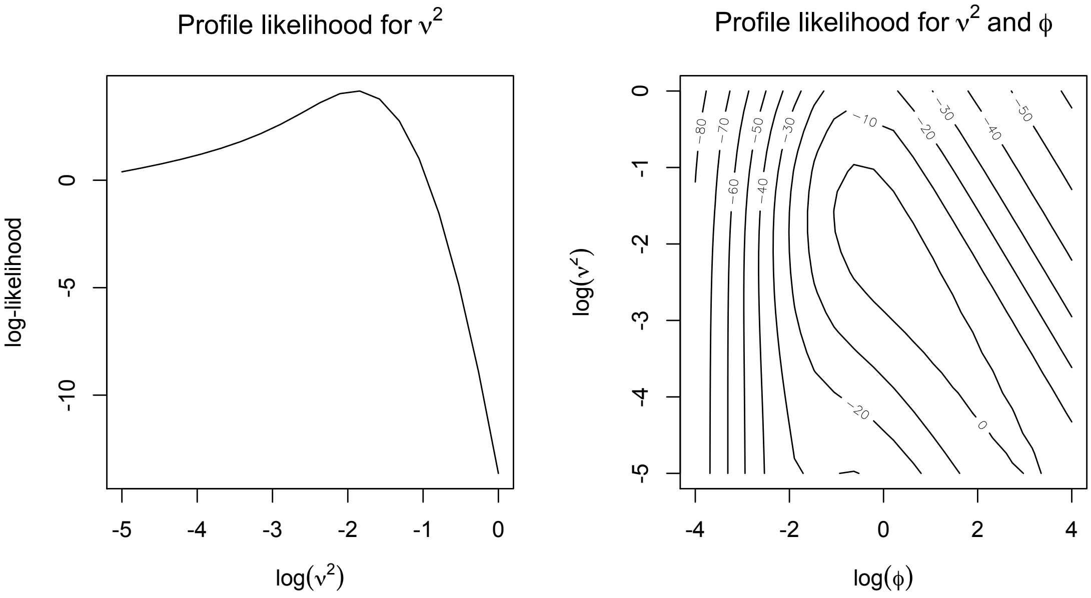

FIGURE 4.3: Profile log-likelihood for ν^2 (left panel) and (ν^2, ϕ) (right panel) obtained using the function `loglik.linear.model`.

When calling a summary of the fitted model, estimates and standard errors of the covariance parameters are given on the log-scale by default. Setting `log.scale = FALSE` gives estimates and standard errors on the original scale.

The function `loglik.linear.model` can be used either for computation of the profile likelihood for ϕ and/or ν^2 or for evaluation of the likelihood keeping the other parameters fixed. The auxiliary function `control.profile` is used to define the set of values for ϕ and/or ν^2 used in the evaluation of the likelihood, and the fixed values for β and σ^2, if necessary. The shape parameter κ is also fixed at the value defined in the fitted model object that must be specified as first argument of `loglik.linear.model`.

```
R> cp1 <- control.profile(rel.nugget = exp(seq(-5, 0, length = 20)))
```

```
Control profile: parameters have been set for
evaluation of the profile log-likelihood.
```

```
R> cp2 <- control.profile(rel.nugget = exp(seq(-5, 0, length = 20)),
+                 phi = exp(seq(-4, 4, length = 20)))
```

```
Control profile: parameters have been set for
evaluation of the profile log-likelihood.
```

```
R> lp1 <- loglik.linear.model(fit.MLE, cp1, plot.profile = FALSE)
R>
R> lp2 <- loglik.linear.model(fit.MLE, cp2, plot.profile = FALSE)
R>
R> par(mfrow = c(1, 2))
R> plot(lp1, type = "l", log.scale = TRUE,
+        xlab = expression(log(nu^2)),
+        ylab = "log-likelihood",
+        main = expression("Profile likelihood for" ~ nu^2))
R> plot(lp2, log.scale = TRUE, xlab = expression(log(phi)),
+   ylab = expression(log(nu^2)),
+   main = expression("Profile likelihood for" ~ nu^2 ~ "and" ~ phi))
```

The resulting plots of the profile log-likelihood for ν^2 and the profile log-likelihood surface of (ν^2, ϕ) are shown in Figure 4.3. These are generated using the function `plot.profile.PrevMap` as an S3 method, in which the logical argument `log.scale` can be set to `TRUE` in order to plot the profile likelihood on the log-scale of the chosen parameters. Likelihood-based confidence intervals for ϕ or ν^2 can also be obtained by using the `loglik.ci` function. As with the `shape.matern` function, the `loglik.ci` function uses a spline to interpolate the univariate profile likelihood and obtain a confidence interval of coverage specified by `coverage`.

```
R> ci0.95 <- loglik.ci(lp1, coverage = 0.95, plot.spline.profile = FALSE)
```

```
Likelihood-based 95% confidence interval: (0.04460758, 0.2936487)
```

4.3.3 Binomial logistic model

We now show how to fit a BL model to the *Loa loa* data using either the MCML method (Section 4.3.3.1) or a Bayesian approach (Section 4.3.3.2).

4.3.3.1 Likelihood-based analysis

For the MCML method, we set the parameters of the importance sampling distribution, β_0 and θ_0, to the estimates reported in Section 4.3.1 using ordinary logistic regression and a least squares fit to the variogram, respectively.

```
R> fit.glm <- glm(cbind(NO_INF, NO_EXAM - NO_INF) ~ 1, data = loaloa,
+                        family = binomial)
R> par0 <- c(coef(fit.glm), vari.fit$cov.pars, vari.fit$nugget)
R> c.mcmc <- control.mcmc.MCML(n.sim = 10000, burnin = 2000,
+                 thin = 8, h = (1.65)/(nrow(loaloa) ^ (1/6)))
R> fit.MCML1 <- binomial.logistic.MCML(formula = NO_INF ~ 1,
+               units.m = ~ NO_EXAM, par0 = par0,
+               coords = ~ LONGITUDE + LATITUDE, data = loaloa,
+               control.mcmc = c.mcmc,
+               kappa = 0.5,
+               start.cov.pars = c(par0[3], par0[4]/par0[2]))
R> fit.MCML1$log.lik
```

```
[1] 24.24903
```

The above code fits a BL model by simulating 10,000 samples and retaining every eighth sample after a burn-in of 2,000 values to approximate the likelihood integral. The function `control.mcmc.MCML` sets the control parameters of the MCMC algorithm. The argument `h` represents the proposal density of the Langevin-Hastings (see Section 4.2.1). Our suggestion is to set this to $1.65/n^{1/6}$, where n is the sample size, which corresponds to the optimal value for sampling from a standard multivariate Gaussian distribution (Roberts and Rosenthal, 2001).

We now repeat the MCML procedure twice, but with new values for β_0 and θ_0 set as the MCML estimates each time; in the last iteration, we also increase the number of retained simulated samples to 10,000.

```
R> par0 <- coef(fit.MCML1)
R> start <- c(par0[3], par0[4]/par0[2])
```

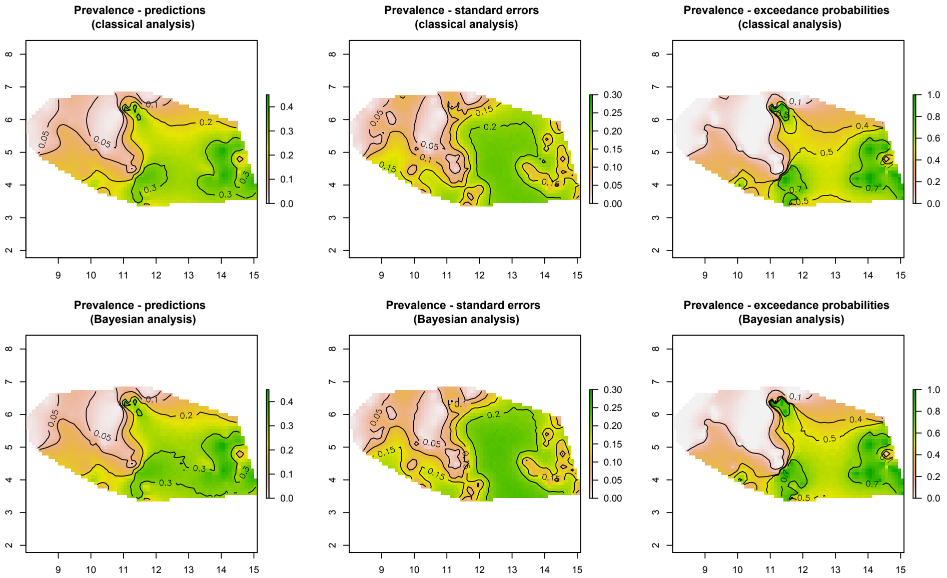

FIGURE 4.4: Plots of the prevalence estimates, standard errors and exceedance probabilities for the *Loa loa* data from the MCML (upper panels) and Bayesian (lower panels) analyses.

```
R> fit.MCML2 <- binomial.logistic.MCML(formula = NO_INF ~ 1,
+                               units.m = ~ NO_EXAM, par0 = par0,
+                              coords = ~ LONGITUDE + LATITUDE, data = loaloa,
+                              control.mcmc = c.mcmc,
+                              kappa = 0.5,
+                              start.cov.pars = c(par0[3], par0[4]/par0[2]))
R> fit.MCML2$log.lik
```

```
[1] 1.287294
```

```
R> c.mcmc <- control.mcmc.MCML(n.sim = 65000, burnin = 5000,
+                           thin = 6, h = (1.65)/(nrow(loaloa)^(1/6)))
R> par0 <- coef(fit.MCML2)
R> fit.MCML3 <- binomial.logistic.MCML(formula = NO_INF ~ 1,
+                          units.m = ~ NO_EXAM,par0=par0,
+                          coords = ~LONGITUDE+LATITUDE,data=loaloa,
+                          control.mcmc = c.mcmc,
+                          kappa = 0.5, start.
+                          cov.pars = c(par0[3],par0[4]/par0[2]))
R> summary(fit.MCML3)
```

```
Binomial geostatistical model
Call:
binomial.logistic.MCML(formula = NO_INF ~ 1, units.m = ~NO_EXAM,
    coords = ~LONGITUDE + LATITUDE, data = loaloa, par0 = par0,
    control.mcmc = c.mcmc, kappa = 0.5, start.cov.pars = c(par0[3],
        par0[4]/par0[2]))

            Estimate   StdErr z.value   p.value
(Intercept) -2.30556  0.51743 -4.4558 8.358e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Objective function: 0.1366855

Covariance parameters Matern function (kappa=0.5)
             Estimate StdErr
log(sigma^2)  0.92408 0.3215
log(phi)     -0.28736 0.3804
log(tau^2)   -3.23648 1.5796

Legend:
sigma^2 = variance of the Gaussian process
phi = scale of the spatial correlation
tau^2 = variance of the nugget effect
```

Note that updating β_0 and θ_0 with the resulting MCML estimates at each iteration results in the maximum value of the approximation to the log-likelihood function approaching zero. This is an indication that the MCML estimates are converging towards the actual maximum likelihood estimates of β and θ, for which the value of the Monte Carlo likelihood is exactly zero. We now carry out spatial predictions over a 0.1 by 0.1 degree regular grid, fixing the model parameters at the MCML estimates, and summarise the predictive distribution of prevalence in each grid cell through its mean, standard deviation and probability that the estimated prevalence is above 20%.

```
R> library("splancs")
```

```
R> poly <- coords[chull(coords),]
R> grid.pred <- gridpts(poly, xs = 0.1, ys = 0.1)
R> pred.MCML <- spatial.pred.binomial.MCML(fit.MCML3, grid.pred,
+                         control.mcmc = c.mcmc, type = "marginal",
+                         scale.predictions = "prevalence",
+                         standard.errors = TRUE, thresholds = 0.2,
+                          scale.thresholds = "prevalence")
R>
R> par(mfrow = c(1,3))
R> plot(pred.MCML, type = "prevalence",
+   summary = "predictions", zlim = c(0,0.45),
+   main = "Prevalence - predictions \n (classical analysis)")
R> contour(pred.MCML, type = "prevalence",
+   summary = "predictions",
+   levels = c(0.05,0.1,0.2,0.3), add = TRUE)
R> plot(pred.MCML, type = "prevalence",
+   summary = "standard.errors", zlim = c(0,0.3),
+   main = "Prevalence - standard errors  \n (classical analysis)")
R> contour(pred.MCML, type = "prevalence",
+   summary = "standard.errors",
+   levels = c(0.05,0.1,0.15,0.2), add = TRUE)
R> plot(pred.MCML, summary = "exceedance.prob",
+   zlim = c(0,1),
+   main = "Prevalence - exceedance probabilities
+              \n (classical analysis)")
R> contour(pred.MCML, summary = "exceedance.prob",
+              levels = c(0.1,0.4,0.5,0.7), add = TRUE)
```

Using the argument `type` in `spatial.pred.binomial.MCML`, we can specify either marginal (`type = "marginal"`) or joint (`type = "joint"`) predictions. Through `scale.predictions`, we can also specify the scale on which predictions are required: `"logit"`, `"prevalence"` or `"odds"`. Exceedance probability thresholds and the scale on which they are provided are specified through the arguments `thresholds` and `scale.thresholds`, respectively. Figure 4.4 shows the images of prevalence estimates, standard errors and exceedance probabilities with associated contours. These plots are obtained using the methods `plot.pred.PrevMap` and `contour.pred.PrevMap`, whose arguments `type` and `summary` can be used to specify which summaries should be displayed.

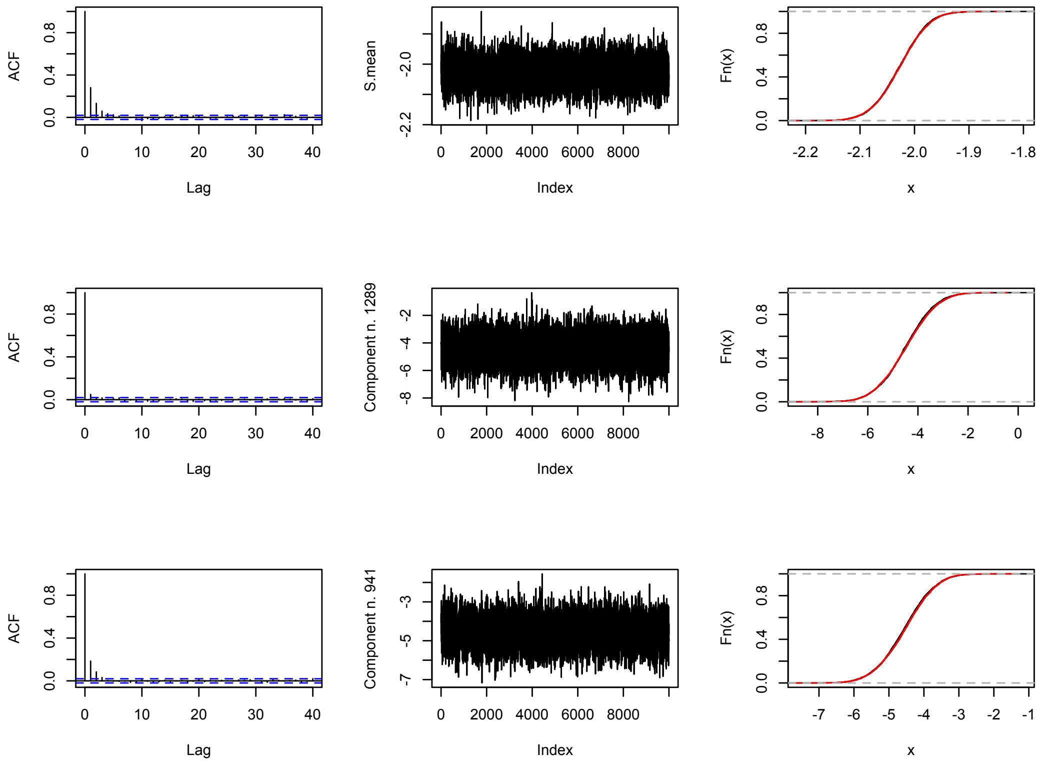

Figure 4.5: Autocorrelation plot of a thinned sequence of 10000 MCMC samples (left panels), trace plot of the same sequence (central panels) and empirical cumulative distribution plots for the first 5000 and second 5000 samples (right panels), for the spatial average of predicted logit-transformed prevalence (first row) and for the predicted logit-transformed prevalence at two randomly selected locations (second and third rows).

The following code generates a set of diagnostic plots, shown in Figure 4.5, that provide checks on convergence of the MCMC.

```
R> par(mfrow=c(3,3))
R> S.mean <- apply(pred.MCML$samples, 2, mean)
R> acf(S.mean,main = "")
R> plot(S.mean,type = "l")
R> plot(ecdf(S.mean[1:5000]), main = "")
R> lines(ecdf(S.mean[5001:10000]), col = 2, lty = "dashed")
+
R> ind.S <- sample(1:nrow(grid.pred), 2)
R> acf(pred.MCML$samples[ind.S[1],], main = "")
R> plot(pred.MCML$samples[ind.S[1], ],
+   ylab = paste("Component n.", ind.S[1]), type = "l")
R> plot(ecdf(pred.MCML$samples[ind.S[1], 1:5000]), main = "")
R> lines(ecdf(pred.MCML$samples[ind.S[1], 5001:10000]),
+   col = 2, lty = "dashed")
+
R> ind.S <- sample(1:nrow(grid.pred), 2)
R> acf(pred.MCML$samples[ind.S[2],], main = "")
R> plot(pred.MCML$samples[ind.S[2], ],
+   ylab = paste("Component n.", ind.S[2]), type = "l")
R> plot(ecdf(pred.MCML$samples[ind.S[2], 1:5000]), main = "")
R> lines(ecdf(pred.MCML$samples[ind.S[2], 5001:10000]),
+   col = 2, lty = "dashed")
```

In the first row of Figure 4.5, the target for prediction is the spatial average of logit-transformed prevalence, in the second and third rows the target is logit-transformed prevalence at each of two randomly sampled location. The three columns show: the autocorrelation plot of a thinned sequence of 10000 MCMC samples; the trace plot of these same 10000 samples; the empirical cumulative distribution functions of the first 5000 and the second 5000 of these 10000 samples. None of these plots show any evidence of non-convergence.

4.3.3.2 Bayesian analysis

For a Bayesian analysis of the *Loa loa* data, we use the following prior specification:

$$\begin{aligned} \phi &\sim \text{Uniform}(0,8), \\ \log(\sigma^2) &\sim N(\cdot;1,25), \\ \log(\tau^2) &\sim N(\cdot;-3,1), \\ \beta|\sigma^2 &\sim N(\cdot;0,\sigma^2 100^2). \end{aligned}$$

In the `PrevMap` package, the `control.prior` function can be used to set a Gaussian prior on β and any required prior distribution for the covariance parameters σ^2, ϕ and τ^2. The arguments `beta.mean` and `beta.covar` are the mean vector and the covariance matrix of the Gaussian prior for β. Log-Gaussian and uniform priors can also be directly defined for each covariance parameter by using the corresponding arguments. For example, `log.normal.sigma2` and `uniform.sigma2` define log-Gaussian and uniform priors, respectively, for σ^2. In both cases a vector of length two must be provided. If the prior is log-Gaussian the two elements are the mean and standard deviation of the distribution on the log scale. If the prior is uniform the two elements are the lower and upper limits of the support of the uniform distribution.

```
R> cp <-control.prior(beta.mean = 0, beta.covar = 100^2,
                      log.normal.sigma2 = c(1,5),
                      uniform.phi  = c(0,8),
                      log.normal.nugget = c(-3,1))
```

If different priors are required for the covariance parameters, user-defined functions of the prior log-density can be specified through the arguments `log.prior.sigma2`, `log.prior.phi` and `log.prior.nugget`.

Control parameters for the MCMC algorithm (see Section 4.2.2) are specified with the function `control.mcmc.Bayes`.

```
R> mcmc.Bayes <- control.mcmc.Bayes(n.sim = 6000,
+                burnin = 1000, thin = 1,
+                h.theta1 = 1, h.theta2 = 0.7, h.theta3 = 0.05,
+                L.S.lim = c(5,50), epsilon.S.lim = c(0.03,0.06),
+                start.beta  = -2.3, start.sigma2 = 2.6,
+                start.phi = 0.8, start.nugget = 0.05,
+                start.S = predict(fit.glm))
```

The arguments `h.theta1`, `h.theta2` and `h.theta3` are the starting values for the standard deviations of the Gaussian proposals; these are then tuned according to the adaptive scheme given by (4.7). The control parameters for the Hamiltonian Monte Carlo procedure, used to update the random effects, are `L.S.lim` and `epsilon.S.lim`. These represent, respectively, the intervals used to randomly generate from a uniform distribution the number of steps and the step size in the *leapfrog* method at each iteration of the MCMC (see Section 4.2.2).

```
R> fit.Bayes <- binomial.logistic.Bayes(formula = NO_INF ~ 1,
+                     units.m = ~ NO_EXAM,
+                    coords = ~ LONGITUDE + LATITUDE,
+                    data = loaloa, control.prior = cp,
+                    control.mcmc = mcmc.Bayes, kappa = 0.5)
R>
R> summary(fit.Bayes, hpd.coverage = 0.95)


Bayesian binomial geostatistical logistic model
Call:
binomial.logistic.Bayes(formula = NO_INF ~ 1, units.m = ~ NO_EXAM,
    coords = ~ LONGITUDE + LATITUDE,
    data = loaloa, control.prior = cp,
    control.mcmc = mcmc.Bayes, kappa = 0.5)

                 Mean   Median      Mode   StdErr HPD 0.025 HPD 0.975
(Intercept) -2.696243 -2.48606 -2.305288 1.827606 -7.253424 0.5964536

Covariance parameters Matern function (kappa = 0.5)

               Mean     Median       Mode     StdErr   HPD 0.025  HPD 0.975
sigma^2 7.66349058 5.27116856 3.28389063 5.86998256 1.650528734 20.5858584
phi     2.58509412 1.79584603 1.04951602 1.98215672 0.440133492  6.9920447
tau^2   0.05250712 0.04516296 0.02365498 0.03371813 0.003049963  0.1190124

Legend:
sigma^2 = variance of the Gaussian process
phi = scale of the spatial correlation
tau^2 = variance of the nugget effect
```

The above code fits a Bayesian BL model and returns summaries of the posterior distribution for each of the model parameters. In the output, high posterior density credible intervals are also computed, with associated coverage specified through the argument `hpd.coverage`.

```
R> par(mfrow = c(2,4))
R> autocor.plot(fit.Bayes, param = "beta", component.beta = 1)
R> autocor.plot(fit.Bayes, param = "sigma2")
```

```
R> autocor.plot(fit.Bayes, param = "phi")
R> autocor.plot(fit.Bayes, param = "tau2")
R> i <- sample(1:nrow(loaloa),4)
R> autocor.plot(fit.Bayes, param = "S", component.S = i[1])
R> autocor.plot(fit.Bayes, param = "S", component.S = i[2])
R> autocor.plot(fit.Bayes, param = "S", component.S = i[3])
R> autocor.plot(fit.Bayes, param = "S", component.S = i[4])
```

Autocorrelation plots can be obtained with the `autocor.plot` function, whose argument `param` specifies the model component for which the autocorrelation plot is required. If `param = "beta"`, then `component.beta` must be used to specify the component of the regression coefficients. To display autocorrelation plots for the random effect, then `param = "S"` and `component.S` must be either a positive integer indicating the component of the random effect, or `"all"` in order to display the autocorrelation for all components in a single plot. Using a similar syntax, the functions `trace.plot` and `dens.plot` are also available for visualization of trace-plots and kernel density estimates based on the posterior samples.

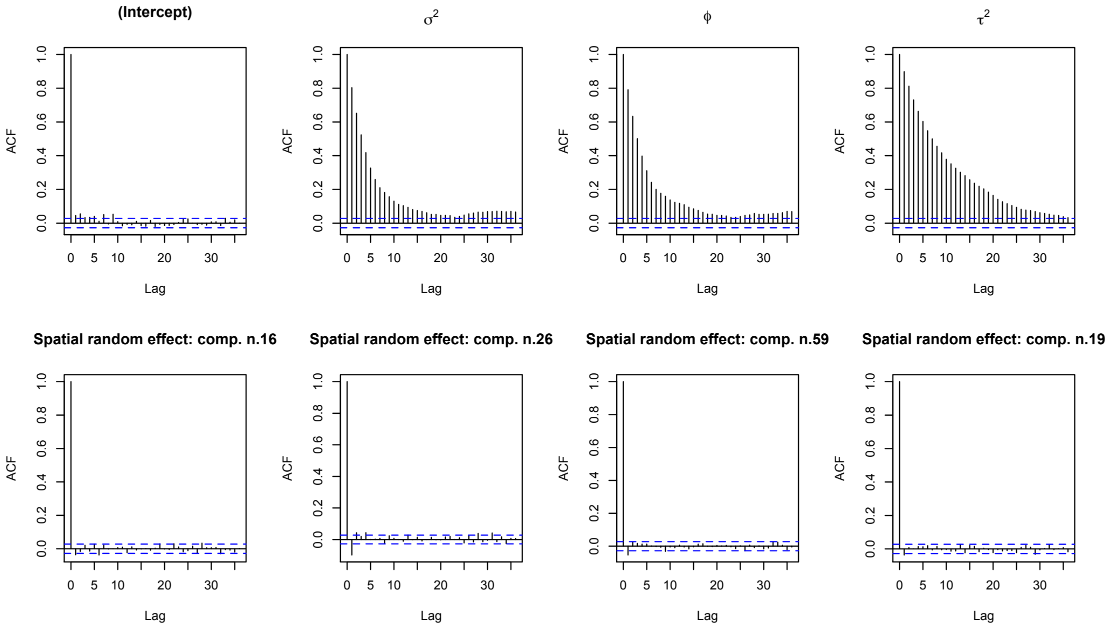

FIGURE 4.6: Autocorrelation plots for the posterior samples of β (the intercept), σ^2, ϕ, τ^2 and four randomly chosen components of the spatial random effect.

```
R> pred.Bayes <- spatial.pred.binomial.Bayes(fit.Bayes, grid.pred,
+                         type = "marginal",
+                         scale.predictions = "prevalence", quantiles = NULL,
+                         standard.errors = TRUE, thresholds = 0.2,
+                         scale.thresholds = "prevalence")
R>
R> par(mfrow = c(1,3))
R> plot(pred.Bayes, type = "prevalence", summary = "predictions",
+       zlim = c(0,0.45),
+       main = "Prevalence - predictions \n (Bayesian analysis)")
R> contour(pred.Bayes, type = "prevalence", summary = "predictions",
+       levels = c(0.05,0.1,0.2,0.3), add = TRUE)
R> plot(pred.Bayes, type = "prevalence", summary = "standard.errors",
+       zlim = c(0,0.3),
+       main = "Prevalence - standard errors \n (Bayesian analysis)")
R> contour(pred.Bayes, type = "prevalence",
+       summary = "standard.errors",
+       levels = c(0.05,0.1,0.15,0.2), add = TRUE)
R> plot(pred.Bayes, type = "exceedance.prob", zlim = c(0,1),
+       main = "Prevalence - exceedance probabilities \n
+       (Bayesian analysis)")
R> contour(pred.Bayes, type = "exceedance.prob",
+       levels = c(0.1,0.4,0.5,0.7), add = TRUE)
```

The function `spatial.pred.binomial.Bayes` generates spatial Bayesian predictions using the same syntax as `spatial.pred.binomial.MCML`. The resulting plots of the prevalence estimates, standard errors and exceedance probabilities are shown in Figure 4.4.

4.4 Example: simulated data

In this example, we use a simulated binomial data-set, available in the package as `data_sim`. For these data, a zero-mean Gaussian process was generated over a 30 by 30 grid covering the unit square, with parameters $\sigma^2 = 1$, $\phi = 0.15$ and $\kappa = 2$; the nugget effect was not included, hence $\tau^2 = 0$. Binomial observations, with 10 trials at each grid point and probabilities given by the anti-logit of the simulated values of the Gaussian process, constitute the variable `y` in the data. To illustrate the accuracy of the low-rank approximation, we analyse these data using three different grids covering the square $[-0.2, 1.2] \times [-0.2, 1.2]$ with 25, 100 and 225 spatial knots, respectively. By letting some knots lie outside of the unit square, we avoid the presence of edge-effects due to the restriction of the integral in (4.9) to a sub-region of the real plane.

```
R> data("data_sim")
R> knots1 <- expand.grid(seq(-0.2,1.2, length = 5),
+                        seq(-0.2,1.2, length = 5))
R> knots2 <- expand.grid(seq(-0.2,1.2, length = 10),
+                        seq(-0.2,1.2, length = 10))
R> knots3 <- expand.grid(seq(-0.2,1.2, length = 15),
+                        seq(-0.2,1.2, length = 15))
```

We use the MCML method to fit a BL model using both exact and approximate approaches. We then use the resulting binomial fits to generate spatial predictions of prevalence at each of the 900 sampling locations.

```
R> par0.exact <- c(0,1,0.15)
R> exact.mcmc <- control.mcmc.MCML(n.sim = 65000,
+          burnin = 5000, thin = 12,
+          h = 1.65/(nrow(data_sim)^(1/6)))
R> system.time(fit.MCML.exact <- binomial.logistic.MCML(
+               y ~ 1, units.m = ~ units.m,
+               coords = ~ x1 + x2,
+               data = data_sim, par0 = par0.exact,
+               start.cov.pars = 0.15,
+               control.mcmc = exact.mcmc,
+               kappa = 2, fixed.rel.nugget = 0,
+               method = "nlminb",
+               plot.correlogram = FALSE))
```

```
    user   system  elapsed
 2401.530  297.871 2714.146
```

```
R> par0.lr <- c(-0.219294,0.97945,0.21393)
R> lr.mcmc <- control.mcmc.MCML(n.sim = 65000,
+              burnin = 5000, thin = 12,
+              h = 1.65/(nrow(knots1)^(1/6)))
R> system.time(fit.MCML.lr1 <- binomial.logistic.MCML(
+              y ~ 1,
+              units.m = ~ units.m, coords = ~ x1 + x2,
+              data = data_sim, par0 = par0.lr,
+              start.cov.pars = par0.lr[3],
+              control.mcmc = lr.mcmc,
+              low.rank = TRUE, knots = knots1, kappa = 2,
+              method = "nlminb",
+              plot.correlogram  = FALSE))
```

```
   user  system elapsed
 72.893   2.785  77.157
```

```
R> lr.mcmc$h <- 1.65/(nrow(knots2)^(1/6))
R> par0.lr <- c(-0.017333,0.16490,0.16971)
R> system.time(fit.MCML.lr2 <- binomial.logistic.MCML(
+                y ~ 1,
+                units.m = ~ units.m, coords = ~ x1 + x2,
+                data = data_sim, par0 = par0.lr,
+                start.cov.pars = par0.lr[3], control.mcmc = lr.mcmc,
+                low.rank = TRUE, knots = knots2, kappa = 2,
+                method = "nlminb", plot.correlogram = FALSE))
```

```
   user  system elapsed
172.864  20.973 194.625
```

```
R> lr.mcmc$h <- 1.65/(nrow(knots3)^(1/6))
R> par0.lr <- c(-0.031759,0.30572, 0.18854)
R> system.time(fit.MCML.lr3 <- binomial.logistic.MCML(
+                y ~ 1,
+                units.m = ~ units.m, coords = ~ x1 + x2,
+                data = data_sim, par0 = par0.lr,
+                start.cov.pars = par0.lr[3], control.mcmc = lr.mcmc,
+                low.rank = TRUE, knots = knots3, kappa = 2,
+                method = "nlminb", plot.correlogram = FALSE))
```

```
   user  system elapsed
407.376  14.397 423.235
```

To fit a low-rank approximation, we only need to specify `low.rank = TRUE` and define the set of spatial knots through the argument `knots`. For parameter estimation, this approach was about 35, 13 and 6 times faster than the exact method when using 5, 100 and 225 knots, respectively.

```
R> par.hat <- coef(fit.MCML.exact)
R> Sigma.hat <- varcov.spatial(coords = data_sim[c("x1","x2")],
+                    cov.pars = par.hat[2:3], kappa = 2)$varcov
R>  mu.hat <- rep(par.hat[1], nrow(data_sim))
R>  system.time(S.cond.sim <- Laplace.sampling(mu = mu.hat,
+               sigma = Sigma.hat,
+               y = data_sim$y,
+               units.m = data_sim$units.m,
+               control.mcmc = exact.mcmc,
+               plot.correlogram = FALSE))
```

```
    user   system  elapsed
 1275.890  134.015 1393.457
```

```
R> prevalence.sim <- exp(S.cond.sim$samples)/
+                                  (1 + exp(S.cond.sim$samples))
R> prevalence.exact <- apply(prevalence.sim,2, mean)
R>
R> lr.mcmc$h <- 1.65/(nrow(knots1)^(1/6))
R> system.time(pred.MCML.lr1 <- spatial.pred.binomial.MCML(
+          fit.MCML.lr1,
+          grid.pred = data_sim[c("x1","x2")],
+          control.mcmc = lr.mcmc,
+          type = "joint", scale.predictions = "prevalence",
+          plot.correlogram = FALSE))
```

```
   user  system elapsed 
 34.571   2.954  37.664 
```

```
R> lr.mcmc$h <- 1.65/(nrow(knots2)^(1/6))
R> system.time(pred.MCML.lr2 <- spatial.pred.binomial.MCML(
+          fit.MCML.lr2,
+          grid.pred = data_sim[c("x1","x2")],
+          control.mcmc = lr.mcmc,
+          type = "joint", scale.predictions = "prevalence",
+          plot.correlogram = FALSE))
```

```
   user  system elapsed 
 75.035   6.008  81.399 
```

```
R> lr.mcmc$h <- 1.65/(nrow(knots3)^(1/6))
R> system.time(pred.MCML.lr3 <- spatial.pred.binomial.MCML(
+          fit.MCML.lr3,
+          grid.pred = data_sim[c("x1","x2")],
+          control.mcmc = lr.mcmc,
+          type = "joint", scale.predictions = "prevalence",
+          plot.correlogram = FALSE))
```

```
   user  system elapsed 
169.352  21.975 192.218 
```

```
R> par(mfrow = c(2,2), mar = c(3,4,3,4))
R> r.exact <- rasterFromXYZ(
+                  cbind(data_sim[, c("x1","x2")],
+                  prevalence.exact))
R> plot(r.exact, zlim = c(0,1), main = "Exact method")
```

```
R> contour(r.exact, levels = seq(0.1,0.9,0.1), add = TRUE)
R>
R> plot(pred.MCML.lr1,"prevalence",
+        "predictions", zlim = c(0,1),
+        main = "Low-rank: 25 knots")
R> contour(pred.MCML.lr1,"prevalence",
+              "predictions", zlim = c(0,1),
+              levels = seq(0.1,0.9,0.1), add = TRUE)
R>
R> plot(pred.MCML.lr2,"prevalence",
+        "predictions", zlim = c(0,1),
+        main = "Low-rank: 100 knots")
R> contour(pred.MCML.lr2,"prevalence","predictions", zlim = c(0,1),
+              levels = seq(0.1,0.9,0.1), add = TRUE)
R>
R> plot(pred.MCML.lr3,"prevalence",
+        "predictions", zlim = c(0,1),
+        main = "Low-rank: 225 knots")
R> contour(pred.MCML.lr3,"prevalence",
+        "predictions", zlim = c(0,1),
+        levels = seq(0.1,0.9,0.1), add = TRUE)
```

The above code generates and plots spatial predictions of prevalence at the 900 sample locations using exact and approximate methods. In the exact case, we first use the function `Laplace.sampling` to sample from the predictive distribution of $T^\top = (T_1, \dots, t_{900})$, where T_i is given by (4.1). The arguments `mu` and `Sigma` of this function represents the mean vector and covariance matrix of the unconditional distribution of T. We post-process the simulation output to obtain estimates of prevalence by using the anti-logit transformation of each simulated sample and taking the average of these values at each sampling location. Figure 4.7 shows the resulting estimates of prevalence. As expected, the accuracy of the low-rank approximation increases as more knots are included: while using 5 knots leads to a computationally fast but poor approximation, 100 and 225 knots give progressive improvements in accuracy which might be considered sufficient in practice.

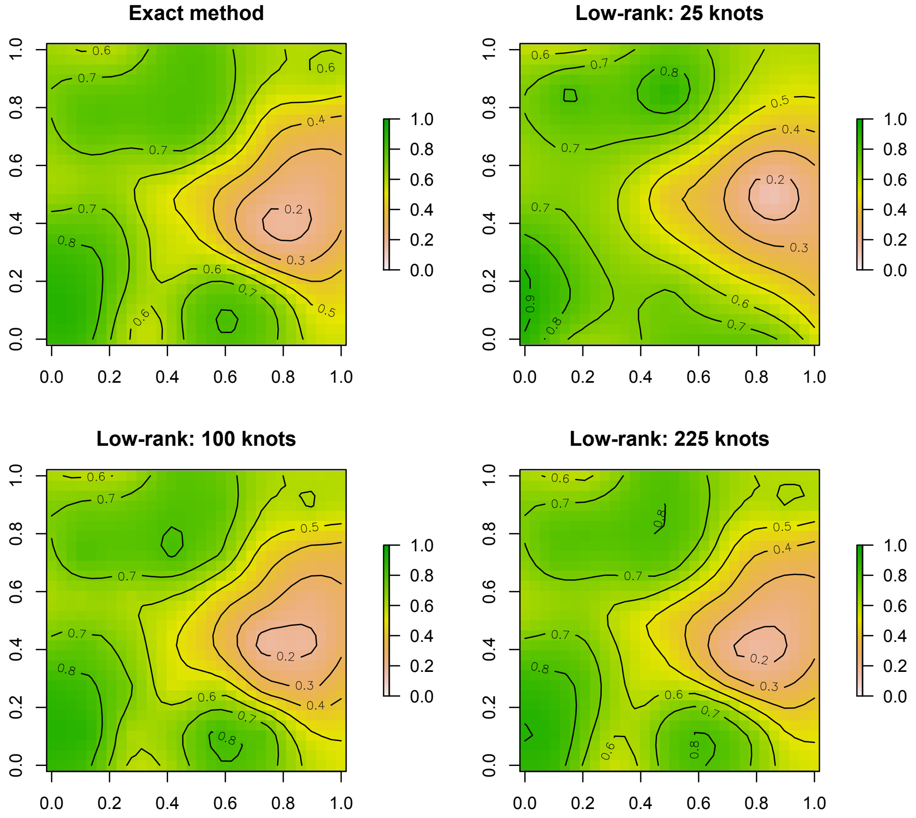

FIGURE 4.7: Images of the estimated surfaces of prevalence obtained for the simulated data using the exact method (upper left panel) and the low-rank approximation using 25 (upper right panel), 100 (lower left panel) and 225 (lower right panel) spatial knots.

4.5 Conclusions and future developments

We have illustrated the use the `PrevMap` package for geostatistical modelling of spatially referenced prevalence data. The package is intended to be compatible with the existing `geoR` and `geoRglm` packages, but with increased functionality.

The package provides more accurate numerical procedures for maximum likelihood estimation of the geostatistical linear and BL models, as well as routines for evaluation of the profile likelihood. Computationally faster approximations of the likelihood function for geostatistical BL models can be obtained using the Laplace approximation (LA). However, the resulting parameter estimates can be substantially biased in the case of binomial observations with small denominators (Joe, 2008), whereas the MCML method delivers asymptotically unbiased estimates.

For likeihood-based inference we have used a Langevin-Hastings MCMC algorithm because the availability of optimal scaling results makes it easier to tune than the Hamiltonian MCMC. However, for Bayesian inference where model parameters are also updated at each iteration, the required computation of the mode of the random effects conditional distribution (see Section 4.2.1) would have been computationally too demanding. For Bayesian analysis, we have therefore implemented an efficient Hamiltonian MCMC scheme that updates the random effects on their original scale and allows a more flexible prior specification for the model parameters.

The package also allows the user to specify whether marginal or joint predictions are required for different predictive targets: logit, prevalence, odds and exceedance probabilities. This overcomes the inherent limitation of methods based on analytical approximations to the marginal predictive distributions, such as INLA (Rue, Martino, and Chopin, 2009) on which the `geostatsp` package is based, which cannot calculate predictive distributions for functionals of the latent field.

The package includes several functions for automatic post-processing of the results, such as the diagnostic plots illustrated in Figure 4.5. As is the case for any MCMC application, these can only reveal non-convergence rather guarantee convergence, but are nevertheless useful as partial checks, and we therefore considered it important to make them easily accessible to users.

The accuracy of the low-rank approximations that are incorporated into the package is context-specific. However, used with care they offer computationally efficient procedures for analysing large data-sets. The `spBayes` package implements a low-rank procedure based on Gaussian predictive process models (Banerjee et al., 2008). In this approach, the latent field $S(x)$ in (4.1) is replaced by the conditional expectation of $S(x)$ given $S(\tilde{x}_i)$ for $i = 1, \ldots, r < n$, where $\tilde{x}_i$ is a set of pre-defined spatial knots. This is particularly useful and computationally advantageous when spatial interpolation is the sole objective of the analysis. In this context, other computationally efficient procedures could also be considered, such as low-rank spline smoothers (Wood, 2003). However, for applications that involve a range of inferential objectives, including both spatial prediction and estimation of covariate effects, it is desirable that the low-rank method approximates the same probabilistic model that would be used were computational burden not an issue, rather than changing the model specification. For this reason, we consider our version of low-rank approximation (Section 4.2.4) to be more suitable for disease mapping applications where, typically, the objectives include inference for regression parameters, both to assess the importance of hypothesised risk-factors and to enable spatial prediction under a range of scenarios. A specific example is the construction of predictive maps for malaria under different climate scenarios, or before and after widespread distribution of insecticide-treated bed-nets.

Another feature not illustrated in the present paper is the possibility of fitting a BL model to prevalence data from household surveys so as to include information at both household and individual level. More specifically, let i and j identify the i-th household and the j-th individual within that household; in this case the linear predictor is

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\} = d_{ij}^{\top}\beta + S(x_i) + Z_i,$$

where the random effects are now defined at household level.

Possible extensions of the package include the implementation of functions for spatio-temporal analyses, for geostatistical modelling of zero-inflated data and for combining data from multiple spatially referenced prevalence surveys (Giorgi et al., 2014). We will report these extensions separately in due course.

Acknowledgements

This work was supported by the Economic and Social Research Council [ES/J500094/1] through the award of a PhD studentship to Emanuele Giorgi.

References

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). “Gaussian Predictive Process Models for Large Spatial Data Sets”. In: *Journal of the Royal Statistical Society B* 70.4, pp. 825–848.

Brown, Patrick E. (2015). “Model-Based Geostatistics the Easy Way”. In: *Journal of Statistical Software* 63.12, pp. 1–24. URL: http://www.jstatsoft.org/v63/i12/.

Christensen, O. F., G. O. Roberts, and M. Sköld (2006). “Robust Markov Chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models”. In: *Journal of Computational and Graphical Statistics* 15.1, pp. 1–17.

Christensen, O.F. and Paulo J. Ribeiro (2002). “geoRglm - A Package for Generalised Linear Spatial Models”. In: *R-NEWS* 2.2. ISSN 1609-3631, pp. 26–28. URL: http://cran.R-project.org/doc/Rnews.

Christensen, Ole F (2004). “Monte Carlo Maximum Likelihood in Model-Based Geostatistics”. In: *Journal of Computational and Graphical Statistics* 13.3, pp. 702–718.

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

Diggle, Peter J. and Paulo J. Ribeiro (2007). *Model Based Geostatistics.* New York: Springer-Verlag.

Diggle, P.J., M.C. Thomson, O.F. Christensen, B. Rowlingson, V. Obsomer, J. Gardon, S. Wanji, I. Takougang, P. Enyong, J. Kamgno, H. Remme, M. Boussinesq, and D.H. Molyneux (2007). "Spatial Modelling and Prediction of Loa Loa Risk: Decision Making Under Uncertainty". In: *Annals of Tropical Medicine and Parasitology* 101.6, pp. 499–509.

Finley, A. O., S. Banerjee, and B. P. Carlin (2007). "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-Referenced Spatial Models". In: *Journal of Statistical Software* 19.4, pp. 1–24. URL: http://www.jstatsoft.org/v19/i04/.

Finley, A. O., S. Banerjee, and A. E. Gelfand (2015). "spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models". In: *Journal of Statistical Software* 63.13, pp. 1–28. URL: http://www.jstatsoft.org/v63/i13/.

Geyer, C. J. (1994). "On the Convergence of Monte Carlo Maximum Likelihood Calculations". In: *Journal of the Royal Statistical Society, Series B* 56, pp. 261–274.

Geyer, C. J. (1996). "Estimation and Optimization of Functions". In: *Markov Chain Monte Carlo in Practice.* Ed. by W. Gilks, S. Richardson, and D. Spiegelhalter. London: Chapman and Hall, 241–âĂŞ258.

Geyer, C. J. (1999). "Likelihood Inference for Spatial Point Processes". In: *Stochastic Geometry, Likelihood and Computation.* Ed. by O. E. Barndorff-Nielsen, W. S.Kendall, and M. N. M. van Lieshout. Boca Raton, FL: Chapman and Hall/CRC, 79–âĂŞ140.

Geyer, C. J. and E. A. Thompson (1992). "Constrained Monte Carlo Maximum Likelihood for Dependent Data". In: *Journal of the Royal Statistical Society, Series B* 54, pp. 657–699.

Giorgi, E., S. S. S. Sesay, D. J. Terlouw, and P. J. Diggle (2014). "Combining Data From Multiple Spatially Referenced Prevalence Surveys Using Generalized Linear Geostatistical Models". In: *Journal of the Royal Statistical Society A* 178.2, pp. 445–464.

Henningsen, A. and O. Toomet (2011). "maxLik: A Package for Maximum Likelihood Estimation in R". In: *Computational Statistics* 26.3, pp. 443–458.

Higdon, D. (1998). "A process-convolution approach to modeling temperatures in the North Atlantic Ocean". In: *Environmental and Ecological Statistics* 5, pp. 173–190.

Higdon, D. (2002). “Space and space-time modeling using process convolutions”. In: *Quantitative methods for current environmental issues*. Ed. by C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi. Springer-Verlag, New York, pp. 37–56.

Joe, H. (2008). “Accuracy of Laplace Approximation for Discrete Response Mixed Models”. In: *Computational Statistics & Data Analysis* 52.12, pp. 5066–5074.

Matérn, B. (1986). *Spatial Variation*. Second. Springer, Berlin.

Neal, R. M. (2011). “MCMC using Hamiltonian Dynamics”. In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Chapman & Hall, CRC Press. Chap. 5, pp. 113–162.

Ribeiro, Paulo J. and Peter J. Diggle (2001). “geoR: a Package for Geostatistical Analysis”. In: *R-NEWS* 1.2. ISSN 1609-3631, pp. 14–18. URL: http://CRAN.R-project.org/doc/Rnews/.

Roberts, G. O. and J. S. Rosenthal (2001). “Optimal Scaling for Various Metropolis-Hastings Algorithms.” In: *Statistical Science* 16.4, pp. 351–367.

Rue, H., S. Martino, and N. Chopin (2009). “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations”. In: *Journal of the Royal Statistical Society B* 71.3, pp. 319–392.

Stanton, M. C. and P. J. Diggle (2013). “Geostatistical Analysis of Binomial Data: Generalised Linear or Transformed Gaussian Modelling?” In: *Environmetrics* 24.3, pp. 158–171.

Wood, Simon N. (2003). “Thin Plate Regression Splines”. In: *Journal of the Royal Statistical Society B* 65.1, pp. 95–114.

Zhang, H. (2004). “Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics”. In: *Journal of the American Statistical Association* 99, pp. 250–261.

Chapter 5

Paper 4. Model-based geostatistics for prevalence mapping in low-resource settings

P. J. Diggle and E. Giorgi

Lancaster Medical School, Lancaster University, Lancaster, UK

Summary

In low-resource settings, prevalence mapping relies on empirical prevalence data from a finite, often spatially sparse, set of surveys of communities within the region of interest, possibly supplemented by remotely sensed images that can act as proxies for environmental risk factors. A standard geostatistical model for data of this kind is a generalized linear mixed model with binomial error distribution, logistic link and a combination of explanatory variables and a Gaussian spatial stochastic process in the linear predictor. In this paper, we first review statistical methods and software associated with this standard model, then consider several methodological extensions whose development has been motivated by the requirements of specific applications. These include: methods for combining randomised survey data with data from non-randomised, and therefore potentially biased, surveys; spatio-temporal extensions; spatially structured zero-inflation. Throughout, we illustrate the methods with disease mapping applications that have arisen through our involvement with a range of African public health programmes.

Keywords: geostatistics; multiple surveys; prevalence; spatio-temporal models; zero-inflation.

5.1 Introduction

The term “geostatistics” is typically used as a convenient shorthand for statistical models and methods associated with analysing spatially discrete data relating to an unobserved spatially continuous phenomenon. The name derives from its origins in the South African mining industry (Krige, 1951) and its subsequent development by the late Georges Matheron and colleagues in L’École des Mines, Fontainebleau, France (Chilés and Delfiner, 2012). Geostatistical methodology has since been applied in a wide range of scientific contexts, and is now widely accepted as one of three main branches of spatial statistics (Cressie, 1993). The descriptive phrase “model-based geostatistics” was coined by Diggle, Tawn, and Moyeed (1998) to mean the embedding of geostatistics within the general framework of statistical modelling and likelihood-based inference as applied to geostatistical problems. In contrast, “classical” Fontainebleau-style geostatistics has its own terminology and self-contained methodology, developed largely independently of the statistical mainstream.

Whether tackled through the model-based or classical approach, a typical feature of most geostatistical problems is a focus on prediction rather than on parameter estimation. The canonical geostatistical problem, expressed in the language of model-based geostatistics, is the following. Data $\{(y_i, x_i) : i = 1, ..., n\}$ are realised values of random variables Y_i associated with pre-specified locations $x_i \in A \subset \mathbb{R}^2$. The Y_i are assumed to be statistically dependent on an unobserved stochastic process, $\{S(x) : x \in \mathbb{R}^2\}$, as expressed through a statistical model $[S, Y] = [S][Y|S]$, where $[\cdot]$ means “the distribution of,” $Y = (Y_1, ..., Y_n)$ and $S = \{S(x_1), ..., S(x_n)\}$. What can be said about the realisation of S? The formal model-based solution is the conditional distribution, $[S|Y]$, which follows as a direct application of Bayes’ theorem,

$$[S|Y] = [S][Y|S]/\int[S][Y|S]dS.$$

By far the most tractable case is the linear Gaussian model, for which S is a Gaussian process and the Y_i given S are conditionally independent, $Y_i|S \sim \mathrm{N}(S(x_i), \tau^2)$. It follows that both the marginal distribution of Y and the conditional distribution of S given Y are multivariate Normal.

Note that in the above formulation, no model is specified for the x_i. The implicit assumption is that the x_i are pre-specified as part of the study-design or are located according to a process that is stochastically independent of S. If $X = (x_1, ..., x_n)$ is stochastic, a complete factorisation is $[S, X, Y] = [S][X|S][Y|X, S]$. Then, if $[X|S] = [X]$ and the properties of $[X]$ are not of interest, it is legitimate to condition on $[X]$ and so recover the previous formulation, $[S, Y] = [S][Y|S]$.

Diggle et al. (2013) argue that the geostatistical label should be applied more generally to scientific problems that involve predictive inference about an unobserved spatial phenomenon $S(x)$ using any form of incomplete information. This includes, for example, predictive inference for the intensity of a Cox process (Cox, 1955), and inference when X is both stochastic and dependent on S.

In this paper, we restrict our substantive scope to the problem of analysing data from spatially referenced prevalence surveys. We also focus on prevalence mapping in low-resource countries where registry data are lacking. We argue that in low-resource settings the sparsity of the available data justifies a more strongly model-based approach than would be appropriate if accurate registries were available.

5.2 The standard geostatistical model for prevalence data

In its most basic form, a prevalence survey consists of visiting communities at locations $x_i : i = 1, ..., n$ distributed over a region of interest A and, in each community, sampling m_i individuals and recording whether each tests positive or negative for the disease of interest. If $p(x)$ denotes prevalence at location x, the standard sampling model for the resulting data is binomial, $Y_i \sim \text{Bin}(m_i, p(x_i))$ for $i = 1, ..., n$. Linkage of the $p(x_i)$ at different locations is usually desirable, and is essential if we wish to make inferences about $p(x)$ at unsampled locations x.

The simplest extension to the basic model is a binary regression model, for example a logistic regression model of the form

$$\log[p(x_i)/\{1 - p(x_i)\}] = d(x_i)'\beta, \tag{5.1}$$

where $d(x_i)$ is a vector of explanatory variables associated with the location x_i. This assumes that the value of $d(x)$ is available not only at the data-locations x_i but also at any other location x that is of interest. When extra-binomial variation is present, two further extensions are possible. Firstly, a standard mixed effects model adds a random effect to the right-hand-side of (5.1), to give

$$\log[p(x_i)/\{1 - p(x_i)\}] = d(x_i)'\beta + Z_i,$$

where the Z_i are independent, $\text{N}(0, \tau^2)$ variates. Secondly, if the context suggests that covariate-adjusted prevalence should vary smoothly over the region of interest, we can add a spatially correlated random effect, to give

$$\log[p(x_i)/\{1 - p(x_i)\}] = d(x_i)'\beta + S(x_i) + Z_i, \tag{5.2}$$

where $\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean zero, variance σ^2 and correlation function $\rho(x, x') = \text{Corr}\{S(x), S(x')\}$. We shall assume that the process $\mathcal{S}$ is stationary and isotropic, hence $\text{Corr}\{S(x), S(x')\} = \rho(||x - x'||)$, where $|| \cdot ||$ denotes the Euclidean distance. The initial focus of inference within this model is the unobserved surface $p(x)$ or specific properties thereof. In general, we call $T = \mathcal{T}(\mathcal{S})$ a *target* for predictive inference. For example, we may wish to delineate sub-regions of A where $p(x)$ is likely to exceed a policy intervention threshold, in which case the target is $T = \{x : p(x) > c\}$ for pre-specified c, and the required output from the analysis is the predictive distribution of the random set T.

Equation (5.2) defines what we shall call the *standard geostatistical prevalence sampling model*. Various approaches to fitting this model to geostatistical data have been suggested in the literature. Diggle, Tawn, and Moyeed (1998) used Bayesian inference for parameter estimation and prediction, implemented by an MCMC algorithm. Rue, Martino, and Chopin (2009) used integrated nested Laplace approximation (INLA) methods. The INLA methodology and its associated software yield accurate and computationally fast approximations to the marginal posterior distributions of model parameters and to the marginal predictive distributions of $S(x)$ at any set of locations x, but not to their joint predictive distribution; this limits INLA's applicability to point-wise targets T. Giorgi and Diggle (2014) provide an `R` package for Monte Carlo maximum likelihood estimation and plug-in prediction with an option to use a low-rank approximation to $\mathcal{S}$ for faster computation with large data-sets. The low-rank method approximates $\mathcal{S}$ by $\mathcal{S}^*$, where

$$S^*(x) = \sum_{k=1}^{r} f(x - x_k) V_k. \tag{5.3}$$

In (5.3), the V_k are independent $\text{N}(0, \tau^2)$ variates associated with a pre-specified set of locations x_k and $f(x)$ is a prescribed function, typically monotone non-increasing in $||x||$. The covariance function of $\mathcal{S}^*$ is

$$\text{Cov}\{S^*(x), S^*(x')\} = \tau^2 \sum_{k=1}^{r} f(x - x_k) f(x' - x_k), \tag{5.4}$$

Low-rank specifications have been proposed as models in their right; see, for example, Higdon (1998) and Higdon (2002). We consider them as approximations to a limiting, full-rank process. Taking the x_k in (5.3) as the points of an increasingly fine regular lattice and scaling the function $f(\cdot)$ commensurate with the lattice spacing gives a limiting, full-rank process with covariance function

$$\text{Cov}\{S^*(x), S^*(x')\} = \tau^2 \int_{\mathbb{R}^2} f(x - u) f(x' - u) du. \tag{5.5}$$

From this perspective, the summation in (5.4) represents a quadrature approximation to the integral in (5.5). Note, however, that this construction admits only a sub-class of the allowable covariance functions for a spatially continuous Gaussian process.

Gotway and Stroup (1997) suggest using generalized estimating equations (Liang and Zeger, 1986) when scientific interest is focused on the regression parameters rather than on prediction of $\mathcal{S}$. However, in this approach the implicit target for inference is not the parameter vector β that appears in (5.2), but rather the marginal regression parameter vector, β^* say. The elements of β^* are smaller in absolute value than those of β by an amount that depends on τ^2, σ^2 and $\rho(u)$.

Diggle et al. (2007) use the standard model, but without the mutually independent random effects Z_i, to construct predictive maps of the prevalence of *Loa loa*, a parasitic infection of the eye, in an area of equatorial west Africa covering Cameroon and parts of its neighbouring countries. Following Thomson et al. (2004) they include two remotely sensed covariates, height above sea-level and the Normalised Digital Vegetation Index (NDVI), as proxies for the ability of the disease vector, a particular species of *Chrysops* fly, to breed at each location. As described in Thomson et al. (2004) and Diggle et al. (2007), *Loa loa* prevalence mapping plays an important role in the implementation of a multi-national prophylactic mass-treatment programme for the control of onchocerciasis (river blindness), the African Programme for Onchocerciasis Control, APOC (WHO, 2012), following the recognition that a generally safe filaricide medication, Ivermectin, could produce severe, occasionally fatal, adverse reactions in people heavily co-infected with onchocerciasis and *Loa loa* parasites. As a result, APOC adopted the policy that in areas where *Loa loa* prevalence was greater than 20%, precautionary measures should be taken before local administration of Ivermectin.

Diggle et al. (2007) mapped the minimum mean square error point predictor, $p(x) = \mathrm{E}[p(x)|Y]$ but also argued that a more useful quantity was the point-wise predictive probability, $q(x)$ say, that $p(x)$ exceeded 0.2, in line with APOC's precautionary policy. In addition to addressing directly the relevant practical problem, a map of $q(x)$ conveys the uncertainty associated with the resulting predictions. This map, here reproduced as Figure 5.1, identifies large areas that almost certainly do and do not meet the policy-intervention criterion, but also delineates large areas where the only honest answer is "don't know," indicating the need for further investigation or, if practicalities dictate, taking an informed risk.

Other prevalence mapping applications of model-based geostatistics include: Claridge et al. (2012) on liver fluke and bovine tuberculosis in the UK cattle herd; Clements et al. (2006) on schistosomiasis in Tanzania; Diggle and Ribeiro (2002) on childhood malaria in the Gambia; Gemperli et al. (2004) on infant mortality in Mali; Gething et al. (2012)

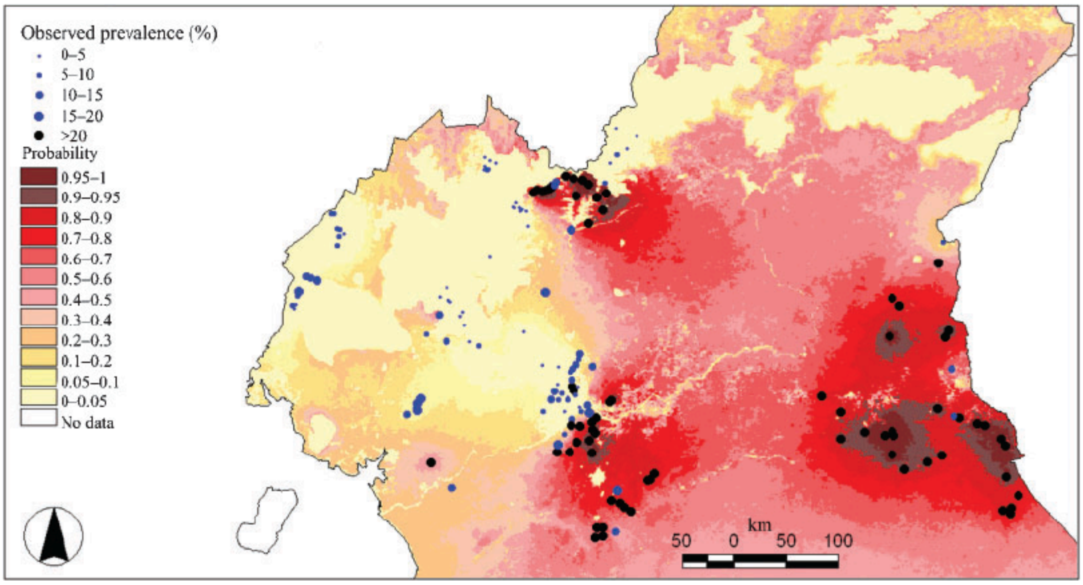

FIGURE 5.1: Predictive probability map of *Loa loa* prevalence in Cameroon and surrounding areas (adapted from Diggle et al. (2007)). Empirical prevalences at surveyed locations are indicated by size and colour coded dots.

on the world-wide distribution of *Plasmodium vivax*; Hay et al. (2009) on the world-wide distribution of *Plasmodium falciparium*; Kleinschmidt et al. (2001) on malaria incidence in Kwazuku Natal, South Africa; Kleinschmidt et al. (2007) on HIV in South Africa; Soares Magalhaes and Clements (2011) on anemia in preschool-aged children in West Africa; Raso et al. (2005) on schistosomiasis in Côte D'Ivoire; Pullan et al. (2011) on soil-transmitted infections in Kenya; Zoure et al. (2014) on river blindness in the 20 participating countries of the African Programme for Onchocerciasis control.

5.3 Combining information from multiple surveys

In order to obtain good geographical coverage of the population of interest, it is often necessary to combine information from multiple prevalence surveys. However, understanding the limitations of the sampling design adopted in each survey is crucial in order to draw valid inferences from a joint analysis of the data. In particular, non-randomized "convenience" surveys in which data are gathered opportunistically, for example at schools, markets or hospital clinics, may reach an unrepresentative sub-population or be biased in other ways. Nonetheless, convenience samples represent a tempting, low-cost alternative to random samples. A combined analysis of data from randomised and convenience samples that estimates and adjusts for bias can be more efficient than an analysis that considers only the data from randomised surveys. In a non-spatial context, Hedt and Pagano (2011) propose a hybrid estimator of prevalence that supplements information from random samples with convenience samples, and show that this leads to more accurate prevalence estimates than those available from using only the data from randomised surveys.

Giorgi et al. (2015) develop a multivariate generalized linear geostatistical model to account for data-quality variation amongst spatially referenced prevalence surveys. They assume that at least one of the available surveys is a "gold-standard" that delivers unbiased prevalence estimates and for which the standard model (5.2) is appropriate. Bias in a "non gold-standard" survey is then modelled using covariate information together with an additional, zero-mean stationary Gaussian process $\mathcal{B} = \{B(x) : x \in \mathbb{R}^2\}$. The resulting model for a non-randomised survey is

$$\log[p(x_i)/\{1 - p(x_i)\}] = d(x_i)'\beta + S(x_i) + Z_i + \{d(x_i)'\delta + B(x_i)\}. \tag{5.6}$$

Data from both the randomised and the non-randomised survey then contribute to inference on the predictive target, $d(x)'\beta + S(x)$.

5.3.1 Application: using school and community surveys to estimate malaria prevalence in Nyanza Province, Kenya

We now show an application to malaria prevalence data from a community survey and a school survey conducted in July 2010 in Rachuonyo South and Kisii Central Districts, Nyanza Province, Kenya. In the community survey, all residents above the age of 6 months were eligible for inclusion. A finger-prick blood sample was collected on each participant and examined for presence/absence of malaria parasites by a rapid diagnostic test (RDT).

In the school survey, 46 out of 122 schools with at least 100 pupils were randomly selected using an iterative process to limit the probability of selecting school with overlapping catchment areas. All eligible children in attendance were included. In the community survey, residential compounds lying within 600 meters of each school were randomly sampled and all eligible residents in each sampled compound examined by the RDT. The design of the community survey delivers an unbiased sample of residents from the catchment area of each school, whereas the school survey is potentially biased by a plausible association between a child's health status and their attendance at school. More details on the survey procedures can be found in Stevenson et al. (2013).

In our analysis, we extracted information on sampled individuals between the ages of 6 and 25 years in both surveys, as some adults have taken advantage of the introduction of free primary education in Kenya. The community survey included 1430 individuals distributed over 740 compounds whilst the school survey included 4852 pupils distributed over 3791 compounds, i.e. averages per compund of approximately 1.9 and 1.3 people, respectively. Figure 5.2 shows the locations of the sampled compounds from both surveys.

For our joint analysis of the data from both surveys, we used exponential correlation functions for both $S(x)$ and $B(x)$, with ϕ and ψ denoting the respective scale parameters. We parameterise the respective variances of $S(x)$, $B(x)$ and Z_i as σ^2, $\nu^2\sigma^2$ and $\omega^2\sigma^2$.

For selection of significant explanatory variables we used ordinary logistic regression, retaining variables with nominal p-values smaller than 5%. Table 5.1 gives the final set of explanatory variables included in the geostatistical model. The "District" indicator variable accounts for a known higher level of malaria risk in Rachuonoyo district. Socio-economic status (SES) is an indicator of household wealth taking discrete values from 1 (poor) to 5 (wealthy).

Table 5.2 reports Monte Carlo maximum likelihood estimates and 95% confidence intervals for the model parameters. The β-parameters reflect the district effect mentioned above as well as confirming a lower risk of malaria associated with higher scores of SES and greater age. The negative estimate of δ_0 and its associated confidence interval indicate a significantly lower malaria prevalence in individuals attending school than in the community at large. The positive estimate and associated confidence interval for δ_1 indicate that for individuals attending school, the negative effect of age is less strong than in the community. Figure 5.3(a) shows point-wise predictions of $B^*(x) = \exp\{B(x)\}$, which represents the unexplained multiplicative spatial bias in the school survey for the odds of malaria at location x. Figure 5.3(b) maps the predictive probability, $r(x)$ say, that $B^*(x)$ lies outside the interval $(0.9, 1.1)$,

$$r(x) = 1 - P\left(0.9 < B^*(x) < 1.1 | y\right). \tag{5.7}$$

The lowest value of $r(x)$ is about 87%, indicating the presence of non-negligible spatially structured bias throughout the study area. The joint analysis of the data from both surveys allows us to remove the bias and so obtain more accurate predictions for $S(x)$ than would be obtained using only the data from the community survey. Figure 5.4(a) shows a scatter plot of the standard errors for $S(x)$ obtained from the joint model for the school and community surveys and from the model fitted to the community data only. Figure 5.4(b) shows that locations for which the joint analysis produces larger standard errors for $S(x)$ correspond to areas where no observations were made.

Table 5.1: Explanatory variables used in the analysis of the Kenya malaria prevalence data.

	Term
β_0	Intercept
β_1	Age in years
β_2	District (=1 if “Rachuonyo”; =0 otherwise)
β_3	Socio-economic status (score from 1 to 5)
δ_0	Survey indicator, 1 if “school,” 0 if “community” (bias term)
δ_1	Age in years (bias term)

Table 5.2: Monte Carlo maximum likelihood estimates and corresponding 95% confidence intervals for the model fitted to the Kenya malaria prevalence data

	Estimate	95% Confidence interval
β_0	-1.412	(-2.303, -0.521)
β_1	-0.141	(-0.174, -0.109)
β_2	2.006	(1.228, 2.785)
β_3	-0.121	(-0.169, -0.072)
δ_0	-0.761	(-1.354, -0.167)
δ_1	0.094	(0.046, 0.142)
$\log(\sigma^2)$	0.519	(0.048, 0.990)
$\log(\nu^2)$	-1.264	(-1.738, -0.790)
$\log(\phi)$	-3.574	(-4.083, -3.064)
$\log(\omega^2)$	-1.408	(-2.267, -0.550)
$\log(\psi)$	-3.366	(-4.178, -2.553)

5.4 Analysing spatio-temporally referenced prevalence surveys

In endemic disease settings where prevalence varies smoothly over time, joint analysis of data from surveys collected at different times can also bring gains in efficiency. The modelling framework in Giorgi et al. (2015) accommodates multiple surveys conducted at different, discrete times. The extension of (5.6) to m surveys conducted at possibly different times is

$$\begin{aligned}\log[p_k(x_i)/\{1-p_k(x_i)\}] &= d(x_{ik})'\beta + S_k(x_i) + Z_{ik} + \\ & I(k \in \mathcal{B})\{d(x_{ik})'\delta + B_k(x_i)\}, k = 1, ..., m \qquad (5.8)\end{aligned}$$

where $\mathcal{B}$ denotes the indices of the non-randomised surveys, $\text{Cov}\{S_k(x), S_{k'}(x')\} = \sigma^2\alpha_{kk'}\rho(x,x')$ and $\alpha_{kk'} = 1$ if surveys k and k' are taken at the same time, $-1 < \alpha_{kk'} < 1$ otherwise.

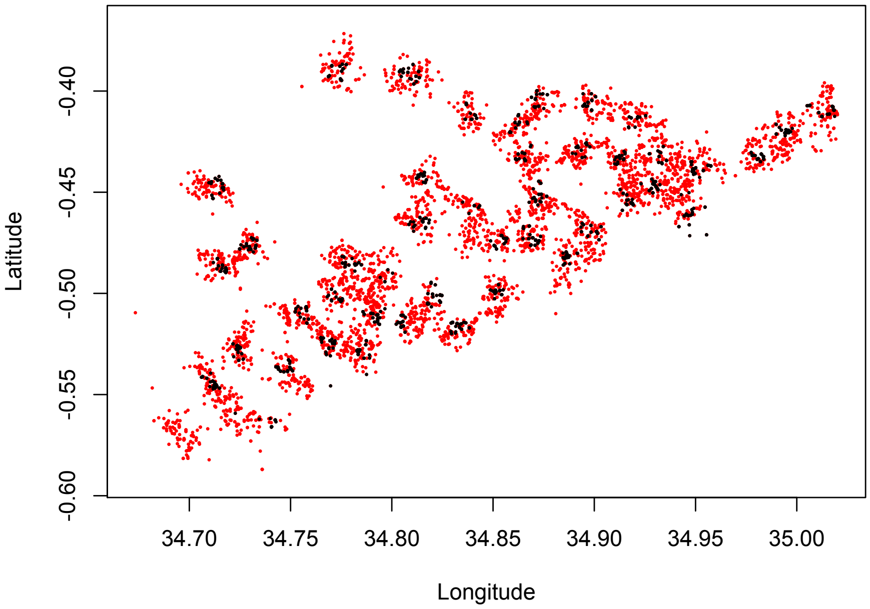

FIGURE 5.2: Geographical coordinates of the sampled compounds in the community (black points) and school (red points) surveys.

A different design for monitoring endemic disease prevalence is the *rolling indicator survey* (Roca-Feltrer et al., 2012). This consists of sampling members of a target population of individuals or households more or less continuously over time, the order of sampling being randomised. A natural model for the resulting data is a spatio-temporal version of (5.2),

$$\log[p(x_i, t_i)/\{1 - p(x_i, t_i)\}] = d(x_i, t_i)'\beta + S(x_i, t_i) + Z_i, \tag{5.9}$$

where now (x_i, t_i) denotes the location and time of the ith sample member. There is an extensive literature on ways of specifying the covariance structure of a spatio-temporal Gaussian process; see, for example, Gneiting and Guttorp (2010). For endemic diseases, a reasonable working assumption is that the relative risk of disease at different times is the same at all locations, and *vice versa*. This implies an additive formulation,

$$S(x, t) = S(x) + U(t), \tag{5.10}$$

where $S(x)$ and $U(t)$ are independent spatial and temporal Gaussian processes, respectively.

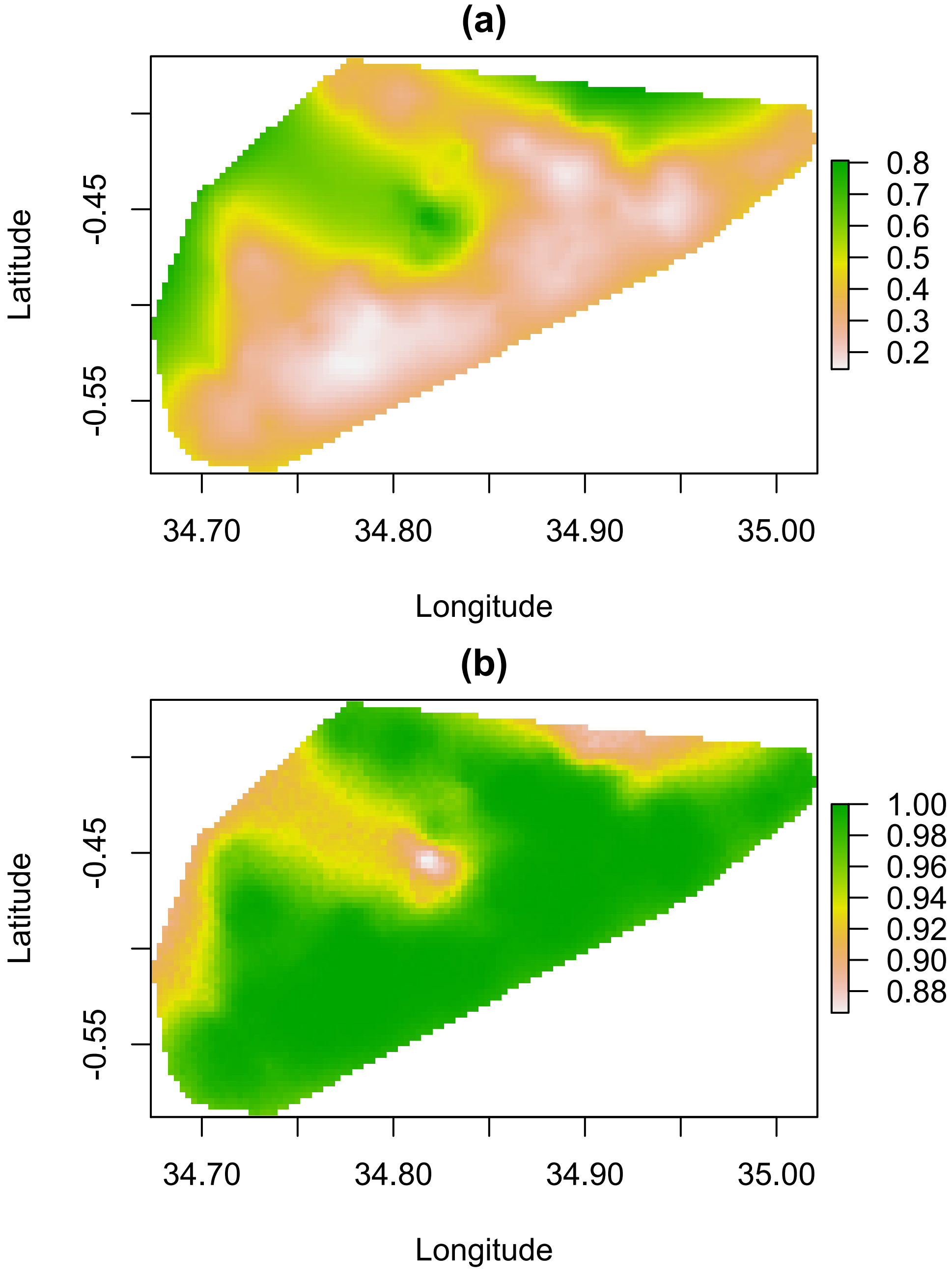

FIGURE 5.3: The predicted surfaces for $B^*(x)$ (a) and $r(x)$ (b).

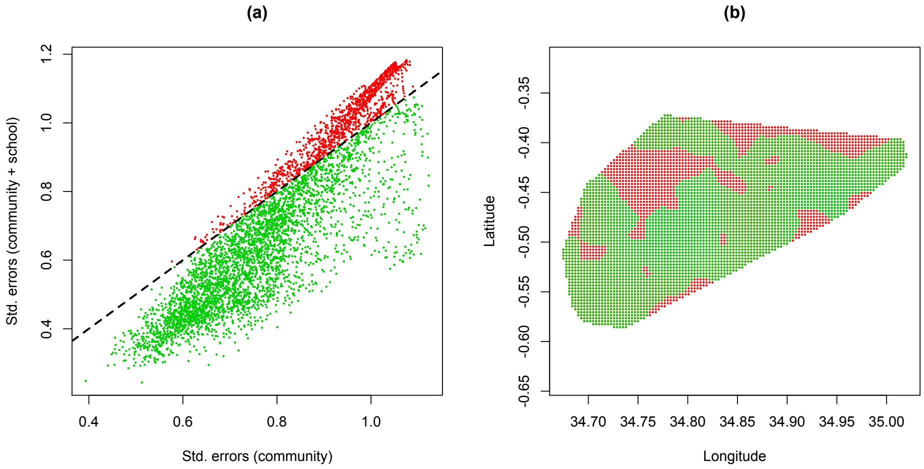

FIGURE 5.4: (a) Scatterplot of the standard errors for $S(x)$ using data from the community survey only (x-axis) and using both community and school survey data. Points coloured green or red lie below or above the identity line $y = x$, respectively. (b) Prediction locations, coloured green or red at locations where the prediction variance for $S(x)$ is smaller or larger, respectively, when using the data from both the community and school surveys.

5.4.1 Application: rolling malaria indicator survey in Chikwawa district, Malawi, May 2010 to June 2013

We now analyse data from a rolling malaria indicator survey (rMIS) conducted in Chikwawa District, Southern Malawi, from May 2010 to June 2013. In this rMIS, children under five years were randomly selected in 50 villages covering an area of approximately 400 km^2. Blood samples were then collected and tested by RDT for malaria. The objectives of the analysis are the following.

(i) Interpolation of the spatio-temporal pattern of malaria prevalence for children under twelve months;

(i) Estimation of the reduction in prevalence and number of infected children through a scale-up in the distribution of insecticide treated nets (ITN) and delivery of indoor residual spraying (IRS), from the actual coverage to 100% coverage in each village.

A practical distinction between these two objectives is that the first can only use explanatory variables that are available throughout the study-region, whereas the second can additionally use explanatory variables associated with the sampled households.

5.4.1.1 Spatio-temporal interpolation of malaria prevalence

Let $p_j(x_i, t_i)$ denote the probability of having a positive RDT outcome for the j-th children in the i-th household at time t_i. Using the model defined by (5.9), the linear predictor assumes the form

$$\begin{aligned}\log[p_j(x_i,t_i)/\{1-p_j(x_i,t_i)\}] \quad = \quad & \beta_0 + \beta_1 d_{ij} + \beta_2 t_i + \beta_3 \sin(2\pi t_i/12) + \beta_4 \cos(2\pi t_i/12) + \\ & \beta_5 \sin(2\pi t_i/6) + \beta_6 \cos(2\pi t_i/6) + S(x_i) + U(t_i), \quad (5.11)\end{aligned}$$

where d_{ij} is a binary indicator that takes value 1 if the child is under twelve months and 0 otherwise. The linear combination of sine and cosine functions with periodicities of one year and six months is used to model the seasonality of malaria. For both $S(x)$ and $U(t)$, we use isotropic exponential correlation functions with scale parameters ϕ and ψ, respectively. We use σ^2 and $\nu^2\sigma^2$ to denote the variance of $S(x)$ and $U(t)$, respectively.

Table 5.3 (Model 1) reports the MCML estimates of the model parameters; for the positive-valued parameters σ^2, ϕ, ν^2 and ψ we applied a log-transformation to improve the quadratic approximation to the log-likelihood. As expected, the estimate of β_1 indicates a significantly lower risk of having a positive RDT outcome for children in the first year of life, as newborns benefit from maternally acquired immunity that gradually fades.

Table 5.3: Monte Carlo Maximum Likelihood estimates for the spatio-temporal models fitted to the Malawi malaria prevalence data. Model 1 is defined at equation (5.11). Model 2 includes three additional explanatory variables: ITN, IRS and SES.

	Model 1		Model 2	
Term	Estimate	95% Confidence interval	Estimate	95% Confidence interval
β_0	4.210	(3.815, 4.605)	4.644	(4.099, 5.189)
β_1	-5.380	(-5.914, -4.847)	-5.428	(-6.066, -4.789)
β_2	-0.067	(-0.083, -0.051)	-0.072	(-0.090, -0.054)
β_3	-0.749	(-0.978, -0.521)	-0.693	(-0.922, -0.465)
β_4	0.361	(0.134, 0.588)	0.160	(-0.070, 0.389)
β_5	-0.099	(-0.307, 0.109)	-0.260	(-0.475, -0.045)
β_6	-0.168	(-0.391, 0.055)	-0.062	(-0.286, 0.162)
β_7 *	-	-	-0.188	(-0.492, 0.117)
β_8 **	-	-	-0.181	(-0.503, 0.141)
β_9 ***	-	-	-0.079	(-0.505, 0.347)
$\log(\sigma^2)$	0.899	(-0.011, 1.808)	0.971	(0.035, 1.906)
$\log(\phi)$	-3.624	(-4.852, -2.397)	-4.463	(-5.769, -3.157)
$\log(\nu^2)$	-3.282	(-4.199, -2.365)	-3.118	(-4.059, -2.177)
$\log(\psi)$	0.882	(-0.170, 1.934)	1.118	(0.017, 2.218)

* ownership of at least one ITN; ** presence of IRS; *** SES (score from 1 to 5).

We now generate prevalence predictions for five of the 50 villages in Chikwawa District. We chose these five villages selectively to include areas of low and high risk for malaria.

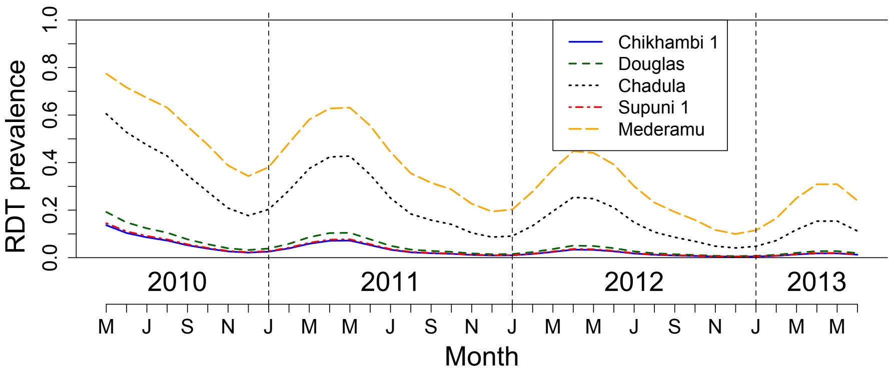

FIGURE 5.5: Estimated temporal trend of RDT prevalence for five villages in Chikwawa District. Figure 5.6 shows the location of each of these five villages.

Let A_i denote the convex hull obtained from the sampled locations of the i-th village. For a fixed time t, we computed

$$p_i(t) = |A_i|^{-1} \int_{A_i} \hat{p}(x,t)\, dx, \text{ for } i = 1, \dots, 5 \tag{5.12}$$

where d_{ij} is fixed at 1 for all $x \in A_i$ and $t = 1, 2, \dots, 38$, where each integer identifies a month, from May 2010 to June 2013. Also, $\hat{p}(x,t)$ is the mean of the predictive distribution of prevalence at location x and month t. For each village, i, we approximated the intractable integral in (5.12) using a quadrature method based on a regular grid covering the corresponding A_i. The results are shown in Figure 5.5, where each $p_i(t)$ is plotted against t; a declining trend of RDT prevalence can be seen, with seasonal troughs and peaks around December-January and April-May, respectively.

For any specified policy-relevant prevalence threshold $\tilde{p}$, a quantity of interest is the predictive probability that the estimated prevalence $\hat{p}(x,t)$ exceeds $\tilde{p}$. In Figure 5.6, we map the exceedance probabilities in June of each year for $\tilde{p} = 0.2$. Two areas of high and low prevalence are clearly identified. The former corresponds approximately to a flooding area where the the presence of local ponds also favours mosquito breeding.

5.4.1.2 Estimating the impact of scaling-up control interventions

The model (5.11) that we used to predict malaria prevalence throughout the study-region necessarily excluded any covariate that was only available at the sampled locations. We now propose a procedure to estimate community-wide prevalence and number of infected children under a pre-defined control scenario, focusing on the effects of ownership of ITN and presence of IRS, and adjusting for a measure of each household's socio-economic status (SES, scored from 1 to 5). We first fit a model with linear predictor of the same form in (5.11), but including these three additional explanatory variables. The resulting

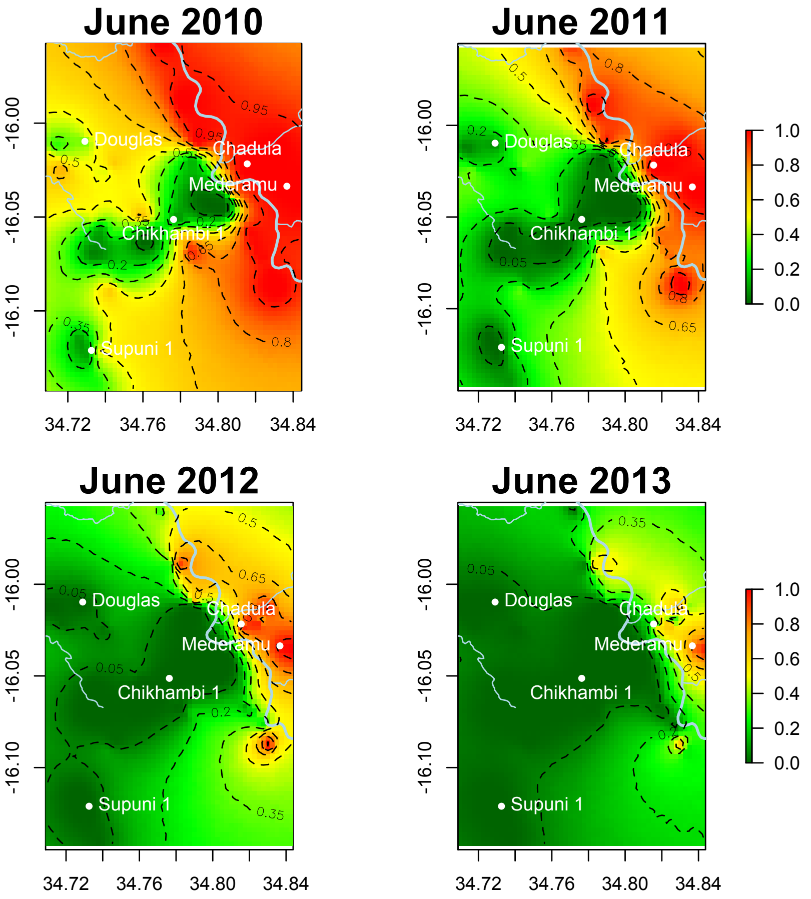

FIGURE 5.6: Maps of the predictive exceedance probabilities for a 20% malaria prevalence threshold in Chikwawa district; light blue lines correspond to waterways, with the Shire river represented by the thicker line.

parameter estimates are shown in Table 5.3 (Model 2). We then use enumeration data to obtain, for each village, the total number of children under five years and the number of households with at least one child under five years, and proceed as follows.

(i) Allocate the number of children in each household.

(ii) Impute geographical coordinates, ownership of ITN, presence of IRS and remaining explanatory variables for all unsampled children under the pre-defined control scenario.

(iii) Generate values for all the model parameters using the asymptotic distribution of the maximum likelihood estimator, i.e.

$$\hat{\theta} \sim N\left(\theta, I_{\text{obs}}^{-1}\right)$$

where θ is the vector of model parameters and I_{obs} is the observed Fisher information as estimated by the negative Hessian of the Monte Carlo likelihood.

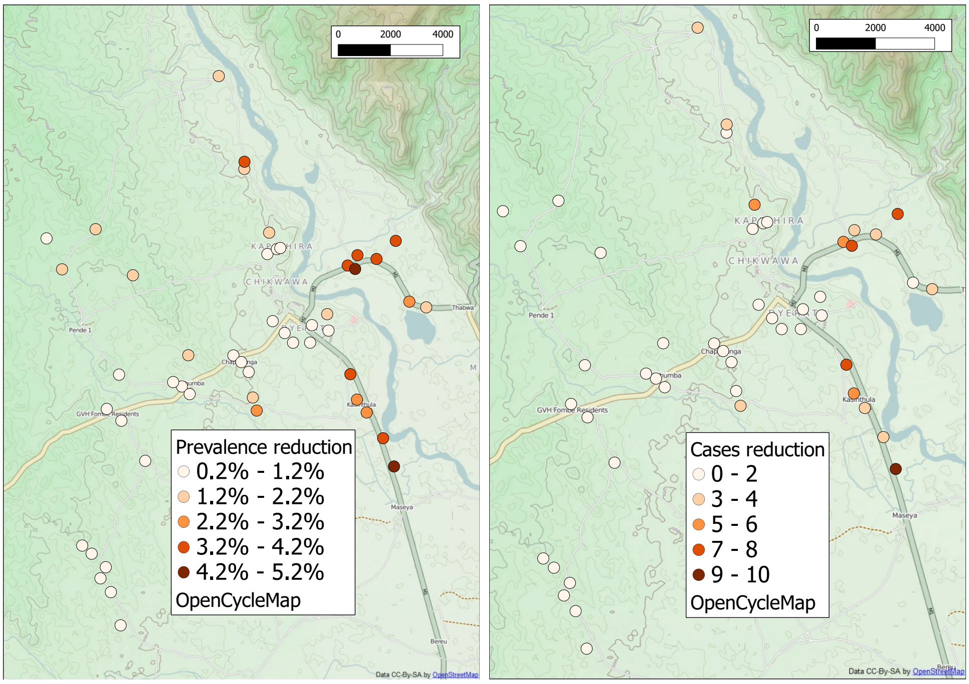

FIGURE 5.7: Estimated reduction in prevalence (left panel) and number of infected children (right panel) for each of the 50 villages in Chikwawa District, assuming a scale-up in the distribution of ITN and IRS to 100% coverage.

(iv) Generate predictive samples for each child's infection status and compute the mean of each sample as a point-estimate of the probability of infection for that child.

(v) For each village, estimate of the number of infected children as the sum of the estimated child-specific probabilities of infection, and average these to estimate the village-level prevalence.

We then repeat this process N times and, for each village, compute summary statistics of the N samples of estimated numbers of infected children and village-level prevalence. We applied this procedure under two different scenarios for April 2013, the most recent peak in RDT prevalence within the period covered by the data, as follows.

S1. Households having IRS and at least one ITN are equally distributed among sampled and unsampled households.

S2. Every household, whether sampled or unsampled, has IRS and at least one ITN.

In Step (ii), we imputed the locations of unsampled children by independent random sampling from the uniform distribution over each village area A_i, defined as the convex hull of the sampled households' locations.

In scenario S1, we imputed age, ITN, IRS and SES by random sampling from the empirical villlage-level distribution of the sampled households. In scenario S2, only SES and age need to be imputed as ITN and IRS are both present in every household. The differences between estimated prevalences and between numbers of infected children under S2 and S1 are reported in Figure 5.7. The main gains achieved by scenario S2 are in villages situated in the high prevalence area to the east of the Shire river.

5.5 Spatially structured zero-inflation

The standard geostatistical model for prevalence data in (5.2) assumes binomial sampling variation around the true prevalence, with a latent risk surface that approaches, but does not exactly reach, zero. However, empirical prevalence data often show an excess of zeros, i.e. zero-inflation. For diseases that are environmentally driven, one explanation for this is that some areas are fundamentally unsuitable for disease transmission. Hence, a zero prevalence estimate in a particular community can be either a chance finding, or a necessary consequence of the community being disease/infection-free. Ways of handling spatially structured zero-inflation have been proposed in ecology (Agarwal, Gelfand, and Citron-Pousty, 2002) and in specific epidemiological applications (Amek et al., 2011; Giardina et al., 2012). These approaches assume that the zero-inflation can be explained by regressing on a limited set of measured risk factors. In this extension to the standard geostatistical model (5.2) for spatially varying prevalence, $p(x)$, the distribution for the prevalence data Y conditional on S now takes the form of a mixture,

$$P(Y_i = y|S(x_i)) = \begin{cases} [1-\pi(x_i)] + \pi(x_i)\mathrm{Bin}(0; m_i, p(x_i)) & \text{if } y = 0 \\ \pi(x_i)\mathrm{Bin}(y; m, p(x_i)) & \text{if } y > 0 \end{cases} \tag{5.13}$$

where $\pi(x_i) \in (0,1)$ denotes the probability that x_i is suitable for transmission of the disease, $\log[\pi(x_i)/\{1-\pi(x_i)\}] = d(x_i)'\gamma$ and $\mathrm{Bin}(y; m, p)$ denotes the probability mass function of a binomial distribution with probability of success p and number of trials m. The modelled prevalence at location x is $p^*(x) = \pi(x)p(x)$.

An alternative way of specifying the conditional distribution of Y given S is given by the so called "hurdle" model (Mullahy, 1986). In this case the mixture distribution for Y assumes the form

$$P(Y_i = y|S(x_i)) = \begin{cases} 1-\pi(x_i) & \text{if } y = 0 \\ \dfrac{\pi(x_i)\mathrm{Bin}(y; m, p(x_i))}{1-\mathrm{Bin}(0; m, p(x_i))} & \text{if } y > 0 \end{cases}. \tag{5.14}$$

In our view, (5.14) is unsuitable for diseases mapping for the two following reasons. Firstly, the model does not distinguish between observing no cases amongst sampled individuals as a chance finding or as a necessary consequence of the entire community being disease-free. Secondly, the model can generate unnatural patches of low prevalence around each sampled location for which no cases are observed amongst sampled individuals.

A natural extension of the models in (5.13) and (5.14) that allows zero-inflation to depend on both measured and unmeasured covariates can obtained as follows. Define an additional stationary Gaussian process $T(x)$ such that

$$\log[\pi(x_i)/\{1-\pi(x_i)\}] = d(x_i)'\gamma + T(x_i). \tag{5.15}$$

The spatial processes $S(x)$ and $T(x)$ can also be further decomposed as

$$\begin{aligned} S(x) &= U_1(x) + V(x), \\ T(x) &= U_2(x) + V(x) \end{aligned}$$

where $U_1(x)$, $U_2(x)$ and $V(x)$ are independent Gaussian proccesses. In this formulation, $V(x)$ accounts for unmeasured factors that jointly affect the risk of the disease at a location x that is suitable for transmissionof the disease and the risk that x is itself suitable for transmission. However, identification of all of the resulting parameters requires a large amount of data. A pragmatic response is to assume that $V(x) = 0$ for all x, i.e. that $S(x)$ and $T(x)$ are independent processes.

5.5.1 Application: river-blindness prevalence mapping

We now show an application to river-blindness prevalence data, previously analysed in Zoure et al. (2014). Here, we restrict our analysis to three of the twenty APOC countries, namely Mozambique, Malawi and Tanzania. Figure 5.8 shows the locations of the sampled villages in the three countries. Red dots identify the 513 villages with no cases of river-blindness amongst sampled individuals, black dots the 397 villages with at least one case.

We fit the model with conditional distribution for Y given by (5.13), and logistic link functions (5.2) and (5.15) for $p(x)$ and $\pi(x)$, respectively. We also assume that $S(x)$ and $T(x)$ are independent processes with ciovariance functions $\sigma^2 \exp(-u/\phi)$ and $\sigma^2\omega^2 \exp(-u/\psi)$, respectively; we denote the variance of the nugget effect Z by $\sigma^2\nu^2$. We do include covariates, but simply fit constant means μ_1 and μ_2 on the logit-scale of $p(x)$ and $\pi(x)$, respectively.

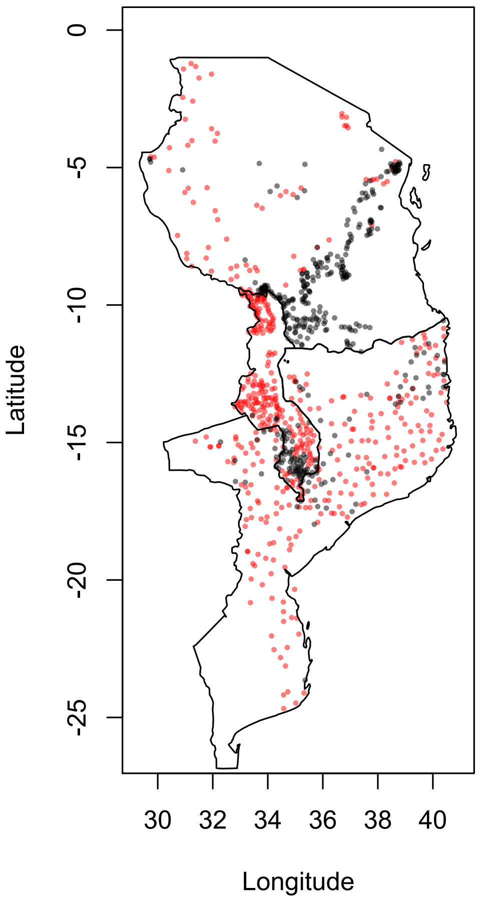

FIGURE 5.8: Sampled villages in Mozambique, Malawi and Tanzania, with balck and red dots corresponding to villages with no observed case and at least one observed case of river-blindness, respectively.

TABLE 5.4: MCML estimates of the parameters in the zero-inflated geostatistical model and associated 95% confidence intervals.

Term	Estimate	95% confidence interval
μ_1	-5.812	(-8.746, -2.877)
μ_2	2.287	(1.361, 3.213)
$\log(\sigma^2)$	-3.138	(-4.075, -2.200)
$\log(\nu^2)$	1.579	(0.615, 2.543)
$\log(\phi)$	-2.899	(-6.162, 0.363)
$\log(\omega^2)$	2.390	(1.425, 3.354)
$\log(\psi)$	1.679	(0.704, 2.654)

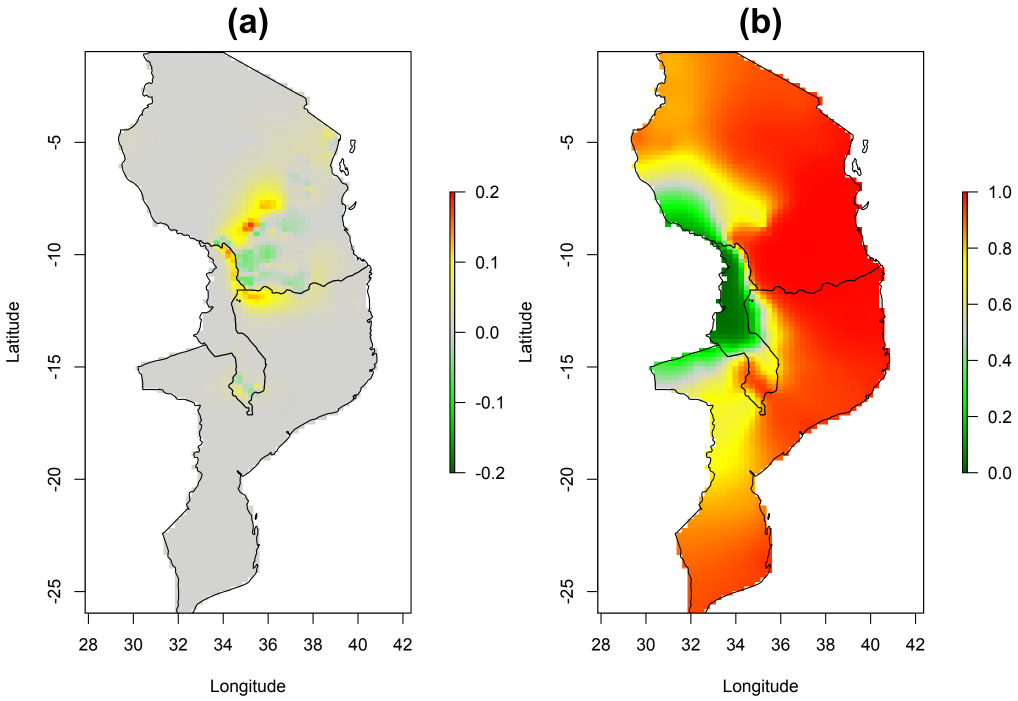

FIGURE 5.9: (a) Difference between predicted prevalences using the standard and zero-inflated geostatistical models. (b) Predicted surface of $\pi(x)$.

Table 5.4 shows the MCML estimates of the model parameters. The estimated scale of the spatial correlation of $T(x)$ is much larger than that of $S(x)$. Also, the estimate of the noise-to-signal ratio ν^2 is substantial.

Figure 5.9(a) shows the difference between estimates of prevalence $\hat{p}_s(x)$ and $\hat{p}_z(x)$ based on the standard and zero-inflated, geostatistical models, respectively; these range between plus and minus 0.2. Figure 5.9(b) shows the estimated surface of $\pi(x)$, and indicates that the central and northern parts of Malawi are disease-free, whereas most of the reported zero cases in Mozambique and Tanzania are more likely to be attributable to binomial sampling error.

5.6 Discussion

We have discussed four important issues that arise in prevalence mapping of tropical diseases, namely: combining data from multiple surveys of different quality; spatio-temporal interpolation of disease prevalence; assessment of the impact of control interventions; and accounting for zero-inflation in empirical prevalences. For each issue we have presented an extension of the standard geostatistical model and have described an application that we have encountered through our involvement with public health programmes in Africa.

In each application, we have used the MCML method for parameter estimation. This fitting procedure can be used under a very general modelling framework. Let W_i for $i = 1, \dots, n$ denote a set of random effects associated with Y_i, following a joint multivariate Normal distribution with mean μ and covariance matrix Σ. Assume that Y_i conditionally on W_i are mutually independent random variables with distributions $f(\cdot|W_i)$. The likelihood function for the vector of model parameters θ is given by

$$\begin{aligned} L(\theta) &= \int_{\mathbb{R}^{\mathrm{dim(W)}}} g(W, y; \theta)\, dW \\ &= \int_{\mathbb{R}^{\mathrm{dim(W)}}} N(W; \mu, \Sigma) \prod_{i}^{n} f(y_i|W_i)\, dW, \end{aligned}$$

where $\dim(W)$ denotes the dimension of W. Note, for example, that in the model used in Section 5.5.1, the random effect associated with village i is a bivariate random variable, $W_i = \{S(x_i) + Z_i, T(x_i)\}$, hence $\dim(W) = 2n$ with $f(\cdot|W_i)$ given by (5.13). Monte Carlo methods are then used in order to approximate the above intractable integral using importance sampling. As discussed in Giorgi and Diggle (2014), a convenient choice for the importance sampling distribution is $g(W, y; \theta_0)$ for some fixed θ_0, which can be iteratively updated. With this choice, a Markov chain Monte Carlo (MCMC) algorithm is then required for simulation of W_i conditionally on y_i under θ_0. We used a Langin-Hastings algorithm that updates the transformed vector of random effects $\hat{\Sigma}^{-1/2}(W - \hat{W})$, where $\hat{W}$ and $\hat{\Sigma}$ are the mode and the inverse of the negative Hessian at $\hat{W}$ of $g(W, y; \theta_0)$. In each of the applications, diagnostic plots based on the resulting samples of W_i showed fast convergence of the MCMC algorithm; details are available from the authors.

In the applications of Section 5.3.1 and Section 5.4.1, we considered extra-binomial variation at household-level but not at individual-level within households. An extension of the standard geostatistical model (5.2) that accounts for within-household random variation is

$$\log\{p_{ij}/(1 - p_{ij})\} = \alpha + [c_{ij}'\delta + U_{ij}] + [d(x_i)'\beta + S(x_i) + Z_i],$$

where i denotes household, j denotes individual within household, c_{ij} is a vector of individual-specific explanatory variables with associated regression parameters δ and the V_{ij} are mutually independent, zero-mean, Normally distributed random effects. However, when the data consists of empirical prevalences with small denominators, it is generally difficult to disentangle the effects of Z_i and U_{ij}. For this reason we used the more pragmatic approach of setting $U_{ij} = 0$ for all i and j.

The results of Section 5.4.1.2 on the impact of scaling-up the distribution of ITN and IRS to a 100% coverage should be interpreted cuatiously. The procedure that we used to obtain estimates of prevalence and number of infected children under different scenarios does not deal with the issue of causation. The control scenarios S1 and S2 represent virtual scenarios under which coverage of ITN and IRS is assumed to follow a pre-defined pattern without having any impact on other risk factors for malaria. In reality, a scale-up of ITN and IRS coverage may influence other features of the process, for example the extent to which ITNs are used correctly.

Under model (5.13) that accounts for zero-inflation, the risk surface can approach, but not reach, zero. We are are currently working on two further extensions of the standard geostatistical model. In the first of these, prevalence can reach zero but is constrained to do so smoothly. The second allows discontinuities in risk between suitable and unsuitable areas of transmission. Spatial discontinuities may seem artificial but can give a better fit to the data, especially when the pattern of risk is highly non-linear. Statistical tools for automatic choice between non-nested models are available from both frequentist and Bayesian perspectives, but our preference would be to reach agreement with a subject-matter expert on what qualitative features of the model best reflect the behaviour of the underlying process.

Acknowledgements

We thank the following people for providing the data analysed in the paper and helpful discussions: Dr. Gillian Stresman and Dr. Jennifer Stevenson (community and school-children data); Dr. Anja Terlouw (rMIS data); Dr. Hans Remme (river blindness data).

Support was provided by the award of a UK Economic and Social Research Council PhD studentship (ESRC Grant No.: ES/J500094/1) to Emanuele Giorgi and by The Farr Institute@HeRC. The Farr Institute@HeRC is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates) and the Wellcome Trust (MRC Grant No.: MR/K006665/1).

References

Agarwal, D. K., A. E. Gelfand, and S. Citron-Pousty (2002). “Zero-inflated models with application to spatial count data”. In: *Environmental and Ecological Statistics* 9, pp. 341–355.

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

Chilés, J.-P. and P. Delfiner (2012). *Geostatistics: Modelling Spatial Uncertainty.* Second. Wiley, New York.

Claridge, J., P. J. Diggle, C. M. McCann, G. Mulcahy, R. Flynn, J. McNair, S. Strain, M. Welsh, M. Baylis, and D. J. L. Williams (2012). “Fasciola hepatica is associated with the failure to detect bovine tuberculosis in dairy cattle”. In: *Nature Communications* 3, p. 853. DOI: 10.1038/ncomms1840.

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

Cox, D. R. (1955). “Some Statistical Methods Connected with Series of Events”. In: *Journal of the Royal Statistical Society, Series B* 17, pp. 129–164.

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Diggle, P. J. and P. J. Ribeiro (2002). “Bayesian inference in Gaussian model-based geostatistics”. In: *Geographical and Environmental Modelling* 6, pp. 129–146.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). “Model-based geostatistics (with discussion)”. In: *Applied Statistics* 47, pp. 299–350.

Diggle, P. J., P. Moraga, B. Rowlingson, and B. M. Taylor (2013). “Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm”. In: *Statistical Science* 28, pp. 542–563.

Diggle, P.J., M.C. Thomson, O.F. Christensen, B. Rowlingson, V. Obsomer, J. Gardon, S. Wanji, I. Takougang, P. Enyong, J. Kamgno, H. Remme, M. Boussinesq, and D.H. Molyneux (2007). “Spatial Modelling and Prediction of Loa Loa Risk: Decision Making Under Uncertainty”. In: *Annals of Tropical Medicine and Parasitology* 101.6, pp. 499–509.

Gemperli, A., P. Vounatsou, I. Kleinschmidt, M. Bagayoko, C. Lengeler, and T. Smith (2004). “Spatial Patterns of Infant Mortality in Mali: The Effect of Malaria Endemicity”. In: *American Journal of Epidemiology* 159, pp. 64–72.

Gething, P. W., I. R. F. Elyazar, C. L. Moyes, D. L. Smith, K. E. Battle, C. A. Guerra, A. P. Patil, A. J. Tatem, R. E. Howes, M. F. Myers, D. B. George, P. Horby, H. F. L. Wertheim, R. N. Price, I. MÃijeller, J. K. Baird, and S. I. Hay (2012). "A Long Neglected World Malaria Map: *Plasmodium vivax* Endemicity in 2010". In: *PLoS Neglected Tropical Diseases* 6, e1814. DOI: 10.1371/journal.pntd.0001814.

Giardina, F., L. Gosoniu, L. Konate, Mame Birame Diouf, R. Perry, O. Gaye, O. Faye, and P. Vounatsou (2012). "Estimating the Burden of Malaria in Senegal: Bayesian Zero-Inflated Binomial Geostatistical Modeling of the MIS 2008 Data". In: *PLoS ONE* 7.3, e32625. DOI: 10.1371/journal.pone.0032625.

Giorgi, E. and P. J. Diggle (2014). "PrevMap: an R package for prevalence mapping". Submitted.

Giorgi, E., S. S. S. Sesay, D. J. Terlouw, and P. J. Diggle (2015). "Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models". In: *Journal of the Royal Statistical Society, Series A* 178, pp. 445–464.

Gneiting, T. and P. Guttorp (2010). "Continuous parameter spatio-temporal processes". In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp. Boca Raton: Chapman and Hall/CRC Press, pp. 427–436.

Gotway, C. A. and W. W. Stroup (1997). "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction". In: *Journal of Agricultural, Biological, and Environmental Statistics* 2, pp. 157–178.

Hay, S. I., C. A Guerra, P. W. Gething, A. P Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H Manh, I. R. F Elyazar, S. Brooker, D. L. Smith, R. A Moyeed, and R. W. Snow (2009). "A World Malaria Map: *Plasmodium falciparum* Endemicity in 2007". In: *PLoS Medicine* 6, e1000048. DOI: 10.1371/journal.pmed.1000048.

Hedt, B. L. and M. Pagano (2011). "Health indicators: Eliminating bias from convenience sampling estimator". In: *Statistics in Medicine* 30, pp. 560–568.

Higdon, D. (1998). "A process-convolution approach to modeling temperatures in the North Atlantic Ocean". In: *Environmental and Ecological Statistics* 5, pp. 173–190.

Higdon, D. (2002). "Space and space-time modeling using process convolutions". In: *Quantitative methods for current environmental issues*. Ed. by C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi. Springer-Verlag, New York, pp. 37–56.

Kleinschmidt, I., B. L. Sharp, G. P. Y. Clarke, B. Curtis, and C. Fraser (2001). "Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-Area Malaria Incidence Rates in Kwazulu Natal, South Africa". In: *American Journal of Epidemiology* 153, pp. 1213–1221.

Kleinschmidt, I., A. Pettifor, N. Morris, C. MacPhail, and H. Rees (2007). “Geographic Distribution of Human Immunodeficiency Virus in South Africa”. In: *The American journal of tropical medicine and hygiene* 77, pp. 1163–1169.

Krige, D. G. (1951). “A statistical approach to some basiv mine valuation problems on the Witwatersrand”. In: *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52, pp. 119–139.

Liang, K. and S. L. Zeger (1986). “Longitudinal Data Analysis Using generalized linear models”. In: *Biometrika* 73, pp. 13–22.

Mullahy, J. (1986). “Specification and testing of some modified count data models”. In: *Journal of Econometrics* 33, pp. 341–365.

Pullan, R. L., P. W. Gething, J. L. Smith, C. S. Mwandawiro, H. J. W. Sturrock, C. W. Gitonga, S. I. Hay, and S. Brooker (2011). “Spatial Modelling of Soil-Transmitted Helminth Infections in Kenya: A Disease Control Planning Tool”. In: *PLoS Neglected Tropical Diseases* 5, e958. DOI: 10.1371/journal.pntd.0000958.

Raso, G., B. Matthys, E. K. N’goran, bM. Tanner, P. Vounatsou, and J. Utzinger (2005). “Spatial risk prediction and mapping of Schistosoma mansoni infections among schoolchildren living in western Côte d’Ivoire”. In: *Parasitology* 131, pp. 97–108. DOI: 10.1017/S0031182005007432.

Roca-Feltrer, A., D. J. Lalloo, K. Phiri, and D. J. Terlouw (2012). “Rolling Malaria Indicator Surveys (rMIS): A Potential District-Level Malaria Monitoring and Evaluation (M & E) Tool for Program Managers”. In: *American Journal of Tropical Medicine and Hygiene* 86, pp. 96–98.

Rue, H., S. Martino, and N. Chopin (2009). “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations”. In: *Journal of the Royal Statistical Society B* 71.3, pp. 319–392.

Soares Magalhaes, R. J. and A: C. A. Clements (2011). “Mapping the Risk of Anaemia in Preschool-Age Children: The Contribution of Malnutrition, Malaria, and Helminth Infections in West Africa”. In: *PLoS Medicine* 8, e1000438. DOI: 10.1371/journal.pmed.1000438.

Stevenson, J. C., G. H. Stresman, C. W. Gitonga, J. Gillig, C. Owaga, E. Marube, W. Odongo, A. Okoth, P. China, R. Oriango, S. J. Brooker, T. Bousema, C. Drakeley, and J. Cox (2013). “Reliability of School Surveys in Estimating Geographic Variation in Malaria Transmission in the Western Kenyan highlands”. In: *PLoS ONE* 8.10, e77641. DOI: 10.1371/journal.pone.0077641.

Thomson, M., V. Obsomer, J. Kamgno, J. Gardon, S. Wanji, I. Takougang, P. Enyong, J. Remme, D. Molyneux, and M. Boussinesq (2004). “Mapping the distribution of Loa loa in Cameroon in support of the African Programme for Onchocerciasis Control”. In: *Filaria Journal* 3.1. DOI:10.1186/1475-2883-3-7, p. 7.

WHO (2012). “African Programme for Onchocerciasis Control: meeting of national onchocerciasis task forces”. In: *Weekly epidemiological record* 87, pp. 493–508.

Zoure, Honorat, Mounkaila Noma, Afework Tekle, Uche Amazigo, Peter Diggle, Emanuele Giorgi, and Jan Remme (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”. In: *Parasites & Vectors* 7.1. doi: 10.1186/1756-3305-7-326, p. 326.

Chapter 6

Conclusions and future developments

In this chapter we further discuss each of the papers separately. As each paper already contains its own discussion, we focus on some specific issues and future research aimed at improving and broadening the range of applicability of the developed methodology.

6.1 Paper 1

In this paper we proposed a multivariate geostatistical model for the combined analysis of data from multiple spatially referenced prevalence surveys. We focused our attention on temporal variation, when surveys are repeated over time, and data-quality variation, as in the case of randomised and non-randomised surveys. Through simulation studies and an application to malaria prevalence data from Chikwawa District, Malawi, we showed that using a joint model for the data can lead to more accurate prevalence estimates. This approach finds wide applicability in low-resource settings, where sampling of the community of interest is essential in the absence of digitalized disease registries.

However, our approach can also be applied for disease mapping in developed countries, where disease registries are in fact available. For example, official surveys conducted by government agencies or other public bodies usually make use of sampling methodologies that generate “gold-standard” data. Commercial surveys instead often use biased sampling strategies but are more frequently updated than official surveys. In Manzi et al. (2011), local authority (LA) level smoking prevalences from multiple surveys conducted in the United Kingdom are analysed using hierarchical Bayesian models. In their analysis, residual spatial correlation between LA is not taken into account and the bias of some of the surveys is modelled using unstructured survey-specific random effects at LA level. However, if we consider smoking prevalences in each LA to be the realization of a spatially discrete process, the model in (2.1) can then be modified as follows. Let i and j denote the indices corresponding the i-th survey and the j-th LA; use $\mathcal{B}$ to denote the index set of potentially biased surveys, and let d_{ij} be a vector of explanatory variables for the j-th LA in the i-th survey. The linear predictor in (2.1) now assumes the form

$$\begin{aligned} \eta_{ij} &= d_{ij}^\top\beta_1 + S_{ij} + I(i \in \mathcal{B})[B_{ij} + d_{ij}^\top\beta_i] + Z_{ij}, \\ & \quad j = 1, \dots, n_i; i = 1, \dots, r. \end{aligned} \tag{6.1}$$

where $S_i(x_{ij})$ and $B_i(x_{ij})$ have been replaced by S_{ij} and B_{ij}, respectively, which are modelled as independent spatially discrete processes. In this context, a possible choice would be to use conditional autoregressive (CAR) models (Besag, 1974). The distribution of S_{ij} conditionally on S_{ik} for all $k \neq j$ is then given by

$$S_{ij}|S_{ik} : k \in \delta_j \sim N\left(\rho\frac{\sum_{k\in\delta_j} w_{ik}S_{ik}}{\sum_{k\in\delta_j} w_{ik}}, \frac{\sigma^2}{\sum_{k\in\delta_j} w_{jk}}\right)$$

where δ_j is the index set of neighbouring LAs to the j-th LA, w_{ik} are pre-defined weights and ρ regulates the strength of the spatial correlation between LAs. Likewise, the B_{ij} for $i = 1, \dots, r$ would then be modelled as CAR processes with spatial correlation parameter ρ_{B_i} and marginal variance $\sigma^2_{B_i}$. The unstructured random effects Z_{ij} would still be modelled as zero-mean Gaussian noise with variance τ_i^2 and, in this case, would represent extra-binomial variation within LA. To account for temporal variation across surveys, an approach analogous to (2.1) can be adopted by assuming separability of the temporal and spatial correlation functions. We conjecture that using this modelling framework would lead to very similar results to those shown in Chapter 2.

6.2 Paper 2

In the analysis of river-blindness prevalence across 20 APOC countries a standard geostatistical model was fitted to the data. Due to the high-dimensionality of the spatial random effects, a low-rank approximation was used so as to make parameter estimation and spatial prediction feasible. One of the initial issues in the analysis was that estimated prevalence was judged by a subject-matter expert to be too high in the proximity of onchocerciasis-free areas. We then accommodated this by placing a set of fictitious zero-cases in those areas. This turned out to be effective without affecting the estimated prevalence in onchocerciasis endemic areas. An alternative solution to this problem would be to carry out parameter estimation and spatial prediction separately in each APOC country in order to allow for country-specific means and covariance structures. However, we considered this approach to be unsatisfactory since it does not exploit the spatial correlation across country boundaries and does not make the most efficient use of all the available data.

A more general and sophisticated approach would be to model the spatial stochastic process $S(x)$ assuming a spatially varying mean, say $T(x)$, and variance, say $V(x)$. A spatially varying scale parameter of the spatial correlation could also be assumed but we expect this to be very difficult to recover from the data, as well as adding more complexity that would not give significant additional flexibility to a model already including $T(x)$ and $V(x)$. Forms of non-stationarity in the covariance structure can be taken into account by allowing, for example, for directional effects. An example of directional effect is given by *geometrical anisotropy* where stationarity of the covariance structure is transformed by a differential stretching and rotation of the coordinate axes; see Diggle and Ribeiro (2007, pp. 58-60) for more details.

In the context of our analysis of river-blindness prevalence, the spatial process $T(x)$ could be potentially useful to account for geopolitical differences between countries that may induce residual long-range spatial variation. The introduction of the spatial process $V(x)$ would account for spatial heteroscedasticity. Indeed, large values of $V(x)$ would identify regions of Africa where observations tend to be relatively far away from the estimated trend surface. By modelling $T(x)$ and $\log\{V(x)\}$ as independent stationary and isotropic Gaussian processes, the linear predictor in (1.1) is then modified as

$$\begin{aligned}\log\{p_i/(1-p_i)\} &= d(x_i)^\top\beta + S(x_i)\\ &= d(x_i)^\top\beta + T(x_i) + \sqrt{V(x_i)}S^*(x_i),\end{aligned} \tag{6.2}$$

where the spatial process $S^*(x_i)$ has the same properties of the stationary version of $S(x_i)$ in (1.1). The introduction of the spatial process $V(x)$ has also been proposed by Palacios and Steel (2006) in order to generate mean-square continuous fields with tails heavier than the Gaussian model. Additionally, by defining the variance and mean of $\log\{V(x)\}$ as ν^2 and $-\nu^2/2$, respectively, where ν^2 is a positive real parameter to estimate, we have that the expected value of $V(x)$ is 1 and as ν^2 tends to 0 we recover the standard geostatistical model with stationary variance.

It is clear that a large amount of data would be required to recover all the components of the model in (6.2). Then, a restrictive but practical assumptions would be, for example, that $\log\{V(x_i)\}$ and $S^*(x)$ share the same covariance function. Identifying $T(x)$ can also be very difficult. However, since $T(x)$ accounts for large scale spatial variation, this implies that the scale parameter of the correlation function for $T(x)$, say δ, is larger than that of $S^*(x)$, say ϕ. We can then approximate $T(x)$ with its country-level spatial average, i.e.

$$T(x) \approx T_j = |C_j|^{-1} \int_{C_j} T(x)\, dx, x \in C_j \tag{6.3}$$

where C_j is the region delimited by the borders of the j-th country and $|C_j|$ is the area of C_j. Indeed, it can be shown that for large values of δ, T_j is a good approximation to $T(x)$. Let $\tau^2\rho(\cdot;\delta)$ denote the covariance function of the process $T(x)$; the approximated covariance structure is given by the following expression

$$\begin{aligned} \text{cov}(T(x), T(x')) &\approx \text{cov}(T_h, T_k) \\ &= |C_h|^{-1}|C_k|^{-1}\tau^2 \int_{C_h}\int_{C_k} \rho(\|y-y'\|;\delta)\, dy\, dy', x \in C_h, x' \in C_k, \end{aligned}$$

where the above intractable integral can be approximated using a quadrature procedure. An alternative approach would be to approximate $T(x)$ with a spatially discrete CAR process introduced in the previous section. However, one of the issues of this approach is that, unlike (6.3), it does not provide a direct and unique way to account for the different sizes of the countries.

6.3 Paper 3

In this paper, we presented some of the features of the `PrevMap` package and illustrated how these can be used to conduct Bayesian and likelihood-based analysis of spatially referenced prevalence data. Future extensions of the package will be the following.

- Fitting of geostatistical Poisson-models. This is sometimes an appropriate model in its own right but also a useful approximation to the binomial model in the case of large binomial denominators and a very small probability of having a positive test.

- Fitting of multivariate geostatistical models to combine data from multiple surveys (Chapter 2), spatio-temporal models (Section 5.4) and geostatistical zero inflated binomial models (Section 5.5).

- Faster computational procedures for fitting binomial models based on covariance *tapering* (Kaufman, Schervish, and Nychka, 2008). Covariance *tapering* techniques approximate a spatial covariance matrix, say Σ, with $\tilde{\Sigma} = \Sigma \circ \Sigma_0$ where $\circ$ is the Hadamard product and Σ_0 is a covariance matrix based on a spatial correlation function with compact support (e.g. the spherical correlation function). Since $\tilde{\Sigma}$ is sparse, matrix calculations, such as inversion, Cholesky factorization and multiplication, are computationally more efficient than with the dense matrix Σ.

- Faster computational procedures for fitting geostatistical models to binary data using auxiliary variables techniques.

We now focus on the last extension, describing a computational procedure that we are currently implementing.

Let Y_{ij} denote a binary indicator that takes value 1 if the test for the disease for the j-th individual in the i-th household is positive and 0 otherwise. We assume that conditionally on the random effects $S(x_i)$ and Z_i, Y_{ij} are mutually independent Bernoulli variables with probit-link function given by

$$\Phi^{-1}(p_{ij}) = d_{ij}^\top \beta + S(x_i) + Z_i,$$

where $\Phi^{-1}(\cdot)$ is the quantile function of a standard Gaussian variable and d_{ij} is a vector that includes both individual specific and location-specific explanatory variables. An auxiliary variable V_{ij} can then be introduced such that

$$V_{ij}|\beta, S(x_i), Z_i \sim N\left(d_{ij}^\top \beta + S(x_i) + Z_i, 1\right).$$

It then follows that

$$Y_{ij}|V_{ij} = \begin{cases} 1 & \text{if } V_{ij} > 0, \\ 0 & \text{oterhwise} \end{cases}$$

since

$$\begin{aligned} p_{ij} &= P(Y_{ij} = 1) = P(V_{ij} > 0|S(x_i), Z_i) \\ &= 1 - \Phi(-d_{ij}^\top \beta + S(x_i) + Z_i) \\ &= \Phi(d_{ij}^\top \beta + S(x_i) + Z_i). \end{aligned}$$

By using a Gaussian prior for β, say $N(\xi, \Omega)$, one of the advantage of this representation is that the full conditional distribution of the vector $T^\top = (\beta^\top, S(x_1) + Z_1, \ldots, S(x_n) + Z_n)$ is multivariate Gaussian. Let D denote a matrix of covariates and A a binary matrix with entries $[A]_{hk} = 1$ if the h-th individual has been sampled at location x_k and $[A]_{hk} = 0$ otherwise. Let θ denote the vector of covariance parameters and $V^\top = (V_{11}, \ldots, V_{1,m_1}, \ldots, V_{n,1}, \ldots, V_{n,m_n})$; we can then write

$$\text{cov}(T|V, \theta) = \begin{pmatrix} \Omega^{-1} + D^\top D & D^\top A \\ A^\top D & \Sigma^{-1} + A^\top A \end{pmatrix}^{-1},$$

where Σ is the spatial covariance matrix including the nugget variance. The mean of the full conditional of T is then given by

$$\operatorname{cov}(T|V,\theta)\begin{pmatrix}\Omega^{-1}\xi + D^{\top}V \\ A^{\top}V\end{pmatrix}.$$

Finally, the full conditional distribution of V is a set of mutually independent right and left half-truncated Gaussian variables. More specifically, V_{ij} is a half-truncated Gaussian variable with support in $(0,+\infty)$ if $Y_{ij}=1$ or in $(-\infty,0)$ if $Y_{ij}=0$. A Gibbs sampler can then be used to update T and V in turn. To simulate from a half-truncated Gaussian variable, we can either use the method of inversion or some rejection-sampling procedure which prevents the issue of numerical errors in the case of a large negative (positive) mean for a right (left) half-truncated Gaussian variable.

This approach can be extended to the logistic link function by introducing an additional set of mutually independent auxiliary variables λ_{ij} that follow a Kolmogorov-Smirnov distribution (Stefanski, 1991) such that

$$V_{ij}|\beta, S(x_i), Z_i, \lambda_{ij} \sim N\left(d_{ij}^{\top}\beta + S(x_i) + Z_i, 1/\lambda_{ij}\right).$$

However, simulating from the full conditional distribution of λ_{ij} is not trivial, since the distribution of a Kolmogorov-Smirnov distribution is only known as an infinite series. For a more detailed discussion see Section 4.3 in Rue and Held (2005).

6.4 Paper 4

In this paper, we presented three extensions of the standard geostatistical model to address the statistical issues of combining data from multiple surveys, spatio-temporal interpolation of disease prevalence, estimation of the impact of control interventions and accounting for an excess of no reported cases. As discussed in Section 5.6, the model presented in Section 5.5 can only approach but does not exactly reach zero-prevalence in disease-free areas. We now describe two further extensions that can be used to model spatially structured zero-inflation and allow prevalence to reach zero exactly. Using the same notation as Section 5.5, we model $\pi(x)$ as a binary indicator taking value 1 if location x is suitable for transmission and 0 otherwise.

Assume that prevalence $p^{*}(x) = p(x)\pi(x)$ is constrained to approach zero smoothly as the location x approaches the boundary of the unsuitable area for transmission, i.e. allocations for which $\pi(x)=0$. This can be achieved by defining a threshold $\theta \in \mathbb{R}$ with $\pi(x)=0$ if $d(x)'\beta + S(x) < \theta$ and $\pi(x)=1$ otherwise. The conditional distribution of

Y_i given $S(x_i)$ then takes the form

$$P(Y_i = y|S(x_i)) = \begin{cases} 1 & \text{if } y = 0 \text{ and } d(x_i)'\beta + S(x_i) < \theta \\ \text{Bin}(y; m, p(x_i)) & \text{if } d(x_i)'\beta + S(x_i) > \theta \end{cases} \tag{6.4}$$

where now

$$p(x_i) = 2\frac{\exp\{d(x_i)'\beta + S(x_i) - \theta\}}{1 + \exp\{d(x_i)'\beta + S(x_i) - \theta\}} - 1.$$

If, instead, we want to allow for discontinuities between suitable and unsuitable regions for transmission, this can be achieved as follows. Partition the area of interest A using a Voronoi tessellation with cells given by the observed set of locations, i.e. $A = \bigcup_{i=1}^{n} R_i$ where

$$R_i = \{x \in A : ||x - x_i|| \leq ||x - x_j||, \text{for all } j \neq i\}.$$

Let $w = (w_1, \ldots, w_n)$ denote a binary random Markov field (Rue and Held, 2010). Hence, $w_i \in \{0, 1\}$ with w_i having conditional distribution

$$f(w_i|w_j, \text{for all } j \neq i) = f(w_i|w_j, \text{for all } j \in \partial_i), \tag{6.5}$$

where ∂_i is the set of all regions R_j that are neighbours to site R_i. One possibility would be to consider the following form for (6.5)

$$f(w_i|w_j, \text{for all } j \in \partial_i) = \frac{\exp\{\alpha[(1 - w_i)n_0 + w_i n_1]\}}{\exp\{\alpha n_0\} + \exp\{\alpha n_1\}}, \alpha > 0$$

where n_0 and n_1 are the numbers of neighbours to w_i that are zero and one, respectively. The parameter α regulates the interaction between neighbouring tiles, with large values of α associated with long-range interactions.

Assume that w and S are independent; the distribution of Y_i conditioned on $S(x_i)$ and w_i is given by

$$P(Y_i = y|S(x_i)) = \begin{cases} 1 & \text{if } y = 0 \text{ and } w_i = 0 \\ \text{Bin}(y; m, p(x_i)) & \text{if } w_i = 1 \end{cases}, \tag{6.6}$$

where $p(x_i)$ retains its form given by (5.2). Preliminary results show that model (6.6) gives very similar results in prevalence estimates with respect to the model of Section 5.5. The main disadvantage of this approach is that (6.6) is an artificial construction that does not allow for the use of explanatory variables to model the binary field w. One of the advantages, instead, is that allowing for discontinuities in prevalence between neighbouring tiles can potentially give a better fit to the data in the case of highly non-linear patterns of prevalence.

References

Besag, J. (1974). “Spatial interaction and the statistical analysis of lattice systems (with discussion)”. In: *Journal of the Royal Statistical Society, Series B* 36, pp. 192–236.

Diggle, P. J. and P. J. Ribeiro (2007). *Model-based geostatistics.* Springer Science+Business Media, New York.

Kaufman, C. G., M. J. Schervish, and D. W. Nychka (2008). “Covariance tapering for likelihood-based estimation in large spatial data sets”. In: *Journal of the American Statistical Association* 103, p. 1545.

Manzi, G., D. J. Spiegelhalter, R. M. Turner, J. Flowers, and S. G. Thompson (2011). “Modelling bias in combining small area prevalence estimates from multiple sruveys”. In: *Journal of the Royal Statistical Society, Series A* 174, pp. 31–50.

Palacios, M. B. and M. F. J. Steel (2006). “Non-Gaussian Bayesian geostatistical modeling”. In: *Journal of the American Statistical Association* 101.474, pp. 604–618.

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

Rue, H. and L. Held (2010). “Discrete spatial variation”. In: *Handbook of spatial statistics.* Ed. by P. Guttorp M. Fuentes A. E. Gelfand P. J. Diggle. CRC press, pp. 171–200.

Stefanski, L. A. (1991). “A normal scale mixture representation of the logistic distribution”. In: *Statistics & Probability Letters* 11.1, pp. 69 –70.