Johns Hopkins University, Dept. of Biostatistics Working Papers

3-21-2006

BIVARIATE BINOMIAL SPATIAL MODELLING LOA loa PREVALENCE IN TROPICAL AFRICA

Ciprian M. Crainiceanu
Johns Hokins Bloomberg School of Public Health, Department of Biostatistics, ccrainic@jhsph.edu

Peter J. Diggle
Department of Mathematics and Statistics Fylde College, Lancaster University, United Kingdom

Barry Rowlingson
Department of Mathematics and Statistics, Lancaster University, United Kingdom

Suggested Citation

Crainiceanu, Ciprian M.; Diggle, Peter J.; and Rowlingson, Barry, "BIVARIATE BINOMIAL SPATIAL MODELLING LOA loa PREVALENCE IN TROPICAL AFRICA" (March 2006). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 103.
http://biostats.bepress.com/jhubiostat/paper103

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Bivariate Binomial Spatial Modelling of *Loa loa* Prevalence in Tropical Africa

Ciprian M. Crainiceanu* Peter J. Diggle † Barry Rowlingson ‡

March 21, 2006

Abstract

We present a state-of-the-art application of smoothing for dependent bivariate binomial spatial data to *Loa loa* prevalence mapping in West Africa. This application is special because it starts with the non-spatial calibration of survey instruments, continues with the spatial model building and assessment and ends with robust, tested software that will be used by the field scientists of the World Health Organization for online prevalence map updating. From a statistical perspective several important methodological issues were addressed: (a) building spatial models that are complex enough to capture the structure of the data but remain computationally usable; (b) reducing the computational burden in the handling of very large covariate data sets; (c) devising methods for comparing spatial prediction methods for a given exceedance policy threshold.

Keywords: Geostatistics, low–rank, thin–plate splines

1 Introduction

The African Programme for Onchocerciasis Control (APOC) administers mass-treatment with the drug Ivermectin, which is highly effective in eliminating onchocerciasis parasites from the blood of infected individuals. However, in areas of high *Loa loa* prevalence, some individuals treated with Ivermectin develop severe, and occasionally fatal, adverse reactions to the drug. Hence, APOC policy is that before implementing mass-treatment in areas with high *Loa loa* prevalence, precautionary measures should be put in place to enable prompt treatment of any cases of serious adverse reaction to the drug. The official policy is that precautionary measures should be taken in areas where *Loa loa* prevalence exceeds 20%. Hence, there is a need to estimate the spatial distribution of *Loa loa* prevalence in potential treatment areas, which include a large part of central Africa.

Prevalence is traditionally estimated by parasitological sampling, i.e. by taking blood-samples from selected village communities and using the observed proportion of positive

*Assistant Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: ccrainic@jhsph.edu

†Professor, Department of Mathematics and Statistics Fylde College, Lancaster University, LA1 4YF, UK. E-mail: p.diggle@lancaster.ac.uk

‡Research Associate, Department of Mathematics and Statistics Fylde College, Lancaster University, LA1 4YF, UK. E-mail: B.Rowlingson@lancaster.ac.uk

Johns Hopkins University, Dept. of Biostatistics Working Papers
Research Archive

Hosted by The Berkeley Electronic Press

results as an estimate of the local prevalence. However, it is not feasible to undertake parasitological sampling in every community where Ivermectin treatment is envisaged.

Geostatistical modelling provides a way of using community-level results to estimate continuous spatial variation in prevalence and express the results as an "exceedance map," i.e. a map of the probability that local prevalence exceeds the 20% policy intervention threshold. Diggle, Moyeed, and Tawn (1998) and Diggle et al. (2006) proposed the following univariate binomial geostatistical model for describing village level parasitology data

$$\begin{cases} Y(x)|P(x) & \sim & \text{Binomial}\{N(x), P(x)\} \\ \text{logit}\{P(x)\} & = & \mu(x) + S(x) \end{cases} . \tag{1}$$

In (1) $Y(x)$ is the number of positive blood-test results out of $N(x)$ people sampled in the village identified by location x. $P(x)$ denotes prevalence, $\mu(x)$ is a function of elevation and greenness of vegetation as determined from satellite data and $S(x)$ is a stationary Gaussian process.

Collecting additional prevalence data is needed to improve the accuracy of prediction, but parasitological sampling is expensive and resources are scarce. Therefore, World Health Organization (WHO) researchers have developed a questionnaire instrument, RAPLOA, which for a given total cost enables many more communities to be surveyed than would be possible using parasitological sampling (Tako et al. (2002)). In order to validate the RAPLOA methodology, surveys were carried out in which both methods of determining prevalence were used. In this paper, we formulate a class of bivariate geostatistical models for data of this kind and describe a method for fitting a sub-class of these models using random coefficient thin-plate splines to represent a bivariate counterpart of the unobserved spatial process $S(\cdot)$ in (1). We consider two inferential approaches: Bayesian predictive inference implemented via Markov chain Monte Carlo (MCMC) and a computationally fast approximation to Bayesian inference. The rationale for this dual approach is that when a new survey is undertaken, field-workers may need to construct a local exceedance map quickly, whereas on completion of each survey the authoritative, region-wide exceedance map can be updated off-line by incorporating the new data in an optimal manner.

In Section 2 we present a non-spatial exploratory analysis of the validation data, which demonstrates the potential value of the RAPLOA instrument as a low-cost alternative to parasitological sampling. Section 3 describes the formulation of the bivariate geostatistical model which forms the basis for our proposed solution of the *Loa loa* mapping problem. Sections 4 and 5 give the the results obtained using Bayesian predictive inference, and our proposed computationally fast approximation, respectively. Section 6 contains a realistic simulation study comparing the Bayesian predictive inference with the simpler frequentist approximation. Section 7 discusses practical problems related to software implementation and testing and Section 8 provides the conclusion.

2 Exploratory analysis of the validation data

The validation data relate to a series of surveys conducted with the specific purpose of calibrating estimates of community-level *Loa loa* prevalence obtained by two different methods, RAPLOA and parasitological sampling. In the RAPLOA methodology, each person in the survey is classified as a positive case if they answer "yes" to all three of the following questions. Have you ever suffered from eye-worm? Did it look like this photograph? Did it

Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper103

Survey	Location	Number of villages	Subjects/village min	mean	max
0	Cameroon	74	24	117.3	268
1	DRC West	49	47	81.8	102
2	DRC East	50	46	81.8	96
3	Congo	50	27	66.5	100

Table 1: Location and size of each of the four calibration surveys.

last less than one week? In parasitological sampling, each person in the survey provides a finger-prick sample of blood, the blood-sample is smeared onto a glass slide, and positive cases are those whose blood-samples contain visible microfilariae at 10 times magnification. Data from four surveys are available, each including a sample of villages within a defined area. Table 1 summarizes the amount of data available.

For a preliminary assessment of the calibration relationship between prevalence as assessed by parasitology and by RAPLOA, we analyzed the data as follows. We assume that, after applying an empirical logit transformation, the data within each of the four surveys can be regarded as a random sample from a bivariate Gaussian distribution. We then compute the sample mean vector and covariance matrix of each sample, and derive the principal axis of each fitted bivariate Gaussian distribution as the eigenvector associated with the larger of the two eigenvalues of the sample covariance matrix. Finally, we back-transform the principal axis onto the prevalence scale.

For each datum, if n denotes the number of persons surveyed and y the number of positives, the raw estimated prevalence is y/n and the empirical logit is $\log\{(y+0.5)/(n-y+0.5)\}$. The two panels of Figure 1 show a strong, direct relationship between the results obtained by the two methods. This relationship is approximately linear on the logit scale, with correlation 0.83, a pronounced shift between the two means (-0.77 for RAPLOA, -2.41 for parasitology) but approximately equal variances (2.53 for RAPLOA, 2.76 for parasitology). Results from the four surveys show the same general pattern, with the Congo survey deviating somewhat from the other three in presenting a shallower slope for the fitted principal axis.

Figure 1 also shows the calibration relationships obtained as described above. Note in particular that on the prevalence scale, the calibration curves obtained from the four surveys agree closely over the range of parasitological prevalences between zero and 20%. This is the relevant range with respect to the declared policy regarding precautionary measures to be taken in advance of mass treatment with Ivermectin.

3 Bivariate geostatistical modelling

3.1 A bivariate Binomial geostatistical model

To enable predictive mapping of both parasitological and RAPLOA prevalence, we fit the following bivariate Binomial model for village level numbers of positive indications according

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

to RAPLOA and parasitology.

$$\begin{cases} Y_1(x)|P_1(x) & \sim & \text{Binomial}\{P_1(x), N_1(x)\} \\ Y_2(x)|P_2(x) & \sim & \text{Binomial}\{P_2(x), N_2(x)\} \\ \text{logit}\{P_1(x)\} & = & L_1(x) \\ \text{logit}\{P_2(x)\} & = & L_2(x) \\ L_2(x)|L_1(x) & \sim & \text{Normal}(\alpha_0 + \alpha_1 L_1(x), \sigma_\epsilon^2) \\ L_1(x) & = & \mu + C(x)^T\beta + x^T\gamma + S(x) \end{cases} \tag{2}$$

Here $Y_1(x)$, $Y_2(x)$ denote the numbers of positive indications according to parasitology and RAPLOA sampling, respectively, for the village at geographical location x, whilst $N_1(x)$ and $N_2(x)$ denote the corresponding numbers of people sampled. Conditional on the spatial prevalence processes $P_1(x)$ and $P_2(x)$, the count responses $Y_1(x)$ and $Y_2(x)$ are assumed to follow independent binomial distributions. For our application, $P_1(x)$, the *Loa loa* parasitological prevalence process, is the focus of interest and our specific objective is to identify geographic areas where $P_1(x) > 0.2$ with high probability.

The unobserved processes $L_1(x)$ and $L_2(x)$ represent the spatially varying log-odds of *Loa loa* prevalence according to the parasitological and RAPLOA surveys respectively, and are linked through a calibration relationship described by the fifth equation of (2). This equation plays an important role when only RAPLOA data become available at new locations and we wish to use these data to update our exceedance map for parasitological prevalence. A key assumption is that the parameters α_1, α_2 and σ_ϵ^2 do not depend on the location x. The results of the exploratory analysis reported in Section 2 indicate that this assumption is reasonable. The model is completed by specifying the spatial model for $L_1(x)$ in the final equation of (2). Here, μ is the overall mean, whilst $C(x)^T\beta$ describes the spatial variation in the mean attributable to the effects of covariates observed at location x. In the current application, we include as covariates altitude, mean greenness and standard deviation of greenness, where greenness is derived from repeated satellite scans over a period of one year. Finally, $S(x)$ is a zero-mean stationary process representing any residual spatial variation which is not explained by the available covariates.

Our bivariate binomial geostatistical model (2) is of necessity complex because the structure of the data is complex. However, the complexity is built using a series of individually simple conditional relationships, making the model easy to understand. With regard to the process $S(x)$, the standard approach would be to use a stationary Gaussian process, as in Diggle et al. (2006). However, this would be computationally burdensome for the current application because of very large number of prediction locations. A second practical consideration is the need to fit models to increasingly large data-sets as new data become available. In the following section we describe a model for $S(x)$ based on low-rank thin-plate splines, which provides a computationally efficient alternative to conventional Gaussian processes or full-rank thin-plate splines, without serious loss of flexibility.

3.2 Full-rank and low-rank thin-plate spline smoothing

The very widely used geostatistical method known as (ordinary) kriging is a linear smoothing method which is formally equivalent to minimum mean-square prediction for a Gaussian process, e.g. Chilès and Delfiner (1999), Cressie (1993). Both kriging and thin-plate spline smoothing, e.g. Green and Silverman (1994), are *full-rank* smoothers which fall within the family of *general radial smoothers*. Good discussions of the formal connection between these

Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper103

two important methods are provided by Cressie (1993) and Nychka (2000). In kriging, the covariance structure of the unobserved process is specified directly, usually from one of several standard parametric families. Thin-plate splines can also be identified as Gaussian spatial processes, although from this perspective their covariance structure may seem unnatural, e.g. Wahba (1990) 1990 and Nychka (2000).

Here, we discuss smoothing in two dimensions, although the extension to more than two dimensions is straightforward. In its simplest form, two–dimensional smoothing operates by fitting a model of the form

$$Y_i = f(x_i) + \epsilon_i$$

to data $(Y_i, x_i) : i = 1, ..., n$ under the assumption that the ϵ_i are mutually independent $N(0, \sigma_\epsilon^2)$ variables. Writing $x_i = (x_{1i}, x_{2i})$ so as to make explicit its two-dimensional character, the thin-plate spline smoother $\hat{f}(\cdot)$ is the solution to the following optimization problem,

$$\min_{f(\cdot,\cdot)} \left[\sum_{i=1}^{n} \{Y_i - f(x_{1i}, x_{2i})\}^2 + \lambda \int\int \left\{ \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right\} dx_1 dx_2 \right] . \quad (3)$$

To characterize the solution $\hat{f}(\cdot)$, consider the radial basis functions $\mathcal{C}(r) = r^{2(M-1)} \log(r)$ where the integer M controls the smoothness of the correlation function $\mathcal{C}(\cdot)$. Denote by X the matrix with ith row $X_i = (1, x_i)$ and by Z_R the matrix with (i, j)th entry equal to $\mathcal{C}(||x_i - x_j||)$. Write Z_i^R for the ith row of the matrix Z_R. Then, the solution of (3) is of the form $\hat{f}(x_i) = X_i\hat{\beta} + Z_i^R\hat{u}$ where $(\hat{\beta}, \hat{u})$ are the solutions of the quadratic minimization problem

$$\min_{\beta,u} \left(||Y - X\beta - Z_R u||^2 + \lambda u^T Z_R u \right) . \quad (4)$$

For any fixed value of the smoothing parameter λ, the thin-plate spline smoother has the explict form of a ridge regression estimator,

$$(\widehat{\beta}_\lambda, \widehat{u}_\lambda)^T = (C^T C + \lambda D)^{-1} C^T Y , \quad (5)$$

where C is the n by $n + 3$ matrix, $C = [X; Z]$ and D is the $n + 3$ by $n + 3$ diagonal matrix with diagonal elements $(0, 0, 0, Z_R)$.

Many criteria have been suggested for choosing the smoothing parameter λ from the data. These include CV or GCV (Craven and Wahba (1979)), C_p (Mallows (1973)), AIC (Akaike (1973)) and restricted or unrestricted maximum likelihood (Ruppert, Wand, and Carroll (2003)). We focus now on the restricted maximum likelihood (REML) criterion. Dividing (4) by the error variance σ_ϵ^2 gives

$$\min_{\beta,u} \left(\frac{1}{\sigma_\epsilon^2} ||Y - X\beta - Z_R u||^2 + \frac{\lambda}{\sigma_\epsilon^2} u^T Z_R u \right) . \quad (6)$$

If we now write $\sigma_u^2 = \sigma_\epsilon^2/\lambda$ and treat the u's as a set of Normally distributed random coefficients with mean zero, variance σ_u^2 and $\mathrm{Cov}(u) = \sigma_u^2 Z_R^{-1}$, the solution of (6) is equivalent to the best linear unbiased predictor (BLUP) in the linear Gaussian mixed model,

$$Y = X\beta + Z_R u + \epsilon , \ \mathrm{E} \begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0_{n\times 1} \\ 0_{n\times 1} \end{pmatrix} , \ \mathrm{Cov} \begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_u^2 Z_R^{-1} & 0_{n\times n} \\ 0_{n\times n} & \sigma_\epsilon^2 I_n . \end{pmatrix} \quad (7)$$

For a formal proof of this result see, for example, Ruppert, Wand and Carroll (2003). A further simplification can be obtained by reparameterization of the random coefficients to

Hosted by The Berkeley Electronic Press

$b = Z_R^{1/2} u$, where $Z_R^{1/2}$ is the principal square root of Z_R. Writing $Z = Z_R Z_R^{-1/2}$, model (7) becomes equivalent to

$$y = X\beta + Zb + \epsilon\ ,\ \ \mathrm{E}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0_{n\times 1} \\ 0_{n\times 1} \end{pmatrix}\ ,\ \ \mathrm{Cov}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_n & 0_{n\times n} \\ 0_{n\times n} & \sigma_\epsilon^2 I_n\ . \end{pmatrix} \qquad (8)$$

The above equivalence between thin-plate spline smoothers and Gaussian linear mixed models provides a statistically natural way to introduce the computationally convenient device of low-rank thin-plate spline smoothing. We first look more closely at the structure of the matrix Z_R. Since the (i,j)th entry of Z_R is $\mathcal{C}\left(||x_i - x_j||\right)$, Z_R can be interpreted as the correlation matrix of an isotropic spatial process with correlation function $\mathcal{C}(r) = r^{2(M-1)}\log(r)$. From a thin-plate spline perspective, every observation is treated as a knot and Z_R is the matrix of distances between each observation and each of the n knots in the metric $\mathcal{C}(\cdot)$. For example, the ith row of Z_R is $\{\mathcal{C}\left(||x_i - x_1||\right), \ldots, \mathcal{C}\left(||x_i - x_n||\right)\}$ and represents the distances from the ith observation to the knots, which are the sampling locations $x_1, \ldots, x_n$. More generally, we can consider any set of knots $\kappa_1, \ldots, \kappa_K$ in $\mathbb{R}^2$ and construct the $n \times K$ matrix Z_K with ith row $\{\mathcal{C}\left(||x_i - \kappa_1||\right), \ldots, \mathcal{C}\left(||x_i - \kappa_K||\right)\}$. The basic idea behind low-rank smoothing is that it is usually not necessary to consider as many knots as the sample size n, because when n is large and the underlying spatial structure is smooth, $K << n$ knots is usually sufficient to give the desired flexibility for the fitted surface in $\mathbb{R}^2$.

To preserve the nice statistical interpretation of the full-rank thin-plate spline, we would like to have a mixed model representation similar to (7), but where the Z_R matrix is replaced by Z_K. This can be done directly, but with one important difference. The matrix Z_K has dimension $n \times K$ and defines a non-invertible linear transformation from the k-dimensional random vector u to the n-dimensional space of the data. Therefore $\sigma_u^2 Z_K^{-1}$ does not exist, and cannot be the covariance matrix of u. However, in the full-rank model the matrix Z_R can also be viewed as the matrix of distances between the knots $x_1, \ldots, x_n$. Adapting this idea to any set of knots $\kappa_1, \ldots, \kappa_K$ we can assume that $\mathrm{Cov}(u) = \sigma_u^2 \Omega_K^{-1}$ where the (k,l)th entry of Ω_K is $w_{k,l}^K = \mathcal{C}(||\kappa_k - \kappa_l||)$. Using the same strategy as for full rank thin–plate splines we define $Z = Z_K \Omega_K^{-1/2}$ and obtain the low-rank thin-plate spline fit as the BLUP in the mixed model

$$y = X\beta + Zb + \epsilon\ ,\ \ \mathrm{E}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0_{K\times 1} \\ 0_{n\times 1} \end{pmatrix}\ ,\ \ \mathrm{Cov}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_K & 0_{K\times n} \\ 0_{n\times K} & \sigma_\epsilon^2 I_n \end{pmatrix}\ , \qquad (9)$$

which is now a direct analogue of (8).

Despite this simple mixed model formulation of $S(x)$, existing statistical software cannot handle models like (2). Therefore, a reasonable strategy is to use Bayesian inference based on MCMC simulations. However, in this framework, MCMC simulations based on full rank approaches are extremely expensive computationally and can be unstable as the complexity of the algorithm increases substantially with additional data. In contrast, the computational complexity of low-rank smoothers is determined by the number of knots, K.

More details on geostatistical modeling using low-rank thin-plate splines can be found in Kammann and Wand (2003). Good discussions of their computational advantages can be found in Nychka (2000) and in Ruppert, Wand and Carroll (2003).

3.3 Number and location of knots

The number of knots, K, in a low-rank smoother limits the maximum complexity of the model, whilst the smoothing parameter λ controls the fit to the data. Ruppert, Wand and

http://biostats.bepress.com/jhubiostat/paper103

Carroll (2003) suggest $K = \max\{20, \min(n/4, 150)\}$ as a default. In the *Loa loa* study with $n = 223$ village locations, we used $K = 50$ to model the spatial process component, $S(x)$, of the *Loa loa* prevalence. Together with covariate terms, this implies a maximum of 57 degrees of freedom for modelling the logit of the prevalence surface for parasitological sampling, $L_1(\cdot)$.

To determine the knot locations we used the space-filling design of Nychka and Saltzman (1998), as implemented in the R-package `FIELDS` Nychka (2004). The algorithm to obtain the design is fast for our sample sizes, but can be slow when n and K are large. A simple solution is to apply the algorithm to a random sub-set of the sample locations, x_i. The intuitive idea behind the space-filling algorithm is as follows.

As discussed in section 3.2, for a given set of locations $x_1, \ldots, x_n$ and function $\mathcal{C}(\cdot)$ a full rank smoother is the BLUP in the mixed model (8), where $Z = Z_R Z_R^{-1/2}$ with (i, j)th entry $\mathcal{C}(||x_i - x_j||)$. For most configurations of sample locations x, the eigenvalues of the $Z^T Z$ matrix decay very quickly to zero, indicating that the effective dimensionality of the space spanned by the columns of Z is much smaller than n. Suppose that we want to identify K n-element vectors of the form $[\mathcal{C}(||x_i - \kappa_k||)]_{1 \le i \le n}$ to define a subspace which best approximates the subspace spanned by the columns of Z_R. A reasonable strategy is to choose the knots κ_k, $k = 1, \ldots, K$, so that most are placed in regions of the space in which sample locations x_i are relatively dense. Now suppose that the observation locations form several clusters. Then, we need to place a number of knots within each cluster, but to best approximate the subspace spanned by the data–values y_i a natural strategy to avoid redundancy is to maximise the average spacing between knots. In one dimension, the resulting design reduces to choosing the knots at the sample quantiles of the x_i corresponding to probabilities $k/K+1$, as recommended by Ruppert (2002) and Ruppert et al. (2003). Other ways of choosing the knots have been suggested. For example, Ganguli and Wand (2005) use the `clara` algorithm of Kaufman and Rousseeuw (1990) implemented in the R–package `cluster` as the default for their low–rank thin–plate spline bivariate smoother implemented in the R–package `SemiPar`. However, a general property of low-rank thin-plate splines is that their fit to data is not strongly dependent on the exact locations of the knots.

4 Bayesian predictive inference

An important advantage of penalized low–rank thin plate splines is that they can be readily extended to more complex models such as (2). Indeed, the only component that still needs to be defined in model (2) is the parameterization of parasitological *Loa loa* logit prevalence, $L_1(x)$. In fact, $L_1(x)$ includes a spatially varying mean $\mu + C(x)^T\beta + x^T\gamma$ and residual spatial variation $S(x)$. The spatially varying mean is a standard linear function in the parameters μ, β, γ, where $C(x)$ are the, possibly transformed, observed covariates at location x. Using ideas from Section 3 we model

$$S(x) = Z(x)b,$$

where $Z(x)$ is the row of the matrix $Z = Z_K \Omega_K^{-1/2}$ corresponding to location x, and b is a $K \times 1$ vector of random coefficients. Model (2) is fully defined by specifying the prior distribution on the b coefficients that controls the amount of spatial smoothing. As in Section 3 this is

$$b \sim N(0, \sigma_b^2),$$

UCLA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where the shrinkage parameter σ_b^2 is estimated from the data. Note that, once inference is conducted at the sampling locations, model (2) provides a simple recipe for interpolation of the logit prevalence function at any new location, x_0, because

$$L_1(x_0) = \mu + C(x_0)^T\beta + x_0^T\gamma + Z(x_0)b.$$

If a Bayesian analysis is used then the joint posterior distribution of (μ, β, γ, b) is known and the posterior distribution of $L_1(x_0)$ at any location is also known. In practice, the posterior distribution is not available in closed form, but a correlated chain from the joint posterior distribution is usually available using Gibbs sampling. Such a chain can then be used to obtain a correlated chain from the posterior distribution of $L_1(x_0)$. This has the very nice practical property that to make inference about L_1 at any location or cluster of locations one need only use the output from the simulation algorithm using the original sampling locations.

The parameters of the model (2) are μ, β, γ, $\alpha = (\alpha_0, \alpha_1)$, σ_ϵ^2 and σ_b^2. For all parameters that were not variance components we used independent Gaussian priors with mean zero and standard deviation $1,000$. This choice was made by first doing a simplified frequentist analysis as described in detail in Section 5. The priors were then chosen with a standard error roughly 100 times larger than the largest standard deviation of individual parameters. We conducted a limited simulation study and, as expected, centering the priors at zero did not affect posterior inference.

As discussed by Crainiceanu, Ruppert, and Wand (2005b), the prior distributions on the variance components should be treated carefully, since a poor choice of prior can have serious effect on the smoothing function. To better understand this we show how critically the choice of Gamma prior $\tau_b = 1/\sigma_b^2$ may depend upon the scaling of the variables. Consider the simple case of Gaussian smoothing described in Section 3.2. If $[\tau_b] \sim \text{Gamma}(A_b, B_b)$ where $\text{Gamma}(A, B)$ has mean A/B and variance A/B^2, then

$$[\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon] \sim \text{Gamma}\left(A_b + \frac{K}{2}, B_b + \frac{||\boldsymbol{b}||^2}{2}\right) \tag{10}$$

The prior does not influence the posterior distribution of τ_b when both A_b and B_b are small compared to $K/2$ and $||\boldsymbol{b}||^2/2$ respectively. Since the number of knots is $K \geq 50$ it is safe to choose $A_b \leq 0.001$. When $B_b << ||\boldsymbol{b}||^2/2$ the posterior distribution is practically unaffected by the prior assumptions. When B_b increases compared to $||\boldsymbol{b}||^2/2$, the conditional distribution is increasingly affected by the prior assumptions. In our application the posterior distribution of σ_b^2 was essentially supported by $[0.2, \infty]$ and the value $B_b = 0.001$ did not influence the posterior inference. A similar discussion holds for σ_ϵ^2. Thus, we used $\text{Gamma}(0.001, 0.001)$ priors both for σ_b^2 and σ_ϵ^2.

4.1 Application to the *Loa loa* mapping problem

Our model (2) for the logit of *Loa loa* prevalence according to parasitology sampling, $L_1(x)$, includes a spatially varying mean $\mu + C(x)^T\beta + x^T\gamma$ and residual spatial variation represented by the Gaussian process, $S(x)$. Also, $C(x)$ contains 4 covariates. The first two covariates are the mean and standard deviation of the annual Normalized Difference Vegetation Index, which is a continuous measure of greenness derived from repeated satellite scans during the year 2000. Both measures are calculated at the pixel level, where one pixel is roughly equivalent to 1km^2.

Johns Hopkins Bloomberg School of Public Health
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

The last two covariates are elevation and elevation truncated at 800 meters, which together define a linear spline with one knot at 800 meters. We modelled the residual spatial variation $S(x)$ as a thin-plate spline with $k = 50$ knots placed according to the space-filling algorithm of Nychka, Haaland, O'Connell, and Ellner (1998).

As discussed in Section 3, one consequence of using a low-rank model is that the whole of the spatially continuous surface $S(x)$ is determined by a finite number of parameters and the progress of the MCMC can be monitored accordingly. This has important computational advantages because the number of parameters is very small compared to the numbers of locations where *Loa loa prevalence* is predicted. In particular, the exceedance probability at every location x within a geographical region of interest can be obtained from the monitored model parameters. The MCMC sample from the posterior distribution of the parameters induces a sample from the posterior distribution of the prevalence surface $P_1(x)$ in (2), and the required posterior exceedance probability can be obtained as the frequency with which $P_1(x) > 0.2$ in this sample.

For inference we used Gibbs sampling to simulate the joint posterior distribution of the parameters given the data. We used $20,000$ burn–in simulations and an additional $500,000$ simulations from the target distribution. Figure 2 displays every $1,000$th sample for 9 parameters of interest indicating reasonable sampling properties. A similar, but much less clear plot was obtained using every 100th sample. Three chains with initial parameter values dispersed with respect to the posterior densities were used to assess convergence. Visual inspection of these chains revealed that convergence to the target distribution occurs before $10,000$ simulations.

While the logit prevalence function, $L_1(x)$, has very good mixing properties, the mixing of several parameters was not as good, probably due to the lack of information about them in the data. The best mixing properties were exhibited by $\alpha_0, \alpha_1, \sigma_\epsilon$, which are the parameters of the spatial calibration, $L_2(x)|L_1(x)$, between the logit of the prevalence according to the RAPLOA and parasitological survey. The difference between these results and the nonspatial calibration models in Section 1 is that the posterior distributions of parameters depend on all of the data, including covariates, as well as on the structure of the spatial model $S(x)$. Another parameter with good MCMC mixing properties is σ_b (chain not shown) which controls the amount of shrinkage of the coefficients of the radial basis.

Table 2 gives the posterior mean and 95% credible intervals for several model parameters. The columns labeled "Whole data" correspond to analyses of the entire data set, while the columns labeled "2 observations removed" correspond to analyses of a subset of the original data set, where two obvious outliers were removed. More precisely, we removed observations for two villages with elevations of 1804 and 1806 meters and empirical parasitological prevalences of 0.51 and 0.33, respectively.

These results indicate that greeness was not statistically significant in either data set. Increased elevation up to 800 meters was postively associated with increased *Loa Loa* prevalence when the whole data set was used but became negatively associated when the two obvious outliers were removed. Moreover, in the reduced data set elevation above 800 meters is negatively associated with prevalence, with the probability of a negative change in slope above 800 meters being 0.97. The findings in the reduced data set tend to agree more closely with the findings of previous studies, e.g. Diggle et al. (1998). Interestingly, the slope of the calibration equation was estimated to be 0.95 and 0.91, both being statistically indistinguishable from 1. Thus, taking into account the spatial variation, the a–posteriori difference between logit parasitology and RAPLOA prevalence is entirely contained in the

COBRA
UC BERKELEY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

	Whole data		2 observations removed	
Parameter	Mean	SD	Mean	SD
Intercept cal.	1.43	.11	1.39	0.11
Slope cal.	0.95	.05	0.91	0.46
Greenness	0.36	.96	0.41	1.04
Std. Greenness	0.61	2.15	3.13	2.56
Elevation $\times 10^{-3}$	.80	.31	-.60	0.32
Elevation$_{>800} \times 10^{-3}$	$-.10$	0.15	-.30	0.16

Table 2: Posterior means and standard deviations for several parameters of interest. The columns "Whole data" correspond to all the 223 sampling locations. The columns "2 observations removed" correspond to 221 sampling locations, with 2 outliers removed.

calibration equation intercept.

Of course, the quality of prediction will depend effectively on which conclusion is supported by data at new locations. An advantage of the Bayesian method is that prior knowledge, such as "Greener areas correspond to larger prevalence" can be easily embedded into the model even if it is not supported by the current data by specifying a prior Uniform$(0, 2\times10^3)$ instead of Normal$(0, 10^6)$.

Figure 3 displays the estimated parasitology prevalence obtained from the Bayesian analysis of the bivariate Binomial model (2). The sampling locations are concentrated in three areas of the map roughly defined by the longitude/latitude rectangles $[8, 16]\times[3,7]$, $[12, 16] \times [-6, -3]$, and $[27, 31] \times [1.5, 4]$. To better view the details of the map the lower plot in Figure 3 shows a zoom in on the rectangle $[8.3, 16] \times [3, 7]$. In this graph we also plotted the actual sampling locations color–coded according to the empirical prevalence estimate: black $\widehat{P}(x) > 0.3$, red $0.25 \leq \widehat{P}(x) < 0.3$, magenta $0.20 \leq \widehat{P}(x) < 0.25$, cyan $0.18 \leq \widehat{P}(x) < 0.20$ and blue $\widehat{P}(x) < 0.18$. An important characteristic of the bottom panel is the smooth shape of the prevalence map, at least some of which is attributable to the sparsity of the data. The data were collected with the primary aim of validating the calibration between parasitological and RAPLOA estimates of prevalence over a wide geographical area; the resulting sampling design is not well suited for estimating spatial variation in prevalence.

5 A fast method for approximate predictive inference

Bayesian estimation has proved to be an effective inferential tool for the bivariate binomial spatial model (2) describing the complex joint distribution of the village level parasitology and RAPLOA sampling outcomes. However, in our context we have identified the following limitations of Bayesian inference based on MCMC sampling.

1. *Slow mixing.* The complex structure of model (2) combined with data sparsity induces large posterior correlations between weakly identified parameters which in turn leads to poor mixing of the Markov chains.

2. *Long updating time.* In our implementation one update of all parameters takes roughly 0.3 seconds on a PC (3.6GHz CPU, 3GB RAM). This, combined with the necessity of running long chains to overcome the slow mixing, leads to simulation times of several hours. Simulation times are likely to be even longer on computers used by field–workers.

COBRA Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper103

3. *Limited testing.* Long simulation times have restricted our ability to do extensive testing of our Bayesian methodology.

4. *Need for expert supervision.* All these limitations require the expert supervision of a statistician. Such expertise is typically not available when new data become available and predictive maps need to be updated.

In section 5.1 we present a computationally fast methodology that avoids these problems and provides a simple and robust basis for software development. In section 5.2 we use this simpler method for making inference about prediction maps of *Loa loa*. In section 6 we provide a realistic simulation study comparing the performance of this new calibration model with the bivariate binomial model.

5.1 A fast calibration methodology

The first step in constructing the calibration model is to ignore the Binomial variability and re–define the outcome. For those locations where parasitological sampling was conducted the logit of the parasitological prevalence can be approximated by

$$\begin{array}{rcl} \widehat{L}_1(x) & = & \text{logit}\{\widehat{P}_1(x)\} \\ \widehat{P}_1(x) & = & Y_1(x)/N_1(x), x \in O \end{array} \tag{11}$$

where $Y_1(x)$ is the number of parasitology positive samples among $N_1(x)$ subjects sampled at location $x \in O$. Here O denotes the set of all locations where parasitological sampling was conducted.

For locations where only RAPLOA sampling is available we simulate independently C data sets from the calibration model

$$\widehat{L}_1^c(x) \;\sim\; \text{Normal}(\widehat{\alpha}_{0,1|2} + \widehat{\alpha}_{1,1|2}\widehat{L}_2(x), \widehat{\sigma}^2_{\epsilon,1|2}),\ c = 1, \ldots, C, x \in M \tag{12}$$

where $\widehat{L}_2(x) = \text{logit}\{\widehat{P}_2(x)\}$, $\widehat{P}_2(x) = Y_2(x)/N_2(x)$ and M is the set of locations where RAPLOA but not parasitology sampling was conducted. The parameters $\alpha_{0,1|2}, \alpha_{1,1|2}, \sigma^2_{\epsilon,1|2}$ are estimated by a standard linear regression of $\widehat{L}_1(x)$ on $\widehat{L}_2(x)$ using those locations where both parasitology and RAPLOA sampling were conducted.

Thus, we obtain C data sets by keeping fixed the sampling locations and covariates and defining the outcome

$$L^c(x) = \widehat{L}_1(x)I(x \in O) + \widehat{L}_1^c(x)I(x \in M) \tag{13}$$

where $I(\cdot)$ is the indicator function. Denote now by L^c the outcome vector with entries $L^c(x)$ for all locations $x \in O \cup M$ and by X the design matrix of fixed effects with the row corresponding to location x equal to

$$X(x) = [1\ x^T\ \text{g}(x)\ \text{s}(x)\ \text{el}(x)\ \text{el}_{\{>800\}}(x)],$$

where x^T is the location expressed as (longitude, latitude), $\text{g}(x)$ is the greeness, $\text{s}(x)$ is the standard deviation of greeness, $\text{el}(x)$ is the elevation and $\text{el}_{\{>800\}}(x)$ is elevation truncated at 800 meters. Of course, other covariates could be included in X when they become available. Denote by Z the low–rank thin plate spline design matrix of random effects corresponding

Hosted by The Berkeley Electronic Press

to the set of n survey locations and a fixed set of K knots obtained as described in Section 3.2. For each data set $c = 1, \ldots, C$ we fit the following mixed model using REML estimation of variance components

$$L^c = X\beta + Zb + \epsilon, \;\; E\begin{bmatrix} b \\ \epsilon \end{bmatrix} = \begin{bmatrix} 0_K \\ 0_n \end{bmatrix}, \;\; \text{Cov}\begin{bmatrix} b \\ \epsilon \end{bmatrix} = \begin{bmatrix} \sigma_b^2 I_K & 0_{K\times n} \\ 0_{n\times K} & \sigma_\epsilon^2 I_n \end{bmatrix}. \tag{14}$$

Here 0_a, $0_{a\times b}$ are the $a \times 1$ vector and $a \times b$ matrix with zero entries respectively, and I_a is the $a \times a$ identity matrix.

Denote by $(\widehat{\beta}_c^T, \widehat{b}_c^T)^T$ the Best Linear Unbiased Predictor (BLUP) of $(\beta^T, b^T)^T$ from model (14) using the cth simulated outcome vector L^c. Suppose that one is interested in producing a predictive map at a particular location x_0. Denote by

$$X_0 = [1 \; x_0^T \; \text{g}(x_0) \; \text{s}(x_0) \; \text{el}(x_0) \; \text{el}_{\{>800\}}(x_0)],$$

and by

$$Z_0 = [||x_0 - \kappa_k||^2 \log ||x_0 - \kappa_k||]_{1\le k\le K}\Omega_K^{-1/2}$$

where $\kappa_1, \ldots, \kappa_K$ are K knot locations and Ω_K is the thin–plate distance matrix between knots defined in Section 3.2. With these notations the mean logit prevalence at x_0 can be estimated by

$$\widehat{L}^c(x_0) = X_0\widehat{\beta}_c^T + Z_0\widehat{b}_c^T \; .$$

The Monte Carlo variability of the $\widehat{L}^c(x_0)$ can be reduced by taking the average over all simulated data sets

$$\widehat{L}^A(x_0) = \frac{1}{C}\sum_{c=1}^{C}\widehat{L}^c(x_0) \; .$$

The variance of the $\widehat{L}^A(x_0)$ has two components. The first component is due to variability of the estimate around its mean while the second component is due to variability of the mean estimator around its mean. We estimate $\text{Var}\{\widehat{L}^A(x_0)\}$ by

$$\widehat{\text{Var}}\{\widehat{L}^A(x_0)\} = \frac{1}{C}\sum_{c=1}^{C}\widehat{\text{Var}}\{\widehat{L}^c(x_0)\} + \frac{1}{C-1}\sum_{c=1}^{C}\{\widehat{L}^c(x_0) - \widehat{L}^A(x_0)\}^2 \; .$$

Since $\{\widehat{L}^c(x_0)\}$ is a linear transformation of the BLUP estimator $(\widehat{\beta}_c^T, \widehat{b}_c^T)^T$ its variance is simple to estimate using mixed model results. In particular,

$$\widehat{\text{Var}}\{\widehat{L}^c(x_0)\} = \widehat{\sigma}_{c,\epsilon}^2 S_{x_0}\left(S^TS + \frac{\widehat{\sigma}_{c,\epsilon}^2}{\widehat{\sigma}_{c,b}^2}D\right)S_{x_0}^T, \tag{15}$$

where the subscript c indicates that parameter estimates are obtained from the c simulated sample, $S_{x_0} = [X_0|Z_0]$, $S = [X|Z]$ and

$$D = \begin{bmatrix} 0_{(p+1)\times(p+1)} & 0_{(p+1)\times K} \\ 0_{K\times(p+1)} & I_{K\times K} \end{bmatrix}.$$

Here p is the number of covariates including longitude and latitude and in our application $p = 6$. Estimator (15) is called the bias–adjusted variability estimate in Chapter 6 of Ruppert et al. (2003) and uses the marginal variance of the BLUP over the random effects.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

The prevalence exceedance probability of any probability threshold p_0 at a particular location x_0 can thus be estimated by

$$\widehat{E}(x_0, p_0) = 1 - \Phi\left(\frac{\text{logit}(p_0) - \widehat{L}^A(x_0)}{\sqrt{\widehat{\text{Var}}\{\widehat{L}^A(x_0)\}}}\right) .$$

For policy reasons, in our application $p_0 = 0.2$ corresponding to $\text{logit}(p_0) = -1.386$.

In summary, the model described in this section approximates model (2) by

1. Replacing the spatial binomial model for parasitology counts with a low–rank thin–plate spline approximation of a Gaussian random field for the logit of the empirical village level prevalence estimates.

2. Using the calibration model between the logit of RAPLOA and parasitology prevalence estimates to predict parasitology prevalence at those locations where only RAPLOA sampling was conducted. In contrast, the Bayesian methodology simulates imputations of missing parasitology observations conditional on all available data and model (2).

3. Combining inferences from C different inferences corresponding to the C simulated outcome vectors.

An important advantage of the methodology described in this section is that implementation is fast. Indeed, one data set is fit almost instantaneously because it only requires the fit of a Linear Mixed Model (LMM) with K random effects. The most delicate part of the estimation procedure is obtaining the REML estimates of the variance components. This was done by maximizing the profile likelihood corresponding to $\lambda = \sigma_b^2/\sigma_\epsilon^2$ over a grid. For the grid we used 1000 equally spaced values on the log scale between $[-10, 10]$. In multiple simulations we noticed that $C = 10$ is generally sufficient to produce reliable and reproducible results. The resulting fitting procedure is so fast that the computational bottlenecks shifted from model fitting to data loading and processing and prevalence map updating.

Because we used several approximations of model (2) it is reasonable to ask how much is actually lost during this approximation process. We address this question in Section 6 using a comparative simulation study in three realistic contexts.

5.2 Application to the *Loa loa* data

We applied the fast calibration methodology described in the previous section to the logit of the empirical parasitology prevalence. Basically, we fitted the mixed model (14) with the difference that all parasitological information is available at all locations, thus avoiding the calibration step. We used $K = 50$ knots for the P–spline and REML estimation of the smoothing parameter.

Table 3 displays the point estimate and 95% confidence intervals for the fixed effects for the whole data set and the reduced data set obtained by removing two outliers. None of the parameters was statistically significant at level $\alpha = 0.05$ when the whole data set was used. However, when the two outliers were removed the effect of greeness and increased elevation above 800 meters became statistically significant (p–value=0.018 and 0.004 respectively). These results agree with prior scientific knowledge that greenness and elevation are reasonable proxies for the density of day–biting Chrysops flies, which are the main agent of transmission of the filarial nematode *Loa loa* to humans.

Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

	Whole data		2 observations removed	
Parameter	Estimator	95% CI	Estimator	95%
greenness	3.71	1.99	4.21	1.78
Std. greenness	6.80	5.06	7.78	4.48
elevation$\times 10^3$	$-.1$	.66	$-.6$	.52
Elevation$_{>800} \times 10^3$	$-.7$	.41	$-.9$	.31

Table 3: Point estimator and standard error for several parameters of interest. The columns "Whole data" correspond to all the 223 sampling locations. The columns "2 observations removed" correspond to 221 sampling locations, with 2 outliers removed.

The estimated number of degrees of freedom of the regression was d.f. = 26.8, down from a maximum of 57 degrees of freedom allowed by the model. The degrees of freedom were partitioned into 4 for fixed effects, 3 for intercept, longitude and latitude, and 22.7 for the random coefficient component. This indicates serious departures from linear spatial effects. In fact, testing for linear effects versus a general nonparametric alternative is equivalent to testing

$$H_0 : \sigma_b^2 = 0 \text{ vs. } H_A : \sigma_b^2 > 0$$

where σ_b^2 is the variance of the random coefficients b in the model (14). Theory developed by Self and Liang (1987) for likelihood ratio tests of zero variance does not apply in this context because the response vector cannot be partitioned into independent subvectors. Instead we used the finite sample distribution of the RLRT derived by Crainiceanu and Ruppert (2004) and Crainiceanu, Ruppert, Claeskens, and Wand (2005a) for testing H_0 and obtained a p–value less than 0.001. Thus, the null hypothesis of linear spatial dependence is rejected against a nonparametric fit.

Figure 4 displays the estimated parasitology prevalence using the fast calibration analysis of the empirical village level parasitology prevalence estimates. The sampling locations are located in three areas of the map roughly defined by the longitude/latitude rectangles $[3, 7] \times [8, 16]$, $[12, 16] \times [-6, -3]$, and $[27, 31] \times [1.5, 4]$. To better view the details of the map the lower plot in Figure 4 shows a zoom in on the rectangle $[3.5, 6.5] \times [13, 15.5]$. In this graph we also plotted the actual sampling locations color–coded according to the empirical prevalence estimate: black $\widehat{P}(x) > 0.3$, red $0.25 \leq \widehat{P}(x) < 0.3$, magenta $0.20 \leq \widehat{P}(x) < 0.25$, cyan $0.18 \leq \widehat{P}(x) < 0.20$ and blue $\widehat{P}(x) < 0.18$.

6 Simulation study

Three simulation studies have been used to compare the performance of the Bayesian inference of the Bayesian binomial spatial model (2) with the frequentist analysis of the fast calibration model described in Section 5.1.

The first simulation study uses the same 223 sampling locations as the ones from the parasitology/RAPLOA sampling locations from West Africa. The underlying logit parasitological prevalence is fixed for all locations and is set equal to

$$L_1 = X\beta^* + Zb^*$$

where (β^*, b^*) are the posterior means of the (β, b) based on the Bayesian inference of model (2), and X and Z are the design matrices described in Section 5.1. Parasitology counts were

http://biostats.bepress.com/jhubiostat/paper103

then simulated independently from

$$\begin{cases} Y_1(x) & \sim & \text{Binomial}(N_1(x), P_1(x)) \\ P_1(x) & = & \frac{\exp\{L_1(x)\}}{1+\exp\{L_1(x)\}}, \end{cases}$$

where $N_1(x)$ is the sample size in the original study and $L_1(x)$ is the fixed logit prevalence at location x. The logit prevalence according to RAPLOA sampling is simulated independently from the model

$$L_2(x) \sim \text{Normal}(\alpha_0^* + \alpha_1^* L_1(x), \sigma^{*2}_\epsilon)$$

where $\alpha_0^*, \alpha_1^*, \sigma^{*2}_\epsilon$ are the posterior means of the calibration model parameters from model (2). RAPLOA counts are simulated from the model

$$\begin{cases} Y_2(x) & \sim & \text{Binomial}(N_2(x), P_2(x)) \\ P_2(x) & = & \frac{\exp\{L_2(x)\}}{1+\exp\{L_2(x)\}} \end{cases}$$

To compare results we focused on the region of interest (ROI) situated between 8.3 and 16 degrees longitude and 3 and 7 degrees latitude which contains 74 sampling locations. There were 100 simulated data sets. Bayesian analysis of model (2) for each data set was based on $20,000$ samples from the joint posterior distribution of the parameters given the data after $20,000$ discarded burn–in samples. Because in the first simulation study there were no missing parasitology observations the frequentist analysis fits only one data set ($C = 1$).

To better characterize the differences between the two methods in terms of prevalence estimation we calculated the MSE for each simulated data set. More precisely, for a particular region, R, simulated data set and fitting method, we calculated

$$MSE = \frac{1}{|R|} \sum_{x \in R} \{\widehat{L}_1(x) - L_1(x)\}^2, \tag{16}$$

where $|R|$ denotes the number of locations in region R and $\widehat{L}_1(x)$ denotes a generic estimator of the logit prevalence function. We focused on two regions R, the first being the entire ROI and the second containing just the sampling locations. Figure 5 displays boxplots for the frequentist calibration versus the Bayesian bivariate methodologies for the ROI (two leftmost boxplots) and for the sampling locations (two rightmost boxplots). Remarkably, the two methods perform almost identically in terms of MSE, with the bivariate method marginally outperforming the fast calibration method. This indicates that if the object of inference were the prevalence function itself it would not practically matter which method is used. In this case it would make sense to use the calibration method which is much faster and provides a more robust software platform.

However, in our application the focus is on predicting locations where the prevalence exceeds 20%. Because in our simulation study the true prevalence function, $P_1(x)$, is known the truly positive ($P_1(x) \geq 0.2$) and truly negative ($P_1(x) < 0.2$) locations are also known. Either inferential procedure produces an estimate $\widehat{L}_1(x)$ of the true logit prevalence function $L_1(x)$ and an estimate of its variability $\widehat{\text{Var}}\{\widehat{L}_1(x)\}$. The exceedance probability of the $p_0 = 0.2$ threshold at location x can be estimated by

$$\widehat{E}(x, p_0) = 1 - \Phi\left(\frac{\text{logit}(p_0) - \widehat{L}_1(x)}{\sqrt{\widehat{\text{Var}}\{\widehat{L}_1(x)\}}}\right) \tag{17}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Once an estimate of the exceedance probability of the threshold of interest is available at every location x, a reasonable decision rule is to fix a particular probability threshold, T, and declare positive all locations x with $\widehat{E}(x, p_0) > T$. For a given set of locations R we define the sensitivity of the inferential procedure as

$$\text{Sens.}(R, T) = \frac{1}{R} \sum_{x \in R} I \left\{ \widehat{E}(x, p_0) > T, P_1(x) \geq 0.2 \right\}, \tag{18}$$

which represents the frequency with which the procedure correctly identifies truly positive locations in region R using the probability threshold T. Here $I(\cdot)$ denotes the indicator function.

Similarly, we define the specificity of a given procedure as

$$\text{Spec.}(R, T) = \frac{1}{R} \sum_{x \in R} I \left\{ \widehat{E}(x, p_0) < T, P_1(x) < 0.2 \right\} \tag{19}$$

and represents the frequency with which the procedure correctly identifies truly negative locations in region R using the probability threshold T. The threshold value T could be anything between 0 and 1, but some insight into reasonable values can be obtained using simulations. Figure 6 displays the sensitivity (left two panels) and specificity (right two panels) functions for the bivariate Bayesian (top two panels) and calibration (bottom two panels) models for each of the 100 simulated data sets. These functions are specific to the ROI and are obtained using an equally spaced grid for threshold values between $[0.0225, 0.975]$.

The trade–off between sensitivity and specificity is clear in Figure 6 because sensitivity is a decreasing function while specificity is an increasing function of the probability threshold. One could, of course, have perfect sensitivity by setting $T = 0$ or close to zero. The problem with such an approach is that it would result in abysmal specificity results. Indeed, for low values of T both procedures correctly identify truly negative locations with probability less than 0.5 for many data sets, meaning that both procedures would be worse than a coin toss. A similar discussion holds for values of T close to 1. The large variability of the specificity and sensitivity functions is most probably due to lack of information at locations that are far from sampling locations.

While very informative, it would be hard to use the plots in Figure 6 for choosing a certain decision threshold. Figure 7 displays the average $\pm 2se$ for the sensitivity and specificity functions. The average and the standard errors were obtained using the 100 values of the function at a fixed threshold corresponding to simulated data sets. This graph shows that with a threshold $T = 0.7$ one would obtain roughly 0.8 average sensitivity with both methods and 0.9 average specificity for the specific ROI. In this specific ROI, under our model there were $231,815$ truly positive locations and $128,502$ truly negative locations.

Note that both inferential methods are much more accurate at the actual sampling locations, as shown by the MSE plot in Figure 5. Similarly, both methods have much better predictive properties at these locations. The average sensitivity curves for sampling locations shown in Figure 8 are much improved over the average sensitivity curves in Figure 7. In fact, the bivariate binomial model has 0.89 average sensitivity for the probability threshold $T = 0.975$. The calibration model behaves reasonably well for thresholds between 0.7 and 0.8, but exhibits a rapid decrease in sensitivity for higher thresholds. In our simulation there were 37 truly positive and 37 truly negative locations.

The second and third simulation studies were designed to mimic a possible sampling scenario, in which RAPLOA sampling is conducted at some locations without conducting

Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper103

the parasitological sampling. This was achieved by following the simulation recipe described above for the first simulation, except that at each simulation we did not use parasitology count data simulated at some specific locations in the ROI. The second simulation study considered 15 specific locations while the third considered all 74 sampling locations within the ROI. Instead, these data were treated as missing and were simulated from their joint posterior distribution in the bivariate Binomial Bayesian model. In the frequentist model we used the calibration and pooling algorithm described in Section 5.1.

Figure 9 presents the same type of results as Figure 5 comparing the MSE for the bivariate Bayesian methodology with that of the calibration model for the case when 15 locations were not parasitologically surveyed. Note that the Bivariate methodology produces larger MSE than its calibration counterpart over the entire ROI and slightly smaller MSE over just the sampling locations. This may, in part, be due to the slow mixing of the Markov Chains combined with the inherent computational limitations of our simulation study. Our choice of number of burn–in/simulation samples was $20,000/20,000$ because this was enough for the case without missing parasitology data. Moreover, increasing the number of simulations to obtain reasonable results would result in un–reasonably long simulation times. Not surprisingly, both methods perform better at the sampling locations than over the entire region. However, it is surprising that there is a serious loss of estimation efficiency from the case when parasitology sampling is actually conducted at sampling locations. If one compares results from Figure 9 with those in Figure 5 one can see the much larger MSE in the case when parasitology is missing at some locations (note the different scales). Indeed, average MSE increases roughly 30% from 0.25 to 0.32 for the ROI and from 0.08 to 0.11 for the sampling locations using the frequentist calibration methodology.

While the results in Section 2 seem to indicate good calibration between the parasitological and RAPLOA sampling, using only RAPLOA sampling may result in serious losses of efficiency in estimating the prevalence function. As expected, the loss of information is also reflected in the loss of prediction properties of both methods. This can be seen by comparing the average sensitivity and specificity curves for the ROI corresponding to missing parasitology sampling in Figure 10 with the ones corresponding to full data analysis in Figure 6. Similar losses were observed for the sampling locations.

Figures 11 and 12 present the same type of results when parasitology was not conducted at all 74 locations of the ROI. The Bayesian bivariate method performs better than the calibration methodology both in terms of MSE and prediction properties both at sampling location and over the entire region. However, these properties are farther degraded when compared to the case when only 15 observations were missing. Indeed, average MSE increases roughly 63% from 0.32 to 0.52 for the ROI and 255% from 0.11 to 0.39 for the sampling locations using the frequentist calibration methodology.

7 Software Implementation and Testing

The statistical methods described here were implemented using the R software package [R Development Core Team (2004)]. Code for computing predictions of exceedence probabilities was written into an R *package* and called `arlat`, standing for 'A Raploa Analysis Tool'.

The robustness of the code was tested by simulations. We first generate a fairly smooth Gaussian random field over our possible study area. Then at a number of locations we compute simulated parasitology and RAPLOA prevalences. We run the exceedence computation using some subset of the points as calibration data (using both parasitology and RAPLOA

Hosted by The Berkeley Electronic Press

data) and the rest as new RAPLOA data. This process was repeated many times with various variations on the numbers and positions of locations. This kind of testing reveals shortcomings with the code coping with certain edge-case situations, such as having one or no locations in the calibration or survey data, but cannot test whether the results are correct!

In order to enable workers in the field to use our statistical methodology we needed to develop an easy-to-use user interface. We could have developed something completely within R but instead chose to write an add-on to an existing geographic information system (GIS). Our choice of GIS was constrained by the need for interaction with R, and we chose a freely-available multi-platform (Windows, Linux, Solaris) package for ease of development and lack of licensing issues.

After studying several alternatives we settled on OpenEV [G Walter (2002)]. Large amounts of this system are written using the Python [van Rossum and Fred L. Drake (2003)] language and it has the facility to write add-on modules that integrate with the menu system, can query geographic data, and create new data layers.

In order to communicate between Python and R we use the Rserve facility. This runs R in the background as a server, and client programs connect to it via a socket interface. Client libraries for Rserve are available for Java and C++ languages, but it was relatively easy to develop a library for Python to talk to Rserve. With this in place we now have the components for OpenEV to use R for calculations and OpenEV for display and manipulation of the data.

The end-user experience is quite simple: they supply a dataset of new survey locations with the number of people tested for Raploa and the number of positive tests at each location. This is imported into OpenEV (from a spreadsheet file) and can then be displayed with other map data, such as country or region boundaries, village maps, roads, topography etc. The user then starts the Arlat dialog from a drop-down menu. Here they select the map layer containing the survey data and define a rectangle over which to produce the predicted exceedance probability. The resolution of the output grid is also chosen.

On clicking the 'ok' button, OpenEV uses the Python-Rserve code to use the `arlat` package in R to compute the exceedence probability over the specified area. This is then loaded into OpenEV as a new raster grid layer.

Figure 13 provides a screen shot of our software showing sampling locations (in red) and corresponding exceedance probabilities in a small area containing the sample locations. By default the grid is coloured such that exceedences below 0.7 are invisible, and from 0.7 to 1.0 are coloured from green through yellow and orange up to red. The table in the upper right hand corner of the screen allows simple data updating, while the arlat software can rapidly produce exceedance probabilities based on the methodology described in Section 5.

8 Discussion

Our paper describes a challenging application of spatial statistical methodology to tropical disease epidemiology. The general area of application is the spatial mapping of disease prevalence in settings where registry data are unavailable, and the only feasible way to collect prevalence data is by binomial sampling within a relatively small number of scattered village communities. The specific goal in our case is to map the continuous spatial variation in the predictive probability that local prevalence exceeds a pre–determined policy intervention threshold. In our data, empirical prevalence in each sampled village is assessed by two methods: a traditional, parasitological method based on the microscopic examination of

Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

blood-smears; and a rapid, questionnaire-based method, RAPLOA Tako et al. (2002). In our data, both methods are used in all sampled villages, and the data were collected primarily to establish a calibration relationship which, to a good approximation, holds over a wide geographical area of central Africa. Because of resource limitations, it is likely that future surveys in many areas where the current data give very imprecise predictions of prevalence will use only the questionnaire-based method. How best to combine the existing and future data is therefore a problem in bivariate spatial modelling. A further practical consideration is that field-workers need a computationally simple method for the initial inspection of data obtained in local surveys.

From a methodological perspective, our approach has been to adapt and extend the methods of model–based geostatistics as proposed by Diggle et al. (1998), in which an unobserved, stationary Gaussian process $S(x)$ is added to the linear predictor in a generalised linear model. One limitation of the methods used in Diggle et al. (1998) is that they are, at the time of writing, computationally impractical for our data, for which we need to make predictions at approximately millions of locations. Our response to this has been to replace the stationary process $S(x)$ by a low-rank, random-coefficient two-dimensional spline smoother. One implication of this is that the computational load is essentially independent of both the number of sampling locations and the number of prediction locations, but rather is determined primarily by the number of knots specified for the spline smoother. Our method is similar in spirit to the geoadditive models of Kammann and Wand (2003) or the thin-plate regression splines of Wood (2003). A different approach to the computational problems posed by the need to make predictions at a large number of locations would be to approximate the spatially continuous process $S(x)$ by a spatially discrete Markov random field. See, for example, Rue and Tjelmeland (2002) or Besag and Mondal (2005).

We believe our paper is also the first to address the problem of formulating and fitting a geostatistical model for bivariate binomial data. We use a method of construction previously proposed by Gelfand, Schmidt, Banerjee, and Sirmans (2004) for Gaussian data, in which there is a natural asymmetry between the two components of $S(x) = \{S_1(x), S_2(x)\}$. This justifies modelling $S_1(x)$ marginally, and $S_2(x)$ conditional on $S_1(x)$.

Finally, we have taken account of the practical problems of implementing sophisticated spatial statistical analyses, especially those which rely on careful tuning of a Monte Carlo Markov chain algorithm, routinely under field-conditions by comparing the formal Bayesian analysis of our bivariate model with a much simpler, albeit approximate, analysis which fits a Gaussian model on the empirical logit scale. This is a version of what Cressie (1993) calls"trans-Gaussian kriging" which, as far as we are aware, has not previously been used in a bivariate setting, nor have its resulting predictions been compared with those made by the theoretically superior generalised linear modelling approach. The extent to which, in general, generalised linear geostatistical modelling out-performs transformed Gaussian geostatistical modelling remains an open question.

Returning to the *Loa loa* application, we have demonstrated the feasibility of both the full Bayesian analysis and the simpler transformed Gaussian analysis for prediction problems on the required geographical scale. We have also shown that the calibration relationship between parasitological and questionnaire-based methods is consistent across widely separated areas of central Africa. However, the sampling design for the current data is not well suited to spatial prediction because of its strongly clustered nature, and the data therefore do not give an authoritative solution to the prediction problem. This same point was emphasized in Diggle et al. (2006). They argued that an important feature of probabilistic prediction in

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

this context was that it can identify areas where more data are required, to fill-in gaps where no surveys have been conducted and environmental covariates do not allow an unequivocal conclusion that for the area in question the local prevalence lies below the policy-relevant threshold.

Acknowledgements

The support for this research was provided by the World Health Organization through grant H060132 "Calibration and Mapping for Parasitological and Raploa Estimates for Loa Loa Prevalence Parasitological disease mapping in Africa".

References

H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 1973.

J. Besag and D. Mondal. First-order intrinsic autoregressions and the de wijs process. *Biometrika*, 2005.

J-P Chilès and P. Delfiner. *Geostatistics*. New York : Wiley, 1999.

C.M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society – Series B*, 2004.

C.M. Crainiceanu, D. Ruppert, G. Claeskens, and M.P. Wand. Likelihood ratio tests in linear mixed models with one variance component. *Biometrika*, 2005a.

C.M. Crainiceanu, D. Ruppert, and M.P. Wand. Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software*, 2005b.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 1979.

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

P.J. Diggle, R.A. Moyeed, and J.A. Tawn. Model–based geostatistics (with discussions). *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 1998.

P.J. Diggle et al. Spatial modelling and prediction of *loa loa* risk: decision making under uncertainty. *under review*, 2006.

P Farris-Manning G Walter, F Warmerdam. An open source tool for geospatial image exploitation. In *Proceedings of the IGARSS 2002 Conference*, volume 6, pages 3522–3524, 2002. DOI: 10.1109/IGARSS.2002.1027236.

B. Ganguli and M.P. Wand. *SemiPar 1.0 Users' Manual*, 2005. URL `http://www.maths.unsw.edu.au/ wand/papers.html`. R package version 1.0.

A.E. Gelfand, A.M. Schmidt, S. Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion). *Test*, 2004.

http://biostats.bepress.com/jhubiostat/paper103

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall Ltd (London; New York), 1994.

Kammann and M.P. Wand. Geoadditive models. *Applied Statistics*, 2003.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley: New York), 1990.

C. L. Mallows. Some comments on c_p. *Technometrics*, 1973.

Doug Nychka. *fields: Tools for spatial data*, 2004. URL `http://www.cgd.ucar.edu/stats/Software/Fields`. R package version 2.0.

D.W. Nychka. *Spatial process estimates as smoothers. In M. Schimek (Ed.), Smoothing and regression*. Heidelberg: Springer–Verlag, 2000.

D.W. Nychka, P. Haaland, M. O'Connell, and S. Ellner. *FUNFITS, data analysis and statistical tools for estimating functions. In D. Nychka, W.W. Piegorsch, and L.H. Cox (Eds.), Case studies in Environmental Statistics (Lecture Notes in Statistics)*, volume 132. New York: Springer–Verlag, 1998.

D.W. Nychka and N. Saltzman. *Design of air quality monitoring networks. In D. Nychka, W.W. Piegorsch, and L.H. Cox (Eds.), Case studies in Environmental Statistics (Lecture Notes in Statistics)*, volume 132. New York: Springer–Verlag, 1998.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

H. Rue and H. Tjelmeland. Fitting gaussian markov random fields to gaussian fields. *Scandinavian Journal of Statistics*, 2002.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 2002.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge University Press: UK, 2003.

S.G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 1987.

I. Tako et al. Rapid assessment method for prevalence and intensity of *l. loa* infection. *Bulletin of the World Health Organisation*, 2002.

Guido van Rossum and Jr. (Editor) Fred L. Drake. *The Python Language Reference Manual*. Network Theory Ltd, 2003. ISBN 0954161785.

G. Wahba. *Spline models for observational data*. SIAM [Society for Industrial and Applied Mathematics], 1990. ISBN 0-89871-244-0.

S.N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 2003.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

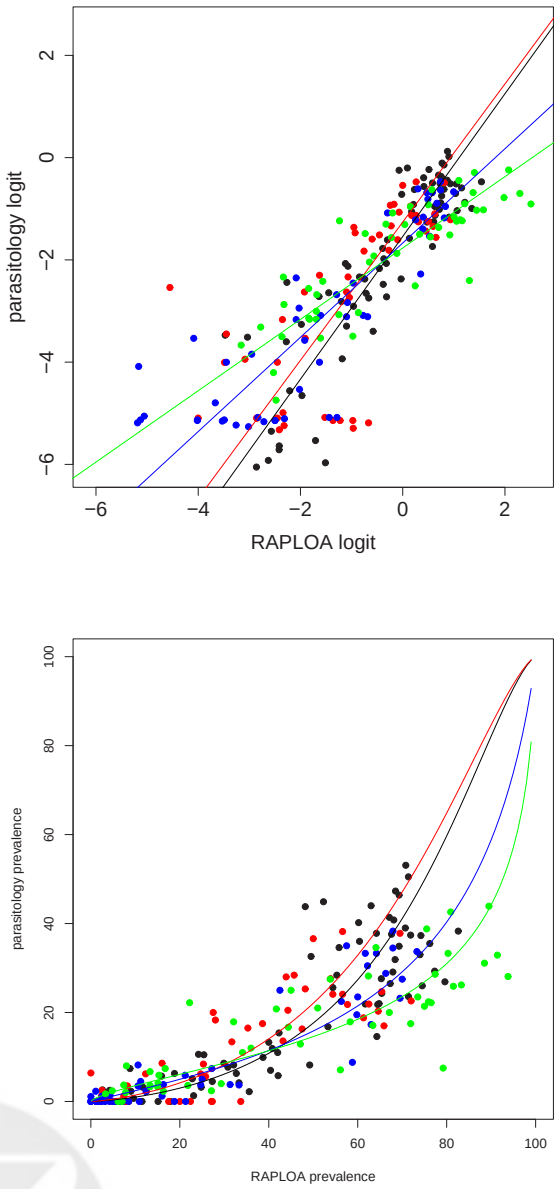

Figure 1: Calibration relationships between RAPLOA–based and parasitology–based estimates of prevalence from four surveys. The top plot shows results on the empirical logit scale. The bottom plot shows results back–transformed to the prevalence scale. The four surveys are distinguished by the plotting colors: (Cameroon); (DRC West); (DRC East); (Congo).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

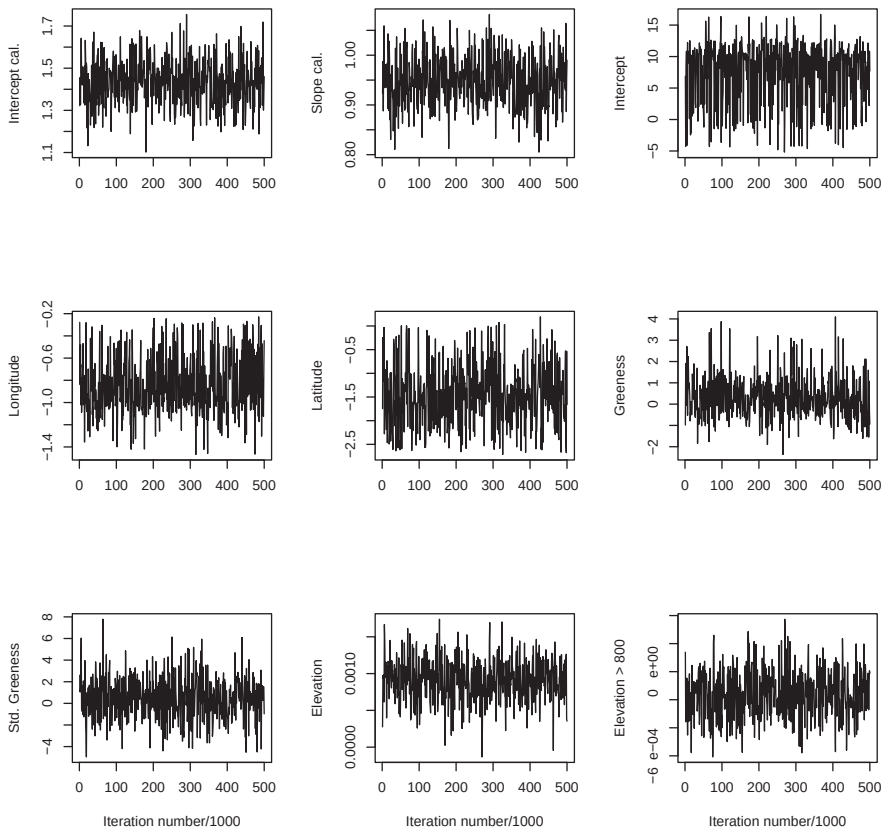

Figure 2: Posterior simulations from the joint distribution of some of the parameters of model (2). MCMC sampling was used to produce 500, 000 correlated samples from the target distribution after an initial 20, 000 burn–in simulations. For clarity, only every 1000th sample is displayed.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

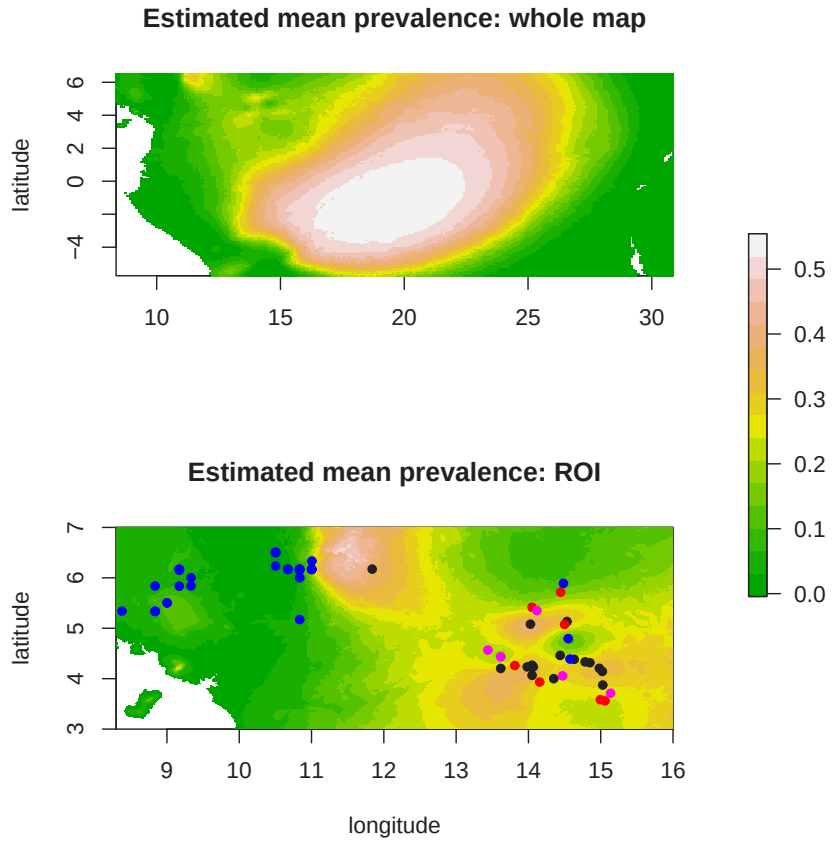

Figure 3: Posterior predictive surface for the *Loa loa* parasitological prevalence based on the Bayesian analysis of the Bivariate Binomial spatial model (2). Top panel represents the results as they are extrapolated to a very large region containing all sampling locations. The bottom panel is the inference in a smaller region that contains 74 sampling locations color–coded according to the observed (empirical) parasitology prevalence: black $\widehat{P}(x) > 0.3$, red $0.25 \leq \widehat{P}(x) < 0.3$, magenta $0.20 \leq \widehat{P}(x) < 0.25$, cyan $0.18 \leq \widehat{P}(x) < 0.20$ and blue $\widehat{P}(x) < 0.18$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

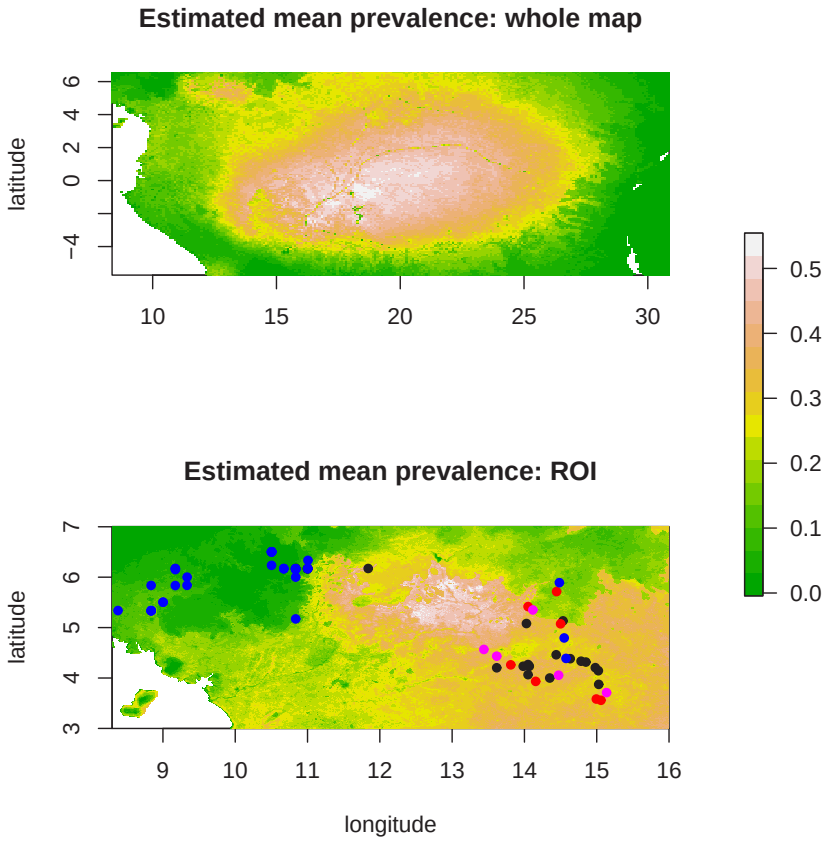

Figure 4: Posterior predictive surface for the *Loa loa* parasitological prevalence based on the calibration and pooling method described in Section 5.1 of the calibration Gaussian spatial model (2). Top panel represents the results as they are extrapolated to a very large region containing all sampling locations. The bottom panel is the inference in a smaller region that contains sampling locations and contains 74 sampling locations color–coded according to the convention from Figure 3.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

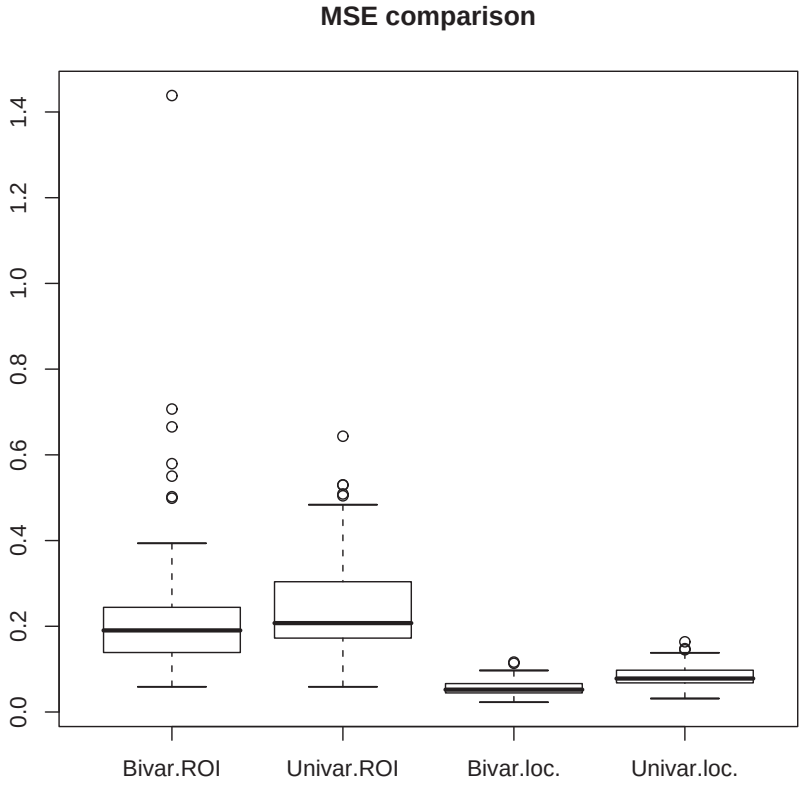

Figure 5: Mean square error comparison of the Bayesian bivariate method calculated over the ROI (Bivar.ROI) and at the sampling location (Bivar.loc.) with the frequentist calibration method calculated over the ROI (Univar.ROI) and at the sampling locations (Univar.loc.). Results are calculated over 100 simulated data sets according to the first simulation study described in Section 6. All data sets contained parasitology and RAPLOA sampling at every location.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

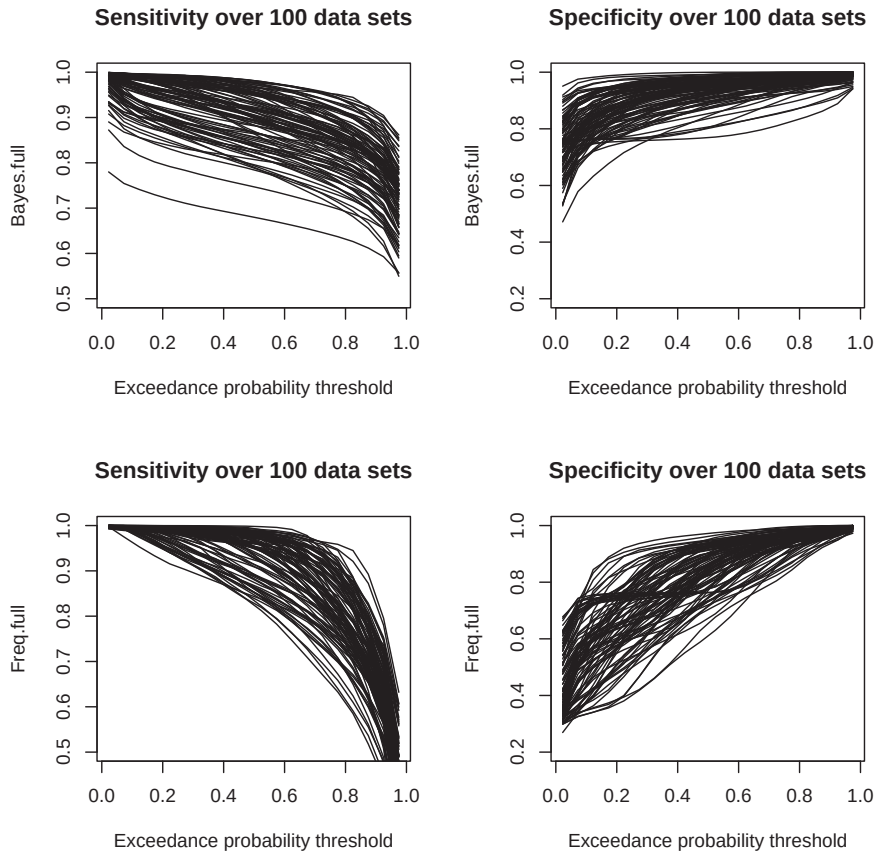

Figure 6: Sensitivity (left panels) and specificity (right panels) functions calculated over the ROI for each of the 100 simulated data sets as a function of the probability threshold. The Bayesian bivariate method is represented by the top panels and the frequentist calibration method by the bottom panels. All data sets contained parasitology and RAPLOA sampling at every location. Sensitivity represents the proportion of truly positive locations identified by a method for a given exceedance probability threshold.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

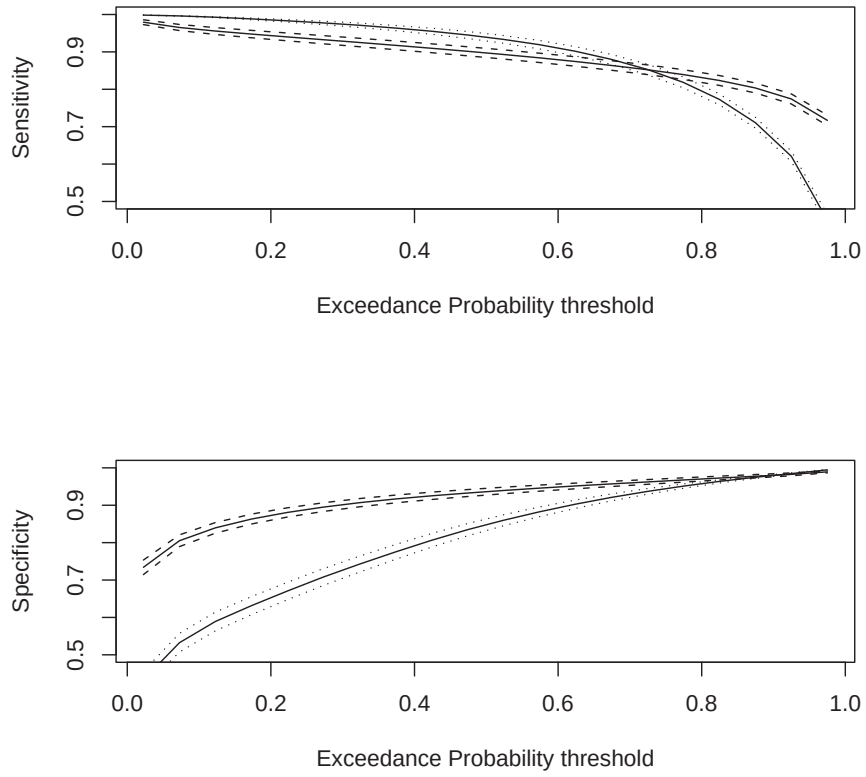

Figure 7: Average sensitivity and specificity functions with 95% confidence intervals based on sensitivity and specificity results shown in Figure 6. In particular, for every exceedance probability threshold the mean and the standard error is calculated based on the 100 results corresponding to that particular threshold. The bivariate binomial model results are presented as a solid line with dashed lines for confidence intervals. The calibration model results are presented as a solid line with dotted lines for confidence intervals. Results correspond to the ROI.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

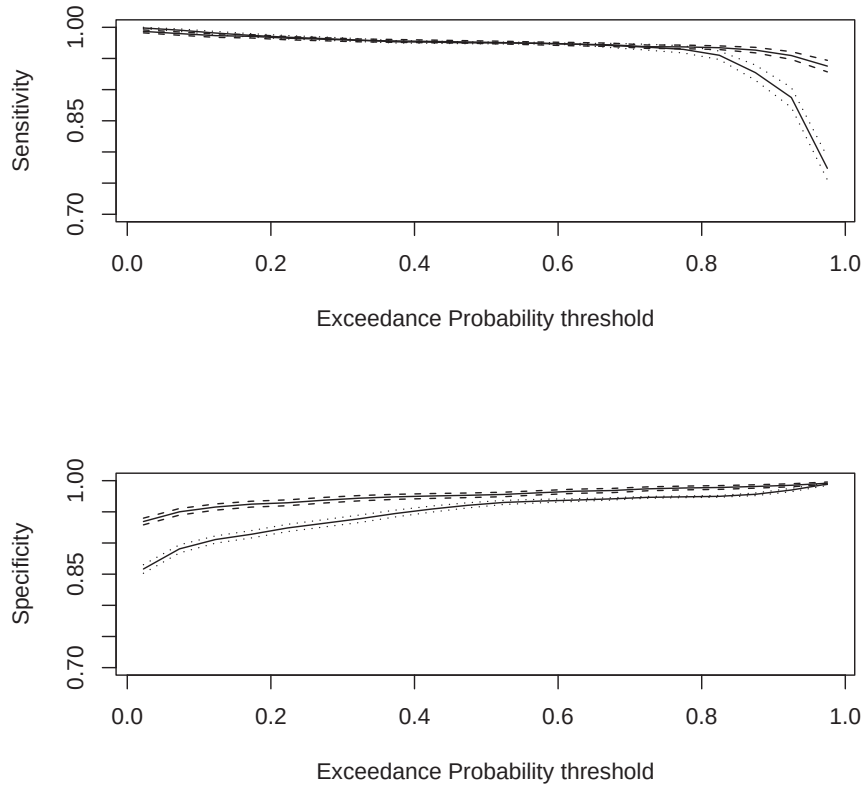

Figure 8: Same type of results as Figure 7 with the difference that results correspond to sampling locations only. The bivariate binomial model results are presented as a solid line with dashed lines for confidence intervals. The calibration model results are presented as a solid line with dotted lines for confidence intervals.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

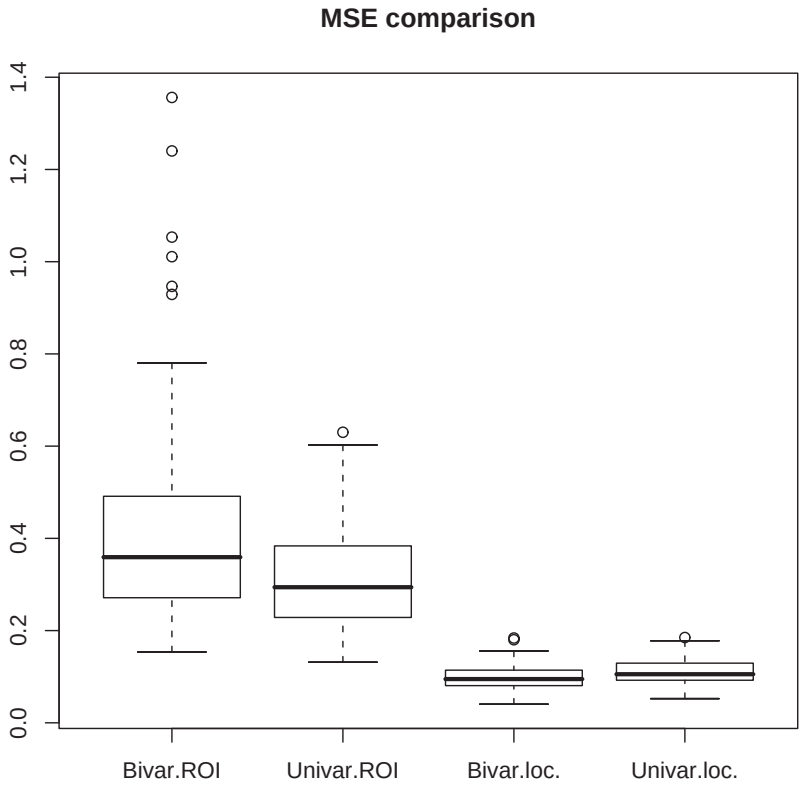

Figure 9: Mean square error comparison of the Bayesian bivariate method calculated over the ROI (Bivar.ROI) and at the sampling location (Bivar.loc.) with the frequentist calibration method calculated over the ROI (Univar.ROI) and at the sampling locations (Univar.loc.). Results are calculated over 100 simulated data sets according to the first simulation study described in Section 6. All data sets contained RAPLOA sampling at every location. Parasitology samples were simulated at all but 15 specific locations in the ROI

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

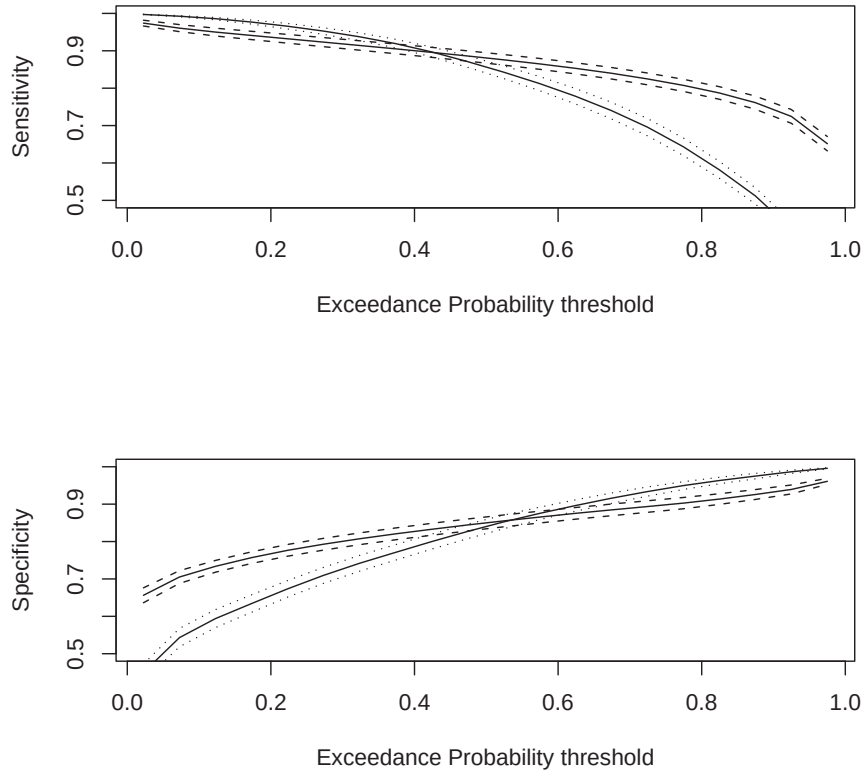

Figure 10: Average sensitivity and specificity functions with 95% confidence intervals based on 100 simulated data sets. Parasitology samples were simulated at all but 15 specific locations in the ROI. For every exceedance probability threshold the mean and the standard error is calculated based on the 100 results corresponding to that particular threshold. The bivariate binomial model results are presented as a solid line with dashed lines for confidence intervals. The calibration model results are presented as a solid line with dotted lines for confidence intervals. Results correspond to the ROI.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

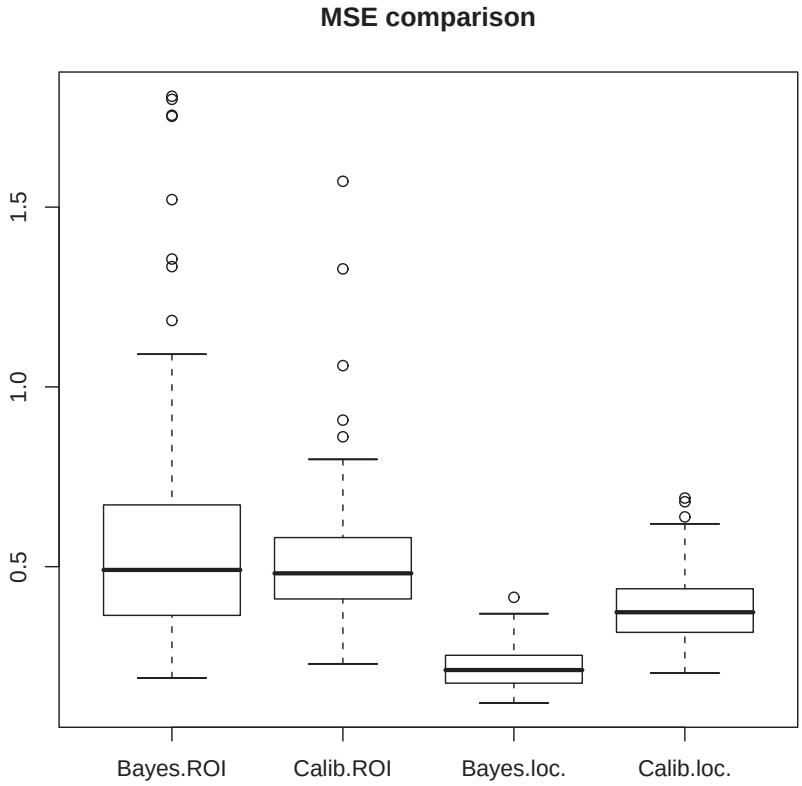

Figure 11: Mean square error comparison of the Bayesian bivariate method calculated over the ROI (Bivar.ROI) and at the sampling location (Bivar.loc.) with the frequentist calibration method calculated over the ROI (Univar.ROI) and at the sampling locations (Univar.loc.). Results are calculated over 100 simulated data sets according to the first simulation study described in Section 6. All data sets contained RAPLOA sampling at every location. Parasitology samples were simulated at all locations with the exception of the 74 locations within the ROI.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103

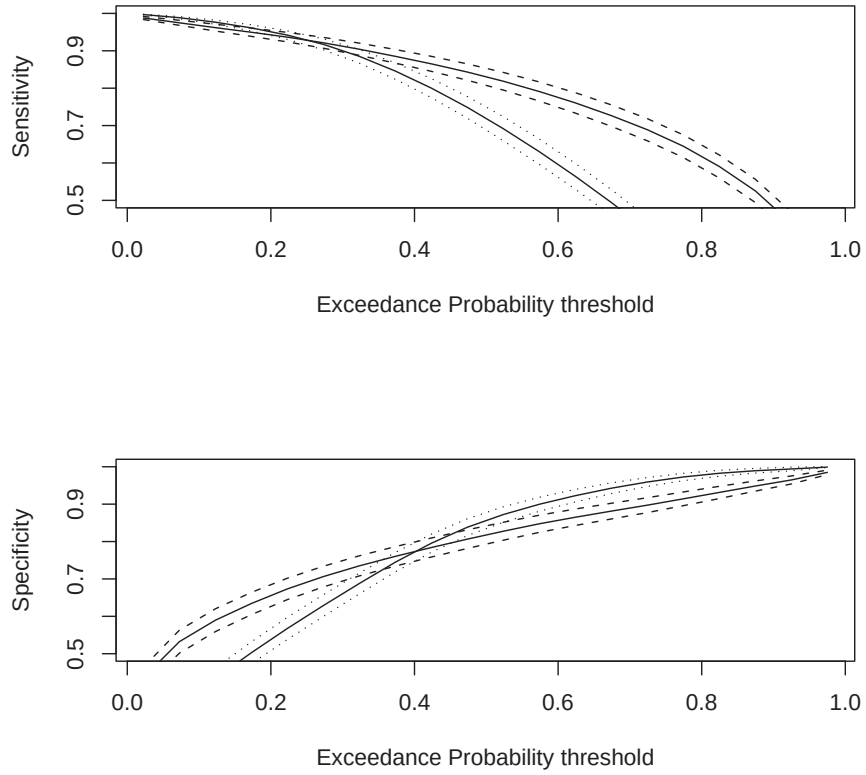

Figure 12: Average sensitivity and specificity functions with 95% confidence intervals based on 100 simulated data sets. Parasitology samples were simulated at all locations with the exception of the 74 locations within the ROI. For every exceedance probability threshold the mean and the standard error is calculated based on the 100 results corresponding to that particular threshold. The bivariate binomial model results are presented as a solid line with dashed lines for confidence intervals. The calibration model results are presented as a solid line with dotted lines for confidence intervals. Results correspond to the ROI.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

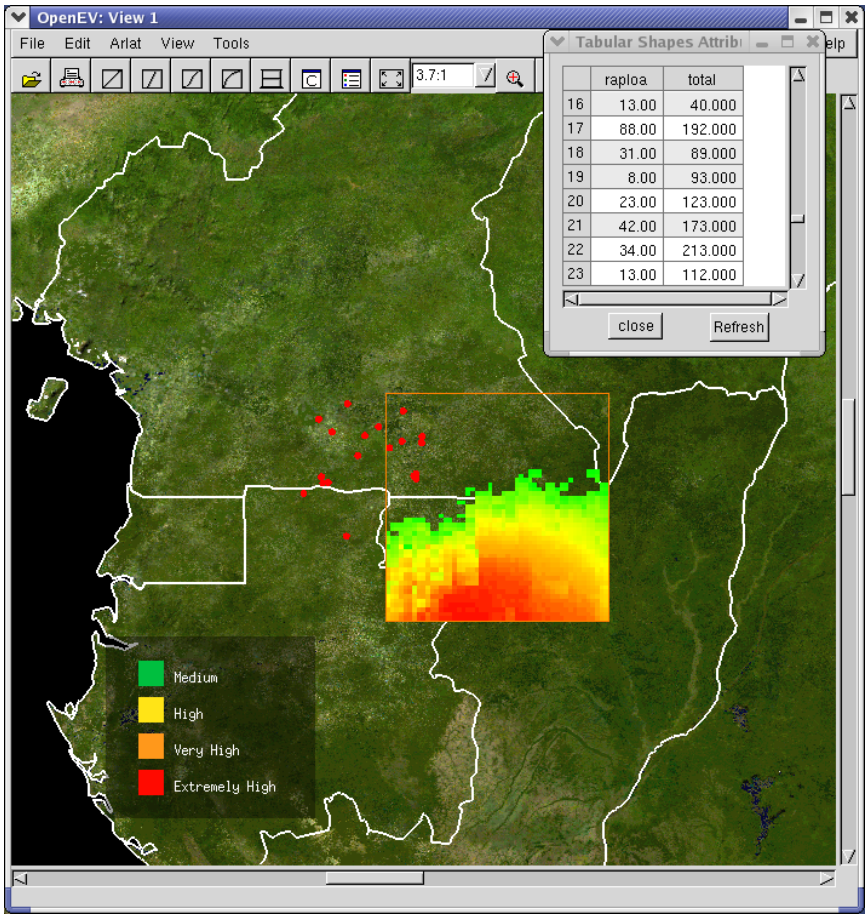

Figure 13: Screen shot of our software showing sampling locations (in red) and corresponding exceedance probabilities in a small area containing the sample locations. By default the grid is coloured such that exceedences below 0.7 are invisible, and from 0.7 to 1.0 are coloured from green through yellow and orange up to red. The table in the upper right hand corner of the screen allows simple data updating, while the arlat software can rapidly produce exceedance probabilities based on the methodology described in Section 5.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper103