Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Bivariate Binomial Spatial Modeling of Loa loa Prevalence in Tropical Africa

Ciprian M Crainiceanu, Peter J Diggle & Barry Rowlingson

To cite this article: Ciprian M Crainiceanu, Peter J Diggle & Barry Rowlingson (2008) Bivariate Binomial Spatial Modeling of Loa loa Prevalence in Tropical Africa, Journal of the American Statistical Association, 103:481, 21-37, DOI: 10.1198/016214507000001409

To link to this article: https://doi.org/10.1198/016214507000001409

Published online: 01 Jan 2012.

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

Ciprian M. CRAINICEANU, Peter J. DIGGLE, and Barry ROWLINGSON

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 starts with the nonspatial calibration of survey instruments, continues with the spatial model building and assessment, and ends with robust, tested software intended for use by field workers for online prevalence map updating. From a statistical perspective, we address several important methodological issues: building spatial models that are sufficiently complex to capture the structure of the data but remain computationally usable, reducing the computational burden in the handling of very large covariate data sets, and devising methods for comparing spatial prediction methods for a given exceedance policy threshold.

KEY WORDS: Geostatistics; Low rank; Thin-plate splines.

1. INTRODUCTION

Loaisis, or eyeworm, is an endemic disease of the wet tropics, caused by *Loa loa*, a filarial parasite transmitted to humans by the bite of an infected *Chrysops* fly. The public health importance of this disease has recently increased because of its impact on the African Programme for Onchocerciasis Control (APOC). Onchocerciasis, or river blindness, is also endemic in many tropical areas and is of major direct public health importance because of its enormous impact on infected populations. Onchocerciasis is caused by the filarial parasite *Onchocerca volvulus*, which is transmitted by the bite of an infected *Simulium* fly. Onchocerciasis causes blindness if untreated; hence its common name.

The APOC, launched in 1995, administers mass treatment with the drug ivermectin, which is highly effective in eliminating onchocerciasis parasites from the blood of infected individuals (World Health Organization 1995; Seketeli et al. 2002). By the end of 2005, the program had administered approximately 50 million treatments (Alleman, Twum-Danso, and Thylefors 2006). Unfortunately, some individuals who are highly infected with *Loa loa* parasites develop severe, and occasionally fatal, adverse reactions when treated with ivermectin (Boussinesq et al. 1998). Boussinesq et al. (2001) established a close relationship between village-level prevalence of *Loa loa* and the existence of highly affected individuals within the village. Thus, APOC policy is that before implementing mass treatment for onchocerciasis in areas where *Loa loa* prevalence is thought to exceed 20%, precautionary measures should be put in place to enable prompt treatment of any cases of serious adverse reaction to the drug. Thus, 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 parasitologic sampling, that is, by taking blood-samples from selected village communities and using the observed proportion of positive results as an estimate of the local prevalence. However, it is not feasible to undertake parasitologic sampling in every community where ivermectin treatment is envisaged.

Geostatistical modeling provides a way of using community-level results to estimate continuous spatial variation in prevalence and express the results as an "exceedance map," that is, a map of the probability that local prevalence exceeds the 20% policy intervention threshold. Following Diggle, Moyeed, and Tawn (1998), Thompson et al. (2007) proposed the following univariate binomial geostatistical model for describing village-level parasitology data:

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

where $Y(x)$ 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 measured by the normalized difference vegetation index (NDVI) 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 parasitologic 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, allows the sampling of many more communities than would be possible using parasitologic sampling (Takoungang et al. 2002). To validate the RAPLOA methodology, surveys were carried out in which both methods of determining prevalence were used. In this article, we formulate a class of bivariate geostatistical models for data of this kind and describe a method for fitting a subclass 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 through Markov chain Monte Carlo (MCMC) simulations, 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.

Ciprian M. Crainiceanu is Assistant Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ccrainic@jhsph.edu*). Peter J. Diggle is Professor, Department of Medicine, Lancaster University, Lancaster LA1 4YF, U.K. and Adjunct Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *p.diggle@lancaster.ac.uk*). Barry Rowlingson is Research Officer, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K. (E-mail: *B.Rowlingson@lancaster.ac.uk*). This work was supported by World Health Organization grant F30/181/215 (Calibration and Mapping for Parasitological and RAPLOA estimates of *Loa loa* Prevalence). Diggle's work also was supported by the U.K. Engineering and Physical Sciences Research Council through the award of a Senior Fellowship (grant GR/S48059/01), and by the National Institute of Environmental Health Science grant 1 R01 ES012054 (Statistical Methods for Environmental Epidemiology).

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Applications and Case Studies
DOI 10.1198/016214507000001409

Section 2 presents a nonspatial exploratory analysis of the validation data, which demonstrates the potential value of the RAPLOA instrument as a low-cost alternative to parasitologic sampling. Section 3 describes the formulation of the bivariate geostatistical model, which forms the basis for our proposed solution to the *Loa loa* mapping problem. Sections 4 and 5 give the results obtained using Bayesian predictive inference and our proposed computationally fast approximation. 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 presents some conclusions.

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 parasitologic sampling. In the RAPLOA methodology, each person in the survey is classified as a positive case if he or she answers "yes" to all three of the following questions: Have you ever suffered from eye worm?; did it look like this photograph?; did it last less than 1 week? In parasitologic 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 validation 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 assumed 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 computed the sample mean vector and covariance matrix of each sample, and derived 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-transformed the principal axis onto the prevalence scale.

For each datum, if n denotes the number of persons surveyed and y the number of positives, then the raw estimated prevalence is y/n. We define the empirical logit of prevalence as $\log\{(y+.5)/(n-y+.5)\}$. Figure 1 shows a strong, direct relationship between the results obtained by the two methods. This relationship is approximately linear on the empirical logit scale,

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

Survey	Location	Number of villages	Subjects/village		
			Minimum	Mean	Maximum
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

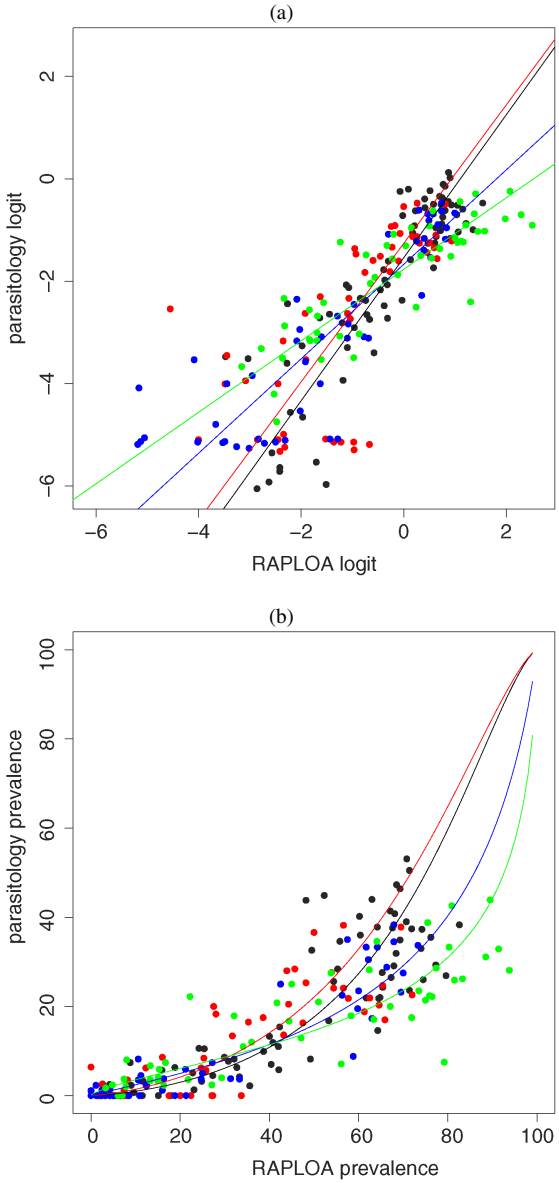

Figure 1. Calibration relationships between RAPLOA-based and parasitology-based estimates of prevalence from four surveys. (a) Results on the empirical logit scale. (b) Results back-transformed to the prevalence scale. The four surveys are distinguished by the plotting colors: black (Cameroon); red (DRC West); blue (DRC East); green (Congo).

with correlation .83, a pronounced shift between the two means ($-.77$ for RAPLOA and -2.41 for parasitology) but approximately equal variances (2.53 for RAPLOA and 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 just described. Note in particular that on the prevalence scale, the calibration curves obtained from the four surveys agree closely over the range of parasitologic prevalences between 0 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 MODELING

The results presented in Section 2 suggest that a bivariate geostatistical analysis may well enable more precise spatial predictions of parasitologic prevalence by exploiting the relationship between parasitologic and RAPLOA prevalence rather than by analysing only the parasitologic data. More importantly, a bivariate model would allow us to incorporate data from additional RAPLOA surveys, which are more cost-effective than would be additional parasitologic surveys and can be targeted at areas where the current predictions are least precise.

We initially used the same data for bivariate geostatistical modeling as we used for the exploratory analysis. However, we found that the data from survey 2, which were located far to the east of the other three surveys, did not help the spatial modeling. In addition, we obtained an additional data set from Cameroon, which improved the spatial coverage of western equatorial Africa. Thus, for the geostatistical analysis, we used the calibration data from surveys 0, 1, and 3, together with the new data. This gave a total of 275 locations within a study region demarcated by longitude 3 to 16 degrees east and latitude 5 degrees south to 15 degrees north (Fig. 3 in Sec. 4 and Fig. 4 in Sec. 5).

3.1 A Bivariate Binomial Geostatistical Model

To enable predictive mapping of both parasitologic and RAPLOA prevalence, we fit the following bivariate binomial model for village level numbers of positive indications according 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, for the village at geographical location x, and $N_1(x)$ 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* parasitologic prevalence process, is the focus of interest, and our specific objective is to identify geographic areas where $P_1(x) > .2$ 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 parasitologic and RAPLOA surveys and are linked through a calibration relationship described by the fifth equation in (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 parasitologic 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, whereas $C(x)^T\beta$ describes the spatial variation in the mean attributable to the effects of covariates observed at location x. Finally, $S(x)$ is a mean-0 stationary process representing any residual spatial variation not explained by the available covariates. Note that our sequence of conditional models provides a joint model for parasitology and RAPLOA prevalence. Whereas both indicators could have been viewed as error-prone markers, the WHO policy is designed for the prevalence estimated according to parasitologic sampling, which is considered the gold standard. For this reason, we wanted to explicitly model the parasitologic prevalence, which is the object of our statistical inference.

During our study, we considered simpler models including fixed and random subregional effects. However, a subregional fixed-effects model could not be used to predict prevalence at new locations, and a random-effects model with three or four subregional effects would fail to borrow strength across regions. Defining a subregion would be relatively easy with the original data, because each of the four surveys was associated with a particular subregion, but would became increasingly problematic as new data became available at sampling locations between the existing ones. Thus geographic partitioning into subregions would become an increasingly difficult problem in a dynamic data acquisition framework. We also considered more complex models, including random subject effects and overdispersion for the binomial variability, but these models exceed the scope of the present study.

Our bivariate binomial geostatistical model (2) is of necessarily 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 done by Thompson et al. (2007). However, this would be computationally burdensome for the current application because of the 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 next 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 widely used geostatistical method known as (ordinary) kriging is a linear smoothing method that is formally equivalent to minimum mean squared 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 that are part of the family of *general radial smoothers*. Good discussions of the formal connection between these two important methods have been 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 also can be identified as Gaussian spatial processes, although from this perspective, their covariance structure may seem unnatural (e.g., Wahba 1990; 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 $(Y_i, x_i) : i = 1, \ldots, n$ under the assumption that the ϵ_i's are mutually independent $\mathrm{N}(0, \sigma_\epsilon^2)$ 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 $\mathcal{C}(\|x_i - x_j\|)$. Let Z_i^R be 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

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

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 cross-validation (CV) or generalized CV (GCV) (Craven and Wahba 1979), C_p (Mallows 1973), Akaike information (Akaike 1973) and restricted or unrestricted maximum likelihood (REML) (Ruppert, Wand, and Carroll 2003; Kammann and Wand 2003) criteria. In particular, Kammann and Wand (2003) and Ruppert et al. (2003) showed that the penalized spline solution using REML estimation is equivalent to the best linear unbiased predictor (BLUP) in the linear Gaussian mixed model

$$\begin{gathered} y = X\beta + Zb + \epsilon, \\ \mathrm{E}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0_{n\times 1} \\ 0_{n\times 1} \end{pmatrix}, \qquad \mathrm{cov}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_n & 0_{n\times n} \\ 0_{n\times n} & \sigma_\epsilon^2 I_n \end{pmatrix}, \end{gathered} \quad (6)$$

where $Z = Z_R Z_R^{-1/2}$. This 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, whose (i, j)th entry is $\mathcal{C}(\|x_i - x_j\|)$. 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 $\{\mathcal{C}(\|x_i - x_1\|), \ldots, \mathcal{C}(\|x_i - x_n\|)\}$ 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 $\{\mathcal{C}(\|x_i - \kappa_1\|), \ldots, \mathcal{C}(\|x_i - \kappa_K\|)\}$. 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, $K \ll n$ knots usually are sufficient to give the desired flexibility to 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 (6), but with the $n \times n$ matrix Z_R replaced by an $n \times K$ matrix Z_K. This can be done directly, but with one important difference. The matrix Z_K has dimension $n \times K$ and defines a noninvertible 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. However, in the full-rank model, the matrix Z_R also can 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 define Ω_K to be the $K \times K$ matrix with (k, l)th entry $w_{k,l}^K = \mathcal{C}(\|\kappa_k - \kappa_l\|)$. 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

$$\begin{gathered} y = X\beta + Zb + \epsilon, \\ \mathrm{E}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0_{K\times 1} \\ 0_{n\times 1} \end{pmatrix}, \qquad \mathrm{cov}\begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_b^2 I_K & 0_{K\times n} \\ 0_{n\times K} & \sigma_\epsilon^2 I_n \end{pmatrix}, \end{gathered} \quad (7)$$

which is now a direct analog of (6).

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 computationally expensive 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.

3.3 Number and Location of Knots

The number of knots, K, in a low-rank smoother limits the maximum complexity of the model, whereas the smoothing parameter λ controls the fit to the data. Rupert et al. (2003) suggested $K = \max\{20, \min(n/4, 150)\}$ as a default. In the *Loa loa* study with $n = 275$ 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 modeling the logit of the prevalence surface for parasitologic 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 for obtaining 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 subset 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

(6), 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 that 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 to best approximate the subspace spanned by the data values y_i. A natural strategy to avoid redundancy is to maximize 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 Rupert et al. (2003). Other ways of choosing the knots have been suggested; for example, the R package of Ganguli and Wand (2005) uses 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 the 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 remaining to be defined in model (2) is the parameterization of parasitologic *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, and γ, where $C(x)$ are the (possibly transformed) observed covariates at location x. Using ideas and 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

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

where the shrinkage parameter σ_b^2 is estimated from the data. 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 also is 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 then can be used to obtain a correlated chain from the posterior distribution of $L_1(x_0)$. This has the nice practical property that to make inference about L_1 at any location or cluster of locations, we 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 0 and standard deviation 1,000. This choice was made by first running 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; as expected, centering the priors at 0 did not affect posterior inference. As discussed by Crainiceanu, Ruppert, and Wand (2005b), the prior distributions on the variance components should be treated carefully, because a poor choice of prior can seriously affect the smoothing function. To better understand this, we show how the choice of gamma prior $\tau_b = 1/\sigma_b^2$ may interact critically with the scaling of the variables. Consider the simple case of Gaussian smoothing described in Section 3.2. If $[\tau_b] \sim \mathrm{gamma}(A_b, B_b)$, where $\mathrm{gamma}(A, B)$ has mean A/B and variance A/B^2, then

$$[\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \mathbf{b}, \tau_\epsilon] \sim \mathrm{gamma}(A_b + K/2, B_b + \|\mathbf{b}\|^2/2). \quad (8)$$

The prior does not influence the posterior distribution of τ_b when both A_b and B_b are small compared with $K/2$ and $\|\mathbf{b}\|^2/2$. Because the number of knots is $K \ge 50$, it is safe to choose $A_b \le .001$. When $B_b \ll \|\mathbf{b}\|^2/2$, the posterior distribution is practically unaffected by the prior assumptions. When B_b increases compared with $\|\mathbf{b}\|^2/2$, the conditional distribution is increasingly affected by the prior assumptions. In our application, the posterior distribution of σ_b^2 was essentially supported by $[.2, \infty]$, and the value $B_b = .001$ did not influence the posterior inference. A similar discussion holds for σ_ϵ^2. Thus we used gamma(.001, .001) priors both for σ_b^2 and σ_ϵ^2.

Alternatives to gamma priors have been discussed by, for example, Natarajan and Kass (2000) and Gelman (2006). These have the advantage of requiring less care in the choice of the hyperparameters. However, we find that with reasonable care, the conjugate gamma priors can be used in practice. Nonetheless, exploration of other prior families for penalized splines would be well worthwhile, though beyond the scope of this article.

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, $S(x)$. In addition, $C(x)$ contains four environmental covariates. The first two covariates are the mean and standard deviation of the mean NDVI, a continuous measure of greenness of vegetation (Gates 1980), derived from repeated satellite scans during the year 2000. The third and fourth covariates are elevation and elevation truncated at 800 meters, which together define a linear spline with one knot at 800 meters; this choice was consistent with earlier analyses of spatial variation in *Loa loa* (e.g., Thompson et al. 2007) and with subject matter knowledge that the transmission vector does not breed at high elevations. All four covariates are calculated at

the pixel level, where one pixel is roughly equivalent to 1 km^2. Moreover, as is essential for spatially continuous prediction, all four are available throughout the study-region. We modeled the residual spatial variation $S(x)$ as a thin-plate spline with $k = 50$ knots placed according to the space-filling algorithm of Nychka and Saltzman (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 with 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 $P_1(x) > .2$ 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 1,000th 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 suggests that convergence to the target distribution occurs before 10,000 simulations.

Although the logit prevalence function, $L_1(x)$, had very good mixing properties, the mixing of the spline parameters, **b**, was not as good, likely due to the lack of information about them in the data. The best mixing properties were exhibited by α_0, α_1, and σ_ϵ. These are the parameters of the conditional distribution of $L_2(x)|L_1(x)$, which models the calibration relationship between the logit of the prevalence according to RAPLOA and parasitologic sampling methods. 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. These results indicate that the annual mean NDVI was positively associated with increased *Loa loa* prevalence, whereas increased elevation up to 800 meters was negatively associated with increased *Loa loa* prevalence. These findings agree with the findings of previous studies (e.g., Thompson et al. 2007).

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 $[3.5, 7] \times [7, 15]$, $[8, 16] \times [3, 7]$, and $[12, 16] \times [-6, -3]$.

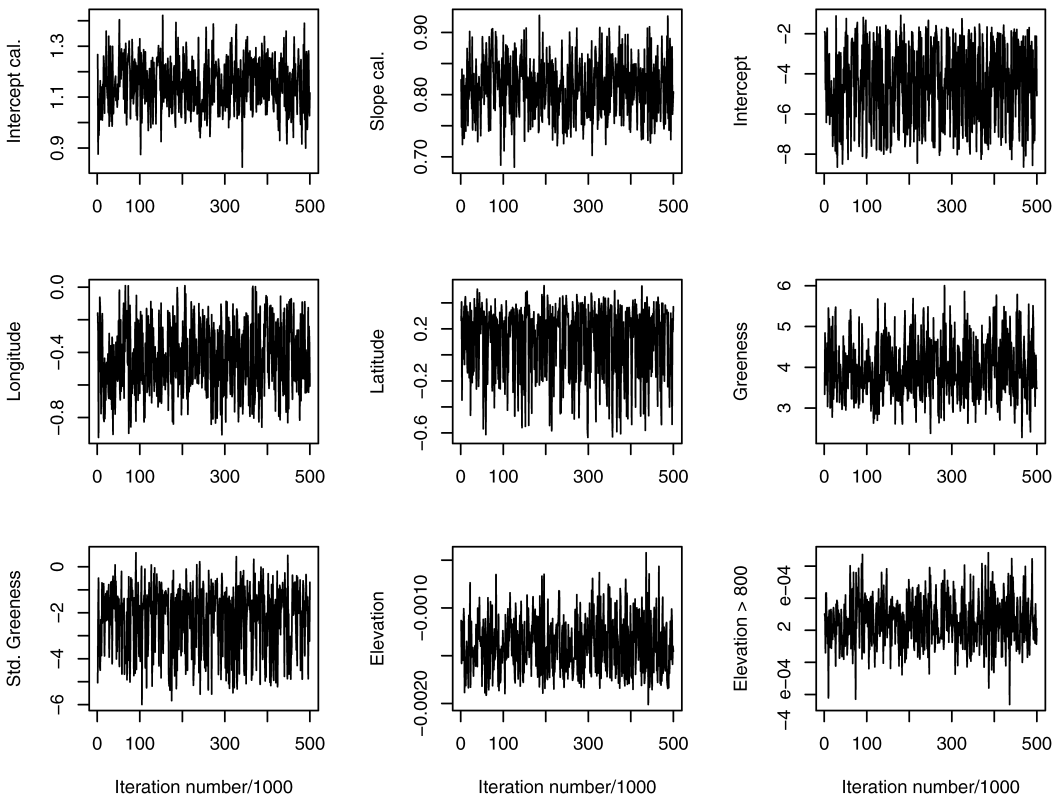

Figure 2. Posterior simulations from the joint distribution of some of the parameters of model (2). MCMC sampling was used to produce 500,000 correlated samples from the target distribution after an initial 20,000 burn-in simulations. For clarity, only every 1,000th sample is displayed.

Table 2. Posterior means and standard deviations for several parameters of interest based on Bayesian MCMC

Parameter	Mean	Standard deviation
Intercept cal.	1.15	.10
Slope cal.	.81	.05
Mean NDVI	3.96	.73
Std. NDVI	−2.22	1.48
Elevation ($\times 10^{-3}$)	−1.38	.29
Elevation > 800 ($\times 10^{-3}$)	− .06	.16

To better view the details of the map, the lower plot in the figure shows a zoom in on the rectangle $[8, 16] \times [3, 7]$. This graph plots the sampling locations color-coded according to the empirical prevalence estimate: black, $\widehat{P}(x) > .3$; red $.25 \le \widehat{P}(x) < .3$; magenta, $.20 \le \widehat{P}(x) < .25$; cyan, $.18 \le \widehat{P}(x) < .20$, and blue, $\widehat{P}(x) < .18$. 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.

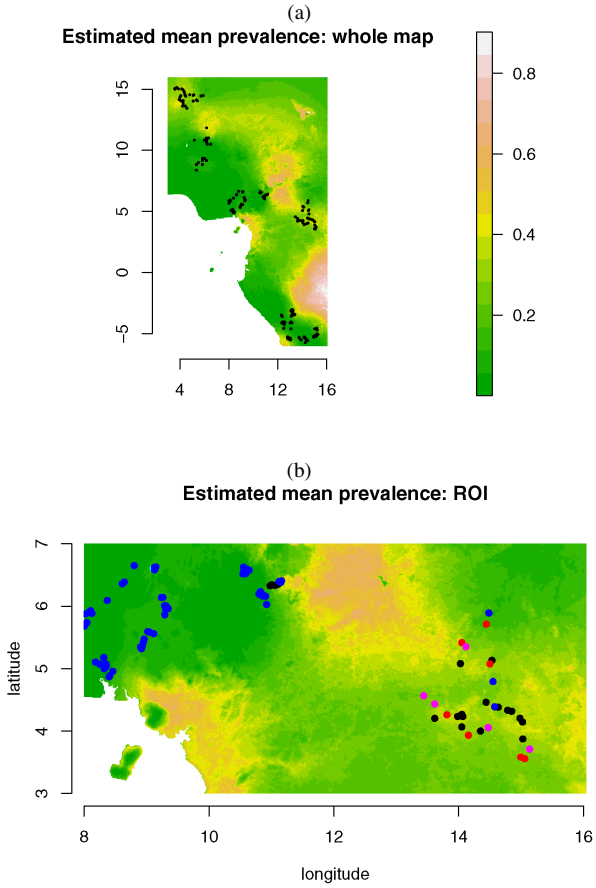

Figure 3. Posterior predictive surface for the *Loa loa* parasitologic prevalence based on the Bayesian analysis of the bivariate binomial spatial model (2). (a) The estimated prevalence within a large region containing all sampling locations with 275 sampling locations (black dots). (b) The estimated prevalence in a smaller region that contains 102 sampling locations, color-coded according to the observed (empirical) parasitology prevalence: black, $\widehat{P}(x) > .3$; red, $.25 \le \widehat{P}(x) < .3$; magenta, $.20 \le \widehat{P}(x) < .25$; cyan, $.18 \le \widehat{P}(x) < .20$, and blue, $\widehat{P}(x) < .18$.

5. A FAST METHOD FOR APPROXIMATE PREDICTIVE INFERENCE

Bayesian estimation has proven 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. But 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, leading to poor mixing of the Markov chains.
2. Long updating time. In our implementation, one update of all parameters takes roughly .1 seconds on a PC (3.6 GHz CPU, 3 GB RAM). This, combined with the need to run 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.
3. Limited testing. Long simulation times have restricted our ability to do extensive testing of our Bayesian methodology.

All these limitations require the expert supervision of a statistician. Such expertise typically is not available when new data become available, and predictive maps need to be updated in the field. 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 fast calibration model is to use a Normal approximation to the sampling distribution of empirical logits. Thus we define

$$\begin{cases} \widehat{L}_1(x) = \text{logit}\{\widehat{P}_1(x)\} \\ \widehat{P}_1(x) = Y_1(x)/N_1(x), \qquad x \in O, \end{cases} \tag{9}$$

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 parasitologic sampling was conducted. In the special cases when $Y_1(x) = 0$ and $Y_1(x) = N_1(x)$, we defined the empirical prevalence as $\widehat{P}_1(x) = 1/N_1(x)$ and $\widehat{P}_1(x) = 1 - 1/N_1(x)$.

To incorporate the prediction variability of the calibration model, we independently simulate C data sets for locations where only RAPLOA sampling is available from the calibration model

$$\widehat{L}_1^c(x) \sim \text{Normal}\big(\widehat{\alpha}_{0,1|2} + \widehat{\alpha}_{1,1|2}\widehat{L}_2(x), \widehat{\sigma}^2_{\epsilon,1|2}\big), \qquad c = 1, \ldots, C, x \in M, \tag{10}$$

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 $\alpha_{0,1|2}, \alpha_{1,1|2}$, and $\sigma^2_{\epsilon,1|2}$ 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

$$L^c(x) = \widehat{L}_1(x) I(x \in O) + \widehat{L}_1^c(x) I(x \in M), \tag{11}$$

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

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

where x^T is the location expressed as (longitude, latitude), $\mathrm{g}(x)$ is the greenness, $\mathrm{s}(x)$ is the standard deviation of greenness, $\mathrm{el}(x)$ is the elevation, and $\mathrm{el} > 800(x)$ 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 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:

$$\begin{aligned} L^c &= X\beta + Zb + \epsilon, \\ E\begin{bmatrix} b \\ \epsilon \end{bmatrix} &= \begin{bmatrix} 0_K \\ 0_n \end{bmatrix}, \qquad \operatorname{cov}\begin{bmatrix} b \\ \epsilon \end{bmatrix} = \begin{bmatrix} \sigma_b^2 I_K & 0_{K\times n} \\ 0_{n\times K} & \sigma_\epsilon^2 I_n \end{bmatrix}. \end{aligned} \tag{12}$$

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

Denote by $(\widehat{\beta}_c^T, \widehat{b}_c^T)^T$ the BLUP of $(\beta^T, b^T)^T$ from model (12) using the cth simulated outcome vector L^c. Suppose that we are interested in producing a predictive map at a particular location x_0. Write

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

and

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

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 this notation, 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 due to variability of the estimate around its mean and the second due to variability of the mean estimator around its mean. Thus we estimate $\operatorname{var}\{\widehat{L}^A(x_0)\}$ by

$$\begin{aligned} &\widehat{\operatorname{var}}\{\widehat{L}^A(x_0)\} \\ &\quad = \frac{1}{C} \sum_{c=1}^{C} \widehat{\operatorname{var}}\{\widehat{L}^c(x_0)\} + \frac{1}{C-1} \sum_{c=1}^{C} \{\widehat{L}^c(x_0) - \widehat{L}^A(x_0)\}^2. \end{aligned}$$

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

$$\widehat{\operatorname{var}}\{\widehat{L}^c(x_0)\} = \widehat{\sigma}_{c,\epsilon}^2 S_{x_0} \left(S^T S + \frac{\widehat{\sigma}_{c,\epsilon}^2}{\widehat{\sigma}_{c,b}^2} D \right) S_{x_0}^T, \tag{13}$$

where the subscript c indicates that parameter estimates are obtained from the cth 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; in our application, $p = 6$. The estimator (13), called the bias-adjusted variability estimate by Rupert et al. (2003, chap. 6), uses the marginal variance of the BLUP over the random effects.

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

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

For policy reasons, in our application $p_0 = .2$ corresponding to $\operatorname{logit}(p_0) = -1.386$.

In summary, the model described in this section approximates model (2) in the following three ways:

1. It replaces 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. It uses 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. It combines inferences from C different inferences corresponding to the C simulated outcome vectors.

An advantage of the methodology described in this section is that implementation is fast. Indeed, one data set can be fitted almost instantaneously, because it requires only 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 1,000 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. Whereas we preferred to implement our own mixed model software for added numerical control and inferential flexibility, the method described here can be easily implemented using standard mixed model software.

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. To do this, we fitted the mixed model (12), with the difference that all parasitologic 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. The results indicate that annual mean NDVI is positively associated, and elevation up to 800 meters is negatively associated, with prevalence. These results agree with previous scientific knowledge that NDVI and elevation are reasonable proxies for the density of day-biting *Chrysops* flies, which are the main agent of transmission of the disease to humans. The results also agree with the results of the Bayesian analysis in Section 4, although the magnitude of the association between NDVI and prevalence is higher using the frequentist analysis. This is likely due to the large uncertainty in the smoothing parameter estimation.

The estimated number of degrees of freedom (df) of the regression was df = 36.4, down from a maximum of 57 df allowed by the model. The df were partitioned into 4 for fixed effects; 3 for intercept, longitude, and latitude; and 29.4 for the random coefficient component. An important issue is whether simpler models that include only covariates can capture the complex stochastic nature of the spatial data and have good predictive power. Table 4 shows the results of likelihood ratio tests for models of the logit empirical parasitologic prevalence. All models contain longitude and latitude as linear predictors and covariates added sequentially in the following order: mean NDVI, and elevation, standard deviation of NDVI, and elevation truncated at 800 meters. The column labelled "Δ-LRT" provides twice the log-likelihood difference for models with one added covariate; for example, the value of 11.83 represents the increase in twice the log-likelihood by adding elevation to a linear model incorporating longitude, latitude, and mean NDVI. The last row, labelled "+thin-plate spline," corresponds to the model in which we add a thin-plate spline with 50 knots to capture the residual spatial correlation. The last test is nonstandard because it is designed to test linear effects versus a general nonparametric alternative. It can be shown that this is equivalent to testing

$$H_0 : \sigma_b^2 = 0 \quad \text{versus} \quad H_A : \sigma_b^2 > 0,$$

Table 3. Point estimator and standard deviation for several parameters of interest based on mixed model inference

Parameter	Estimator	95% SD
Mean NDVI	7.71	1.94
Standard NDVI	1.28	4.56
Elevation ($\times 10^3$)	−1.59	.66
Elevation > 800 ($\times 10^3$)	−.35	.55

Table 4. Likelihood ratio tests for an increasingly complex sequence of models for the logit empirical parasitologic prevalence

Model	Δ-LRT	p value
+Mean NDVI	26.26	<.001
+Elevation	11.83	<.001
+Standard NDVI	4.37	.037
+Elevation > 800	3.73	.053
+Thin-plate spline	127.47	<.001

NOTE: All models contain longitude and latitude as linear predictors and covariates added sequentially in the following order: mean NDVI, elevation, standard deviation of NDVI, and elevation truncated at 800 meters. The last model is obtained by adding a low rank thin-plate spline.

where σ_b^2 is the variance of the random coefficients b in the model (12). Here we use a restricted likelihood ratio test instead of the likelihood ratio test. 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

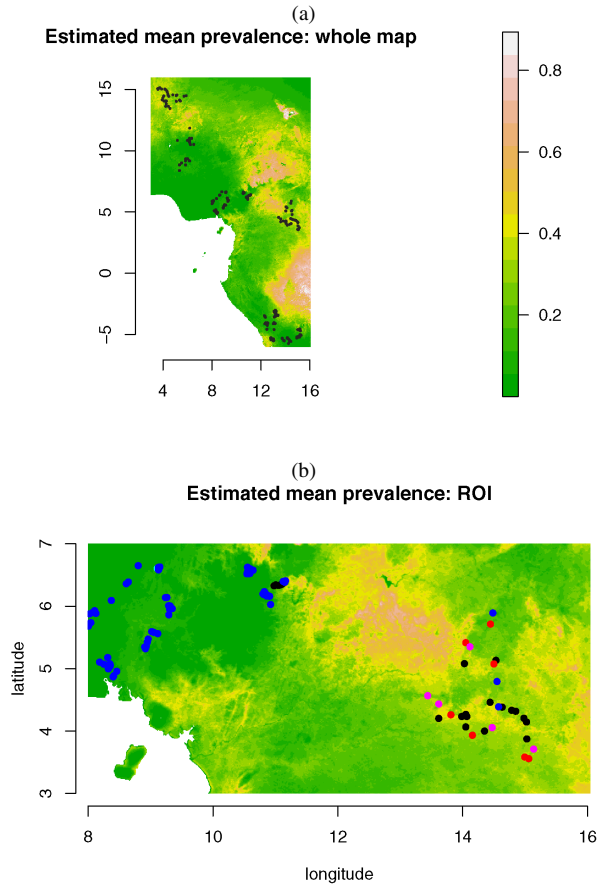

Figure 4. Estimated parasitological prevalence surface for the *Loa loa* parasitological prevalence based on the calibration and pooling method described in Section 5.1 of the fast calibration model (2). Because parasitologic data are observed at all locations, a simpler univariate model was used. (a) The results as they are extrapolated to a very large region containing all sampling locations. (b) The inference in a smaller region that contains sampling locations and contains 102 sampling locations color-coded according to the convention in Figure 3.

Wand (2005a) for testing H_0 and obtained a p value $<.001$. 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. Although covariate parameter estimates are different from the Bayesian analysis, it is reassuring to observe that the prevalence prediction map is quite similar to the one shown in Figure 3.

6. SIMULATION STUDY

We conducted three simulation studies to compare the performance of Bayesian inference for the binomial spatial model (2) with frequentist analysis of the fast calibration model described in Section 5.1. The first simulation study uses the same 275 sampling locations as those from the parasitology/RAPLOA sampling locations from west Africa. The underlying logit parasitologic 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) obtained from the Bayesian analysis of model (2), and X and Z are the design matrices described in Section 5.1. Parasitology counts were then simulated independently from

$$\begin{cases} Y_1(x) \sim \text{Binomial}\{N_1(x), P_1(x)\} \\ P_1(x) = \exp\{L_1(x)\}/\big[1 + \exp\{L_1(x)\}\big], \end{cases}$$

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 α_0^*, α_1^*, and σ^{*2}_ϵ are the posterior means of the calibration model parameters from model (2). RAPLOA counts are simulated from the model

$$\begin{cases} Y_2(x) \sim \text{Binomial}\{N_2(x), P_2(x)\} \\ P_2(x) = \exp\{L_2(x)\}/\big[1 + \exp\{L_2(x)\}\big]. \end{cases}$$

To compare results, we focused on the region of interest (ROI) situated between 8 and 16 degrees longitude and 3 and 7 degrees latitude, which contains 102 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 there were no missing parasitology observations in the first simulation study, 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 mean squared error (MSE) for each simulated data set. More precisely, for a particular region R, simulated data set, and fitting method, we calculated

$$MSE = \frac{1}{|R|} \sum_{x \in R} \{\widehat{L}_1(x) - L_1(x)\}^2, \tag{14}$$

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 only 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, then it would not matter in practice which method were used. In this case it would make sense to use the calibration method, because it is much faster and provides a more robust software platform.

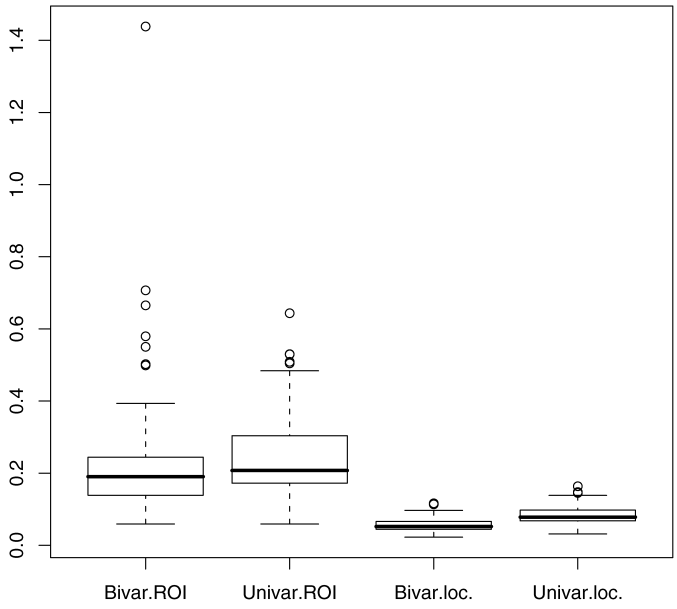

Figure 5. Mean squared error comparison of the Bayesian analysis calculated over the ROI (Bivar.ROI) and at the sampling location (Bivar.loc.) with the fast calibration method calculated over the ROI (Univar.ROI) and at the sampling locations (Univar.loc.). Results are calculated over 100 simulated data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

But 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 $[P_1(x) \geq .2]$ and truly negative $[P_1(x) < .2]$ locations also are 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 $p_0 = .2$ threshold at location x can be estimated by

$$\widehat{E}(x, p_0) = 1 - \Phi\left\{\frac{\text{logit}(p_0) - \widehat{L}_1(x)}{\sqrt{\widehat{\text{var}}\{\widehat{L}_1(x)\}}}\right\}. \tag{15}$$

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

$$\text{Sensitivity}(R, T) = \frac{1}{R} \sum_{x \in R} I\{\widehat{E}(x, p_0) > T, P_1(x) \geq .2\}, \tag{16}$$

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

$$\text{Specificity}(R, T) = \frac{1}{R}\sum_{x \in R} I\{\widehat{E}(x, p_0) < T, P_1(x) < .2\}, \quad (17)$$

which 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 of the procedures based on 100 simulated data sets. Results corresponding to the Bayesian method are shown in the top two panels, whereas those corresponding to the calibration model are shown in the bottom two panels. These functions are specific to the ROI and are obtained using an equally spaced grid for threshold values between [.025, .975].

The trade-off between sensitivity and specificity is clearly shown in Figure 6, where sensitivity is a decreasing function while specificity is an increasing function of the probability threshold. We could of course have perfect sensitivity by setting $T = 0$ or close to 0, but at the cost of abysmal specificity results. Indeed, for low values of T, both procedures correctly identify truly negative sampling locations with probability $<.5$ 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 likely due to lack of information at locations that are distant from all sampling locations.

Figure 6 is informative in its own right, but could not be easily used for choosing a 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 $T = .7$ we would obtain roughly .8 average sensitivity with both methods and .9 average specificity for the specific ROI. In this specific ROI, under our model there were 245,846 truly positive locations and 123,455 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 .89 average sensitivity for the probability threshold $T = .975$. The calibration model behaves reasonably well for thresholds between .7 and .8 but exhibits a rapid decrease in sensitivity for higher thresholds. In our simulation, there were 38 truly positive locations and 64 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 the

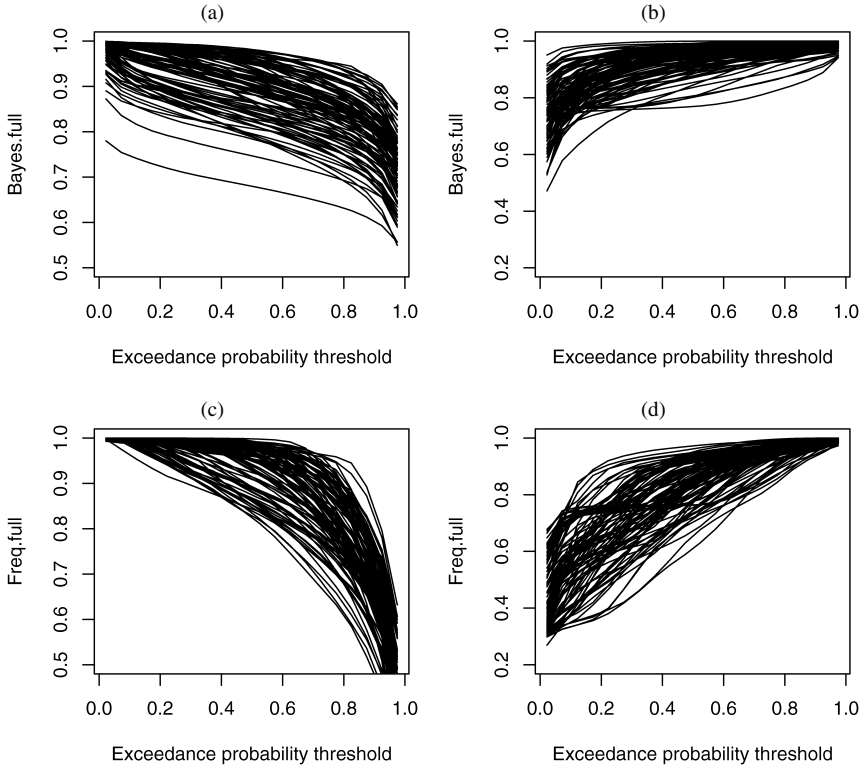

Figure 6. Sensitivity [(a) and (b)] and specificity [(b) and (d)] functions calculated over the ROI for each of the 100 simulated data sets as a function of the probability threshold. The Bayesian analysis is represented by (a) and (b), and the fast calibration method by (b) and (d). All data sets contained 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.

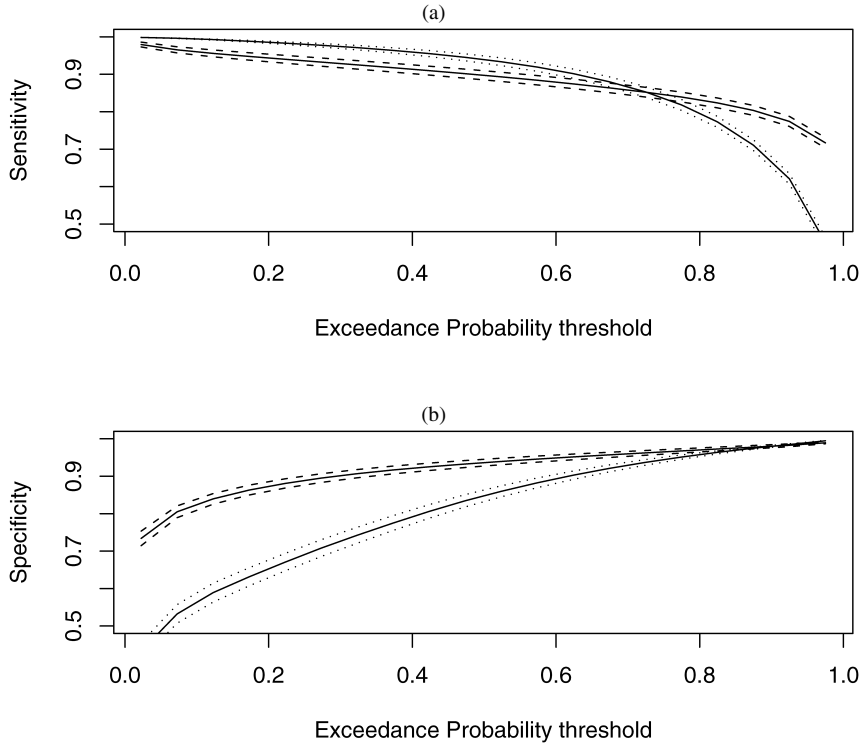

Figure 7. Average sensitivity (a) and specificity (b) functions with 95% confidence intervals based on sensitivity and specificity results shown in Figure 6. In particular, for every exceedance probability threshold, the mean and the standard error is calculated based on the 100 results corresponding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

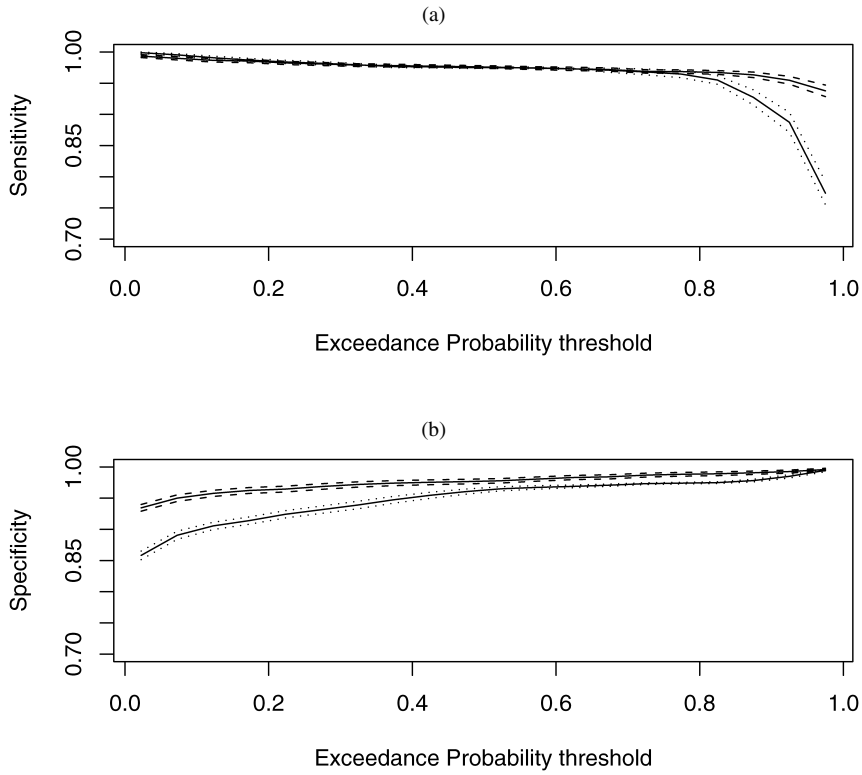

Figure 8. The same type of results as in Figure 7 [(a) sensitivity; (b) specificity], with the difference that results correspond to sampling locations only. The bivariate binomial model results are presented as a solid line, with dashed lines for confidence intervals. The fast calibration model results are presented as a solid line, with dotted lines for confidence intervals.

parasitologic sampling. This was achieved by following the simulation recipe described earlier for the first simulation, except that at each simulation we did not use parasitology count data simulated at some specific locations in the ROI, 50 in the second study and all 102 in the third study. Instead, these data were treated as missing and were simulated from their joint posterior distribution as obtained from the Bayesian analysis of the bivariate binomial model. For the frequentist method, we used the calibration and pooling algorithm described in Section 5.1.

Figure 9 presents the same type of results as in Figure 5. It compares the MSE for the Bayesian methodology with that of the fast calibration method for the case in which a number of locations are not parasitologically surveyed. The four leftmost boxplots correspond to the second simulation study, and the four rightmost boxplots correspond to the third simulation study. Note that the Bayesian methodology produces larger MSEs than its fast calibration counterpart over the entire ROI and slightly smaller MSEs over just the sampling locations. This may, be due in part to the slow mixing of the Markov chains combined with the inherent computational limitations of our simulation study. We chose 20,000/20,000 burn-in/simulation samples because this was sufficient for the case without missing parasitology data. Moreover, increasing the number of simulations to obtain reasonable results would result in unreasonably 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 in which parasitology sampling is actually conducted at sampling locations. Comparing results from Figure 9 with those in Figure 5 demonstrates the much larger MSE in the case in which parasitology is missing at some locations (note the different scales). Indeed, average MSE increases by roughly 30%, from .25 to .32 for the ROI and from .08 to .11 for the sampling locations, using the frequentist calibration methodology.

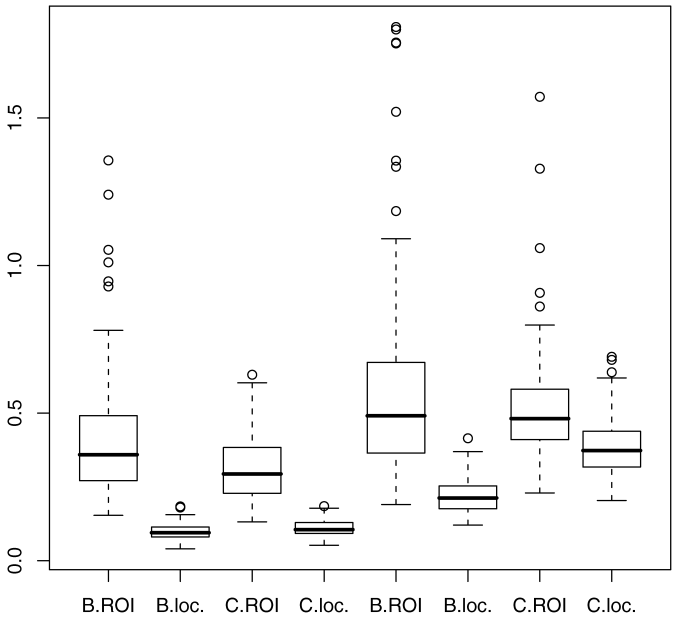

Figure 9. Mean squared error comparison of the Bayesian analysis calculated over the ROI (B.ROI) and at the sampling location (B.loc) with the fast calibration method calculated over the ROI (C.ROI) and at the sampling locations (C.loc). The results reported in the four boxplots on the left correspond to the second simulation study described in Section 6. The four boxplots on the right correspond to the third simulation study described in Section 6. All data sets contained RAPLOA sampling at every location. Parasitology samples were simulated at all but 50 specific locations in the ROI for the second study and at all but 102 locations in the third study. Each study used 100 simulated data sets.

Whereas the results in Section 2 seem to indicate good calibration between the parasitologic and RAPLOA sampling, using only RAPLOA sampling may result in serious loss of efficiency when estimating the prevalence function. As expected, the loss of information also is 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 corresponding to full data analysis in Figure 6. Similar losses were observed for the sampling locations.

Figure 11 presents the same type of results when parasitology was not conducted at all 102 locations of the ROI (third simulation study). The Bayesian methodology now delivers a better performance, both at sampling locations and over the entire region, but its performance is affected by the increased number of locations without parasitologic sampling. Moreover, the average MSE increases by roughly 63% (from .32 to .52) for the ROI and by 255% (from .11 to .39) for the sampling locations using the fast 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` (a RAPLOA analysis tool).

We tested the robustness of the code by simulations. We first generated a fairly smooth Gaussian random field over our possible study area and computed simulated parasitology and RAPLOA prevalences at a number of locations within the study region. We then ran the exceedence computation using a subset of the points as calibration data (i.e., using both parasitology and RAPLOA data) and the rest as new RAPLOA data. We repeated this process many times with variations in the numbers and positions of locations. This type of testing is intended to reveal shortcomings in the code's ability to cope with anomalous data configurations, for example, having one or no locations in the calibration or survey data, without crashing.

To allow 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 a freely available multiplatform (Windows, Linux, Solaris) package that interacts with R.

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

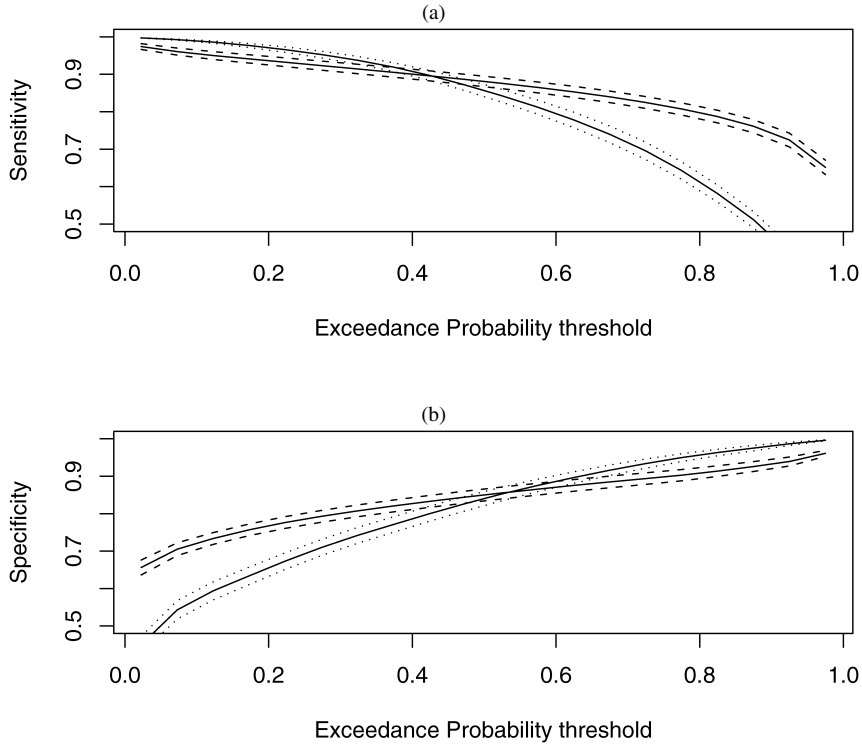

Figure 10. Average sensitivity (a) and specificity (b) functions with 95% confidence intervals based on 100 simulated data sets. Parasitology samples were simulated at all but 50 specific locations in the ROI. For every exceedance probability threshold, the mean and standard error are calculated based 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 fast calibration model results are presented as a solid line, with dotted lines for confidence intervals. The results correspond to the ROI.

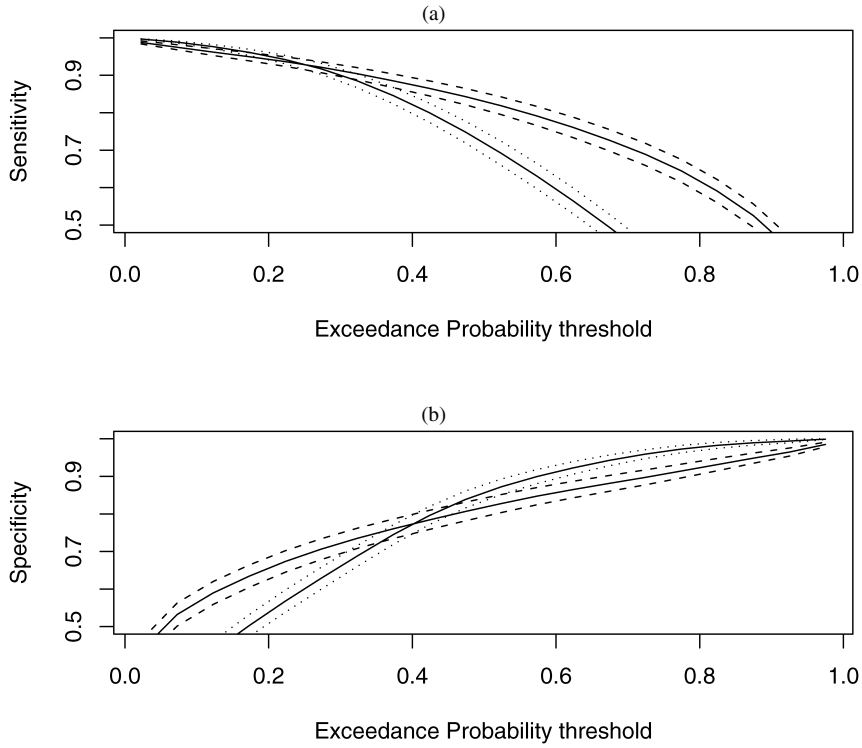

Figure 11. Average sensitivity (a) and specificity (b) functions with 95% confidence intervals based on 100 simulated data sets. Parasitology samples were simulated at all locations with the exception of the 102 locations within the ROI. For every exceedance probability threshold, the mean and the standard 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 fast calibration model results are presented as a solid line, with dotted lines for confidence intervals. The results correspond to the ROI.

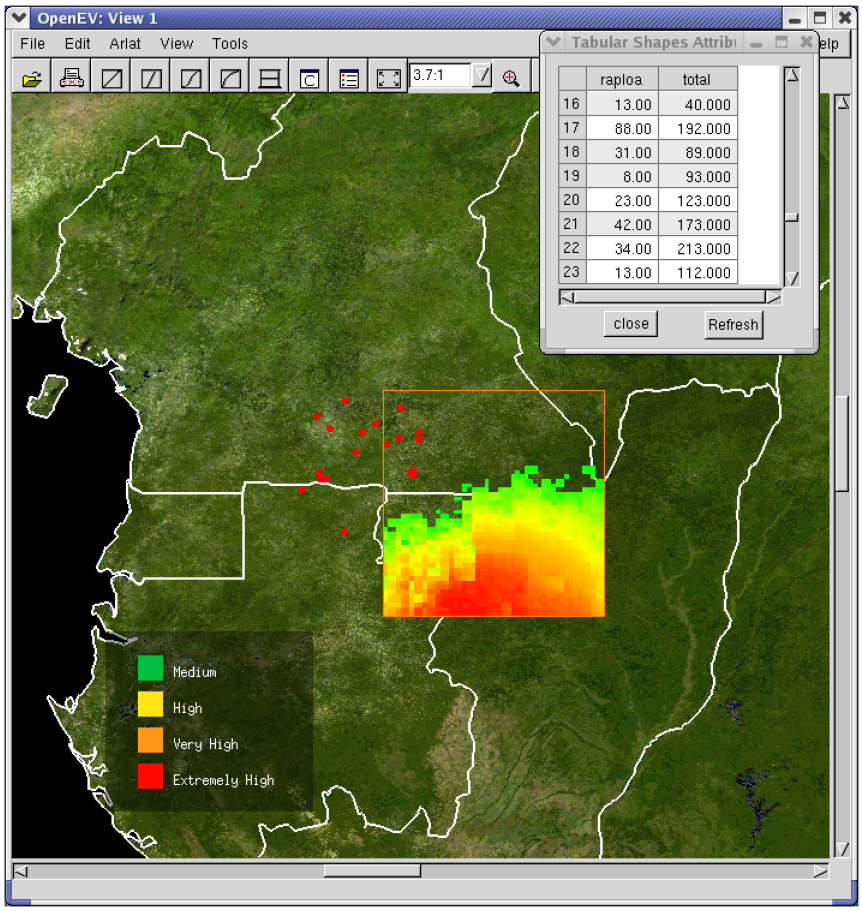

Figure 12. Screen shot of our software showing sampling locations (in red) and corresponding exceedance probabilities in a small area containing the sample locations. By default the grid is colored such that exceedance probabilities below .7 are invisible, and from .7 to 1.0 are colored from green through yellow and orange up to red. The table in the upper right hand corner of the screen allows simple data updating, whereas the `arlat` software can rapidly produce exceedance probabilities based on the methodology described in Section 5.

To communicate between Python and R, we used the Rserve facility. This runs R in the background as a server, and client programs connect to it through 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 had 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 data set 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 be displayed with other map data, such as country or region boundaries, village maps, roads and topography. The user then starts the Arlat dialog from a drop-down menu, from which he or she can select the map layer containing the survey data, define a rectangle over which to produce the predicted exceedance probability, and choose the resolution of the output grid. On clicking the "ok" button, OpenEV uses the Python-Rserve code to call the `arlat` package in R which computes the exceedence probability over the specified area. This is then loaded into OpenEV as a new raster grid layer.

Figure 12 provides a screen shot of our software showing sampling locations (in red) and corresponding exceedance probabilities over a small region containing the sample locations. By default, the grid is colored such that exceedance probabilities below .7 are invisible, whereas probabilities from .7 to 1.0 are colored from green through yellow and orange up to red. The table in the upper right-hand corner of the screen allows simple data updating, whereas the `arlat` software can rapidly produce exceedance probabilities based on the methodology described in Section 5. The `arlat` software is available on request from the third author.

8. DISCUSSION

In this article we have described 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 predetermined policy intervention threshold. In our data, empirical prevalence in each sampled village is assessed by two methods: a traditional, parasitologic method based on the microscopic examination of blood smears

and a rapid questionnaire-based method, RAPLOA (Takoungang et al. 2002). In our data, both methods were used in all sampled villages, and the data were collected primarily to establish a calibration relationship that, to a good approximation, holds over a wide geographical area of equatorial 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. Thus, how best to combine the existing and future data is a problem in bivariate spatial modeling. 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, Moyeed, and Tawn (1998), in which an unobserved stationary Gaussian process $S(x)$ is added to the linear predictor in a generalized linear model. One limitation of the methods used by Diggle, Moyeed, and Tawn (1998) is that they are, at the time of this writing, computationally impractical for our data, for which we need to make predictions at several million 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 (e.g., Rue and Tjelmeland 2002 or Besag and Mondal 2005).

We believe that the present article also is 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 modeling $S_1(x)$ marginally, and $S_2(x)$ conditionally, on $S_1(x)$.

Finally, we have taken into account the practical problems of implementing sophisticated spatial statistical analyses, especially those that rely on careful tuning of a MCMC algorithm, routinely under field conditions by comparing the formal Bayesian analysis of our bivariate model with a much simpler (albeit approximate) analysis that 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 generalized linear modeling approach. The extent to which in general generalized linear geostatistical modeling outperforms transformed Gaussian geostatistical modeling 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 also have shown that the calibration relationship between parasitologic 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 thus the data do not give an authoritative solution to the prediction problem. This same point was emphasized by Thompson et al. (2007), who argued that an important feature of probabilistic prediction in this context was that it can identify areas in which 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.

[Received March 2006. Revised January 2007.]

REFERENCES

Akaike, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255–265.

Alleman, M. L., Twum-Danso, N. A. Y., and Thylefors, B. I. (2006), "The Mectizan Donation Programme—Highlights From 2005," *Filaria Journal*, 5: 11.

Besag, J., and Mondal, D. (2005), "First-Order Intrinsic Autoregressions and the de Wijs Process," *Biometrika*, 92, 909–920.

Boussinesq, M., Gardon, J., Gardon-Wendel, N., Kamgno, J., Ngoumou, P., and Chippaux, J. P. (1998), "Three Probable Cases of *Loa loa* Encephalopathy Following Ivermectin Treatment for Onchocerciasis," *American Journal of Tropical Medicine and Hygiene*, 58, 461–469.

Boussinesq, M., Gardon, J., Kamgno, J., Pion, S. D. S., Gardon-Wendel, N., and Chippaux, J. P. (2001), "Relationships Between the Prevalence and Intensity of *Loa loa* Infection in the Central Province of Cameroon," *Annals of Tropical Medicine and Parasitology*, 95, 495–507.

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

Crainiceanu, C. M., and Ruppert, D. (2004), "Likelihood Ratio Tests in Linear Mixed Models With One Variance Component," *Journal of the Royal Statistical Society*, Ser. B, 66, 165–185.

Crainiceanu, C. M., Ruppert, D., Claeskens, G., and Wand, M. P. (2005a), "Likelihood Ratio Tests in Linear Mixed Models With One Variance Component," *Biometrika*, 92, 91–103.

Crainiceanu, C. M., Ruppert, D., and Wand, M. P. (2005b), "Bayesian Analysis for Penalized Spline Regression Using Winbugs," *Journal of Statistical Software*, 14 (14).

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403.

Cressie, N. A. C. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley.

Diggle, P. J., Moyeed, R. A., and Tawn, J. A. (1998), "Model-Based Geostatistics" (with discussion), *Journal of the Royal Statistical Society*, Ser. C, 47, 299–350.

Diggle, P. J., Thompson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., et al. (2007), "Spatial Modelling and Prediction of *Loa loa* Risk: Decision Making Under Uncertainty," *Annals of Tropical Medicine and Parasitology*, 101, 499–509.

Ganguli, B., and Wand, M. P. (2005), *SemiPar 1.0 Users' Manual*, available at *http://www.maths.unsw.edu.au/wand/papers.htm*.

Gates, D. M. (1980), *Biophysical Ecology*, New York: Springer.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary Multivariate Modeling Through Spatially Varying Coregionalization" (with discussion), *Test*, 13, 263–312.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman & Hall.

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Applied Statistics*, 52, 1–18.

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

Mallows, C. L. (1973), "Some Comments on c_p," *Technometrics*, 15, 661–675.

Natarajan, R., and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237.

Nychka, D. W. (2000), "Spatial Process Estimates as Smoothers," in *Smoothing and Regression*, ed. M. Schimek, Heidelberg: Springer-Verlag, pp. 393–424.

——— (2004), *Fields: Tools for Spatial Data*, available at *http://www.cgd.ucar.edu/stats/Software/Fields*.

Nychka, D. W., and Saltzman, N. (1998), "Design of Air Quality Monitoring Networks," in *Case Studies in Environmental Statistics*, eds. D. Nychka, W. W. Piegorsch, and L. H. Cox, New York: Springer-Verlag, pp. 51–76.

R Development Core Team (2004), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, available at *http://www.R-project.org*.

Rue, H., and Tjelmeland, H. (2002), "Fitting Gaussian Markov Random Fields to Gaussian Fields," *Scandinavian Journal of Statistics*, 29, 31–50.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Seketeli, A. et al. (2002), "The Achievements and Challenges of the African Programme for Onchocerciasis Control (APOC)," *Annals of Tropical Medicine and Parasitology*, 96, 15–28.

Self, S. G., and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

Takoungang, I. et al. (2002), "Rapid Assessment Method for Prevalence and Intensity of *L. loa* Infection," *Bulletin of the World Health Organisation*, 80, 852–858.

van Rossum, G., and Drake, F. L. Jr. (Eds.) (2003), *The Python Language Reference Manual*, *http://docs.python.org/ref/ref.html*, Network Theory Ltd.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

Walter, G., Warmerdam, F., and Farris-Manning, P. (2002), "An Open Source Tool for Geospatial Image Exploitation," in *Proceedings of the IGARSS 2002 Conference*, Vol. 6, pp. 3522–3524.

World Health Organization (1995), *Onchocerciasis and Its Control: Report of a WHO Expert Committee on Onchocerciasis Control*, World Health Organisation Technical Report Series 852, Geneva: Author.

Wood, S. N. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society*, Ser. B, 65, 95–114.

Comment

Christopher J. PACIOREK

Crainiceanu, Diggle, and Rowlingson (henceforth CDR) provide a thorough analysis of a Bayesian spatial model and simplified spatial mixed model to model *Loa loa* parasite prevalence and develop a prediction methodology for field use to guide public health practice. The study nicely links hierarchical Bayesian modeling to practical implementation using the mixed model with an R library and a user front end for use in the field. As a Bayesian, but also an applied statistician who harbors some suspicions about the transparency of and marginal benefit from complicated models, I liked the comparison of the full model and the calibration model, as well as the careful discussion of development and assessment of the Bayesian model. My comments focus on the clustering of the data and the role of spatial statistics in the analysis (Sec. 1), the comparison of Bayesian and fast calibration models (Sec. 2), MCMC mixing in spatial models (Sec. 3), and variance component priors (Sec. 4).

1. SAMPLING AND THE SPATIAL STATISTICIAN

As CDR note, the spatial distribution of the observed villages is highly clumped. The model extrapolates to make predictions far from the observations, a major concern in a spatial analysis. The clumping has potentially serious implications for fitting and predicting from the model, with the implications depending on the nature of future data collection and on the relative importance of the covariates and the spatial term to the predictive ability of the model.

In light of the simulation results that show importance differences in predictive ability depending on location and a serious loss in efficiency when parasitology data are missing, could CDR tell us more specifically how sampling may proceed in the future, and thus to what data will future model fitting and prediction have access? Will the calibration part of the model rely on the data provided in the article, or will additional parasitology sampling be done? Will RAPLOA sampling be done in many or most of the villages in which a prediction is desired, or at least in nearby villages, or will there be large spatial gaps, as in the current data set? This will help provide insight into whether the simulation results over the ROI or those at the locations are most relevant for understanding the success of the method in the field in the future. One would hope that the parasitology data would be sufficient and generalizable to new locations such that the results in the first simulation would pertain, rather than the results of the second and third simulations that show the loss in efficiency.

As a reader, I am faced with several seemingly contradictory pieces of information in understanding the relative importance of the spatial term and the covariates. First, the spatial term uses 29.4 df and is highly significant with a huge change in the likelihood when it is left out of the model, which suggests its importance in the model, as does the loss of predictive power in going from the training locations to the ROI, indicating the related inability of the model to extrapolate outside of the clusters. On top of this, whereas mean NDVI and elevation contribute something to the likelihood, the effect on the likelihood is not nearly as large as that from the spatial term. And yet the ROC results indicate that the model still seems to do relatively well throughout the ROI, so the covariates presumably have sufficient influence so as to provide some predictive inference, as seen in the simulations. Furthermore, I am surprised that $L_1(x)$ mixes well despite one of its components not mixing. Could it be that $S(x) = \mathbf{Z}(\mathbf{x})^T\mathbf{b}$ may not explain much of the variability in $L_1(x)$ relative to the covariates, and that the poor mixing of $\mathbf{b}$ is hidden within the good mixing of the other components of $L_1(x)$ because the magnitude of $S(x)$ is small relative to that of $L_1(x)$? The high predicted surface in locations with no data (e.g., 13E, 5.5N) also indicates that covariates drive the surface. A spatial smooth usually will not be larger than the observations themselves, because spatial smoothers generally can

Christopher J. Paciorek is Assistant Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (E-mail: *paciorek@hsph.harvard.edu*).

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Applications and Case Studies
DOI 10.1198/016214507000001418