Geostatistical methods for disease mapping and visualization using data from spatio-temporally referenced prevalence surveys

Emanuele Giorgi[1], Peter J. Diggle[1], Robert W. Snow[2,3], Abdisalan M. Noor[2]

[1] *Lancaster Medical School, Lancaster University, Lancaster, UK*
[2] *Population and Health Theme, Kenya Medical Research Institute - Wellcome Trust Research Programme, Nairobi, Kenya*
[3] *Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK*

December 14, 2017

Abstract

In this paper we set out general principles and develop geostatistical methods for the analysis of data from spatio-temporally referenced prevalence surveys. Our objective is to provide a tutorial guide that can be used in order to identify parsimonious geostatistical models for prevalence mapping. A general variogram-based Monte Carlo procedure is proposed to check the validity of the modelling assumptions. We describe and contrast likelihood-based and Bayesian methods of inference, showing how to account for parameter uncertainty under each of the two paradigms. We also describe extensions of the standard model for disease prevalence that can be used when stationarity of the spatio-temporal covariance function is not supported by the data. We discuss how to define predictive targets and argue that exceedance probabilities provide one of the most effective ways to convey uncertainty in prevalence estimates. We describe statistical software for the visualization of spatio-temporal predictive summaries of prevalence through interactive animations. Finally, we illustrate an application to historical malaria prevalence data from 1334 surveys conducted in Senegal between 1905 and 2014.

Keywords: disease mapping; Gaussian processes; geostatistics; parameter uncertainty; parsimony; prevalence; spatio-temporal models.

1 Introduction

Model-based geostatistics (MBG) (Diggle et al., 1998) is a sub-branch of spatial statistics that provides methods for inference on a continuous surface using spatially discrete, noisy

data. MBG is increasingly being used in disease mapping applications (e.g. Hay et al. (2009); Gething et al. (2012); Diggle & Giorgi (2016)), with a particular focus on low-resource settings where disease registries are geographically incomplete or non-existent.

We consider data obtained by sampling from a set of potential locations within an area of interest A, repeatedly at each of a sequence of times $t_1, \ldots, t_N$. At each sampled location, individuals are then tested for the disease under investigation. The data-format can be formally expressed as

$$\mathcal{D} = \{(x_{ij}, t_i, y_{ij}, n_{ij}) : x_{ij} \in A, j = 1, \ldots, m_i, i = 1, \ldots, N\}, \tag{1}$$

where x_{ij} is the location of the jth of m_i sampling units at time t_i, n_{ij} is the number of tested individuals at x_{ij} and y_{ij} is the number of positively identified cases.

The methodology described in this paper can be equally applied to longitudinal or repeated cross-sectional designs. For this reason, we re-write (1) as

$$\mathcal{D} = \{(x_i, t_i, n_i, y_i) : x_i \in A, i = 1, \ldots, N^*\},$$

where $N^* = \sum_{i=1}^{N} m_i$ and either or both of the x_i and t_i may include replicated values.

An essential feature of the class of problems that we are addressing in this paper is that the locations x_i are a discrete set of sampled points within a spatially continuous region of interest. Another possible format for prevalence data, which we do not consider in the present study, is a small-area data-set. In this case, locations x_i are reference locations associated with a partition of A into n sub-regions. Disease registries in relatively well developed countries often use this format, both for administrative convenience and, in associated publications such as health atlases, to preserve individual confidentiality; see, for example, (López-Abente et al., 2007) or (Hansell et al., 2014). In low-resource settings, this is also often the format of data from demographic surveillance systems, such as Demographic and Health Surveys (`dhsprogram.com`), which are nationally representative surveys conducted about every five years to collect information on population, health and nutrition indicators; see, for example, Mercer et al. (2015) for an analysis of data of this kind.

A geostatistical model for data of the kind specified by (1) is that, conditionally on a spatio-temporal process $S(x, t)$ and unstructured random effects $Z(x, t)$, the outcomes Y are mutually independent binomial distributions with number of trials n and probability of being a case $p(x, t)$. Using the conventional choice of a logistic link function, although other choices are also available, we can then write

$$\log\left\{\frac{p(x_i, t_i)}{1 - p(x_i, t_i)}\right\} = d(x_i, t_i)^\top \beta + S(x_i, t_i) + Z(x_i, t_i), \tag{2}$$

where $d(x_i, t_i)$ is a vector of spatio-temporally referenced explanatory variables with associated regression coefficients β. The spatio-temporal random effects $S(x_i, t_i)$ can be interpreted as the cumulative effect of unmeasured spatio-temporal risk factors. These are modelled as a Gaussian process with stationary variance σ^2 and correlation function

$$\mathrm{corr}\{S(x, t), S(x', t')\} = \rho(x, x', t, t'; \theta), \tag{3}$$

where θ is a vector of parameters that regulate the scale of the spatial and temporal correlation, the strength of space-time interaction and the smoothness of the process $S(x,t)$. Finally, the unstructured random effects $Z(x_i,t_i)$ are assumed to be independent zero-mean Gaussian variables with variance τ^2, to account for extra-binomial variation within a sampling location. In particular applications, this can represent non-spatial random variation, such as genetic or behavioural variation between co-located individuals, spatial variation on a scale smaller than the minimum observed distance between sampled locations, or a combination of the two.

The model (2) can be used to address two related, but different, research questions.

Estimation: what are the risk factors associated with disease prevalence? In this case the focus of scientific interest is on the regression coefficients β.

Prediction: how to interpolate the spatio-temporal pattern of disease prevalence? The scientific focus is, in this case, on $d(x,t)^\top\beta + S(x,t)$ at both sampled and unsampled locations $\mathcal{X}$ and times $\mathcal{T}$. In some cases, the scientific interest may be more narrowly focused on $S(x,t)$, in order to identify areas of relatively low and high spatio-temporal variation that is not explained by the available explanatory variables.

Modelling of the residual spatio-temporal correlation through $S(x,t)$ is crucial in both cases: in the first case, in order to deliver valid inferences on the regression relationships by accurately quantifying the uncertainty in the estimate of β (Thomson et al., 1999); in the second case, to borrow strength of information across observations y_i by exploiting their spatial and temporal correlation.

The use of explanatory variables $d(x,t)$ can also be beneficial in two ways: a simpler model for $S(x,t)$ can be formulated by explaining part of the spatio-temporal variation in prevalence through $d(x,t)$; more precise spatio-temporal predictions between data-locations also result from exploiting the association between disease prevalence and $d(x,t)$.

Here, we focus our attention on spatio-temporal prediction of disease prevalence. Our aim is to provide a general framework that can be used as a tutorial guide to address some of the statistical issues common to any spatio-temporal analysis of data from prevalence surveys, especially when sampling is carried out over a large geographical area or time period, or both. More specifically, we provide answers to each of the following research questions. How can we specify a parsimonious spatio-temporal model while taking account of the main features of the underlying process? How can we extend model (2) in order to account for non-stationary patterns of prevalence? What are the predictive targets that we can address using our model for disease prevalence? How can we effectively visualise the uncertainty in spatio-temporal prevalence estimates? These issues have only partly been addressed in current spatio-temporal applications of model-based geostatistics for disease prevalence mapping. Some of these are: Clements et al. (2006) on schistosomiasis in Tanzania; Gething et al. (2012) on the world-wide distribution of *Plasmodium vivax*; Hay et al. (2009) and Noor et al. (2014) on the world-wide and Africa-wide distributions of *Plasmodium falciparium*; Snow et al. (2015b) on historical mapping of malaria in the Kenyan Coast area; Bennett et al. (2013) on the mapping of malaria transmission intensity in Malawi; Kleinschmidt et al. (2001) on malaria incidence in KwaZulu Natal, South Africa; Kleinschmidt et al. (2007) on HIV in South Africa; Soares Magalhaes & Clements (2011) on anemia in preschool-aged children in West Africa; Raso et al. (2005) on

schistosomiasis in Côte D'Ivoire; Pullan et al. (2011) on soil-transmitted infections in Kenya; Zouré et al. (2014) on river blindness in the 20 participating countries of the African Programme for Onchocerciasis control. In almost all of these cases, the adopted spatio-temporal model is only assessed with respect to its predictive performance, using ROC curves and prediction error summaries. In our view, a validation check on the adopted correlation structure in the analysis should precede geostatistical prediction, as misspecification of the spatio-temporal structure of the field $S(x,t)$ can potentially lead to an inaccurate quantification of uncertainty in the prevalence estimates and, therefore, to invalid inferences. In this paper, we describe the different stages of a spatio-temporal geostatistical analysis and provide tools that directly address the issue of specifying a spatio-temporal covariance structure that is compatible with the data.

The paper is structured as follows. Section 2 is a review on geostatistical sampling design, where we show how this might affect our analysis of the data. In Section 3 we describe principles and provide statistical tools for each of the stages of a spatio-temporal geostatistical analysis. In Section 3.1, we define the objectives of an exploratory geostatistical analysis and show how to pursue these using the empirical variogram. In Section 3.2, we outline and contrast likelihood-based and Bayesian methods of inference. In Section 3.3, we propose a general Monte Carlo procedure based on the empirical variogram, in order to check the validity of the assumed spatio-temporal correlation function for $S(x,t)$. In Sections 3.4 and 3.5, we discuss how to define and visualize predictive targets. In Section 4 we illustrate an application to historical mapping of malaria using data from prevalence surveys conducted in Senegal between 1905 and 2014. Section 5 is a concluding discussion.

2 Geostatistical sampling design

Different design scenarios can give rise to data of the kind expressed by (1). A good choice of design depends both on the objectives of the study and on practical constraints.

In a longitudinal design, data are collected repeatedly over time from the same set of sampled locations. This is an appropriate strategy when temporal variation in the outcome of primary interest dominates spatial variation, and more obviously when the scientific goal is to understand change over time at a set of sentinel locations. A longitudinal design is also cost-effective when setting up a sampling location is expensive but subsequent data-collection is cheap.

In a repeated cross-sectional design, a different set of locations is chosen on each sampling occasion. This sacrifices direct information on changes in disease prevalence over time in favour of more complete spatial coverage. Repeated cross-sectional designs can also be adaptive, meaning that on any sampling occasion, the choice of sampling locations is informed by an analysis of the data collected on earlier occasions. Adaptive repeated cross-sectional designs are therefore particularly suitable for applications in which temporal variation either is dominated by spatial variation or can be well explained by available covariates; see Chipeta et al. (2016) and Kabaghe et al. (2017).

To explain how the sampling design might affect our geostatistical analysis of the data, let $\mathcal{X} = \{x_i \in A : i = 1, \ldots, n\}$ denote the set of sampling locations arising from the sampling design, $\mathcal{S} = \{S(x) : x \in A\}$ the signal process and $\mathcal{Y} = \{Y_i : 1 = 1, \ldots, n\}$ the outcome data.

A sampling design is deterministic if it consists of a set of pre-defined sampling locations, and stochastic if the locations are a probability-based selection from a set of candidate designs. In the latter case $\mathcal{X}$ is a finite point process on the region of interest A. Let $[\cdot]$ denote "the distribution of." Our model for the outcome data is then obtained by integrating out $\mathcal{S}$ from the joint distribution $[\mathcal{X}, \mathcal{S}, \mathcal{Y}]$, i.e.

$$[\mathcal{X}, \mathcal{Y}] = \int [\mathcal{X}, \mathcal{S}, \mathcal{Y}] \, d\mathcal{S}. \tag{4}$$

From a modelling perspective, the most natural factorization of the integrand in the above equation is as

$$[\mathcal{X}, \mathcal{S}, \mathcal{Y}] = [\mathcal{S}][\mathcal{X}|\mathcal{S}][\mathcal{Y}|\mathcal{X}, \mathcal{S}]. \tag{5}$$

The design is *non-preferential* if $[\mathcal{X}|\mathcal{S}] = [\mathcal{X}]$, in which case (4) becomes

$$[\mathcal{X}, \mathcal{Y}] = [\mathcal{X}] \int [\mathcal{S}][\mathcal{Y}|\mathcal{X}, \mathcal{S}] \, d\mathcal{S}. \tag{6}$$

Hence, under non-preferential sampling schemes, inference about $\mathcal{S}$ and/or $\mathcal{Y}$ can be conducted legitimately by simply conditioning on the observed set of locations, $\mathcal{X}$.

The simplest example of a probabilistic sampling design is completely random sampling. This can be interpreted, according to context, either as a random sample from a finite, pre-specified set of potential sampling locations or as an independent random sample from the continuous uniform distribution on A. Other examples include spatially stratified random sampling designs, which consist of a collection of completely random designs, one in each of a number of subdivisions of A, and systematic sampling designs, in which the sampled locations form a regular (typically rectangular) lattice to cover A, strictly with the first lattice-point chosen at random, although in practice this is often ignored.

Here as in other areas of statistics, the choice of sampling design affects inferential precision. If, for example, the inferential target is the underlying spatially continuous prevalence surface, $p(x, t^*)$ at a future time t^*, a possible design goal for geostatistical prediction would be to minimise the spatial average of the mean squared error,

$$\int_A \mathrm{E}[\{\hat{p}(x, t^*) - p(x, t^*)\}^2] dx,$$

where $\hat{p}(x, t^*)$ is a predictor for $p(x, t^*)$ obtained from (2). In contrast, a possible design goal for estimation of the relationship between a covariate $d(x, t)$ and disease prevalence would be to minimise the variance of the estimated regression parameter, $\hat{\beta}$.

Efficient sampling designs for spatial prediction generally require sampled locations to be distributed more evenly over A than would result from completely random or stratified random sampling; see, for example, Matérn (1986).

Stratified sampling often provides a more cost-effective design than simple random sampling from the general population. In cases where the strata correspond to sub-populations associated with different disease risk levels, a geostatistical model should account for the stratification through the use of an appropriate explanatory variable. To illustrate this, consider, for example, a population consisting of K strata which correspond to a partition of the region of interest, A, into non-overlapping regions $\mathcal{R}_k$ for $k = 1, \ldots, K$. We then take a random sample from each region $\mathcal{R}_k$ so that each location $x \in \mathcal{R}_k$ has probability of being selected proportional to the population of $\mathcal{R}_k$. If it is known that each of the strata $\mathcal{R}_k$ is associated with different levels in disease risk, this can be accounted for by including a factor variable in (2) with $K - 1$ levels or, if K is large, using random effects at stratum-level. In some cases the strata can also be grouped into sub-populations which are known to differ in their exposure to the disease. For example, let us assume that each stratum can be classified as being urban or rural and that these two types of areas are associated with different risk levels, i.e.

$$\log\left\{\frac{p(x_i, t_i)}{1 - p(x_i, t_i)}\right\} = \beta + \alpha u(x_i) + S(x_i, t_i) + Z(x_i, t_i), \tag{7}$$

where $u(x_i)$ is an indicator function that takes value 1 if $x_i \in \mathcal{R}_k$ and $\mathcal{R}_k$ is urban, and 0 otherwise. Under this model, it follows that

$$[\mathcal{Y}, \mathcal{S}, \mathcal{X}] = [\mathcal{X}][\mathcal{S}][\mathcal{Y}|\mathcal{S}, \mathcal{X}]$$

hence (7) does not constitute an instance of preferential sampling. This shows that variables used in the design should be included in the model when these are associated with the outcome of interest so as to ensure that the sampling is non-preferential. For a wider discussion on this issue in the context of standard regression models, we refer to Skinner & Wakefield (2017) and Lumley & Scott (2017).

Another common design in practice is the opportunistic sampling design (Hedt & Pagano, 2011), in which data are collected at convenient places, for example from presentations at health clinics, a market or a school. The limitations of this are obvious: opportunistic samples may not be representative of the target population and so not deliver unbiased estimates of $p(x, t)$. Also, as unmeasured factors relating to the disease in question are likely to affect an indivudual's decision to present, the assumption of non-preferential sampling is questionable. For example, areas with atypically high or low levels of $p(x, t)$ may have been systematically oversampled; see Diggle et al. (2010) and Pati et al. (2011) for a discussion and formal solution to the problem of geostatistical inference under preferential sampling.

Giorgi et al. (2015) address the issue of combining data from multiple prevalence surveys, with a mix of random and opportunistic sampling designs. By developing a multivariate geostatistical model that enables estimation of the bias from opportunistic samples, they show that combining information from multiple studies can lead to more precise estimates of prevalence, provided that at least one of these is known to be unbiased.

In the remainder of this paper, we shall focus our attention on the case of prevalence data obtained from a non-preferential sampling design.

3 Methods

In this Section we provide a general framework for the analysis of data from spatio-temporally referenced prevalence surveys. Figure 1 shows the different stages of the analysis as a cycle that terminates when all the modelling assumptions are supported by the data. In our context, visualization of the results also plays an important role in order to display the spatio-temporal patterns of estimated prevalence and to communicate uncertainty effectively.

3.1 Exploratory analysis: the spatio-temporal variogram

The usual starting point for a spatio-temporal analysis of prevalence data is an analysis based on a binomial mixed model without spatial random effects, i.e. $S(x,t) = 0$ for all x and t. Let $\tilde{Z}(x_i, t_i)$ denote a point estimate, such as the predictive mean or mode, of the unstructured random effects $Z(x_i, t_i)$ from the non-spatial binomial mixed model. We then analyse $\tilde{Z}(x_i, t_i)$ to pursue the two following objectives:

1. testing for presence of residual spatio-temporal correlation;
2. formulating a model for (3) and providing an initial guess for θ.

We make a working assumption that $S(x,t)$ is a stationary and isotropic process, hence

$$\rho(x, x', t, t'; \theta) = \rho(u, v; \theta), \tag{8}$$

where $u = \|x - x'\|$, with $\|\cdot\|$ denoting the Euclidean distance, and $v = |t - t'|$.

The *variogram* can then be used to formulate and validate models for the spatio-temporal correlation in (3). Let $W(x,t) = S(x,t) + Z(x,t)$, where $S(x,t)$ and $Z(x,t)$ are specified as in (2); the spatio-temporal variogram of this process is given by

$$\gamma(u, v; \theta) = \frac{1}{2} E[\{W(x,t) - W(x',t')\}^2] = \tau^2 + \sigma^2[1 - \rho(u, v; \theta)]. \tag{9}$$

We refer to this as the *theoretical* variogram, since it is directly derived from the theoretical model for the process $W(x,t)$.

We use $\tilde{Z}(x_i, t_i)$ to estimate the unexplained extra-binomial variation in prevalence, at observed locations x_i and times t_i. Let $n(u,v)$ denote the pairs (i,j) such that $\|x_i - x_j\| = u$ and $|t_i - t_j| = v$; the *empirical* variogram is then defined as

$$\tilde{\gamma}(u, v) = \frac{1}{2|n(u,v)|} \sum_{(i,j) \in n(u,v)} \{\tilde{Z}(x_i, t_i) - \tilde{Z}(x_j, t_j)\}^2, \tag{10}$$

where $|n(u,v)|$ is the number of pairs in the set.

Testing for the presence of residual spatio-temporal correlation can be carried out using the following Monte-Carlo procedure:

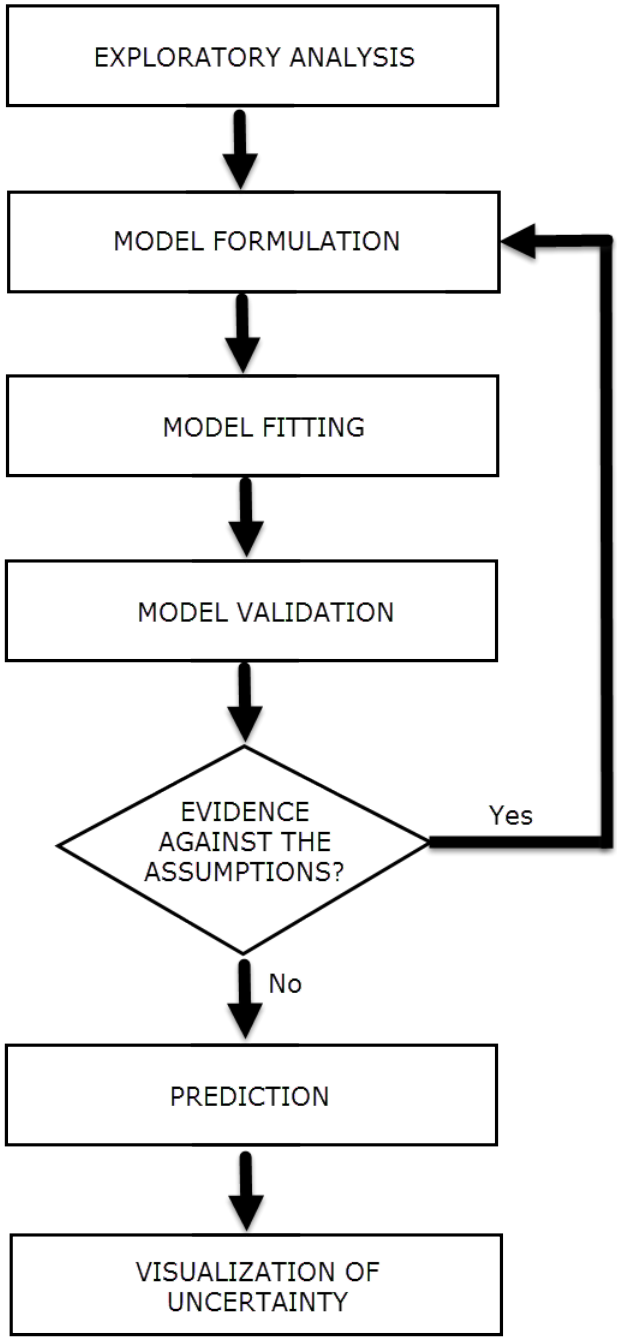

Figure 1: Diagram of the different stages of a statistical analysis.

(Step 1) permute the order of the data, including $\tilde{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;

(Step 2) compute the empirical variogram for $\tilde{Z}(x_i, t_i)$;

(Step 3) repeat(i) and (ii) a large enough number of times, say B;

(Step 4) use the resulting B empirical variograms to generate 95% tolerance intervals at each of the pre-defined distance bins.

If $\tilde{\gamma}(u, v)$ lies outside these intervals, then the data show evidence of residual spatio-temporal correlation. If this is the case, the next step is to specify a functional form for $\rho(u, v)$.

Gneiting (2002) proposed the following class of spatio-temporal correlation functions

$$\rho(u, v; \theta) = \frac{1}{(1 + v/\psi)^{\delta+1}} \exp\left\{-\frac{u/\phi}{(1 + v/\psi)^{\xi/2}}\right\}, \tag{11}$$

where ϕ and (δ, ψ) are positive parameters that determine the rate at which the spatial and temporal correlations decay, respectively. When $\xi = 0$ in (11), $\rho(u, v; \theta) = \rho_1(u)\rho_2(v)$ where $\rho_1(\cdot)$ and $\rho_2(\cdot)$ are purely spatial and purely temporal correlation functions, respectively. Any spatio-temporial correlation function that factorises in this way is called *separable*. In this sense, the parameter $\xi \in [0, 1]$ represents the extent of non-separability. Stein (2005) provides a detailed analysis of the properties of space-time covariance functions and highlights the limitations of using separable families. However, fitting of complex space-time covariance models requires more data than, in our experience, is typically available in prevalence mapping applications. In the application of Section 4, we show that only ψ and ϕ in (11) can be estimated with an acceptable level of precision, whilst the data are poorly informative with respect to the other covariance parameters, in which case the parsimony principle favours a separable model. Note, incidentally, that separability is implied by, but does not imply, that $S(x, t)$ can be factorised as $S_1(x)S_2(t)$, which would be a highly artificial construction.

A spatio-temporal correlation function is separable if

$$\rho(u, v; \theta) = \rho_1(u; \theta_1)\rho_2(v; \theta_2),$$

where θ_1 and θ_2 parametrise the purely spatial and temporal correlation functions, respectively; in the case of (11), this is separable when $\xi = 0$. Separable correlation functions are computationally convenient when joint predictions of prevalence are required at different time points over the same set of prediction locations. Checking the validity of the separability assumption can be carried out using the likelihood-ratio test for models such as (11), where separability can be recovered as a special case.

Once a parametric model has been specified, an initial guess for θ can be used to initialise the maximization of the likelihood function. One way to obtain an initial guess is to choose the value of θ that minimizes the sum of squared differences between the theoretical and empirical variogram ordinates. Section 5.3 of Diggle & Ribeiro (2007) describes the least squares algorithm and other, more refined methods to fit a parametric variogram model to an empirical variogram. However, in our view, variogram-based techniques should only be used for exploratory analysis and diagnostic checking. For parameter estimation and formal inference, likelihood-based and Bayesian methods are more efficient and more objective.

3.2 Parameter estimation and spatial prediction

We now outline likelihood-based and Bayesian methods of parameter estimation for the model in (2).

3.2.1 Likelihood-based inference

Let $\lambda^\top = (\beta^\top, \sigma^2, \theta^\top)$ denote the set of unknown model parameters, including regression coefficients β, the variance σ^2 of $S(x,t)$ and covariance parameters θ. We use $[\cdot]$ as a shorthand notation for "the distribution of". The likelihood function is then obtained from the marginal distribution of the outcome $y^\top = (y_1, \ldots, y_n)$ by integrating out the random effects $W^\top = (W(x_1,t_1), \ldots, W(x_n,t_n))$ to give

$$L(\lambda) = [y|\lambda] = \int [W, y|\lambda]\, dW. \tag{12}$$

In general, the integral in (12) is intractable. However, numerical integration techniques or Monte Carlo methods can be used for approximate evaluation and maximization of the likelihood function, as required for classical inference (Geyer & Thompson, 1992; Geyer, 1994, 1996, 1999). See Christensen (2004) for a detailed description of the Monte Carlo maximum likelihood estimation method in a geostatistical context.

In our application of Section 4, we use the following approach to approximate (12). Let λ_0 represent our best guess of λ. We then rewrite (12) as

$$\begin{aligned} L(\lambda) &= \int \frac{[W, y|\lambda]}{[W, y|\lambda_0]} [W, y|\lambda_0]\, dW \\ &\propto \int \frac{[W, y|\lambda]}{[W, y|\lambda_0]} [W|y, \lambda_0]\, dW \\ &= E\left\{ \frac{[W, y|\lambda]}{[W, y|\lambda_0]} \right\}, \end{aligned} \tag{13}$$

where the expectation in the above equation is taken with respect to $[W|y, \lambda_0]$. Using MCMC algorithms, we then generates B samples from $[W|y, \lambda_0]$, say $w_{(i)}$, and approximate (13) as

$$L_B(\lambda) = \frac{1}{B} \sum_{i=1}^{B} \frac{[w_{(i)}|y, \lambda]}{[w_{(i)}|y, \lambda_0]}.$$

We maximize $L_B(\lambda)$ using a Broyden-Fletcher-Goldfarb-Shanno algorithm (Fletcher, 1987), which incorporates analytical expressions for the first and second derivatives of $L_B(\lambda)$. Let $\hat{\lambda}_B$ denote the Monte Carlo maximum likelihood estimate of λ. We then set $\lambda_0 = \hat{\lambda}_B$ and repeat the outlined procedure until convergence.

To simulate from $[W|y, \lambda_0]$, we first reparametrise the model based on $\tilde{W} = \hat{\Sigma}^{-1/2}(W - \hat{w})$, where $\hat{w}$ is the mode of $[W|y, \lambda_0]$ and $\hat{\Sigma}$ is the inverse of the negative Hessian of $[W|y, \lambda_0]$ at the mode $\hat{w}$. At each iteration of the MCMC, we propose a new value for $\tilde{W}$, given the

current value w, using a Langevin-Hastings algorithm with a Gaussian proposal distribution having mean

$$w + (h/2)\nabla \log[w|y, \lambda_0]$$

and covariance matrix given by hI, where I is the identity matrix and h is tuned so that the acceptance rate is 0.574 (Roberts & Rosenthal, 1998).

Other approaches that have been proposed to maximize (12) are based on the expectation-maximization algorithm (Zhang, 2002) and the Laplace approximation (Bonat & Ribeiro, 2016).

Let W^* denote the vector of values of $W(x,t)$ at a set of unobserved times and locations. The formal solution to the prediction problem is to evaluate the conditional distribution of W^* given the data y. Although the joint predictive distribution of the elements of W^* is intractable, it is possible to simulate samples from this distribution.

If we assume, unrealistically, that λ is known, the predictive distribution of W^* is given by

$$[W^*|y,\lambda] = \int [W^*, W|y,\lambda]dW = \int [W|y,\lambda][W^*|W,y,\lambda]dW = \int [W|y,\lambda][W^*|W,\lambda]dW. \quad (14)$$

See Chapter 4 of Diggle & Ribeiro (2007) for explicit expressions.

If, more realistically, λ is unknown, *plug-in* prediction consists of replacing λ in (14) by an estimate $\hat{\lambda}$, preferably the maximum likelihood estimate. A legitimate criticism of this is that the resulting predictive probabilities ignore the inherent uncertainty in $\hat{\lambda}$. However, this can be taken into account within a likelihood-based inferential framework as follows. Let $\hat{\Lambda}$ denote the maximum likelihood estimator of λ. We define the predictive distribution of W^* as

$$[W^*|y] = \int\int [\hat{\Lambda}][W|y,\hat{\Lambda}][W^*|W,\hat{\Lambda}]\, dW\, d\hat{\Lambda}, \quad (15)$$

where $[\hat{\Lambda}]$ denotes the sampling distribution of the maximum likelihood estimator $\hat{\Lambda}$. Equation (15) acknowledges the uncertainty in $\hat{\Lambda}$ by expressing the predictive distribution $[W^*|y]$ as the expectation of the plug-in predictive distribution (14) with respect to the sampling distribution of $\hat{\Lambda}$. This can then be approximated using a multivariate Gaussian distribution with mean given by the observed MLE, $\hat{\lambda}$, and covariance matrix given by

$$\left[-\frac{\partial^2 \log L(\hat{\lambda})}{\partial^2 \lambda}\right]^{-1}.$$

In our experience, the quality of the Gaussian approximation is improved considerably by applying a log-transformation to each of the covariance parameters. If the Gaussian approximation remains questionable, a more computationally intensive alternative is a parametric bootstrap consisting of the following steps: simulate a number of binomial data-sets using the plug-in MLE for λ; for each simulated data-set, carry out parameter estimation by maximum likelihood. The resulting set of bootstrap estimates for λ can then be used to approximate the distribution of $\hat{\Lambda}$. We give an example of these approaches in the case-study of Section 4.

3.2.2 Bayesian inference

In Bayesian inference, λ is treated as a random variable and must be assigned a prior distribution, $[\lambda]$. Parameter estimation is then carried out through the posterior distribution of λ, which is obtained using Bayes' theorem as

$$[\lambda|y] = \frac{[\lambda][y|\lambda]}{[y]} = \frac{[\lambda]L(\lambda)}{[y]}. \tag{16}$$

All other things being equal, as the sample size increases $L(\lambda)$ becomes more concentrated around the true value of λ, the impact of the prior is reduced and the difference between likelihood-based and Bayesian parameter estimation becomes less important. MCMC algorithms can be used for approximate computation of the posterior in (16). For the Bayesian analysis in the application of Section 4, we develop an MCMC algorithm which separately updates β, σ^2, θ and W. Specifically, we use a Metropolis-Hastings algorithm to update $\log\{\sigma^2\}$ and $\log\{\theta\}$, and a Gibbs sampler to update β. To update the random effect W, we use a Hamiltonian Monte Carlo procedure (Neal, 2011). More computational details on this approach can be found in Section 2.2 of Giorgi & Diggle (2017).

Non-stochastic analytical approximations of (16) can also be obtained using, for example by the use of integrated nested Laplace approximations (Rue et al., 2009). However, their accuracy should be considered carefully in each spefici context. Joe (2008) shows that for binomial mixed models, the smaller the denominator the less accurate is the Laplace approximation. Fong et al. (2010), in a review of computational methods for Bayesian inference in generalized linear mixed models, also report poor performance of the INLA method in the case binary responses

Bayesian predictive inference about W^* uses a second application of Bayes' theorem to give the predictive distribution

$$[W^*|y] = \int\int [\lambda|y][W|y,\lambda][W^*|W,\lambda]\, dW\, d\lambda, \tag{17}$$

where $[\lambda|y]$ is the posterior distribution of θ. Comparison of (17) and (15) shows that both are weighted averages of plug-in predictive distributions. The difference between them is that (17) uses the posterior $[\lambda|y]$ as the weighting distribution whilst (15) uses the sampling distribution $[\hat{\Lambda}]$. In either case, the weights concentrate increasingly around the maximum likelihood estimate of λ as the sample size increases.

In our experience the difference between plug-in prediction using the maximum likelihood estimate $\hat{\lambda}$ and weighted average prediction is often negligible, because the uncertainty in W^* dominates that in λ. An intuitive explanation for this is that for estimation of λ all of the data contribute information, whereas for prediction of $W(x,t)$ only data at locations and times relatively close to x and t contribute materially. However, this is not guaranteed, especially when the predictive target is a non-linear property of W^*; see, for example, Figure 9a of Diggle et al. (2002).

3.3 Diagnostics and novel extensions

In order to check the validity of the chosen spatio-temporal covariance function, we modify the Monte Carlo algorithm introduced in Section 3.1 by replacing (Step 1) with following.

(Step 1) Simulate $W(x_i, t_i)$ at observed locations x_i and times t_i, for $i = 1, \ldots, n$, from its marginal multivariate distribution under the assumed model. Conditionally on the simulated values of $W(x_i, t_i)$, simulate binomial data y_i from (2). Finally, compute the point estimates $\tilde{Z}(x_i, t_i)$ using the simulated data.

In this case, the resulting 95% tolerance band is generated under the assumption that the true covariance function for $S(x, t)$ exactly corresponds to the one adopted for the analysis. If $\tilde{\gamma}(u, v)$ lies outside the intervals, then this indicates that the fitted covariance function is not compatible with the data. To formally test this hypothesis, we can also use the following test statistic

$$T = \sum_{k=1}^{K} |n(u_k, t_k)|[\tilde{\gamma}(u_k, v_k) - \gamma(u_k, v_k; \theta)]^2, \tag{18}$$

where u_k and v_k are the distance and time separations of the variograms bins, respectively, the $n(u_k, t_k)$ are the numbers of pairs of observations contributing to each bin and θ is the true parameter value of the covariance parameters. Since θ is almost always unknown, it can be estimated using either maximum likelihood or Bayesian methods, in which case (18) should be averaged over the posterior distribution of θ using posterior samples $\theta_{(h)}$, i.e.

$$T = \frac{1}{B} \sum_{h=1}^{B} \sum_{k=1}^{K} |n(u_k, t_k)|[\tilde{\gamma}(u_k, v_k) - \gamma(u_k, v_k; \theta_{(h)})]^2. \tag{19}$$

The null distribution of T can be obtained using the simulated values for $\tilde{Z}(x_i, t_i)$ from the modified (Step 1) introduced in this section. Let $T_{(h)}$ denote the h-th sample from the null distribution of T, for $h = 1, \ldots, B$. Since evidence against the adopted covariance model arises from large values of T, an approximate p-value can be computed as

$$\frac{1}{B} \sum_{h=1}^{B} I[T_{(h)} > t],$$

where $I(a > b)$ takes value 1 if $a > b$ and 0 otherwise, and t is the value of the test statistic obtained from the data.

An unsatisfactory result from this diagnostic check could indicate a need for either or both of two extensions to the model: a more flexible family of stationary covariance structures; or non-stationarity induced by parameter variation over time, space or both.

In the former case, we note that the correlation function in (11) can also be obtained a special case of

$$\rho(u, v; \theta) = \frac{1}{(1 + v/\psi)^{\delta+1}} \mathcal{M}\left(\frac{u}{(1 + v/\psi)^{\xi/2}}; \phi, \kappa\right) \tag{20}$$

where $\mathcal{M}(\cdot;\phi,\kappa)$ is the Matérn (1986) correlation function with scale and smoothness parameters ϕ and κ, respectively (Gneiting, 2002). Equation (11) is recovered for $\kappa = 1/2$. However, the additional parameter introduced, κ, is likely to be poorly identified. A pragmatic response is to discretise the smoothness parameter κ in (20) to a finite set of values, e.g. $\{1/2, 3/2, 5/2\}$, over which the likelihood function is maximized.

In the second case, the context of the analysis can provide some insights on the nature of the non-stationary behaviour of the process being studied. For example, if data are sampled over a large geographical area, such as a continent, one may expect the properties of the process $S(x,t)$ to vary across countries. This can then be assessed by fitting the model separately for each country. A close inspection of the parameter estimates for θ might then reveal which of its components show the strongest variation. Furthermore, if these estimates also show spatial clustering, the vector θ, or some of its components, can be modelled as an additional spatial process, say $\Theta(x)$. The process $S(x,t)$ is then modelled as a stationary Gaussian process conditionally on $\Theta(x)$. A similar argument can also be developed if data are collected over a large time period in a geographically restricted area. In this case, θ may primarily vary across time and, therefore, could be modelled as a temporal stochastic process.

3.3.1 Example: a model for disease prevalence with temporally varying variance

We now give an example of how model (2) can be extended in order to allow the nature of the spatial variation in disease prevalence to change over time. We replace the spatio-temporal random effect $S(x,t)$ in the linear predictor with

$$S^*(x,t) = B(t)S(x,t), \tag{21}$$

where $B^2(t)$ represents the temporally varying variance of $S^*(x,t)$. We then model $\log\{B^2(t)\}$ as a stationary Gaussian process, independent of $S(x,t)$, with mean $-\eta^2/2$, variance η^2 and one-dimensional correlation function $\rho_B(\cdot;\theta_B)$, with covariance parameters θ_B. Note that, using this parametrisation, $E[B^2(t)] = 1$ and, therefore, $V[S^*(x,t)] = \sigma^2$. The resulting process $S^*(x,t)$ is a non-Gaussian process with heavier tails than $S(x,t)$ and correlation function

$$\text{corr}\{S^*(x,t), S^*(x',t')\} \;=\; \exp\{\eta^2(\rho_B(v;\theta_B) - 1)\}\rho(u,v;\theta). \tag{22}$$

The likelihood function is obtained as in (12) but now with $W(x_i,t_i) = S^*(x_i,t_i) + Z(x_i,t_i)$.

3.4 Defining targets for prediction

Let $\mathcal{P}(W^*) = \{p(x,t) : x \in A, t \in [T_1,T_2]\}$ denote the set of prevalence surfaces covering the region of interest A and spanning the time period $[T_1,T_2]$. Prediction of $\mathcal{P}$ is carried out by first simulating samples from the the predictive distribution of W^*, i.e. the distribution of W^* conditional on the data y. From each simulated sample of W^*, we then calculate any required summary, $\mathcal{T}$ say, of the corresponding $\mathcal{P}(W^*)$, for example means or selected quantiles at any

(x, t) of interest. By construction, this generates a sample from the predictive distribution of $\mathcal{T}$. Computational details and explicit expressions can be found in Giorgi & Diggle (2017).

Two ways to display uncertainty in the estimates of prevalence are through quantile or exceedance probability surfaces. We define the *α-quantile surface* as

$$\mathcal{Q}_\alpha(W^*) = \{q(x,t) : P(p(x,t) < q(x,t)|y) = \alpha, x \in A, t \in [T_1, T_2]\}. \tag{23}$$

Similarly, we define the *exceedance probability surface* for a given threshold l as

$$\mathcal{R}_l(W^*) = \{r(x,t) = P(p(x,t) > l|y) : x \in A, t \in [T_1, T_2]\}. \tag{24}$$

Values of the point-wise exceedance probability $r(x,t)$ close to 1 identify locations for which prevalence is highly likely to exceed l, and vice-versa.

In public health applications, an exceedance probability surface is a suitable predictive summary when the objective is to identify areas that may need urgent intervention because they are likely to exceed a policy-relevant prevalence threshold, say l. A disease "hotspot" is then operationally defined as the set of locations x, at a given time t, such that $p(x,l) > l$.

In some cases, summaries by administrative areas can be operationally useful. For example, the district-wide average prevalence for a district D at time t is

$$p_t(D) = \frac{1}{|D|} \int_D p(x,t)\, dx, \tag{25}$$

where $|D|$ is its area of D. Incidentally, $p_t(D)$ can also be estimated more accurately than the point-wise prevalence $p(x,t)$, because it uses all the available information within D. Quantile and exceedance probability surfaces can be defined for $p_t(D)$ in the obvious way.

3.5 Visualization

The output from the prediction step consists of a set of N predictive surfaces, whether estimates, quantiles or exceedance probabilities, within the region of interest A at times $t_1 < t_2 < \ldots < t_N$. Animations then provide a useful tool for visualizing the predictive spatio-temporal surfaces and highlighting the main features of the interpolated pattern of prevalence. The R package `animation` (Xie, 2013) provides utilities for writing animations in several video and image formats. However, if interactivity is also desired, web-based "Shiny" applications (SAs) (RStudio, Inc, 2013) represent one of the best alternatives within R.

For the analysis carried out in Section 4, we have developed an SA which can be viewed at

`http://fhm-chicas-apps.lancs.ac.uk/shiny/users/giorgi/mapMalariaSEN/`.

The user-interface of this SA is shown in Figure 2. Any of four panels can be chosen in order to display predictive maps of prevalence ("Prediction maps"), exceedance probabilities with user defined prevalence thresholds ("Exceedance maps"), quantile surfaces ("Quantile maps")

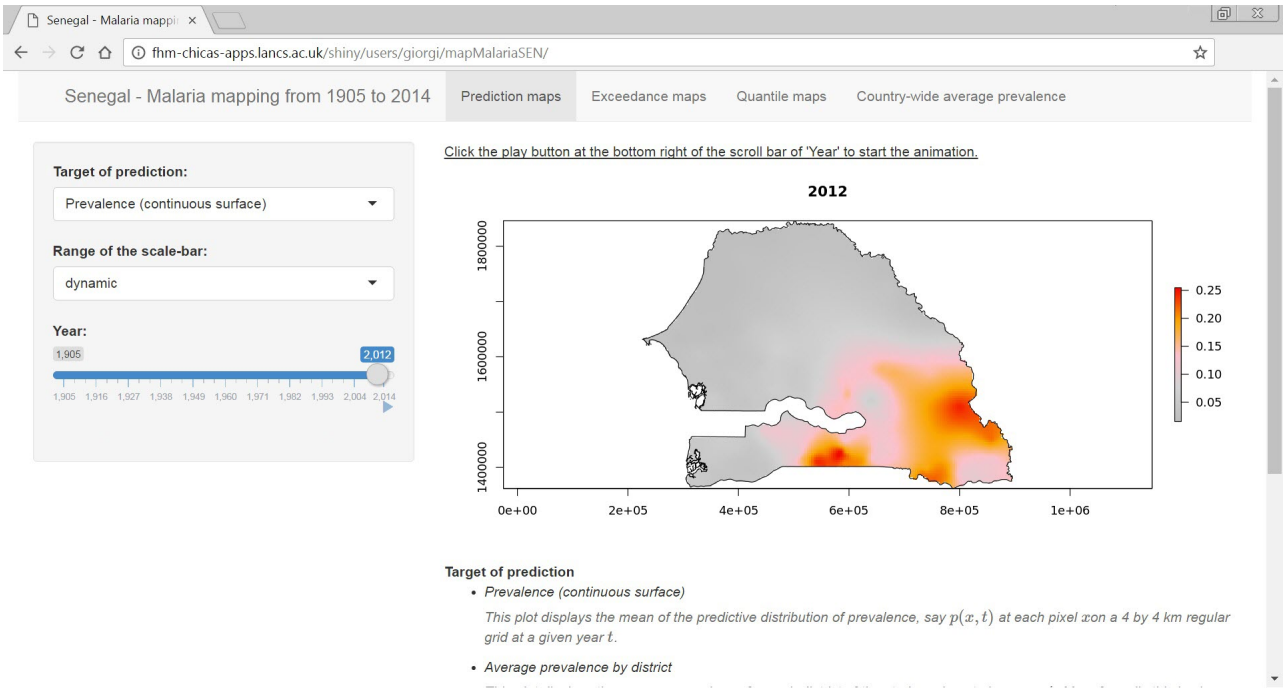

Figure 2: User interface of a Shiny application for visualization of results. The underlying data are described in Section 4.

and country-wide summaries ("Country-wide average prevalence"). In the first three panels, the user can choose which target of prediction to display from a list and select the year on a slide bar. The range of prevalence and exceedance probabilities used to define the colour scale can be set to the observed range across the whole time series ("fixed") or specific to each year ("dynamic"). The former option is convenient for comparisons between years, whilst the latter gives a more effective visualization of the spatial heterogeneity in the predictive target in a given year.

4 Case-study: historical mapping of malaria prevalence in Senegal from 1905 to 2014

We analyse malaria prevalence data from 1,334 surveys conducted in Senegal between 1905 and 2014. The data were assembled from three different data sources: historical archives and libraries of ex-colonial institutes; online electronic databases with data on malaria infection prevalence published since the 1980s; national household sample surveys. In assembling the data for the analysis, we only included locations that were classified as individual villages or communities or a collection of communities within a definable area that does not exceed 5 km^2. For more details on the data extraction, see Snow et al. (2015a).

The outcome of interest is the count y_i of positive microscopy tests out of n_i for *Plasmodium*

falciparum, at a community location x_i and year t_i. Table 1 shows the number of surveys and the average prevalence for each of the indicated time-blocks. These were identified by grouping the data points so that each time-block contains at least 100 surveys. We observe that 649 out of the 1334 surveys were carried out between 2009 and 2014. Also, the empirical country-wide average prevalence steadily declines from the first to the last time-block. Figure 3 displays the sampled community locations within each of the time-blocks. The plot suggests a poor spatial coverage of Senegal in some years. The use of geostatistical methods can therefore be beneficial since it allows us to borrow the strength of information by exploiting the spatio-temporal correlation in the data.

Table 1: Number of surveys and country-wide average *Plasmodium falciparum* prevalence, in each time-block.

Time-block	Number of surveys	Average prevalence
1: 1904 - 1960	180	0.416
2: 1961 - 1966	109	0.384
3: 1967 - 1977	104	0.402
4: 1978 - 1997	101	0.134
5: 1998 - 2008	191	0.111
6: 2009 - 2010	187	0.051
7: 2011	140	0.043
8: 2012 - 2013	157	0.038
9: 2014	165	0.019

Our model for the data is of the form (26), with the following linear predictor

$$\begin{aligned} \log\left\{\frac{p(x_i,t_i)}{1-p(x_i,t_i)}\right\} &= \beta_1 + \beta_2 a(x_i,t_i) + \beta_3[a(x_i,t_i)-5]\times I\{a(x_i,t_i)>5\} + \\ &\quad \beta_4 A(x_i,t_i) + \beta_5[A(x_i,t_i)-20]\times I\{A(x_i,t_i)>20\} + \\ &\quad S(x_i,t_i) + Z(x_i,t_i), \end{aligned} \tag{26}$$

where $a(x_i,t_i)$ and $A(x_i,t_i)$ are the lowest and largest observed ages among the sampled individuals at location x_i and time t_i, respectively. In (26), we use linear splines, each with a single knot, at 5 years for $a(x,t)$ and at 20 years for $A(x,t)$. For the spatio-temporal process $S(x,t)$, we use a Gneiting correlation function, as in (11), with $\delta = \xi = 0$, i.e. a separable covariance function.

Using the predictive mean as a point estimate of the random effects from a non-spatial binomial mixed model, we carry out the test for residual spatio-temporal correlation, as outlined in Section 3.1. The upper panels of Figure 4 show overwhelming evidence against the assumption of spatio-temporal independence. We then initialize the covariance parameters, ϕ and ψ, using a least squares fit to the empirical variogram, as shown by the dotted lines in the lower panels of Figure 4.

We conducted parameter estimation and spatial prediction using both likelihood-based and Bayesian inference. In the latter case, we specifed the following set of independent and vague priors: $\beta \sim MVN(0, 10^4 I)$; $\sigma^2 \sim$ Uniform$(0, 20)$; $\phi \sim$ Uniform$(0, 1000)$; $\tau^2/\sigma^2 \sim$

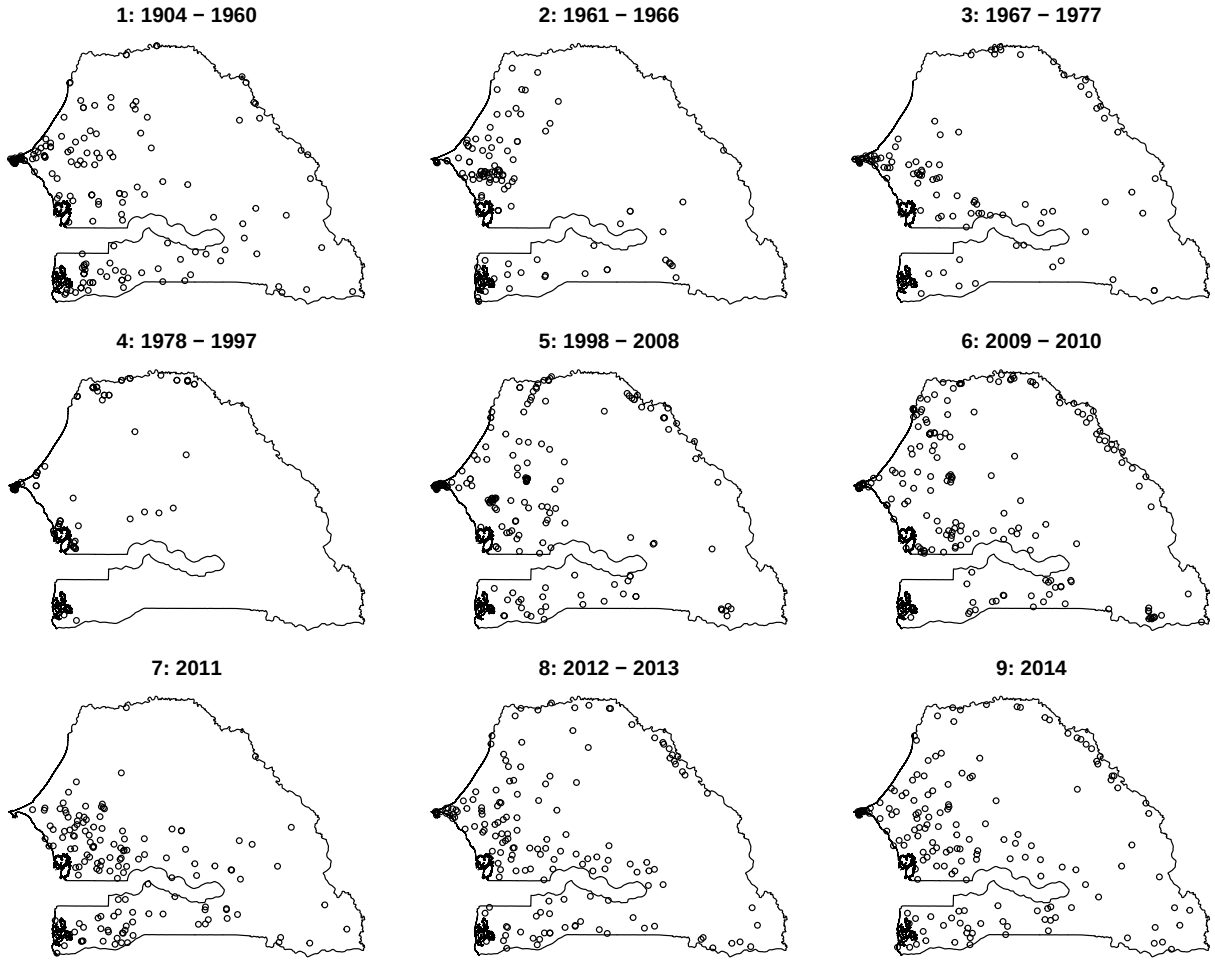

Figure 3: Locations of the sampled communities in each of the time-blocks indicated by Table 1.

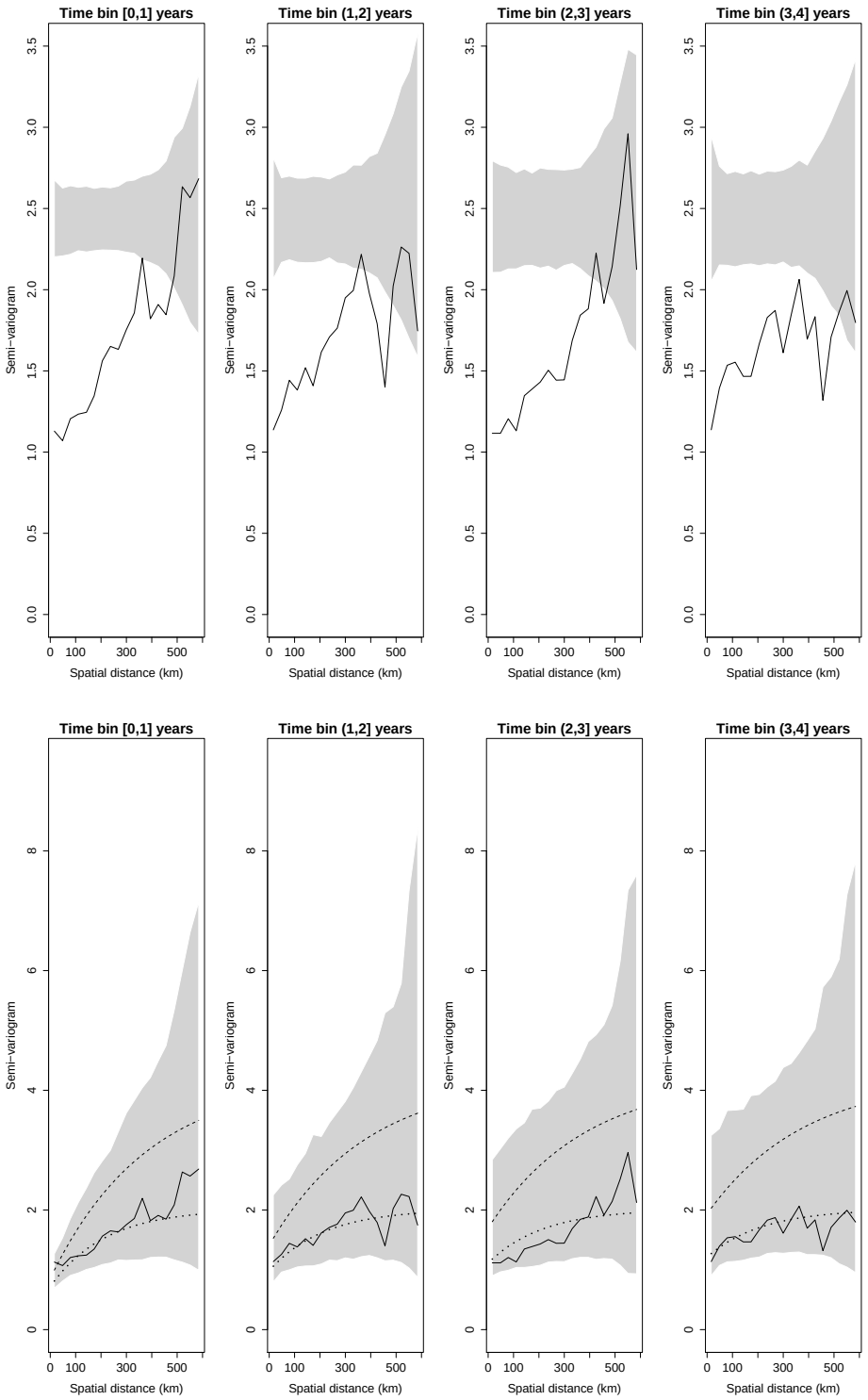

Figure 4: The plots show the results from the Monte Carlo methods used to test the hypotheses of spatio-temporal indepence (upper panels) and of compatibility of the adopted covariance model with the data (lower panels). The shaded areas represent the 95% tolerance region under each of the two hypotheses. The solid lines correspond to the empirical variogram for $\tilde{Z}(x_i, t_i)$, as defined in Section 3.1. In the lower panels, the theoretical variograms obtained from the least squares (dotted lines) and maximum likelihood (dashed lines) methods are shown.

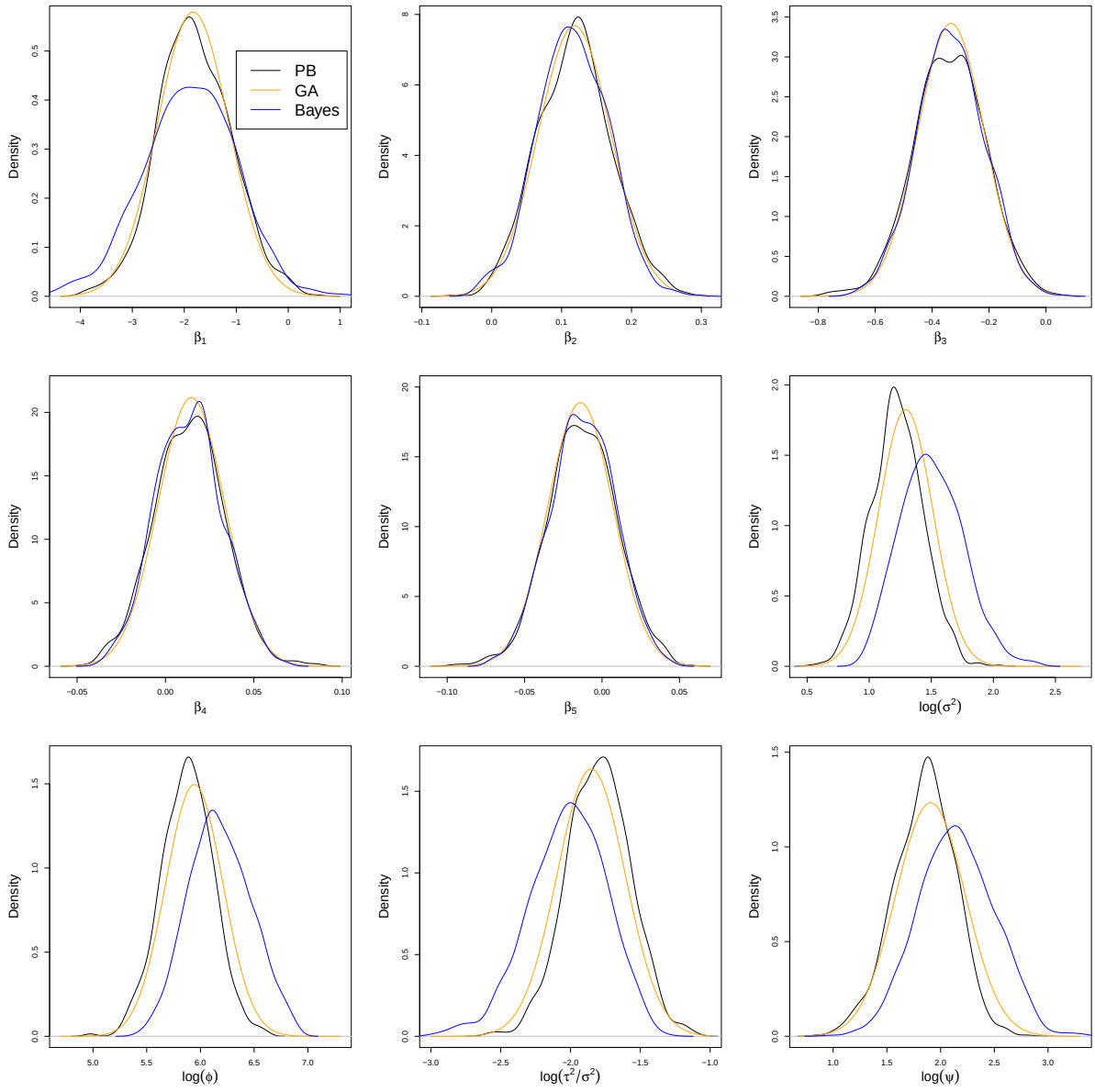

Figure 5: Density functions of the maximum likelihood estimator for each of the model parameters based on parameteric bootstrap (PB), as black lines, and the Gaussian approximation (GA), as orange lines; the blue lines correspond to the posterior density from the Bayesian fit.

Table 2: Maximum likelihood estimates of the model parameters and their 95% confidence intervals (CI) based on the asymptotic Gaussian approximation (GA) and parametric bootstrap (PB).

Parameter	Estimate	95% CI (GA)	95% CI (PB)
β_1	-1.830	(-3.180, -0.480)	(-3.131, -0.367)
β_2	0.118	(0.017, 0.220)	(0.019, 0.226)
β_3	-0.334	(-0.562, -0.105)	(-0.585, -0.103)
β_4	0.015	(-0.022, 0.052)	(-0.025, 0.052)
β_5	-0.014	(-0.055, 0.027)	(-0.056, 0.030)
σ^2	3.650	(2.378, 5.601)	(2.272, 5.222)
ϕ	381.022	(225.948, 642.528)	(220.593, 568.953)
τ^2/σ^2	0.157	(0.097, 0.253)	(0.105, 0.253)
ψ	6.730	(3.571, 12.683)	(3.484, 10.669)

Table 3: Posterior mean and 95% credible intervals of the model parameters from the Bayesian fit.

	Posterior mean	95% credible interval
β_1	-1.899	(-3.746, -0.275)
β_2	0.116	(0.013, 0.212)
β_3	-0.335	(-0.560, -0.115)
β_4	0.013	(-0.023, 0.050)
β_5	-0.013	(-0.054, 0.028)
σ^2	4.649	(2.887, 7.641)
ϕ	504.330	(283.019, 863.198)
τ^2/σ^2	0.137	(0.075, 0.217)
ψ	9.098	(4.443, 16.608)

Uniform$(0, 20)$; $\psi \sim$ Uniform$(0, 20)$. Table 2 shows the maximum likelihood estimates of the model parameters and their corresponding 95% confidence intervals based on the Gaussian approximation (GA) and on parametric boostrap (PB), together with Bayesian esimates (posterior means) and 95% credible intervals. The two non-Bayesian methods give similar confidence intervals; the difference is noticeable, although still small in practical terms, only for the parameter ϕ. The Bayesian method gives materially larger estimates of σ^2 and ϕ . Note that for both of these parameters, the prior means are substantially larger than the maximum likelihood estimates, suggesting that the priors, although vague, have nevertheless had some impact on the estimates.

Figure 5 gives a different perspective on the similarities and differences between the results obtained by the non-Bayesian and Bayesian methods. The Bayesian posterior density of the intercept has heavier tails than the sampling distribution of the maximimum likelihood estimator; the posterior densities of σ^2, ϕ and ψ are shifted to the right of their non-Bayesian counterparts, whilst the posterior density of τ^2/σ^2 is shifted to the left. Finally, there is some residual skewness in the PB distributions of the log-transformed covariance parameters.

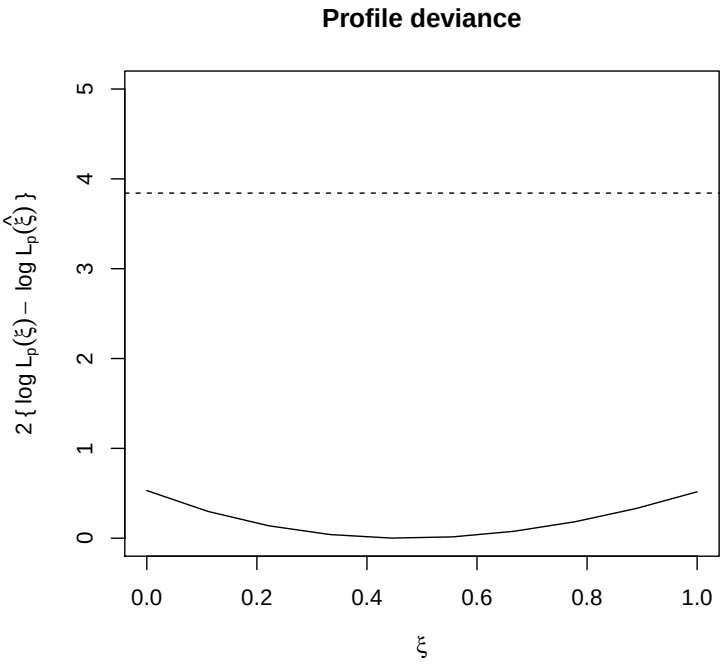

Figure 6: Profile deviance (solid line) for the parameter of spatio-temporal interaction ξ of the Gneiting (2002) family given by (11). The dashed line is the 0.95 quantile of a χ^2 distribution with one degree of freedom.

Using the Monte Carlo methods of Section 3.3, we checked the validity of the assumed covariance model. The lower panels of Figure 4 show that for each of the four time-lag intervals considered, the observed variograms fall within within the 95% tolerance region obtained under the fitted model; the p-value for a Monte Carlo goodness-of fit test using the test statistic (18) is 0.548.

Figure 6 shows the profile deviance function

$$D(\xi) = 2\{\log L_p(\hat{\xi}) - \log L_p(\xi)\},$$

where $L_p(\xi)$ is the profile likelihood for the parameter of spatio-temporal interaction parameter ξ and $\hat{\xi}$ is its Monte Carlo maximum likelihood estimate. The dashed horizontal line is the 0.95 quantile of a χ^2 distribution with one degree of freedom. The flatness of $D(\xi)$ indicates that data give very little information about the non-separability of the correlation structure of $S(x,t)$.

To assess the differences in the spatial predictions obtained using thr GA, PB and Bayesian approaches, we used each method to predict *P. falciparum* prevalence for children between 2 and 10 years of age ($PfPR_{2-10}$) in the year 2014, at each point on a 10 by 10 km regular grid covering the whole of Senegal. Figure 7 shows pairwise scatterplots of the three sets of point predictions and associated standard deviations of $PfPR_{2-10}$. All six scatter plots show only small deviations from the identity line.

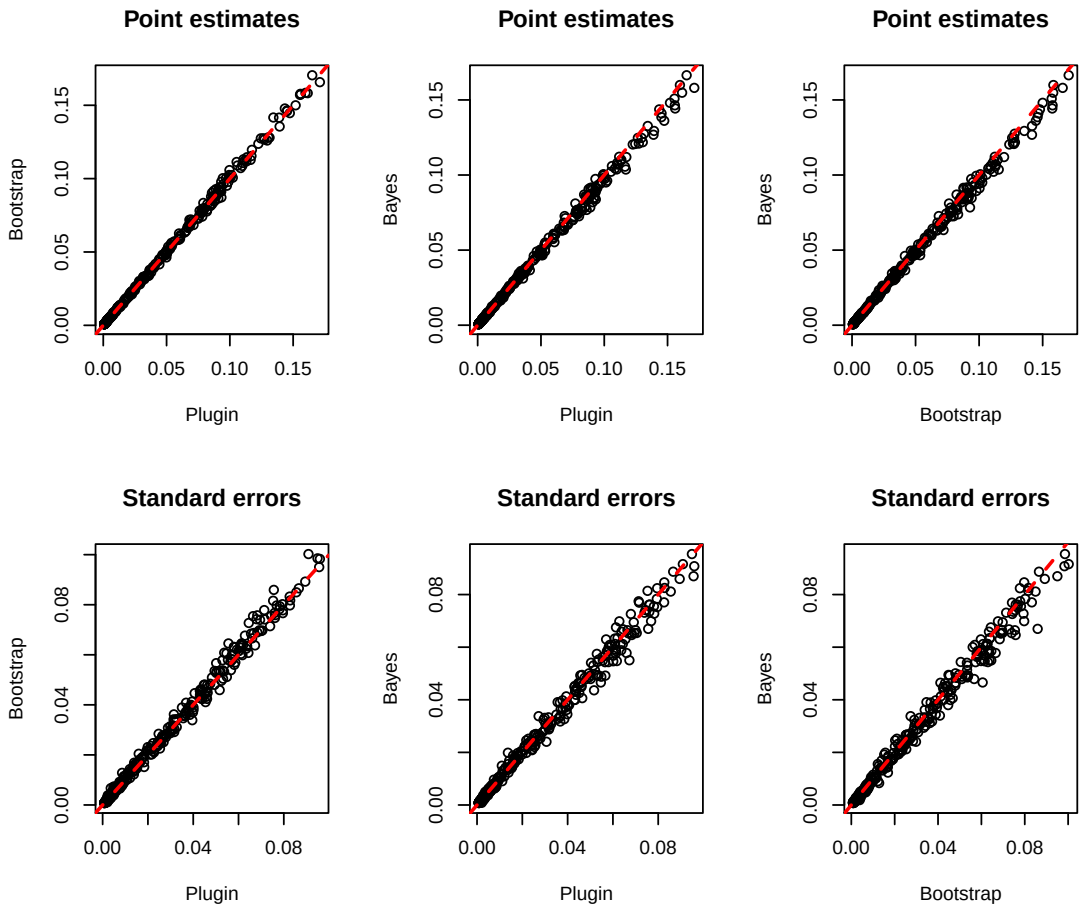

Figure 7: Scatter plots of the point estimates (upper panels) and standard errors (lower panels) of *Plasmodium falciparum* prevalence for children between 2 and 10 years of age, using plugin, parametric bootsptrap and Bayesian methods. The dashed red lines in each panel is the identity line.

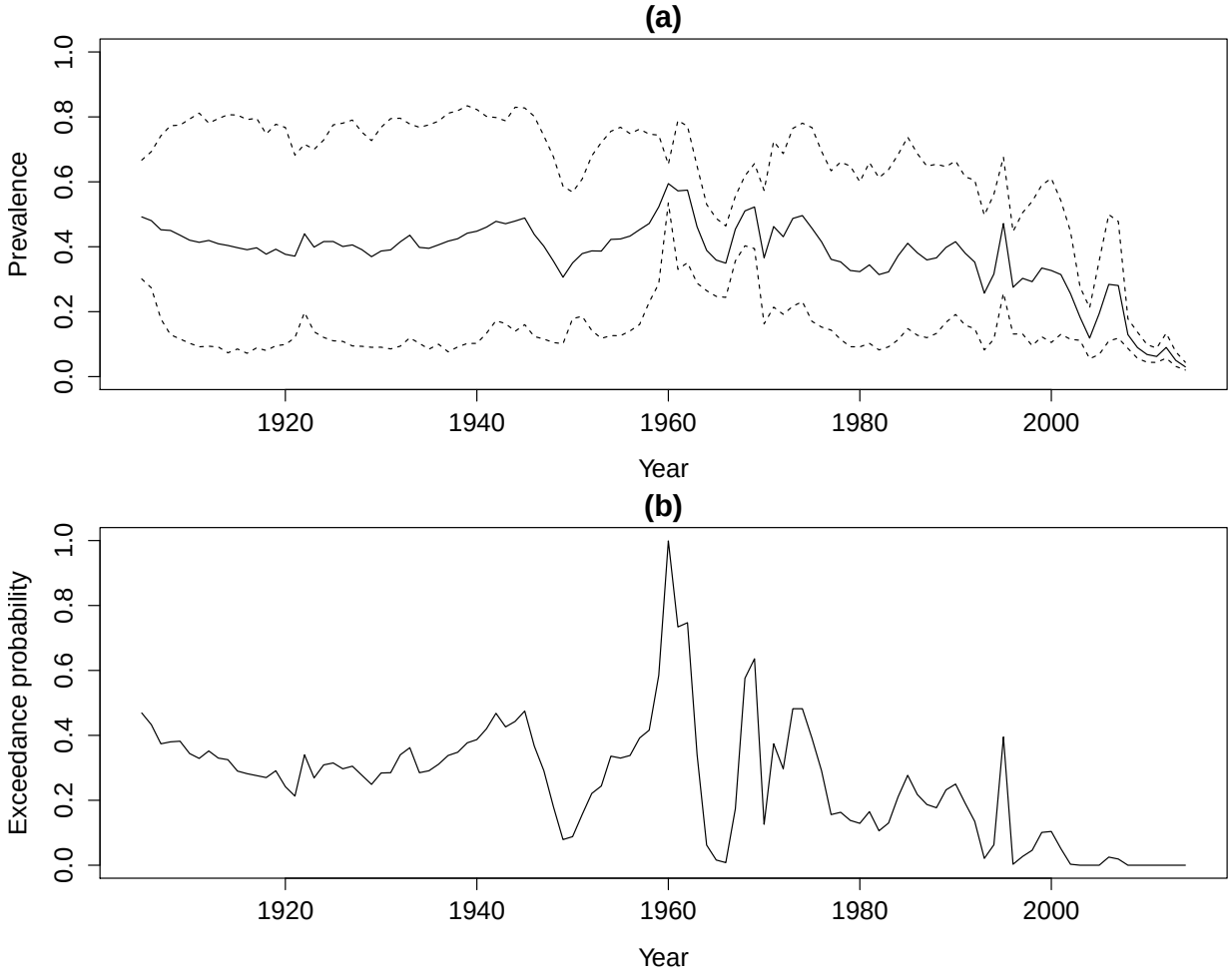

Figure 8: (a) Predictive mean (solid line) of the country-wide average prevalence with 95% predictive intervals. (b) Predictive probability of the country-wide average prevalence exceeding a 50% threshold.

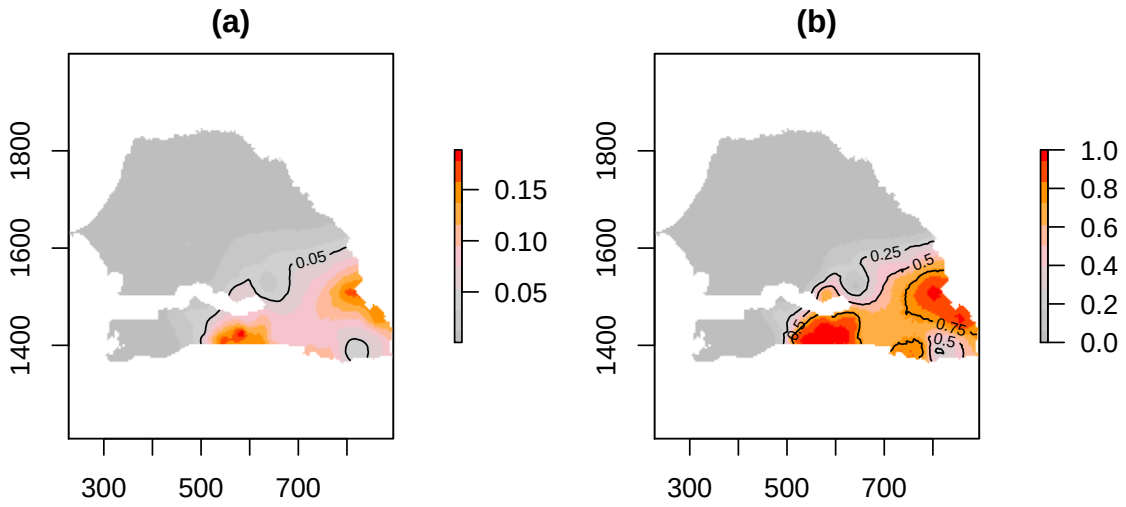

Figure 9: (a) Predictive mean surface of prevalence for children between 2 and 10 ($PfPR_{2-10}$); (b) Exceedance probability surface for a threshold of 5% $PfPR_{2-10}$. Both maps are for the year 2014. The contour lines correspond to 5% $PfPR_{2-10}$, in the left panel, and to 25%, 50% and 75% exceedance probability, in the right panel.

Figure 8(a) shows point and interval predictions of average country-wide $PfPR_{2-10}$. We observe a steady decline in $PfPR_{2-10}$ in the most recent decade. The highest predicted value of $PfPR_{2-10}$ across the whole of the time series occured in 1960, the year in which Senegal gained independence from France. Figure 8(b) shows for each year the predictive probability that average country-wide $PfPR_{2-10}$ exceeded 5%. Figure 9 shows the surfaces of the predictive mean (left panel) and the preditive probability that prevalence exceeds 5% prevalence (right panel), for the year 2014. In the right panel, we can identify two disjoint areas in the south-west of Senegal, where the probability of exceeding 5% $PfPR_{2-10}$ is at least 75%. In areas between the contour of 50% and 75% exceedance probability we are less confident that $PfPR_{2-10}$ exceeds 5%. These aspects relating to the uncertainty about the 5% threshold cannot be deduced from the map of prevalence estimates in the left panel, nor would a map of pointwise prediction variances be of much help.

5 Discussion

We have developed a statistical framework for the analysis of spatio-temporally referenced data from repeated cross-sectional prevalence surveys. Our aim was to provide a set of tools and principles that can be used to identify a parsimonious geostatistical model that is compatible with the data. In our view, model validation should include checking the validity of the specific assumptions made on $S(x,t)$ rather than be focused exclusively on predictive performance, so as to avoid the risk of attaching spurious precision to predictions from an inappropriate model.

The variogram is very widely used in geostatistical analysis. We use it both for exploratory analysis and model validation, but favour likelihood-based methods, whether non-Bayesian or Bayesian, for parameter estimation and formal model comparison; an example of the latter is our use of the profile deviance to justify fitting a model with separable correlation structure to the Senegal malaria data.

In our spatio-temporal analysis of historical malaria prevalence data from Senegal, we have shown how to incorporate parameter uncertainty within a likelihood-based framework by approximation of the distribution of the maximum likelihood estimator using the Gaussian approximation and parametric bootstrap. The results showed that the Gaussian approximation provides reliable numerical inferences for the regression coefficients but was slightly inaccurate for the log-transformed covariance parameters. For this reason, we generally recommend using parametric bootstrap whenever this is computationally feasible. In our view, this gives a viable approach to handling parameter uncertainty in predictive inference without requiring the specification of so-called non-informative priors. Non-Bayesian and Bayesian approaches showed some differences with respect to parameter estimation, but delivered almost identical point predictions and predictive standard deviations for the spatial estimates of prevalence. Our results also illustrate how even large geostatistical data-sets often lead to disappointingly imprecise inferences about model parameters. For this reason, we woild favour Bayesian inference when, and only when, an informative prior can be specified from contextually based expert prior knowledge of the process under investigation.

In Section 3.3, we discussed how to extend the standard model for prevalence data in order to let the model parameters change over time, space or both. However, the use of these models requires a large amount of the data and good spatio-temporal coverage so as to detect non-stationary patterns in prevalence. In the Senegal malaria application application the spatio-temporal sparsity of the sampled locations meant that the data could not be used to reliably detect spatio-temporal variation in the covariance parameters. For this application we also assumed that the sampling locations did not arise from a preferential sampling scheme. The standard geostatistical model for prevalence can also be extended to account for preferentiality in the sampling design, based on the framework developed by Diggle et al. (2010). However, such a model would require a larger amount of data than was available for this application.

Our analysis included data from the Demographic and Health Survey (DHS) conducted in Senegal in 2014. These data were collected using a two-stage stratified sampling design (ANSD, 2015). In the first stage, 200 census districts (CDs) are randomly selected, 79 among urban CDs and 121 among rural CDs, with probability proportional to the population size. In the second stage, an enumeration list from each CD was used to sample households randomly. In the analysis reported above, we could not account for the sampling design of the DHS data because of the lack of information on urban and rural extents for every single year when the surveys were conducted. However, since this variable is available for 2014, we extracted the DHS data and fitted two geostatistical models with and without an explanatory variable that classifies every location as rural or urban. Figure 11 shows the plots for the estimated prevalence and associated standard errors obtained from the two models. The differences both in the point estimates and standard error of prevalence are negligible. Hence, we do not expect the sampling design adopted in the DHS survey to affect the results reported in Section 4.

In model (2), spatial confounding can arise when some of the variation in prevalence due to to the effect of spatially structured risk factors $d(x,t)$ is attributed by the model to the stochastic process $S(x,t)$. This phenomenon affects the interpretation of the regression parameters β; see, for example, Paciorek (2010) and Hodges & Reich (2010). However, the following argument supports our experience that it has a negligible impact on predictive inference for $p(x,t)$. Consider, for simplicity, the following purely spatial model,

$$\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = \beta_0 + \beta_1 D_1(x_i) + \beta_2 D_2(x_i) + S(x_i). \tag{27}$$

If both of $D_1(x)$ and $D_2(x)$ are observed, fitting the model (27) with $D_1(x)$ and $D_2(x)$ as covariates, i.e. conditioning on both $D_1(x)$ and $D_2(x)$, would lead to consistent estimation of β_1 and β_2. If only $D_1(x)$ is observed, we can only condition on $D_1(x)$. Now assume that $D_2(x) = T(x) + D_1(x)$, with $S(x)$ and $T(x)$ independent processes, and re-express (27) as

$$\begin{aligned}\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} &= \beta_0 + \beta_1 D_1(x_i) + \beta_2\{T(x_i) + D_1(x_i)\} + S(x_i) + Z(x_i) \\ &= \beta_0 + \beta_1^* D_1(x_i) + S^*(x_i) \end{aligned} \tag{28}$$

where $\beta_1^* = \beta_1 + \beta_2$ and $S^*(x) = S(x) + \beta_2 T(x)$. Provided that we correctly specify the model for $S^*(x)$, conditioning on $D_1(x)$ will lead to consistent estimation of β^*, which is all that we require for prediction of $p(x)$. Now suppose that $T(x)$ and $S(x)$ are Matérn processes, but we specify $S^*(x)$ to be a Matérn process. This is incorrect, but we conjecture that it is a good approximation. Figure 10 shows an example in which $\beta_2 = 1$ and $S(x)$ and $T(x)$ have Matérn covariance functions with unit variance, scale parameters 0.1 and 0.07 and smoothness parameters 0.5 and 2.5, respectively. The resulting correlation function of $S^*(x)$ is $f_1(u) = 0.5\{\mathcal{M}(u;0.1,0.5) + \mathcal{M}(u;0.07,2.5)\}$, which can be closely approximated by a single Matérn, $f_2(u) = \mathcal{M}(u;0.109,0.774)$, where $\mathcal{M}(\cdot;\phi,\kappa)$ is a Matérn correlation function with scale parameter ϕ and smoothness parameter κ.

For large data-sets, it may be necessary to use an approximation of the spatio-temporal Gaussian process $S(x,t)$ in order to make inference computationally feasible. One such approach is to use a low-rank approximation (Higdon, 1998, 2002) in which $S(x,t)$ is represented as a finite linear combination of basis functions with radom coefficients; see, for example, Rodrigues & Diggle (2010) who develop a class of non-separable spatio-temporal covariance functions using this approach. Another approach is to formulate $S(x,t)$ as the solution to a stochastic partial differential equation (SPDE). Lindgren et al. (2011) develop a general framework for this approach, in which Gaussian Markov random fields are used to obtain a computationally fast solution to a discretised version of the defining SPDE. In the case of binary data, the computational burden can also be reduced by using data augmentation sampling schemes (Holmes & Held, 2006).

Throughout the paper, we have assumed that the process $S(x,t)$ is isotropic. To diagnose anisotropy, a directional version of the variogram can be used, in which inter-point distances u are replaced by vector differences $x_i - x_j$ and the results displayed as a three-dimensional scatterplot at each time-lag. Weller & Hoeting (2016) provides a comprehensive survey of non-parametric diagnostic methods used to test specific deviations from the assumption of

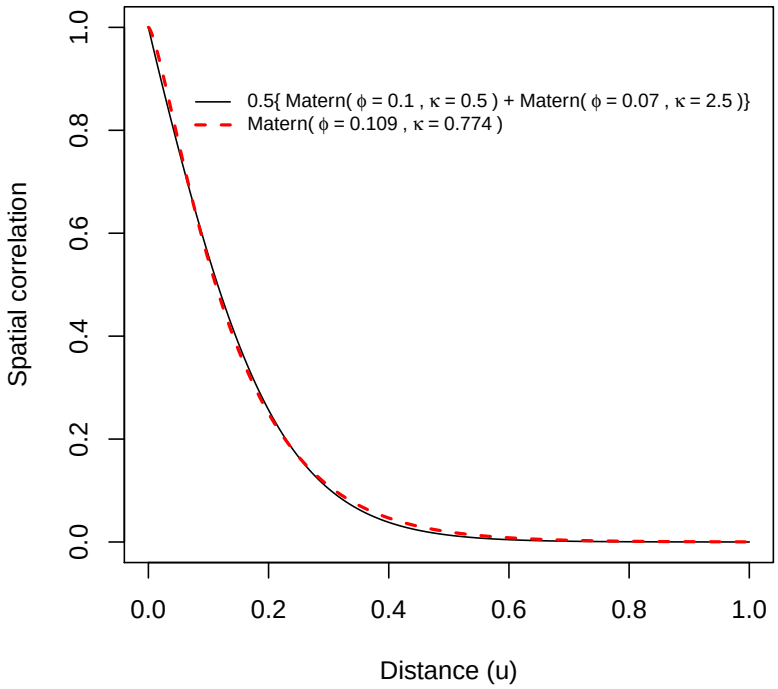

Figure 10: The solid curve corresponds to the function $f_1(u) = 0.5\{\mathcal{M}(u; 0.1, 0.5) + \mathcal{M}(u; 0.07, 2.5)\}$ and the red dashed curve to $\mathcal{M}(u; 0.109, 0.774)$, where $\mathcal{M}(\cdot; \phi, \kappa)$ is a Matérn correlation function with scale parameter ϕ and smoothness parameter κ.

isotropy. A limitation of most of these methods is that they require the spatial process to be observed either on a grid or on a relaisation of a homogeneous Poisson process. Additionally, the properties of these tests have only been investigated when the response is continuous. The sample size required to obtain adequate power is likely to be higher in the case of binomial data.

In addition to the sampling designs that we discussed in Section 2, cluster sampling is another cost-effective alternative to simple random sampling. In households surveys, a cluster might correspond to a geographically restricted area, e.g. a village or group of households, which are randomly selected in a first stage. One of the potential, but still unexplored, uses of this sampling design in disease mapping would be to disentangle the long-range and small-range spatial variation in disease risk. To pursue this objective the nugget component $Z(x_i, t_i)$ in (2) could be modelled as an additional Gaussian process whose scale of spatial correlation is constrained to be smaller than that of $S(x_i, t_i)$. Separating these two spatial scales of correlation would require a large amount of data and would be dependent on the spatial arrangement of the clusters.

We have not considered issues of data-quality variation across multiple surveys. This has been addressed by (Giorgi et al., 2015), who developed a multivariate geostatistical model to combine prevalence data from multiple randomised and non-randomised surveys. Incorporation of this modelling framework into the methods of Section 3 would be straightforward given the required data, since all the different stages of the analysis can still be carried out using the same tools and principles.

Acknowledgements

EG holds an MRC Strategic Skills Fellowship in Biostatistics (MR/M015297/1). RWS is funded as a Principal Fellow by the Wellcome Trust, UK (No. 079080 and 103602) and is grateful to the UKs Department for International Development for their continued support to the project *Strengthening the Use of Data for Malaria Decision Making in Africa first*, funded and piloted in 2013 (DFID Programme Code No. 203155). AMN acknowledges support from the Wellcome Trust as an Intermediary Fellow (No. 095127).

References

ANSD (2015). *Séngal : Enquête Démographique et de SantContinue (EDS-Continue 2014)*. Rockville, Maryland, USA : Agence Nationale de la Statistique et de la Démographie and ICF International.

Bennett, A., Kazembe, L., Mathanga, D., Kinyoki, D., Ali, D., Snow, R. & Noor, A. M. (2013). Mapping malaria transmission intensity in malawi, 2000-2010. *American Journal of Tropical Medicine and Hygiene* **89**, 840–849.

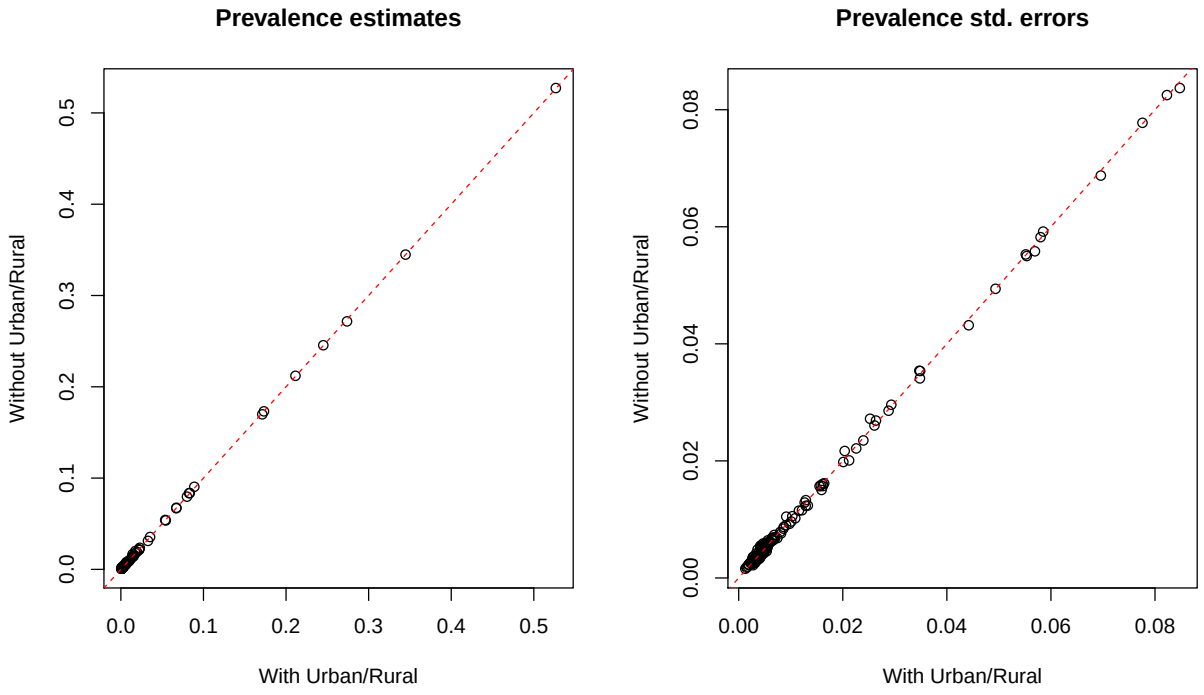

Figure 11: Prevalence estimates (left panel) and standard errors (right panel) based on the Demographic and Health Survey conducted in Senegal in 2014. Those are obtained from a model using a spatial indicator for urban and rural communities (x-axis) and excluding this explanatory variable (y-axis). The dashed line in both graphs is the identity line.

Bonat, W. H. & Ribeiro, P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics* **27**, 83–89. Env.2375.

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

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics* **3**, 702–718.

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

Diggle, P. J. & Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource setting (with discussion). *Journal of the American Statistical Association* DOI: 10.1080/01621459.2015.1123158.

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

Diggle, P. J., Moyeed, R., Rowlingson, B. & Thomson, M. (2002). Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society, Series C* **51**, 493–506.

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

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

Fletcher, R. (1987). *Practical methods of optimization*. John Wiley & Sons, New York, 2nd ed.

Fong, Y., Rue, H. & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11**, 397.

Gething, P. W., Elyazar, I. R. F., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., Patil, A. P., Tatem, A. J., Howes, R. E., Myers, M. F., George, D. B., Horby, P., Wertheim, H. F. L., Price, R. N., Meller, I., Baird, J. K. & Hay, S. I. (2012). A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Neglected Tropical Diseases* **6**, e1814.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B* **56**, 261–274.

Geyer, C. J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds. London: Chapman and Hall, pp. 241–258.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In *Stochastic Geometry, Likelihood and Computation*, O. E. Barndorff-Nielsen, W. S.Kendall & M. N. M. van Lieshout, eds. Boca Raton, FL: Chapman and Hall/CRC, pp. 79–140.

Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B* **54**, 657–699.

Giorgi, E. & Diggle, P. J. (2017). Prevmap: an R package for prevalence mapping. *Journal of Statistical Software* **78**, 1–29. DOI:10.18637/jss.v078.i08.

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

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590–600.

Hansell, A. L., Beale, L. A., Ghosh, R. E., Fortunato, L., Fecht, D., Järup, L. & Elliott, P. (2014). *The Environment and Health Atlas for England and Wales.* Oxford University Press.

Hay, S. I., Guerra, C. A., Gething, P. W., Patil, A. P., Tatem, A. J., Noor, A. M., Kabaria, C. W., Manh, B. H., Elyazar, I. R. F., Brooker, S., Smith, D. L., Moyeed, R. A. & Snow, R. W. (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine* **6**, e1000048.

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

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

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

Hodges, J. S. & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64**, 325–334.

Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis* **52**, 5066–5074.

Kabaghe, A. N., Chipeta, M. G., McCann, R. S., Phiri, K. S., van Vugt, M., Takken, W., Diggle, P. & Terlouw, A. D. (2017). Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural malawi. *PLOS ONE* **12**, 1–14.

Kleinschmidt, I., Pettifor, A., Morris, N., MacPhail, C. & Rees, H. (2007). Geographic distribution of human immunodeficiency virus in South Africa. *The American journal of tropical medicine and hygiene* **77**, 1163–1169.

Kleinschmidt, I., Sharp, B. L., Clarke, G. P. Y., Curtis, B. & Fraser, C. (2001). Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa. *American Journal of Epidemiology* **153**, 1213–1221.

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

López-Abente, G., Ramis, R., Pollán, M., Aragonés, N., Pérez-Gómez, B., Gómez-Barroso, D., Carrasco, J. M., Lope, V., Garciá-Pérez, J., Boldo, E. & García-Mendizábal, M. J. (2007). *Atlas Municipale de Mortalidad por Cancér en Espana 1989-1998.* Madrid: Instituto de Salud Carlos III.

Lumley, T. & Scott, A. (2017). Fitting regression models to survey data. *Statistal Science* **32**, 265–278.

Matérn, B. (1986). *Spatial Variation.* Springer, Berlin, 2nd ed.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H. & Clark, S. (2015). Spacetime smoothing of complex survey data: Small area estimation for child mortality. *Ann. Appl. Stat.* **9**, 1889–1905.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds., chap. 5. Chapman & Hall, CRC Press, pp. 113–162.

Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., Fall, I. S. & Snow, R. W. (2014). The changing risk of plasmodium falciparum malaria infection in africa: 200010: a spatial and temporal analysis of transmission intensity. *The Lancet* **383**, 1739 – 1747.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science* **25**, 107–125.

Pati, D., Reich, B. J. & Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **98**, 35–48.

Pullan, R. L., Gething, P. W., Smith, J. L., Mwandawiro, C. S., Sturrock, H. J. W., Gitonga, C. W., Hay, S. I. & Brooker, S. (2011). Spatial modelling of soil-transmitted helminth infections in Kenya: A disease control planning tool. *PLoS Neglected Tropical Diseases* **5**, e958.

Raso, G., Matthys, B., N'goran, E. K., Tanner, b., Vounatsou, P. & Utzinger, J. (2005). Spatial risk prediction and mapping of schistosoma mansoni infections among schoolchildren living in western Côte d'Ivoire. *Parasitology* **131**, 97–108.

Roberts, G. O. & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 255–268.

Rodrigues, A. & Diggle, P. J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics* **37**, 553–567.

RStudio, Inc (2013). *Easy web applications in R.* `http://www.rstudio.com/shiny/`.

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

Skinner, C. & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science* **32**, 165–175.

Snow, R., Amratia, P., Mundia, C., Alegana, V., Kirui, V., Kabaria, C. & Noor, A. (2015a). Assembling a geo-coded repository of malaria infection prevalence survey data in Africa 1900-2014. Tech. rep. INFORM Working Paper, developed with support from the Department of International Development and Wellcome Trust, UK, June 2015. Avilable at `http://www.inform-malaria.org/wp-content/uploads/2015/07/Assembly-of-Parasite-Rate-Data-Version-1.pdf`.

Snow, R. W., Kibuchi, E., Karuri, S. W., Sang, G., Gitonga, C. W., Mwandawiro, C., Bejon, P. & Noor, A. M. (2015b). Changing malaria prevalence on the kenyan coast since 1974: Climate, drugs and vector control. *PLoS ONE* **10**, 1–14.

Soares Magalhaes, R. J. & Clements, A. C. A. (2011). Mapping the risk of anaemia in preschool-age children: The contribution of malnutrition, malaria, and helminth infections in West Africa. *PLoS Medicine* **8**, e1000438.

Stein, M. L. (2005). Space: Time covariance functions. *Journal of the American Statistical Association* **100**, 310–321.

Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M. & Greenwood, B. (1999). Predicting malaria infection in gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results. *The American Journal of Tropical Medicine and Hygiene* **61**, 2–8.

Weller, Z. D. & Hoeting, J. A. (2016). A review of nonparametric hypothesis tests of isotropy properties in spatial data. *Statistical Science* **31**, 305–324.

Xie, Y. (2013). animation: An R package for creating animations and demonstrating statistical methods. *Journal of Statistical Software* **53**, 1–27.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58**, 129–136.

Zouré, Honorat, G. M., Noma, M., Tekle, Afework, H., Amazigo, U. V., Diggle, P. J., Giorgi, E. & Remme, J. H. F. (2014). The geographic distribution of onchocerciasis in the 20 participating countries of the african programme for onchocerciasis control: (2) pre-control endemicity levels and estimated number infected. *Parasites & Vectors* **7**.