

“Spatial Modeling, with Application to Complex Survey Data”.
Discussion of “Model-based Geostatistics for Prevalence Mapping in
Low-Resource Settings”, by Diggle and Giorgi

Jon Wakefield^{1,2}, Daniel Simpson³ and Jessica Godwin¹

¹Department of Statistics, University of Washington

²Department of Biostatistics, University of Washington

³Department of Mathematical Sciences, University of Bath

1 Introduction

It is a pleasure to discuss the interesting and wide-ranging article of Diggle and Giorgi (2016), henceforth referred to as DG. Prevalence mapping in low resource settings is an increasingly important endeavor to guide policy making and to spatially and temporally characterize the burden of disease. We will focus our discussion on consideration of the complex design when analyzing survey data, and on spatial modeling. With respect to the former, we consider two approaches: direct use of the weights, and a model-based approach using a spatial model to acknowledge clustering. The first of these is considered in Section 2. With respect to spatial modeling we describe, in Section 3, the stochastic partial differential equations (SPDEs, Lindgren et al. 2011) approach to modeling. Throughout, we use the integrated nested Laplace approximation (INLA, Rue et al. 2009) to perform computation. In general, a spatial target of interest may be associated with a point or an area, and in Section 4 we describe how inference can be made for area averages, as well as for probabilities of exceedance of a threshold, both using INLA. A simulation to present the power of the INLA/SPDE approach is provided in Section 5. We conclude with final remarks in Section 6.

2 Surveys with a Complex Design

In the developing world, it is often the case that disease indicators are collected via complex survey designs. For example, Demographic Health Surveys (DHS) are nationally-representative household surveys that are carried out extensively in the developing world and typically use a stratified two- or three-stage cluster design (Corsi et al., 2012). Hence, the data are available with accompanying weights and a randomization-(or design-)based approach to inference is common (for a very readable introduction to the analysis of survey data, including randomization-based inference, see Lohr, 2010). In the model-based approach to the analysis of complex survey data (Gelman, 2007), one accounts for the sampling scheme by including the

design (e.g., stratification) variables in a regression model. Unfortunately, it is not uncommon for these variables to be unavailable. An alternative approach (Chen et al., 2014; Mercer et al., 2014, 2016) takes the (asymptotic) sampling distribution of a weighted estimator, such as the Horvitz-Thompson (Horvitz and Thompson, 1952) or Hájek (Hájek, 1971) estimator, as the likelihood and then smooths across space and time. In practice, often the survey design is ignored in prevalence mapping (DHS, 2014; Bhatt et al., 2015).

In DG, data from a number of different surveys are analyzed. The rolling malaria indicator survey (rMIS) has a design in which households are randomly selected with a household sampled with probability proportional to village size (Roca-Feltrer et al., 2012). School survey data (Stevenson et al., 2013) are also analyzed by DG; these data do not arise from a standard design, with an iterative process being used for school selection, to limit the chance of overlapping school catchment areas.

We briefly outline the hierarchical modeling approach described in Mercer et al. (2014, 2016). Let p_k be the unknown prevalence associated with area k , and let \hat{p}_k be the design-based (weighted) estimator of this prevalence with associated estimated design-based variance \hat{V}_k^* , $k = 1, \dots, K$. The summaries $\{\hat{p}_k, \hat{V}_k^*, k = 1, \dots, K\}$ may be obtained using standard software, for example, in Section 5 we use the **survey** package (Lumley, 2004) in R. The **svyby** function in this package allows the mean prevalence in a region to be estimated, using all of the data collected in that region. The “data” are then taken as $y_k = \log\left(\frac{\hat{p}_k}{1-\hat{p}_k}\right)$ and the asymptotic variance of y_k is obtained, from \hat{V}_k^* , via the delta method, and is denoted \hat{V}_k . If the weighted prevalence estimates are 0 or 1, a fix is required; for example, empirical Bayes may be used. The first stage of the hierarchy is taken as

$$y_k | \eta_k \sim_{iid} N\left(\eta_k, \hat{V}_k\right), \quad (1)$$

and smoothing models over space can then be applied to η_k to alleviate instability due to small samples, a standard approach in small area estimation (SAE). The use of a normal likelihood based on the empirical logit was used, in a non-complex survey setting, by Stanton and Diggle (2013), though with a constant variance. This model is straightforward to fit in R, since we can use INLA with a fixed and known variance at Stage 1 of the hierarchy, as in (1). Mercer et al. (2016) demonstrate the use of this model when modeling under-5 mortality in Tanzania and we present an example in Section 5.

3 Building Appropriate Spatial Models

We now consider a continuous space Gaussian random field (GRF) model at spatial location \mathbf{x}_i , with i indexing points at which responses were measured, $i = 1, \dots, N$, so that N is the number of data points. For the moment we keep things general, and assume a linear predictor of the form

$$\eta_i = \beta_0 + \mathbf{z}(\mathbf{x}_i)\boldsymbol{\beta} + S(\mathbf{x}_i) + \epsilon_i,$$

where $\mathbf{z}(\mathbf{x}_i)$ are covariates measured at the spatial location \mathbf{x}_i , with associated regression coefficients β , $\epsilon_i \sim_{iid} \mathcal{N}(0, \sigma_\epsilon^2)$ is measurement error (aka the nugget term), and $S(\mathbf{x}_i)$ represents a spatial GRF.

Many possibilities are available for the form of the covariance function of the GRF but Stein (1999) (amongst others) makes a strong argument for a Matérn function:

$$\text{cov}[S(\mathbf{x}), S(\mathbf{x}^*)] = \frac{\sigma_s^2}{2^{v-1}\Gamma(v)} (\kappa \|\mathbf{x} - \mathbf{x}^*\|)^v K_v(\kappa \|\mathbf{x} - \mathbf{x}^*\|)$$

where $K_v(\cdot)$ is a modified Bessel function of the second kind, $\sigma_s^2 > 0$ is the marginal variance, $\kappa > 0$ a scale parameter and $v > 0$ a smoothness parameter. When $v + 1$ is an integer, in two spatial dimensions the Matérn fields are Markovian (Rozanov, 1977). Even in this latter case, data analysis that uses the covariance function directly is computationally difficult because of the expensive matrix operations that are required (Rue and Held 2005, Chapter 2).

For modeling the spatial effect, DG use Higdon’s convolution kernel approach (Higdon, 1998). In order to control the computational complexity inherent in the classical spatial model, the general GRF is replaced with a finite dimensional (or, in their terminology, low-rank) model:

$$S_n(\mathbf{x}) = \sum_{i=1}^n w_i \phi_i(\mathbf{x}), \quad (2)$$

where the joint distribution of the weights $\mathbf{w} = [w_1, \dots, w_n]^T$ is multivariate Gaussian and the deterministic basis functions $\{\phi_i(\mathbf{x})\}_{i=1}^n$ may depend on some parameters being inferred. The underlying principle is that these finite dimensional random fields will be reasonable proxies for the true latent spatial surface. The advantage of the finite dimensional representation is that inference costs grow like $\mathcal{O}(n^2N + n^3)$ which, for sufficiently small n , is significantly smaller than the $\mathcal{O}(N^3)$ cost of classical methods. Furthermore, if the basis functions $\phi_i(s)$ are only non-zero in a small part of the domain, the cost is reduced to $\mathcal{O}(N + n^3)$ —or $\mathcal{O}(N + n^{3/2})$ for Markovian models—and the method genuinely grows linearly in the number of basis functions (Simpson et al., 2012a). From this point of view, it is clear that kernel methods (Higdon, 1998), predictive processes (Banerjee et al., 2008), fixed rank Kriging (Cressie and Johannesson, 2008), and the SPDE method (Lindgren et al., 2011) (and for that matter, classical methods like truncated Karhunen-Loève expansions) are all different faces of the same underlying concept. The differences between these methods manifest in the way the basis functions and the weights are chosen. As one would expect, different choices endow these methods with different sets of advantages and disadvantages (Bradley et al., 2015).

A particular point that we want to emphasize is that the choice of the spatial random field model is not an innocuous one and this choice will filter through into estimates of uncertainty (be they constructed in a Bayesian way or not). In the case where we are interested in predictions at a single unmeasured location, a small forest of results exist on the behavior of spatial point predictions for GRFs under the regime in which the data are very close together

(infill asymptotics) or in which the data are collected on an expanding domain (Stein, 1999; Zhang and Zimmerman, 2005). Unfortunately, for the types of models and applications that DG consider, point estimation is not the only summary of interest. In addition, one is interested in estimates of total risk over an area and in locating areas that exceed a threshold; we consider such endeavors in Section 4.

Often, data has both a spatial and a temporal component (such as in the rMIS), and in this case the number of potential asymptotic regimes that we can use to justify our spatial or spatio-temporal model are dizzying. Even more challenging, is the idea that for many models the spatial field is designed to model the “residual” effect after the potentially non-linear effects of covariates are taken into account. To the best of our knowledge, the question of how to select the covariance structure of a GRF for the sorts of geostatistical generalized additive mixed models that are increasingly used in practice is completely unstudied.

In our view, the answer is to look for robustness. A little-appreciated fact is that finite-dimensional Gaussian random fields can be spectacularly robust against misspecification. Why? Because the non-robustness in GRFs is driven by the very fine-scale effects, which finite dimensional models necessarily discard. To see this, imagine there is a true value of the underlying spatial field $S^*(\mathbf{x})$, which we can write as

$$S^*(\mathbf{x}) = \sum_{i=1}^n w_i^* \phi_i(\mathbf{x}) + v^\perp(\mathbf{x}),$$

where $v^\perp(\mathbf{x})$ is orthogonal to $\{\phi_i(\mathbf{x})\}_{i=1}^n$. If the basis functions are chosen appropriately, all of the information in the data can go into estimating the main part of the field, which is modeled by the finite-dimensional GRF, while no assumptions are made about the “fine-scale” effects in $v^\perp(\mathbf{x})$, which are smoothed over. Hence, finite-dimensional GRFs will always get the bulk features right at the expense of the fine-scale ones. This is different to methods like covariance tapering (Furrer et al., 2006), which correctly resolve the fine-scale features necessary for optimal estimation of the field near already observed data points at the expense of resolving the large-scale features (Bolin and Lindgren, 2013).

This discussion gives a lot of insight into how we should choose our basis functions. In the analysis of the rMIS data in DG, for example, the features of interest were village-level prevalence, which suggests that the basis functions $\phi_i(\mathbf{x})$ should be designed to model features on a village scale. We note that this is slightly different from the suggestion of increasing the set of basis functions until inference stabilizes. We are instead suggesting one looks at the basis functions themselves to see if they can resolve the types of questions that are of interest. This will lead to very similar results, but is computationally much easier!

For point-referenced data, our preferred modeling strategy is the SPDE approach to spatial modeling, as originally described by Lindgren et al. (2011), and subsequently elaborated upon in Simpson et al. (2012a,b). Rather than choosing basis functions according to the convolution square root of the covariance function, as DG do, we instead focus on classes of functions with

good approximation properties.

We now give a brief description of the approach introduced by Lindgren et al. (2011) to approximate Matérn Markovian Gaussian random fields (MGRFs). The idea is to set up a fine triangular mesh, with m vertices, over the study area. A set of m piecewise linear basis functions $\phi_i(\mathbf{x})$ is then constructed, taking the value 1 at vertex i and 0 at all other vertices, $i = 1, \dots, m$. This gives a set of pyramids that are the building blocks for the approximation. A key point is that these pyramids are non-zero at only a small number of points. The MGRF is again represented by (2), with random Gaussian weights $\mathbf{w} = [w_1, \dots, w_n]^T$. The spatial prior under this model is therefore, in practice, over functions that are linear combinations of the pyramids (i.e., piecewise linear functions over the mesh). The flexibility in choosing the triangular mesh allows careful control of how well the spatial effect is resolved. The general idea is that features that are more than two triangles large are resolved very well, while those that are smaller than the triangle (such as the value of the field at a point) have a bias of the same order as the triangle size. For very precise versions of these results, we refer the interested reader to the technical appendices of Simpson et al. (2016).

The distribution of \mathbf{w} is still required, and is chosen to provide a good approximation to the MGRF. The primary difference between the SPDE approach and the fixed-rank Kriging approach of Cressie and Johannesson (2008), which also recommends using local basis functions chosen for their approximation properties, is the number of parameters that are allowed. While Cressie and Johannesson (2008) aim for a fully flexible model specified with $n(n-1)/2$ parameters, the SPDE approach focuses instead on a more parsimonious specification with, in the simplest case, only 2 parameters: the scale and the range. There are obviously computational advantages to this choice, as well as the parsimony of allowing more straightforward specification of meaningful prior distributions (Fuglstad et al., 2015b). The disadvantage is that the two-parameter model, which essentially corresponds to the assumption that the underlying model is isotropic, is that it may not be flexible enough to correctly model the residual spatial effect.

The MGRF that is to be approximated arises as the solution to the SPDE

$$(\kappa - \Delta)^{\alpha/2} S(\mathbf{x}) = \sigma_s W(\mathbf{x}), \quad (3)$$

where $\Delta = \frac{\partial}{\partial x_1^2} + \frac{\partial}{\partial x_2^2}$ is the Laplacian and $W(s)$ is white noise. The solution to (3) corresponds to a stationary GRF and if α is an integer the GRF is Markovian (Whittle, 1954), which is the key for implementation. Note that $\nu = \alpha - 1$ in the case of a two-dimensional field and $\alpha = 2$ is the default in INLA. In INLA the parameterization is $\theta_1 = \log \tau$, $\theta_2 = \log \kappa$ where

$$\tau^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)\kappa^{2\nu}\sigma_s^2}.$$

A solution to the SPDE satisfies, for any suitable function $\psi(\mathbf{x})$,

$$\int \psi(\mathbf{x})(\kappa - \Delta)^{\alpha/2} S(\mathbf{x}) d\mathbf{x} = \sigma_s \int \psi(\mathbf{x}) W(\mathbf{x}) d\mathbf{x},$$

with these functions are taken to be $\phi_i(\mathbf{x})$, $i = 1, \dots, m$. The use of these test functions leads to a system of linear equations to solve, and the solution produces the distribution of \mathbf{w} which with a little modification, is a Gaussian Markov random field (GMRF). For the missing details see Simpson et al. (2012b). The GMRF that we obtain for the distribution of the weights comes from two places: the fact that the Matérn form that is used is Markovian and the fact that the basis functions are only non-zero across a small portion of the space.

This prior is combined with the likelihood, with the spatial contribution being evaluated as a piecewise linear function of the GMRF at the data locations. Combining the data \mathbf{y} with the above prior gives a posterior on \mathbf{S} which is again of the form (2), but with the posterior distribution of the “weights” being $\mathbf{w}|\mathbf{y} \sim N(\mathbf{0}, \mathbf{Q}_{S|\mathbf{y}}^{-1})$. In practice, the evaluation of the likelihood at a particular data location turns out to be a weighted sum of the values of the GMRF on the nearest three vertices. The above strategy can be used with a wide range of likelihoods and the SPDE can also be extended to a variety of non-stationary models (Lindgren et al., 2011; Fuglstad et al., 2015a).

4 Area Averages and Excursions

One of the really enjoyable features of DG’s paper is their use of continuously specified Gaussian random fields even when the quantities of interest are areal averages. We broadly think this is a good idea. One concrete reason is that integrating risk over areas allows one to avoid ecological bias, if covariate information is available within areas (Wakefield, 2008). As the authors point out, however, using such fields is a computational challenge. Our favorite engine for overcoming computational challenges is the R-INLA package (Rue et al., 2009; Martins et al., 2013; Lindgren and Rue, 2015a). For three types of problem considered in the paper—estimating area-level prevalence, computing areas where prevalence is above a prescribed level, and using spatially-varying models of zero-inflation—R-INLA can be used to solve two of them (spatially varying models of zero inflation are beyond the functionality of R-INLA for fundamental software design reasons).

When DG say that INLA does not provide the joint predictive distribution for the latent field, they are both right and wrong. By default, INLA computes the univariate predictive distributions and in some cases this is sufficient. It also produces posterior distributions for linear combinations of the latent field, which means that the distribution of the average of the logit prevalence can be obtained. This is, unfortunately, not enough to compute the joint posterior distribution for area-level prevalences. Thankfully, the R-INLA package provides a mechanism for sampling from an approximation to the joint posterior distribution, which allows one to estimate the distribution of any functional of the latent field. The sampler works by noting that the posterior for the latent Gaussian component, which we will denote by $\boldsymbol{\eta}$, can

be approximated by

$$\tilde{\pi}(\boldsymbol{\eta} \mid \mathbf{y}) = \int_{\Theta} \pi_G(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta}) \tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}, \quad (4)$$

where the hyperparameters are denoted $\boldsymbol{\theta}$. In (4), $\pi_G(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta})$ is the Laplace approximation to the full conditional and $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ is the INLA approximation to the posterior of the hyperparameters (Rue et al., 2009). Although this is an, “integrated Laplace approximation”, the full INLA method proceeds from by using another Laplace approximation to approximate the marginal distributions $\pi(\eta_j \mid \mathbf{y})$. The simplest version of the INLA algorithm first constructs a Gaussian approximation to $\pi(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta})$ (alternatives includes simplified Laplace and Laplace approximations). This is then used to construct an approximation to $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, before calculating the approximation

$$\pi(\boldsymbol{\eta} \mid \mathbf{y}) \approx \sum_{i=1}^k w_i \tilde{\pi}(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta}_i) \tilde{\pi}(\boldsymbol{\theta}_i \mid \mathbf{y}),$$

where $\{(w_i, \boldsymbol{\theta}_i)\}_{i=1}^k$ are the weights and points of an integration scheme and $\tilde{\pi}$ stands for the appropriate approximation, that is computed in previous steps of the algorithm (Rue et al., 2009). The `inla.posterior.sample` functions compute a sample from this approximation to $\pi(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta})$ when $\tilde{\pi}(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta}_i)$ is the Gaussian distribution that matches the value and curvature of the true conditional distribution at its mode. To summarize, first the hyperparameters are sampled from the integration points, $\{\boldsymbol{\theta}_i, i = 1, \dots, k\}$, and then for each sampled hyperparameter a sample is taken from the Gaussian approximation to the latent field. The alert reader will note that this will not lead to the exact INLA approximation for the marginal distribution $\pi(\eta_j \mid \mathbf{y})$, which uses a further Laplace approximation to approximate the univariate conditional distribution $\pi(\eta_i \mid \mathbf{y}, \boldsymbol{\theta})$ before integrating out the parameter uncertainty. As these corrected approximations are readily available, they are used to correct the joint mean. A different option, that we have not yet explored, is to build a copula based around the multivariate approximation that matches the INLA marginals. Our experience is that this sampling algorithm is suitable as it stands, but should the user wish, they could use the returned log-density to compute an independence MCMC sampler that will be asymptotically exact.

The second inferential target that DG consider are excursion sets $\mathcal{A}_u^+ = \{\mathbf{x} : p(\mathbf{x}) > u\}$, where $p(\mathbf{x})$ is the probability of a binary event of interest and u is some fixed threshold. Excursion sets are subtle and difficult beasts that have been studied extensively in both the probability (Adler, 1981; Adler and Taylor, 2007) and statistics (Bolin and Lindgren, 2015; French and Hoeting, 2016) literatures. The reason these objects are so hard to study is straightforward: regardless of the statistical philosophy that is being used, the function $p(\mathbf{x})$ is random and hence the set \mathcal{A}_u^+ is a random variable. We would therefore want to find a (say) quantile of \mathcal{A}_u^+ , for example, the set A_u^+ such that

$$\Pr(p(\mathbf{x}) > u \text{ for all } \mathbf{x} \in A_u^+) > 1 - \alpha,$$

for some prescribed level α . DG’s approach to estimation is to construct the set $\tilde{A}_u^+ = \{\mathbf{x} : \Pr(p(\mathbf{x}) > u) > 1 - \alpha\}$. Since the set is constructed pointwise, this is clearly a multiple testing problem and, in general, the set \tilde{A}_u^+ will be too big. This is because the pointwise

tests do not take into account the fact that $p(\mathbf{x})$ is a continuous function and hence there is strong dependence between nearby tests (points): in order for a function value to be above a threshold with high probability (which is the usual case), all of the surrounding points also need to be above that threshold with high probability. This is similar to the reason that care must be taken when simultaneous bands are calculated for unknown functions using splines see, for example, Wakefield (2013, Section 11.2.7).

The situation is even more challenging in the cases that DG have considered due to the complicated sampling design, i.e., where the data are spatially located. In order to say with high probability that a point is above a given threshold, there needs to be a sufficiently large number of observations nearby to narrow down the pointwise uncertainty. Hence, when you have an inhomogeneous sampling design an excursion set isn't really enough to convey the full information about whether or not you are above a specific threshold. It would be more useful to divide the study area into three distinct regions: the upward excursion set A_u^+ ; the downward excursion set A_u^- , which is the set of all points such that $p(\mathbf{x}) < u$ with high probability; and the set of points that are in neither the upward nor the downward sets. This then acknowledges that under imperfect information, there are some areas of the space that you cannot with any certainty say are above or below the threshold. This type of target cannot be directly computed in R-INLA. Fortunately, David Bolin has written the excellent `excursions` package for R (Bolin and Lindgren, 2015), which contains a function for computing these regions using output from the R-INLA package. In the next section we illustrate the calculation of both area averages and exceedence probabilities.

5 Simulation

We now demonstrate the power of the SPDE approach as implemented within INLA. We simulate data within the geography of Kenya, using spatial locations for sampling that correspond to 400 points (enumeration areas) in the 2003 DHS.

For the simulation, we mimic some aspects of the DHS design with enumeration areas (EAs) assumed to be sampled (as first stage clusters) and then households (as second stage clusters) sampled within EAs. We let $i = 1, \dots, n = 400$ index the first stage clusters (the EAs), and $j = 1, \dots, m_i$, represent households sampled within clusters so that m_i is the number of households in first stage cluster i . Let N_{ij} represent the number of household members producing responses in EA i and household j , and Y_{ij} be the number of positive responses, $i = 1, \dots, n = 400$, $j = 1, \dots, m_i$. The sampling model is

$$Y_{ij}|p_{ij} \sim \text{Binomial}(N_{ij}, p_{ij}),$$

with

$$\text{logit } p_{ij} = \beta_0 + S_i + \epsilon_{ij},$$

where β_0 is the intercept (which relates to the overall log odds of prevalence), $S_i = S_i(\mathbf{x}_i)$ arises from a spatial model (which we take as an MGRF with variance and range parameters σ_s^2 and κ , respectively) and $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$ is a random effect that induces dependence between individuals in the same household, in cluster i . In the results we show below, we emphasize that we do not display/include the ϵ_{ij} terms in prevalence surfaces, as these are assumed to be household specific “noise”.

The prevalences were generated from a GRF with mean prevalence of 7%, so that $\beta_0 = \log(0.07/0.93)$. This mean prevalence was chosen based on the national prevalence of HIV in Kenya estimated in the Kenya DHS 2003. The other parameters of the GRF were taken as $\tau = e^{-1/2}$ and $\kappa = e^{1/2}$ with noise variance $\sigma_\epsilon^2 = 0.01$, in order to produce a prevalence field that approximately matched empirical HIV prevalence estimates from the Kenya DHS 2003 AIDS recode. The practical range, is sometimes defined as the distance at which the correlation drops to 0.13, and is given by $\sqrt{8v}/\kappa$ (Lindgren and Rue, 2015b), which equals 1.72 units here. The marginal variance is $\sigma_s^2 = 1/(4\pi) = 0.080$. The number of households in first stage clusters, m_i were taken from the set $(4, 5, \dots, 11)$ which is a truncated version of the range in the Kenya 2013 DHS. Denominators (household sizes) N_{ij} were sampled from a discrete distribution on $(1, 2, \dots, 12)$ also determined by the empirical distribution of the number of people tested per household in the Kenya DHS 2003 AIDS recode. For simplicity, we assume that there are 100 households in each EA.

We let π_i be the probability that cluster i is selected, and $\pi_{j|i}$ be the probability that household j is selected, given PSU i was selected, with $i = 1, \dots, I$ and $j = 1, \dots, m_i$ so that I is the total number of PSUs (which we take as 46,034, see the methods section of Linard et al. 2010) and m_i is the number of SSUs in PSU i . The design weight for all individuals within household j of cluster i are taken as the reciprocal of the two-stage cluster sample selection probabilities which are

$$\pi_{ij} = \pi_i \times \pi_{j|i} = \frac{400}{46034} \times \frac{m_i}{100},$$

where m_i is assumed to be the pre-chosen number of households to select from the 100 households in cluster i .

As an illustration of the calculation of area averages, we make inference for the prevalence at the level of ADM1 in Kenya, whose areas we index by k , $k = 1, \dots, 47$. For comparison, and to link with Section 2, in addition to the SPDE GRF, we also fit the hierarchical model in which the first stage is based on the design-based estimate of logit p_k , with an associated variance, for ADM1 area k . Letting y_k represent the logit of the weighted (Hájek) prevalence, we have $y_k|\eta_k \sim_{iid} N(\eta_k, \hat{V}_k)$, and

$$\eta_k = \beta_0^* + S_k + \epsilon_k, \tag{5}$$

where β_0^* is the area-level intercept, $\epsilon_k \sim_{iid} N(0, \sigma_\epsilon^2)$ are unstructured random effects and S_k are intrinsic conditional autoregressive (ICAR) random effects with variance σ_s^2 . Hence, we are using the popular Besag, York, Mollié (BYM) model Besag et al. (1991)

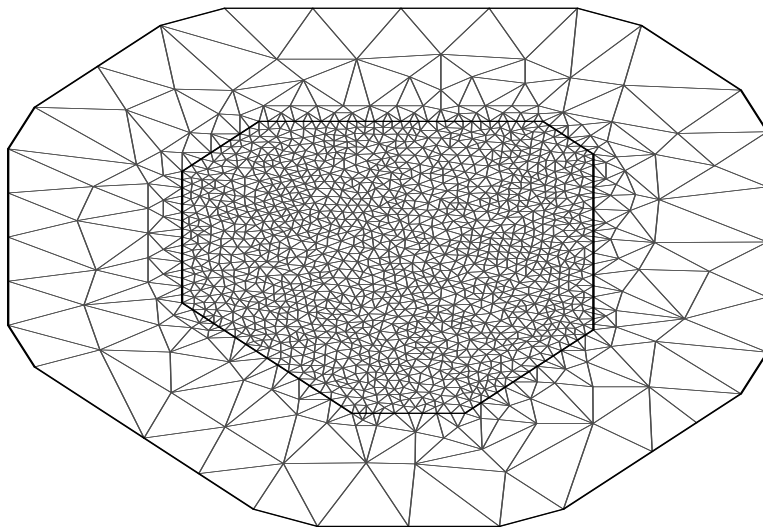


Figure 1: The mesh has two main features. The first is an inner section, in which the triangles are relatively fine. This is the area that we are most interested in. Outside of this inner area, the triangle rapidly become much larger as they get further from the area of interest. This structure mostly eliminates the boundary effects naturally associated with Markovian models.

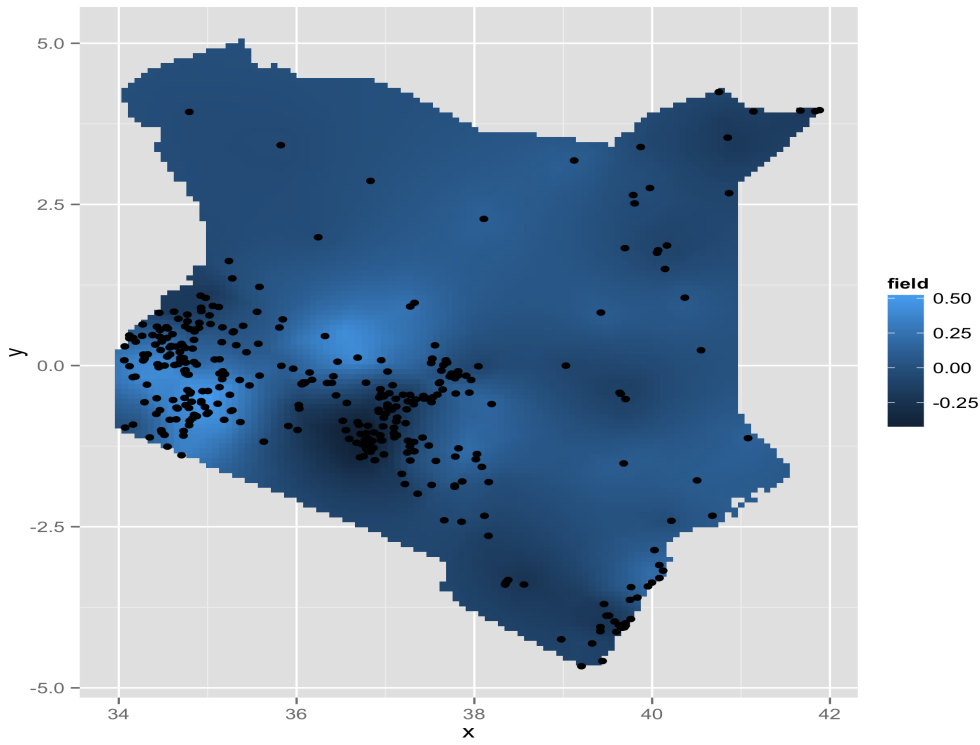


Figure 2: The median posterior spatial effect $S(\boldsymbol{x})$. The smoothness of the field reflects the relatively low amount of information in the data. In this simulation, points more than $\sqrt{8/\kappa} \approx 1.72$ units apart are essentially uncorrelated. Hence, in a large part of the space, the estimated spatial contribution is approximately equal to the prior median 0. At these points, the estimated prevalence is driven entirely by the country-level mean.

The SPDE mesh is shown in Figure 1 and in Figure 2 we display the posterior median of $S(\mathbf{x})|\mathbf{y}$, along with the locations at which samples were obtained. We calculate area-wide summaries of the area level averages,

$$T_k = \int_{A_k} \frac{\exp[\beta_0 + S(\mathbf{x})]}{1 + \exp[\beta_0 + S(\mathbf{x})]} d\mathbf{x}$$

where A_k represent the areas in ADM1, $k = 1, \dots, K$. Posterior means of T_k were constructed in R-INLA using Monte Carlo integration with points \mathbf{x}_{kj} , $j = 1, \dots, 100$, simulated in area A_k . Figure 3(a) displays the true values of T_k and panel (b) the posterior mean estimates obtained from the SPDE model using the Monte Carlo calculation. The posterior mean estimates from the design-based approach (including the independent and ICAR random effects) are displayed in Figure 3(c). Overall, the SPDE and smoothed design-based BYM estimates are quite similar, and display some attenuation as compared to the truth.

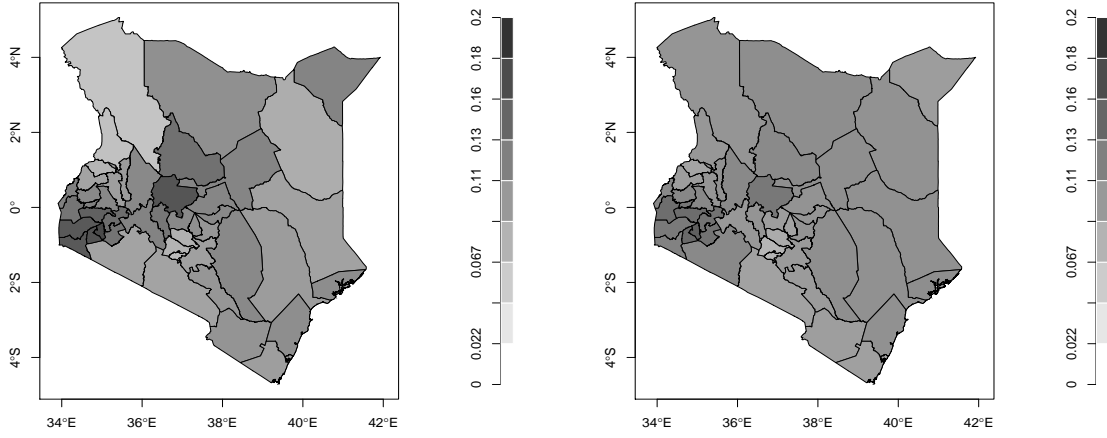
Figure 4 shows the estimate of the 95% excursion sets at the 7% prevalence level calculated using the `excursions.inla` function from the `excursions` package. The large blue area is such that $\Pr(p(\mathbf{x}) < 0.07 \text{ for every blue point}) > 0.95$, while the red areas show the points that simultaneously exceed the threshold. An interesting features of this figure are the black areas, in which there is not enough information to determine with 95% confidence whether the field is above or below the threshold.

6 Concluding Remarks

In Section 5 we considered a very simple situation in which the design was cluster sampling only. Often, stratification is present also (for example, typically in the DHS there is stratification by urban/rural and perhaps on other variables). In the case of a stratified cluster sampling design, then one may add fixed effects for each of the stratification levels. Post-stratification can also be addressed in the model-based framework (Gelman 2007, Gelman and Hill, 2007, Chapter 14). In addition, covariates can be included, though these need to be known at all locations (at least up to the resolution of the grid) for prediction. Code to reproduce the example in Section 5 can be found at <http://faculty.washington.edu/jonno/cv.html>.

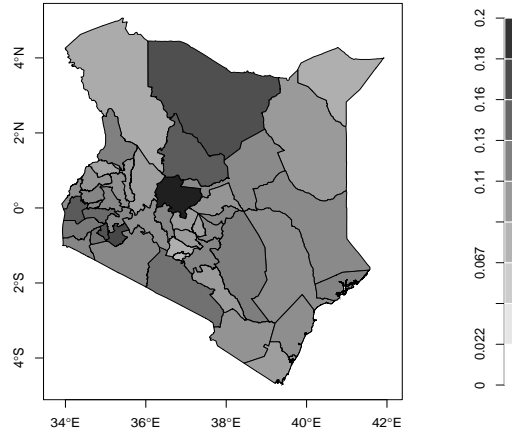
References

- Adler, R. (1981). *The Geometry of Random Fields*. Wiley, New York.
- Adler, R. J. and J. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer.
- Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B* 70, 825–848.



(a) True area average prevalence

(b) Monte Carlo estimate of average, SPDE



(c) Design-based estimate, ICAR

Figure 3: (a) True area-level prevalence averages, (b) estimated area-level prevalences averages from SPDE , (c) estimated area-level prevalences averages from smoothed design BYM model.

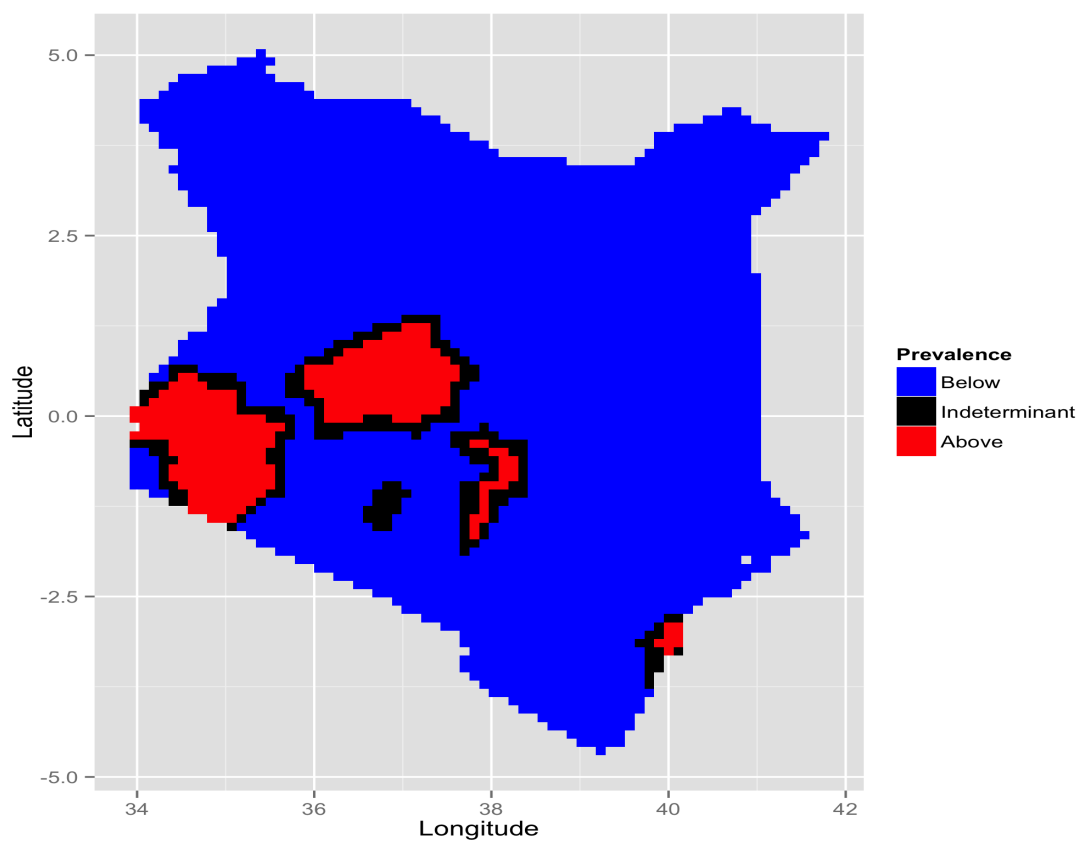


Figure 4: The joint posterior regions where the prevalence is simultaneously estimated to be below (blue) or above (red) 7% at a 95% confidence level. In the black areas, the results are indeterminate.

- Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* 43, 1–59.
- Bhatt, S., D. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. Battle, C. Moyes, A. Henry, P. Eckhoff, et al. (2015). The effect of malaria control on plasmodium falciparum in Africa between 2000 and 2015. *Nature* 526, 207–211.
- Bolin, D. and F. Lindgren (2013). A comparison between Markov approximations and other methods for large spatial data sets. *Computational Statistics and Data Analysis* 61, 7–32.
- Bolin, D. and F. Lindgren (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B* 77, 85–106.
- Bradley, J. R., N. Cressie, and T. Shi (2015). Comparing and selecting spatial predictors using local criteria. *TEST* 24, 1–28.
- Chen, C., J. Wakefield, and T. Lumley (2014). The use of sample weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-Temporal Epidemiology* 11, 33–43.
- Corsi, D. J., M. Neuman, J. E. Finlay, and S. Subramanian (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology* 41, 1602–1613.
- Cressie, N. A. C. and G. Johannesson (2008). Fixed rank Kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B* 70, 209–226.
- DHS (2014). Spatial interpolation with Demographic and Health Survey data: Key considerations. Technical report, ICF International, Rockville, Maryland, USA.
- Diggle, P. and E. Giorgi (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association*.
- French, J. P. and J. A. Hoeting (2016). Credible regions for exceedance sets of geostatistical data. *Environmetrics* 27, 4–14.
- Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015a). Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics* 14, 505–531.
- Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue (2015b). Interpretable priors for hyperparameters for Gaussian random fields. *arXiv preprint arXiv:1503.00256*.
- Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15, 502–523.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* 22, 153–164.

- Gelman, A. and J. Hill (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Hájek, J. (1971). Discussion of, “An essay on the logical foundations of survey sampling, part I”, by D. Basu. In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.
- Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics* 5, 173–190.
- Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663–685.
- Linard, C., V. A. Alegana, A. M. Noor, R. W. Snow, and A. J. Tatem (2010). A high resolution spatial population database of Somalia for disease risk mapping. *International Journal of Health Geographics* 9, 1.
- Lindgren, F. and H. Rue (2015a). Bayesian spatial and spatio-temporal modelling with R-INLA. *Journal of Statistical Software* 63.
- Lindgren, F. and H. Rue (2015b). Bayesian spatial modelling with r-inla. *Journal of Statistical Software* 63.
- Lindgren, F., H. Rue, and J. Linström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B* 73, 423–498.
- Lohr, S. (2010). *Sampling: Design and Analysis, Second Edition*. Boston: Brooks/Cole Cengage Learning.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software* 9.
- Martins, T. G., D. Simpson, F. Lindgren, and H. Rue (2013). Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis* 67, 68–83.
- Mercer, L., J. Wakefield, C. Chen, and T. Lumley (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics* 8, 69–85.
- Mercer, L., J. Wakefield, A. Pantazis, A. Lutambi, H. Mosanja, and S. Clark (2016). Small area estimation of childhood of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*. To appear.
- Roca-Feltrer, A., D. Lalloo, K. Phiri, and D. Terlouw (2012). Rolling malaria indicator surveys (rmis): a potential district-level malaria monitoring and evaluation (M&E) tool for program managers. *The American Journal of Tropical Medicine and Hygiene* 86, 96–98.

- Rozanov, J. A. (1977). Markov random fields and stochastic partial differential equations. *Math. USSR Sb.* 32, 515–534.
- Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Application*. Boca Raton: Chapman and Hall/CRC Press.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* 71, 319–392.
- Simpson, D., J. Illian, F. Lindgren, S. Sørbye, and H. Rue (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika* 103, 49–70.
- Simpson, D., F. Lindgren, and H. Rue (2012a). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* 23, 65–74.
- Simpson, D., F. Lindgren, and H. Rue (2012b). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics* 1, 16–29.
- Stanton, M. and P. Diggle (2013). Statistical analysis of binomial data: generalised linear or transformed Gaussian modeling. *Environmetrics* 24, 158–171.
- Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.
- Stevenson, J., G. Stresman, C. Gitonga, J. Gillig, C. Owaga, E. Marube, W. Odongo, A. Okoth, P. China, R. Oriango, et al. (2013). Reliability of school surveys in estimating geographic variation in malaria transmission in the Western Kenyan highlands. *PloS one* 8, e77641.
- Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health* 29, 75–90.
- Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. New York: Springer.
- Whittle, P. (1954). On stationary processes in the plane. *Biometrika* 41, 434–449.
- Zhang, H. and D. L. Zimmerman (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* 92, 921–936.