Understanding models for spatial structure in small-area estimation

Adam Howes[a], Jeffrey W. Eaton[b], Seth R. Flaxman[c]

[a]*Department of Mathematics Imperial College London*
[b]*MRC Centre for Global Infectious Disease Analysis Imperial College London*
[c]*Department of Computer Science University of Oxford*

Abstract

Background: Spatial correlation between small-areas is typically accounted for within a generalised linear mixed model setting using spatial random effects. However, for some geometries, commonly used models for spatial random effects based upon adjacency relations, like the Besag model, make unrealistic and unsatisfying spatial assumptions. Furthermore, they do not take into account that across many settings areal data is generated by aggregating continuous point data. Defining spatial random effects which instead correspond to aggregated Gaussian processes gives a more intuitively convincing correlation structure, and allows tools from kernel methods to be used in models for areal data.

Methods: In a simulation study, we used a proper scoring rule to evaluate the performance of different spatial random effect specifications. We performed these simulations under model misspecification and across a range of vignette and realistic geometries with different regularities. We also compared model performance in estimating district-level HIV prevalence from PHIA household survey data across four countries in sub-Saharan Africa, using cross-validation and spatial cross-validation.

Results:

Conclusions:

1. Introduction

Spatial random effects are frequently used to model spatial variation in areal data [29, 10]. Gaussian Markov random fields (GMRFs) [58], which combine a Gaussian distribution with Markov conditional independence assumptions between areas, are the most common class of models used to specify spatial random effects. Using a GMRF typically reflects the assumption that observations made in areas close together are correlated, and more distant relationships may be ignored. The simplest GMRF model is arguably that of Besag et al. [3], where information is borrowed equally from each adjacent area. This model requires minimal additional modelling choices and is widely implemented, including as a part of the `R-INLA` package [60], popular within spatial statistics. Use of this model is widespread, including in agriculture [48], ecology [62], epidemiology [39], image analysis [63], neuroscience [28] and public health [17]. However, for irregular geometries like the administrative divisions of a country, the assumptions made by the Besag model about space are unrealistic. In this work, we set out to test the hypothesis that incorporating spatial structure more faithfully to the specific geometry would improve the performance of small-area estimation models, and in doing so provide practical recommendations and further understanding of models for spatial structure.

Our motivating application is the small-area estimation of HIV epidemic indicators in sub-Saharan Africa. Estimates are required to help plan, implement, and evaluate the success of programmes, ensuring that available resources are most effectively used to respond to the epidemic [30, 12]. Household surveys provide nationally represenative data about the general population, but are expensive to carry out and as such have small sample sizes at a district level. Although auxiliary covariates can be used to aid estimation, data on

*Corresponding author
Email address: `ath19@ic.ac.uk` (Adam Howes)

the covariates most strongly associated with HIV, such as sexual risk behaviour or the prevalence of male circumcision, often face similar measurement difficulties as the HIV indicators themselves. Models including covariates have only been found to modestly improve predictions [18, Supplementary Figure 20]. This stands in contrast to other infectious diseases like Malaria where transmission is driven by more predictive and easily-measurable environmental factors [76]. These circumstances foreground the importance of sharing information between districts via spatial smoothing in HIV mapping.

Within the GRMF framework, attempts to more accurately reflect irregular spatial structure have defined weights specifying the extent to which neighbours are related. Duncan et al. [16] compared seventeen methods for specifying these weights, but surprisingly did not find any which outperform the Besag model, though this conclusion was specific to the often-analysed Scottish lip cancer example, and based on the deviance information criteria (DIC) which is recommended against by Vehtari et al. [72]. Another approach is to take into account that areal data is typically produced by aggregating higher resolution data. The problem is then a particular instance of the broader challenge of inference about a variable at a different resolution to that it was observed at, known in geostatistics as the change-of-support problem [22], which dates to foundational work on block kriging [35, 45]. Ideas from change-of-support modelling are applied by Kelsall and Wakefield [34] to consider areal data as coming about by an aggregation of a continuously-indexed Gaussian process, resulting in a covariance structure between two areas given by the average covariance between two points chosen randomly from those areas. This type of model is particularly natural in the log-Gaussian Cox process modelling framework [43, 15, 67, 33]. Inference for aggregated data has recently been advanced by Wilson and Wakefield [77] who consider the SPDE approach of Lindgren et al. [44], using `R-INLA`, and an empirical Bayes approach, using `TMB` [36], and made available as a part of the `disaggregation` package [47].

Recognising that spatial heterogeneity in outcomes commonly reflects a combination of both spatially correlated processes and specific circumstances at a given location, it is usually recommended to use a random effect specification which includes both spatially structured and unstructured components. Examples include the BYM2 model [64] and earlier convolutions such as the BYM [3], Leroux [42] or Dean [14] models. Evidence for this recommendation includes the comparison studies of Lee [41] and Riebler et al. [55]. Improvements to the Besag model are likely transferable to improving these convolution models. The Naomi HIV small-area estimation model [19], which motivates our work, uses both Besag and BYM2 spatial random effects, at different levels of the hierarchical model, and both models are of substantive practical interest.

The remainder of this paper is organised as follows. Section 2 provides background on areal data and the Bayesian hierarchical modelling approach to small-area estimation. In Section 3, we review developments in specifying spatial random effects based on adjacency, before presenting an alternative approach based on kernels in Section 4. In Section 5 we compare models using simulated data, before applying them to mapping HIV prevalence in sub-Saharan Africa in Section 6. Finally, we discuss our conclusions and directions for future research in Section 7.

2. Background

2.1. Areal and point spatial data

Let $\mathcal{S} \subset \mathbb{R}^2$ be the study region, and the disjoint areas $\{A_i\}_{i=1}^n$ be a spatial partition of $\mathcal{S}$. Areal data are a type of spatial data where observations y_i are associated to areas A_i. Examples include the colour of a pixel, the minimum wage in a state, and the number of disease cases in a region. Point data are another type of spatial data where instead observations $y(s)$ can be made at any location $s \in \mathcal{S}$ of a spatially-continuous stochastic process.

Areal data is often aptly conceptualised as arising from aggregation of point data, such that $y_i = \int_{A_i} y(s)\mathrm{d}s$. Exceptions include policies determined at an administrative level, such as the minimum wage or disease control measures, which have a genuinely discrete spatial structure. If the areal data we observe are indeed an aggregation of point data, why not use the higher resolution point data in our models rather than areal data? Explanations include that (a) the underlying point data are unobserved, (b) although point data are collected, they may not be accessible due to privacy constraints, administrative practicality or storage capacity, and (c) the point data are available, but we decide not to use them in the model. In the final case (c), researchers may prefer to use area-level models as models for point data are typically more complex, require more data, are less immediately applicable to area-level policy making, and have higher computational burden. Furthermore, seemly point data may actual be areal, such as data observed at polling stations or health facilities which

individuals from the surrounding area travel to. As such, choosing to model the underlying data generating mechanism as either areal or point is often a matter of pragmatism.

2.2. Small-area estimation

Auto-correlation is an important property of most spatial processes. Usually, outcomes for locations close together in space tend to be more similar than those far apart. This fact is known as "Tobler's first law of geography" [69] and, from a statistical point of view, is both a challenge and a benefit of working with spatial data. Each observation provides less information than it would have had the samples been independent, making it more difficult to estimate global parameters. However, spatial correlation can be used to improve local (indexed by a particular spatial location) parameter estimates particularly in parts of the study region where little to no information is available.

The latter benefit is basis for the statistical task of small-area estimation, which aims to produce reliable local estimates where small sample sizes lead to noisy data [52]. In a spatial setting, this is often in small geographic areas, though the phrase "small-area" is not restricted to geographic areas and may be interpreted more broadly to mean any area where data are insufficient to make accurate local parameter inferences. In the context of multilevel regression and post-stratification [23], small-areas are generated by the intersection of demographic variables like age, gender and race, alongside geographic variables. Due to the cost of gathering samples, a survey may be designed to give reliable estimates at at an aggregated level but not at a small-area level. Although direct estimators of local parameters are unbiased, when data are sparse the total error may be reduced by accepting some bias in exchange for reduced variance using so-called indirect estimators. Smoothing approaches use information from similar units to "borrow strength" from one parameter to another, with determining precisely what is meant by "similar" a central challenge. For a recent review of both design-based and model-based approaches to spatial analysis of health-indicators, including both area and unit-level analysis, see Wakefield et al. [74].

2.3. Bayesian model-based approach and latent Gaussian models

Bayesian hierarchical models are an attractive framework for small-area estimation, whereby areal data y corresponding to the areas $\{A_i\}_{i=1}^n$, are modelled using a three-layer structure [2, 11, 54] given by

$$\text{(Observations)} \qquad y_i \sim p(y_i \,|\, g^{-1}(x_i), \theta), \quad i = 1, \ldots, n, \tag{1}$$

$$\text{(Latent field)} \qquad x \sim p(x \,|\, \theta), \tag{2}$$

$$\text{(Parameters)} \qquad \theta \sim p(\theta), \tag{3}$$

where x is the n-dimensional latent field, θ are the m-dimensional parameters, $m < n$, and g is a link function. Spatial structure is modelled at the level of the latent field, and the observation model, typically either binomial or Poisson, is conditionally independent and identically distributed $p(y \,|\, x, \theta) = \prod_{i=1}^n p(y_i \,|\, x_i, \theta)$. As small-area estimation models are descriptive rather than mechanistic, it is natural [5] to use a flexible and computationally efficient Gaussian distribution $x \sim \mathcal{N}(\mu, \Sigma)$ for the latent field. Three-layer models with Gaussian latent fields comprise a wide array of popular models, including generalised linear mixed models (GLMMs), and have been collectively studied under the title latent Gaussian models (LGMs) [59].

2.4. Spatial random effects

The latent field in a LGM is comprised a constant mean μ, known also as the fixed effects, and spatial random effects $\phi \sim \mathcal{N}(0, \Sigma)$. The mean may be modelled using a linear predictor $\mu_i = \beta_0 + z_i^\top \beta$, where (β_0, β) are regression parameters corresponding to areal covariates $z_i = (z_{i,1}, \ldots, z_{i,p})^\top$. We focus on the spatial random effects ϕ, whose role is to capture the spatial structure between areas not already accounted for in the mean. If no spatial structure remains after inclusion of covariates in the mean, independent and identically distributed (IID) random effects $\phi_i \sim \mathcal{N}(0, \tau_\phi^{-1})$ should be used, where τ_ϕ is a shared precision. Exploratory data analysis techniques such as visual inspection or Moran's I coefficient [9] may be used to determine if there remains any spatial auto-correlation in the data. Specifying ϕ amounts to specifying the entries of a precision or covariance matrix, as we discuss in the following sections.

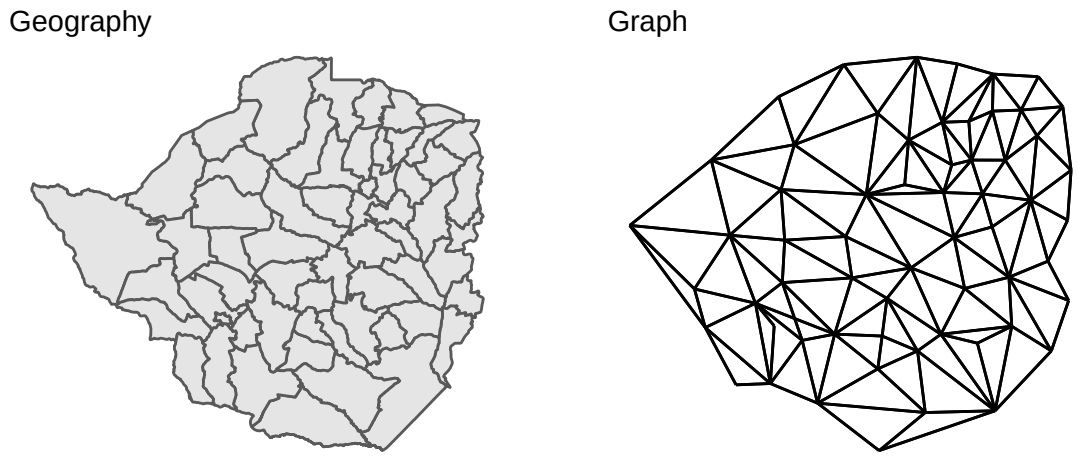

Figure 1: A map of the districts of Zimbabwe together with the corresponding adjacency graph structure $\mathcal{G}$.

3. Spatial random effects using adjacency

3.1. Besag

One way to encode spatial structure between areas is by using a symmetric relation, where $i \sim j$ if the areas A_i and A_j are adjacent or neighbouring. Adjacency is often defined by a shared border, though other choices, such as inclusion of second or higher-degree neighbours [50], are also possible. The Besag model [3] is a type of improper conditional auto-regressive (ICAR) model where the full conditional distribution of the ith spatial random effect is given by

$$\phi_i \,|\, \phi_{-i} \sim \mathcal{N}\left(\frac{1}{n_{\delta i}} \sum_{j:j\sim i} \phi_j, \frac{1}{n_{\delta i}\tau_\phi}\right), \tag{4}$$

where δi is the set of neighbours of A_i with cardinality $n_{\delta i} = |\delta i|$ and $\phi_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)^\top$ is the vector of spatial random effects with the ith entry removed. The conditional mean of the random effect ϕ_i is the average of its neighbours $\{\phi_j\}_{j\sim i}$ and the precision $n_{\delta i}\tau_\phi$ is proportional to the number of neighbours $n_{\delta i}$. By Brook's lemma [58, Chapter 2] the set of full conditionals of the Besag model are equivalent to the Gaussian Markov random field (GMRF) given by

$$\phi \sim \mathcal{N}(0, \tau_\phi^{-1} R^-), \tag{5}$$

where R^- is the generalised inverse of the rank-deficient structure matrix R, so-called because it defines the *structure* of the precision matrix, with entries

$$R_{ij} = \begin{cases} n_{\delta i}, & i = j \\ -1, & i \sim j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The Markov property arises due to the conditional independence structure $p(\phi_i \,|\, \phi_{-i}) = p(\phi_i \,|\, \phi_{\delta i})$ whereby each area only depends on its neighbours. This is reflected in the sparsity of R whereby $\phi_i \perp \phi_j \,|\, \phi_{-ij}$ if and only if $R_{ij} = 0$. The structure matrix R may also be expressed as the Laplacian of the adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $v \in \mathcal{V}$ corresponding to each area and edges $e \in \mathcal{E}$ between vertices i and j when $i \sim j$. Figure 1 shows the adjacency graph for the districts of Zimbabwe alongside the geometry. Shortly, in Section 3.2 we discuss the appropriateness of using the graph rather than the complete geometry.

Rewriting Equation 5, the probability density function of ϕ is

$$p(\phi) \propto \exp\left(-\frac{\tau_\phi}{2}\phi^\top R \phi\right) \propto \exp\left(-\frac{\tau_\phi}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right). \tag{7}$$

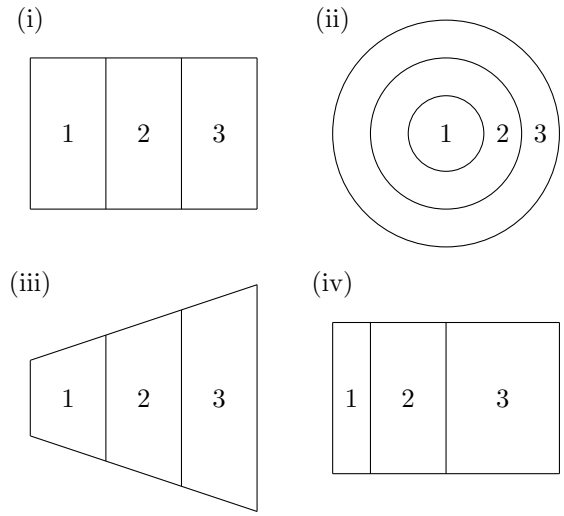

Figure 2: Each of these four geometries corresponds to the same adjacency graph.

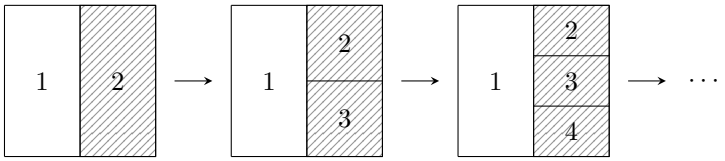

Figure 3: A sequence of geometries where the number of neighbours of area one grows by one at each iteration.

This density is a function of the pairwise differences $\phi_i - \phi_j$ and so is invariant to the addition of a constant c to each entry $p(\phi) = p(\phi + c1)$, leading to an improper uniform distribution on the average of the ϕ_i. If $\mathcal{G}$ is connected, in that by traversing the edges, any vertex can be reached from any other vertex, then there is only one impropriety in the model and $\text{rank}(R) = n - 1$, while if $\mathcal{G}$ is disconnected, and composed of $n_c \geq 2$ connected components with index sets $I_1, \ldots, I_{n_c}$, then the corresponding structure matrix R has rank $n - n_c$ and the density is invariant to the addition of a constant to each of the connected components $p(\phi_I) = p(\phi_I + c1)$ where $I = I_1, \ldots, I_{n_c}$.

3.2. Concerns about the Besag model's representation of space

The Besag model was originally proposed for use in image analysis, where areas correspond to pixels arranged in a regular lattice structure. Since then, it has seen wider use, including in situations where the spatial structure is less regular. There are a number of concerns about the model's applicability to this broader setting, which we highlight below. Closely linked are the modifiable areal unit problem [49] whereby statistical conclusions change as a result of seemingly arbitrary changes in data aggregation, and the challenge of ecological inference and the ecological fallacy [75], which applies particularly to inference about individuals and groups.

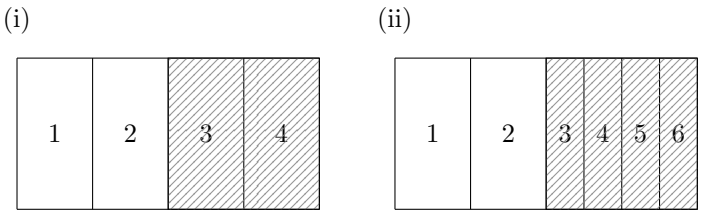

Figure 4: Each of the shaded areas are split into two moving from geometry (i) to (ii).

3.2.1. Adjacency compression

Summarising all spatial information with an adjacency graph represents a loss of information. Many geometries share the same adjacency graph, and therefore the same probability model, as illustrated by Figure 2. This fact is not concerning in itself, though it prompts consideration as to whether the class of geometries with the same adjacency graph is sufficiently similar to merit identical models.[1] It is intuitive that the more regular the spatial structure the less information is lost in compression to an adjacency graph. In image analysis, very little spatial information is lost in compression a lattice structure to an adjacency graph. On the other hand, the regions of a country, determined by political and geographic forces, tend to display greater irregularity. As such, the appropriateness of adjacency compression varies by the type of geometry common to the application setting.

3.2.2. Mean structure

A feature specific to the Besag model is that all adjacent areas count equally. This assumption is unsatisfying: for most geometries, we expect some areas to be more highly correlated to some neighbours than to others. Figure 2 illustrates a number of heuristic features for neighbour importance, including length of shared border in geometry (iii), and the proximity of centers of mass in geometry (iv). This may be remedied using more general weighted ICAR models, as we outline in Section 3.3.

3.2.3. Variance structure

A striking feature of Equation 4 is that the precision of ϕ_i is proportional to its number of neighbours $n_{\delta i}$. It follow that as $n_{\delta i} \to \infty$ then $\mathrm{Var}(\phi_i) \to 0$. This is illustrated by Figure 3 where the area on the right is repeatedly divided such that its number of neighbours increases by one at each step. This property is a consequence of averaging the conditional mean over a greater number of areas, which, in certain situations, can correspond to a greater amount of information. However, if the amount of information in the shaded area remains fixed, it is inappropriate that $\mathrm{Var}(\phi_1)$ should tend to zero just as a result of drawing additional, arbitrary, boundaries. In the image analysis setting this modelling assumption is reasonable: each pixel represents a fixed amount of information and a higher pixel density represents a greater amount of information. On the other hand, in public health and epidemiology, drawing boundaries to create additional regional areas cannot be expected to correspond to a greater amount of information and this assumption is not appropriate.

Suppose we fit a Besag model upon identical data using each of the two geometries in Figure 4. If the spatial variation is relatively smooth, dividing the shaded areas into two will result in a lower estimated variance σ_ϕ^2 in geometry (ii) as compared with geometry (i) because there will appear to be less variation between neighbouring areas. This problem does not only apply locally: since the effect of σ_ϕ^2 applies everywhere, the smoothing will change even in unaltered parts of the study region.

3.3. Weighted ICAR

The Besag model is a special case of a more general class of (zero-mean) ICAR models in which each neighbour may not be weighted equally. These models can be specified in terms of scaled weights $\{b_{ij}\}_{j\sim i}$ and a precision vector $\kappa = (\kappa_1, \ldots, \kappa_n)^\top$ such that $\phi_i \,|\, \phi_{-i} \sim \mathcal{N}(\sum_{j:j\sim i} b_{ij}\phi_j, (\kappa_i\tau_\phi)^{-1})$. The structure matrix R corresponding to the above full conditionals is given by $R = D_\kappa(I - B)$, where B has the entries b_{ij} for $i \sim j$, diagonal entries $b_{ii} = 0$ and $b_{ij} = 0$ for $i \nsim j$ and the matrix D_κ is given by $\mathrm{diag}(\kappa_1, \ldots, \kappa_n)$. As the structure matrix is symmetric we must have the condition $-b_{ij}\kappa_i = -b_{ji}\kappa_j$. To meet this condition, it is often simpler to use an unscaled weights matrix $W = D_\kappa B$ such that $R = D_\kappa - W$. For example, in the Besag model W corresponds to the adjacency matrix. The scaled weights can then be recovered by $b_{ij} = w_{ij}/\kappa_i$ where $\kappa_i = \sum_{k:k\sim i} w_{ik}$. A thorough discussion of methods for specifying W is provided by Duncan et al. [16]. Much of the work in this area focuses on the case where the geometry is a lattice.

[1] The regularity of realistic geometries may help to constrain each class to be more similar than it strictly has to be. In other words, although pathological geometries can be constructed, they are implausible in statistical practise and so not of great concern to us here.

3.4. BYM2

Often, as well as spatial structure, there exists IID over-dispersion in the residuals and it is inappropriate to use purely spatially structured random effects in the model. The Besag-York-Mollié (BYM) model of Besag et al. [3], accounts for this in a natural way by decomposing the spatial random effect $\phi = v + u$ into a sum of an unstructured IID component v and a spatially structured Besag component u, each of which with their own respective precision parameters τ_v and τ_u. The resulting distribution is

$$\phi \sim \mathcal{N}(0, \tau_v^{-1} I + \tau_u^{-1} R^-). \tag{8}$$

Including both v and u is intended to enable the model to learn the relative extent of the unstructured and structured components via τ_v and τ_u. However, in this specification, scaling of the Besag precision matrix Q is not taken into account despite this issue being particularly pertinent when dealing with multiple sources of noise. In particular, placing a joint prior $(\tau_u, \tau_v) \sim p(\tau_u, \tau_v)$ which doesn't privilege either component is more easily accomplished if Q and I have the same scale. Additionally, supposing we have a prior belief that the over-dispersion is primarily IID and v accounts for the majority of the dispersion, then it is not immediately obvious how to represent this belief using $p(\tau_u, \tau_v)$, without inadvertently altering the prior about the overall variation. This highlights identifiability issues of the parameters (τ_u, τ_v) resulting from them not being orthogonal. Building on the models of Leroux et al. [42] and Dean et al. [14] which tackle this the identifiability problem, but do not scale the spatially structured noise, Simpson et al. [64] propose a reparameterisation $(\tau_v, \tau_u) \mapsto (\tau_\phi, \pi)$ of the BYM model known as the BYM2 model and given by

$$\phi = \frac{1}{\tau_\phi} \left(\sqrt{1-\pi}\, v + \sqrt{\pi}\, u^\star \right), \tag{9}$$

where τ_ϕ is the marginal precision of ϕ, $\pi \in [0, 1]$ gives the proportion of the marginal variance explained by each component, and $u^\star$ is a scaled version of u with precision matrix given by the scaled structure matrix $R^\star$ (see Appendix). When $\pi = 0$ the random effects are IID, and when $\pi = 1$ the random effects follow the Besag model. To borrow an analogy, the parameterisation (τ_v, τ_u) is like having one hot water and one cold water tap, whereas the parameterisation (τ_ϕ, π) is like a mixer tap where the amount of water and its temperature can be adjusted separately [57].

4. Spatial random effects using kernels

4.1. Areal kernels

Section 3 reviewed ways to construct spatial random effect precision matrices using an adjacency relation. Another approach is to define the covariance matrix using an areal kernel function $K : \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to \mathbb{R}$ which gives a measure of similarity between two areas, where $\mathcal{P}$ denotes the power set and $\mathcal{P}(\mathcal{S})$ is the space of subsets of the study region. If K is positive semi-definite, then we define areal kernel spatial random effects by

$$\phi \sim \mathcal{N}(0, \frac{1}{\tau_\phi} K), \tag{10}$$

where the $n \times n$ Gram matrix K with entries $K_{ij} = K(A_i, A_j)$ is a valid covariance matrix. Here, we place τ_ϕ outside of the Gram matrix, analogous to the relation of the precision and structure matrices.

Most well-known spatial process models define the correlation structure between a pair of points using a kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. We use a simple method to construct K from k, namely averaging the kernel k computed on some collection of points from within each area. As such, K is described as a kernel on sets in the multiple instance learning literature [21]. Although it may be possible to learn the method for combining kernels from data [26], we do believe that this is feasible in the small-area estimation setting, and instead hope to construct areal kernels which represent our prior beliefs about the spatial process.

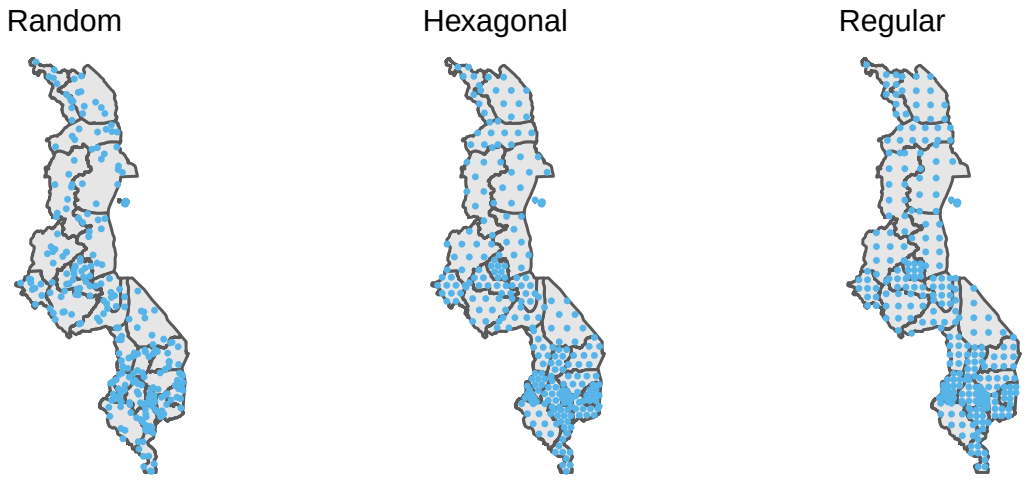

Figure 5: Three ways to choose integration points with $L_i = 10$ for all i in Malawi, as implemented by `sf::st_sample`.

4.2. Centroid kernel

The simplest approach is to use a single point, especially one which is representative of the area, a natural choice being the centroid $c_i \in A_i$ given by the arithmetic mean of the latitude and longitude. This results in the centroid kernel

$$K(A_i, A_j) = k(c_i, c_j). \tag{11}$$

The centroid kernel has been used by Wakefield and Morris [73] in an environmental epidemiology setting and Best et al. [5] in a model comparison study, both choosing the exponential kernel $k(s_i, s_j) = \exp(-|s_i - s_j|/l)$; as well as more recently by Teh et al. [68] as a part of a spatio-temporal Gaussian process model to infer the reproduction number of COVID-19. Wakefield and Morris [73] fix the length-scale l using a semi-variogram and Best et al. [5] place a prior on the inverse length-scale. Best et al. [5, Section 3] simulated data representing heterogeneous exposure to air pollution, including elevated rates of exposure near two hypothetical point source locations, and found that the centroid kernel tended to over-smooth the high-risk areas, though it is unsuprising that a stationary covariance would struggle to recover non-stationary structure.

4.3. Integrated kernel

Rather than attempting to choose a single representative point, an alternative is to represent the whole area by integrating the kernel over the areas of interest. This results in the integrated kernel

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} k(s, s') \mathrm{d}s \mathrm{d}s', \tag{12}$$

which is equivalent to the covariance structure obtained by aggregating a spatially continuous Gaussian process with kernel k over the areal partition. Models of this sort have been studied in the machine learning literature under the name aggregated Gaussian processes [38, 66, 78, 31, 79]. Unlike for the centroid kernel where $K_{ii} = 1$ for all i, the marginal variance of the ith spatial random effect $K_{ii} = K(A_i, A_i)$ varies depending on the area: becoming smaller for more compact areas and larger for areas which are of greater extent or more spread out.

The log-Gaussian Cox Process framework [15] also naturally arrives at the integrated kernel formulation. A Cox process is an inhomogeneous Poisson process with a continuous stochastic intensity function $\{x(s), s \in \mathcal{S}\}$ such that conditional on the realisation of $x(s)$ the number of points in any area A_i follows a Poisson distribution. The rate parameter of this Poisson distribution is explicitly aggregated as follows

$$y_i \,|\, x(s) \sim \text{Poisson}\left(\int_{s \in A_i} x(s) \mathrm{d}s \right). \tag{13}$$

In a LGCP the log intensity $\log x(s) = \eta(s)$ is modelled using a Gaussian process prior $\eta(s) \sim \mathcal{GP}(\mu(s), k(s, s'))$. Johnson et al. [33] obtain Equation 15 by considering a discrete Poisson log-linear mixed model approximation to a continuous LGCP, whereby $\eta(s)$ is approximated by a piecewise constant $\eta_i = \mu_i + \phi_i$ in each area A_i[2]. The ith discrete spatial random effect is then $\phi_i = \int_{A_i} w_i(s)\phi(s)\mathrm{d}s$, with covariance structure

$$\mathrm{Cov}\left(\int_{A_i} w_i(s)\phi(s)\mathrm{d}s, \int_{A_j} w_j(s')\phi(s')\mathrm{d}s'\right) = \int_{A_i}\int_{A_j} w_i(s)w_j(s')k(s,s')\mathrm{d}s\mathrm{d}s', \tag{14}$$

corresponding to an areal integrated kernel with a logarithmic link function and Poisson likelihood.

Disaggregation regression (also known as downscaling or interpolation) is another closely related approach which, rather than using the aggregate nature of areal observations as primarily a route towards better area-level estimates, as is our focus, aims to producing high-resolution or point-level estimates from areal observations [70, 47].

A benefit of the integration approach is that, if available, additional information accounting for heterogeneity over A_i may be incorperated into the model. This can be accomplished using weighting distributions $\{W_i\}$ which represent an unequal contribution of each point to the similarity measure, to give a weighted integrated kernel

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|}\int_{A_i}\int_{A_j} w_i(s)w_j(s')k(s,s')\mathrm{d}s\mathrm{d}s', \tag{15}$$

This may be useful in disease mapping, where we expect regions with populations who live close to their shared border to be more strongly correlated than regions whose populations live far apart, which could be accounted for by weighting according to a high resolution measure of population density.

If areal kernel spatial random effects are created using the integrated kernel, then aggregation occurs at the level of the latent field in the three stage model rather than at the level of the data. This corresponds to a generative model of the form $y_i \sim p(y_i \,|\, g^{-1}(x_i))$, with $x_i = |A_i|^{-1}\int_{A_i} x(s)\mathrm{d}s$. If the link function g is the identity or linear then aggregation at the level of the latent field is equivalent to aggregation at the level of the data. On the other hand, for non-linear link functions g such as the exponential or logistic, the generative model does not match the proposed data generating process $y_i = |A_i|^{-1}\int_{A_i} y(s)\mathrm{d}s$.

Most of the time we do not expect to be able to calculate the integral in Equation 15 analytically. Instead, given n collections of L_i samples $\{s_l^{(i)}\}_{l=1}^{L_i} \sim \mathcal{U}(A_i)$ drawn uniformly from each area then the integral may be approximated using Monte Carlo by the double sum

$$K(A_i, A_j) \approx \frac{1}{L_i L_j}\sum_{l=1}^{L_i}\sum_{m=1}^{L_j} w_i\left(s_l^{(i)}\right) w_j\left(s_m^{(j)}\right) k\left(s_l^{(i)}, s_m^{(j)}\right). \tag{16}$$

Equivalently, samples drawn from W_i may be used without weighting by $w_i(s)$. Integration points may also be selected by some deterministic process, and used in Equation 16 to give a numerical integration estimate of the kernel. As with learning the way in which the kernels should be combined, in more data rich settings it may be possible to learn the integration points [8]. Taking any of these approaches requires $\mathcal{O}(\sum_{i=1}^n \sum_{j=1}^n L_i L_j)$ evaluations of the kernel k to compute the $n \times n$ Gram matrix K. This imposes a significant computational cost if the Gram matrix is often recomputed during inference, as is the case in MCMC when any of the kernel hyperparameters are learnt, placing a limit on the number of samples or integration points it is feasible to use. To make inference more feasible, Kelsall and Wakefield [34] reduce the number of Gram matrix constructions and inversions required, by using a discrete hyperparameter prior.

5. Simulation study

We tested the ability of inferential models with varying spatial random effect specifications to accurately recover small-area quantities. The data and modelling choices were designed with a spatial epidemiology

[2] Johnson et al. [33] use points generated from a class of inhibition processes which combine sequential inhibition with rejection sampling to evaluate the integral.

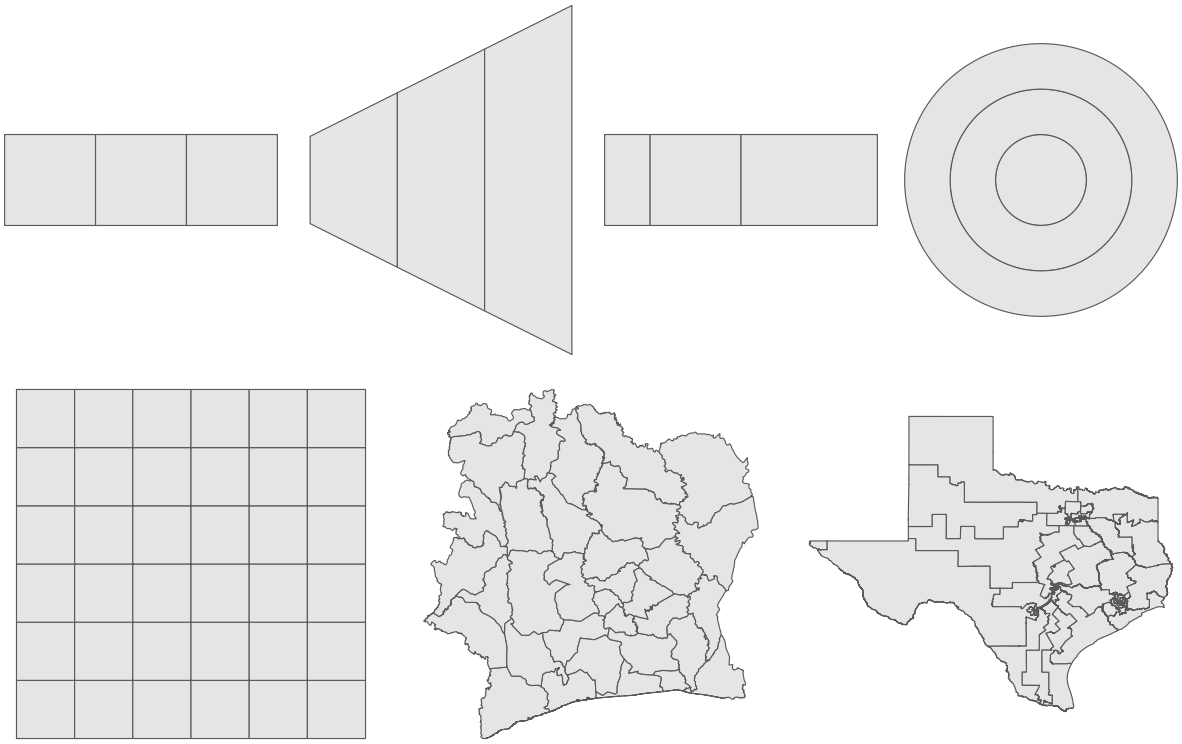

Figure 6: The seven simulation geometries that we considered.

Simulation model	Additional details
S1. *IID*	$\phi \sim \mathcal{N}(0, I_n)$
S2. *Besag*	$\phi \sim \mathcal{N}(0, (R^\star)^-)$ as described in Section 3.1 where $R^\star$ is the scaled structure matrix
S3. *IK*	$\phi \sim \mathcal{N}(0, K^\star)$ as described in Section 4.3 where K_{ij} is calculated via Equation 16 using the Matérn kernel with $\nu = 3/2$, $l = 2.5$ and $L_i = 100$ samples are drawn uniformly from each area. Like the Besag model, we scaled the Gram matrix to have generalised variance one

Table 1: The three simulation models from which we generate ϕ.

setting in mind, and are similar to those used in Section 6 where we considered models for HIV prevalence using data from national household surveys in sub-Saharan Africa. The R [53] code used is available from `github.com/athowes/areal-comparison`. We used the `orderly` package [20] for reproducible research.

5.1. Synthetic data-sets

We studied robustness to spatial misspecification, by simulating synthetic data-sets from three known data generating processes. The spatial random effects ϕ were generated according to IID, Besag and integrated kernel (IK) simulation models (Table 1). We then generated synthetic data $y = (y_1, \ldots, y_n)^\top$ of a form analogous to a household survey whereby $y_i \sim \text{Bin}(m_i, \rho_i)$ where the probabilities $\rho_i \in [0, 1]$ are linked to the latent random variables $x_i \in \mathbb{R}$ via

$$\log\left(\frac{\rho_i}{1-\rho_i}\right) = x_i = \beta_0 + \phi_i, \quad i = 1, \ldots, n. \tag{17}$$

We fixed the sample sizes to $m_i = 25$ for all i, the intercept parameter as $\beta_0 = -2$ and the spatial random effect precision parameter as $\tau_\phi = 1$. We used seven different geometries: the four vignette geometries in Figure 2, which share an adjacency graph, as well as three more realistic geometries, a 6×6 lattice grid, the 33 districts of Côte d'Ivoire and the 36 congressional districts of Texas (Figure 6). These more

Inferential model	Algorithm	Additional details
I1. *Constant*	INLA	No spatial random effects
I2. *IID*	INLA	$\phi \sim \mathcal{N}(0, \tau_\phi^{-1} I_n)$
I3. *Besag*	INLA	$\phi \sim \mathcal{N}(0, (\tau_\phi R^\star)^-)$ following additional recommendations in **??**
I4. *BYM2*	INLA	$\phi = \tau_\phi^{-1}\left(\sqrt{1-\pi}\, v + \sqrt{\pi}\, u^\star\right)$ with a Beta$(0.5, 0.5)$ prior on $\pi \in [0, 1]$
I5. *FCK*	INLA	$\phi \sim \mathcal{N}(0, \tau_\phi^{-1} K)$ with $K_{ij} = k(c_i, c_j)$ where the length-scale l is fixed
I6. *CK*	NUTS	As in model I5, with length-scale prior $l \sim$ Inv-Gamma(a, b)
I7. *FIK*	INLA	$\phi \sim \mathcal{N}(0, \tau_\phi^{-1} K)$ with $K_{ij} = \frac{1}{10^2}\sum_{l=1}^{10}\sum_{m=1}^{10} k\left(s_l^{(i)}, s_m^{(j)}\right)$ with hexagonal integration point spacing and fixed length-scale l
I8. *IK*	NUTS	As in model I7, with length-scale prior $l \sim$ Inv-Gamma(a, b)

Table 2: Each inferential model is implemented as a part of the R package `bsae` available from `github.com/athowes/bsae`.

realistic geometries were chosen to represent variation over spatial regularity, allowing us to test how model performance varies by geometry regularity. For each combination of spatial random effect and geometry we replicated the simulation process above 200 times to generate a total of 1400 synthetic data-sets.

5.2. Inferential models

We fit eight inferential models, each of which differing in its spatial random effect specification (Table 2).

5.2.1. Priors

We placed a half-Gaussian prior on the standard deviation [24] such that $\sigma_\phi \sim \mathcal{N}_+(0, 2.5^2)$. The weakly informative value 2.5 was chosen to avoid placing significant prior density on the region $\sigma_\phi > 5$, which after a logistic transformation facilitates variation on the probability scale very close to either zero or one, neither of which we assumed to be plausible. A weakly informative $\mathcal{N}(-2, 1)$ prior was placed on β_0, setting most of the prior probability density for ρ_i within a range $[0, 0.25]$ typical for a disease prevalence.

5.2.2. Kernel model details

Following Stein [65], for the areal kernel models we used the Matérn kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ given by

$$k(s, s') = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|s - s'|}{l}\right)^\nu B_\nu\left(\frac{\sqrt{2\nu}|s - s'|}{l}\right), \tag{18}$$

where B_ν is the modified Bessel function of the second kind, $|s - s'|$ is the Euclidean distance between s and s', ν is the smoothness hyperparameter and l is the length-scale hyperparameter on the latitude-longitude scale. We fixed the smoothness parameter, which is otherwise typically unidentifiable from data, to be $3/2$, matching that used in simulation model S3, and convenient in that it simplifies Equation 18 to be

$$k(s, s') = \left(1 + \sqrt{3}|s - s'|/l\right)\exp\left(-\sqrt{3}|s - s'|/l\right). \tag{19}$$

Taking a similar approach to Best et al. [4], in models I5 and I7 we fixed the length-scale such that points an average distance apart in the model have 1% correlation a priori, independent of the data. In models I6 and I8, we used a length-scale prior $l \sim$ Inv-Gamma(a, b), with parameters a and b chosen for each model such that 1% of the prior mass is below 0.1 and 1% of the prior mass is above the maximum distance between points in the model [6]. For the integration points we set $L_i = 10$ using a hexagonal spacing structure, as with the central panel of Figure 6. Though note `sf::st_sample` with `type = "hexagonal"` does not guarantee exactly the specified number of samples are returned [51]. We are limited to this relatively small number by the computational costs of Model I8. For Model I6 although it is feasible to use a much larger number of integration points, since the Gram matrix is only computed once, we chose not to for comparability.

5.3. Inference algorithms

In models I1-I5 choosing a Gaussian distribution for β_0 ensures that the latent random variables are Gaussian and the model is a LGM, facilitating use of the integrated nested Laplace approximation (INLA) algorithm [59] for inference, as implemented by the `R-INLA` package. INLA is a deterministic method for approximate Bayesian inference based upon a combination of numerical integration and the Laplace approximation. INLA is significantly faster than MCMC and has been shown to have comparable accuracy for LGMs in realistic, pre-asymptotic settings.

5.4. Model assessment

We assessed each fitted inferential model according to its ability to recover the true underlying value of the probabilities ρ_i at each area in the study region, as well as the calibration of the model's estimates.

5.4.1. Continuous ranked probability score

The continuous ranked probability score (CRPS) [46] gives an overall measure of forecasting performance, and can be seen as a generalisation of the Brier score [7] to distributional forecasts. For a given fitted model and area A_i let ρ_i have posterior marginal $f(\rho_i) = p(\rho_i \,|\, y)$ and ω_i be the true value. Writing F as the cumulative distribution function corresponding to f then

$$\mathrm{CRPS}(f, \omega_i) = \int_{-\infty}^{\infty} (F(\rho_i) - \mathbb{1}\{\rho_i \geq \omega_i\})^2 \mathrm{d}\rho_i = \int_0^1 (F(\rho_i) - \mathbb{1}\{\rho_i \geq \omega_i\})^2 \mathrm{d}\rho_i, \tag{20}$$

where $\mathbb{1}$ denotes the indicator function and the second equality follows from $0 \leq \rho_i \leq 1$. The minimum possible value of the CRPS is zero, corresponding to a a point mass correctly forecast at the true value $f(\rho_i) = \delta_{\omega_i}$. The CRPS is a strictly proper scoring rule such that the uniquely best approach to minimising the score is for the model to report its true probabilistic beliefs about the quantity in question [25]. The logarithmic score [27] given by $\mathrm{LogS}(f, \omega_i) = \log f(\omega_i)$ is another popular strictly proper scoring rule, though in line with the findings of Krüger et al. [37], we found it difficult to compute satisfactorily using kernel density estimation. In contrast, the CRPS can be evaluated directly using samples $\{\rho_i^s\}_{s=1}^S \sim p(\rho_i \,|\, y)$ as

$$\mathrm{CRPS}(f, \omega_i) \approx \frac{1}{S} \sum_{s=1}^{S} |\rho_i^s - \omega_i| - \frac{1}{2S^2} \sum_{s=1}^{S} \sum_{l=1}^{S} |\rho_i^s - \rho_i^l|. \tag{21}$$

5.4.2. Posterior predictive check for coverage

We assessed the coverage of our estimates using posterior predictive checks based on the quantile $q_i = F(\omega_i)$ of the true value within each marginal posterior predictive distribution, computed using numerical integration. For calibrated models, over repeated simulations, $q_i \sim \mathcal{U}[0, 1]$ such that at any given nominal coverage $1 - \alpha$, the proportion of quantile-based credible intervals containing the true value ω_i is also be $1 - \alpha$. To check uniformity we used probability integral transform (PIT) histograms [13] and empirical cumulative distribution function (ECDF) difference plots [1, 61].

5.5. Results

	Inferential model					
Simulation model	Constant	IID	Besag	BYM2	FCK	FIK
Grid						
IID	84.9 (2.95)	33.1 (1.08)	34 (1.14)	34.1 (1.11)	34.3 (1.1)	36.3 (1.18)
Besag	39.1 (1.24)	24.4 (0.758)	22.1 (0.679)	23 (0.705)	22.7 (0.675)	22.8 (0.687)
IK	73.9 (2.53)	33.7 (1.13)	27.4 (0.896)	28.6 (0.93)	28.1 (0.9)	26 (0.819)
Cote d'Ivoire						
IID	84.4 (3.01)	32.5 (1.02)	33.3 (1.07)	33.1 (1.04)	33.7 (1.07)	35.2 (1.14)
Besag	44 (1.54)	26 (0.834)	22.7 (0.712)	23.8 (0.758)	23.7 (0.729)	23.6 (0.728)

	IK	44.9 (1.93)	23.9 (0.807)	19.5 (0.633)	21.2 (0.694)	20.4 (0.644)	19.2 (0.608)
Texas							
	IID	88.5 (2.85)	32.7 (1.01)	33.6 (1.04)	33.3 (0.996)	41.6 (1.4)	42.1 (1.41)
	Besag	44.8 (1.65)	27.1 (0.885)	24.3 (0.773)	25.8 (0.843)	26.9 (0.88)	26.9 (0.89)
	IK	70.4 (2.14)	32.3 (1.1)	26.5 (0.966)	27.4 (0.967)	24.7 (0.881)	23.9 (0.843)
1							
	IID	61.9 (7.11)	27.1 (3.01)	27.5 (2.63)	34.5 (3.24)	28.5 (2.94)	28 (2.89)
	Besag	65.3 (6.46)	36.9 (4.47)	39 (5.21)	39 (4.26)	40.4 (5.03)	40 (5.01)
	IK	29.5 (2.98)	26.8 (2.9)	32 (3.62)	27.5 (2.54)	31.9 (3.44)	31.7 (3.48)
2							
	IID	71.9 (8.25)	29.9 (3.48)	29.6 (3.08)	39.6 (4.59)	NA	29.9 (3.38)
	Besag	64.1 (7.81)	33 (3.7)	35 (3.79)	38.4 (4.07)	NA	35.7 (3.73)
	IK	39 (3.79)	31.3 (3.48)	37.8 (3.85)	35.5 (3.06)	NA	36.1 (3.84)
3							
	IID	70.1 (8.19)	36.8 (4.11)	38.2 (4.55)	43.2 (4.32)	38.4 (4.39)	37.8 (4.31)
	Besag	48 (5.46)	25.2 (2.45)	27.7 (2.65)	29.9 (2.75)	28.1 (2.8)	28 (2.65)
	IK	37.1 (3.49)	30.3 (3.35)	32.9 (3.24)	32.4 (3)	32.1 (3.29)	32.1 (3.38)
4							
	IID	68.1 (7.76)	39.6 (4.55)	41.4 (4.58)	42.5 (4.73)	41.8 (4.49)	42.3 (4.52)
	Besag	56.5 (6.12)	31.3 (3.28)	32.1 (3.63)	35.3 (3.6)	34.5 (3.58)	33.9 (3.57)
	IK	32.4 (3.44)	25 (2.59)	27.1 (2.77)	28 (2.55)	28.3 (3.13)	27.9 (3)

6. HIV prevalence study

We compared model performance in estimating HIV prevalence in adults aged $15-49$ across four countries in sub-Saharan Africa: Côte d'Ivoire (which has $n = 33$ districts), Malawi ($n = 28$), Tanzania ($n = 159$) and Zimbabwe ($n = 60$), again varying the spatial random effect specification. As with the simulation study, R code for this study is available from `https://github.com/athowes/areal-comparison`.

6.1. Household survey data

We used data from the most recent publicly available Population Health Impact Assessment (PHIA) survey in each country. These surveys utilise a complex design, where each individual j in area i has an unequal probability π_{ij} of being included in the sample. A two-stage design in which enumeration areas are first drawn from a stratified sample and then households are chosen using equal probability systematic sampling from within each enumeration area, is common. To account for the survey design we used sampling weights $w_{ij} = 1/\pi_{ij}$ to adjust the raw data in each district, obtaining the Kish effective sample size $m_i^\star = (\sum_k w_{ij})^2 / \sum_k w_{ij}^2$ and effective number of cases $y_i^\star$ which may be thought of as what would have been observed had the survey been a simple random sample.

6.2. Model structure

We used the same eight inferential models (Section 5, Table 2), with a small alteration to the likelihood. Since the effective number of cases and effective sample size may not be integers, we modelled the effective number of cases $y_i^\star \in \mathbb{R}$ according to a generalised binomial distribution $y_i^\star \sim \text{xBin}(m_i^\star, \rho_i)$, where $m_i^\star \in \mathbb{R}$ is the effective sample size and $\rho_i \in [0, 1]$ is the adult (15-49) HIV prevalence. The working likelihood under this model for $m_i^\star \geq y_i^\star$ is given by

$$p(y_i^\star \mid m_i^\star, \rho_i) = \frac{\Gamma(m_i^\star + 1)}{\Gamma(y_i^\star + 1)\Gamma(m_i^\star - y_i^\star + 1)} \rho_i^{y_i^\star} (1 - \rho_i)^{(m_i^\star - y_i^\star)}. \tag{22}$$

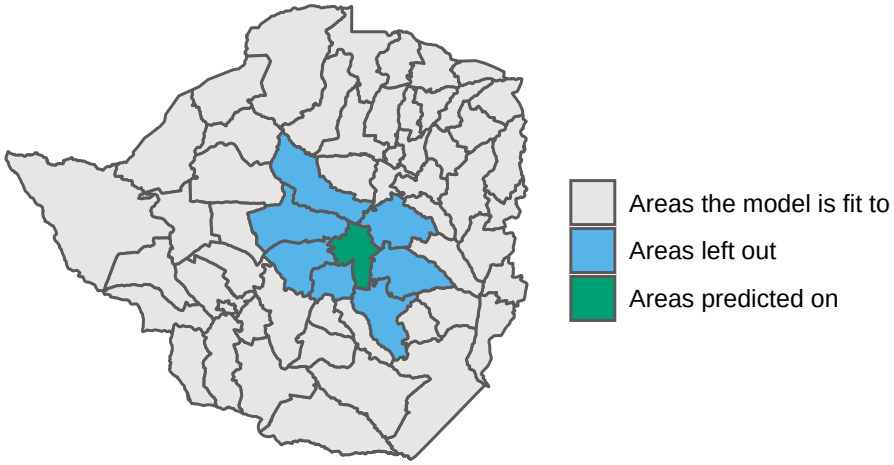

Figure 7: The ith fold in a SLOO-CV in which the model is fitted on the grey areas $A_{-(i,\delta i)}$ with the blue areas $A_{\delta i}$ held-out, to be assessed in predicting the green area A_i.

6.3. Cross-validation

We assessed each model using two cross-validation approaches: (1) a standard leave-one-out cross-validation (LOO-CV), in which at each fold the data from a single district were held-out, and (2) a spatial leave-one-out cross-validation (SLOO-CV), in which at each fold the data from a single district and each of its neighbouring districts were removed. We performed both types of cross-validation manually, as we found approximations that avoid refitting the model, such as the built-in conditional predictive ordinate statistic in `R-INLA` package [32] and the Pareto smoothed importance sampling [71] approach implemented in the `loo` package [72], were inaccurate in this setting. This might be because individual observations in spatial random effect models can exert high influence toward local parameters.

6.3.1. Leave-one-out cross-validation

In the ith LOO-CV fold, the model was fit using the data y_{-i} and assessed according to its prediction on y_i. We assessed forecasting performance using the CRPS, at the level of the data y_i rather than at the level of the prevalence ρ_i. To compute the value of CRPS for the ith fold we use samples $\{y_i^s\}_{s=1}^S$ from the LOO posterior predictive distribution $p(y_i \,|\, y_{-i})$ giving $\text{LOO-CRPS}_i = \frac{1}{S}\sum_{s=1}^S |y_i^s - y_i| - \frac{1}{2S^2}\sum_{s=1}^S\sum_{l=1}^S |y_i^s - y_i|$.

6.3.2. Spatial leave-one-out cross-validation

In the presense of structural dependencies, like spatial dependence, LOO-CV tends to underestimate predictive error [56]. We counteracted this effect by modifying the spatial leave-one-out (SLOO) approach of Le Rest et al. [40]. During the ith fold of SLOO-CV, we leave-out the block $(i, \delta i)$ (Figure 7), increasing the extent to which the training and validation sets are conditionally independent. From the SLOO posterior predictive distribution $p(y_i \,|\, y_{-(i,\delta i)})$ we analogously computed SLOO-CRPS_i.

6.4. Results

	MSE	MAE	CRPS
CIV2017PHIA ($n = 66$)			
Constant	22.6 (3.41)	3.54 (0.265)	1.91 (0.237)
IID	34.7 (4.93)	4.13 (0.297)	1.96 (0.223)
Besag	31.2 (4.13)	4.05 (0.283)	1.99 (0.213)
BYM2	30.9 (4.11)	4.02 (0.281)	1.96 (0.216)

MWI2016PHIA ($n = 64$)			
Constant	1410 (578)	22.5 (3.71)	19.1 (3.57)
IID	2900 (968)	29.2 (4.26)	14.9 (2.54)
Besag	1380 (582)	18 (3.5)	9.55 (2.61)
BYM2	1280 (510)	18.1 (3.26)	9.52 (2.32)
TZA2017PHIA ($n = 356$)			
Constant	30.8 (3.46)	3.82 (0.18)	2.51 (0.165)
IID	64.3 (5.47)	5 (0.22)	2.45 (0.152)
Besag	60.3 (9.33)	4.58 (0.258)	2.53 (0.185)
BYM2	56.7 (5.7)	4.64 (0.223)	2.39 (0.151)
ZWE2016PHIA ($n = 126$)			
Constant	185 (28.2)	9.99 (0.692)	6.99 (0.632)
IID	434 (86.1)	13.6 (0.9)	6.45 (0.487)
Besag	418 (163)	11.1 (1.05)	5.23 (0.493)
BYM2	391 (136)	11.2 (0.963)	5.25 (0.459)

7. Discussion

Funding

AH, HZ were supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). AH, JWE were supported by the Bill and Melinda Gates Foundation (OPP1190661). JWE was supported by UNAIDS and National Institute of Allergy and Infectious Disease of the National Institutes of Health (R01AI136664). SF was supported by the EPSRC (EP/V002910/1). This research was supported by the MRC Centre for Global Infectious Disease Analysis (MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat program and is also part of the EDCTP2 programme supported by the European Union.

Acknowledgements

AH thanks Robert Ashton for help with R package development, and Tim Lucas, Elizaveta Semenova, Marta Blangiardo and Oliver Ratmann for providing feedback on earlier drafts of this manuscript.

Disclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the funding agencies.

References

[1] Sivan Aldor-Noiman, Lawrence D Brown, Andreas Buja, Wolfgang Rolke, and Robert A Stine. The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260, 2013.

[2] L Mark Berliner. Hierarchical Bayesian Time Series Models. In *Maximum Entropy and Bayesian Methods*, pages 15–22. Springer, 1996.

[3] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[4] N Best, N Arnold, A Thomas, L Waller, and E Conlon. Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, volume 6, page 131. Oxford University Press, 1999.

[5] Nicky Best, Sylvia Richardson, and Andrew Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14(1):35–59, 2005.

[6] Michael Betancourt. Robust gaussian processes in stan. 2017. URL `https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html`.

[7] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950.

[8] Trevor Campbell and Tamara Broderick. Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

[9] Andrew David Cliff and J Keith Ord. *Spatial processes: models & applications*. Taylor & Francis, 1981.

[10] SM Cramb, EW Duncan, PD Baade, and KL Mengersen. Investigation of bayesian spatial models. 2018.

[11] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[12] Diego F Cuadros, Jingjing Li, Adam J Branscum, Adam Akullian, Peng Jia, Elizabeth N Mziray, and Frank Tanser. Mapping the spatial variability of hiv infection in sub-saharan africa: Effective information for localized hiv prevention and control. *Scientific reports*, 7(1):1–11, 2017.

[13] A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

[14] CB Dean, MD Ugarte, and AF Militino. Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*, 57(1):197–202, 2001.

[15] Peter J Diggle, Paula Moraga, Barry Rowlingson, Benjamin M Taylor, et al. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

[16] Earl W Duncan, Nicole M White, and Kerrie Mengersen. Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference. *International journal of health geographics*, 16(1):1–16, 2017.

[17] Laura Dwyer-Lindgren, Abraham D Flaxman, Marie Ng, Gillian M Hansen, Christopher JL Murray, and Ali H Mokdad. Drinking patterns in US counties from 2002 to 2012. *American Journal of Public Health*, 105(6):1120–1127, 2015.

[18] Laura Dwyer-Lindgren, Michael A Cork, Amber Sligar, Krista M Steuben, Kate F Wilson, Naomi R Provost, Benjamin K Mayala, John D VanderHeide, Michael L Collison, Jason B Hall, et al. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature*, 570(7760):189–193, 2019.

[19] Jeffrey W. Eaton, Laura Dwyer-Lindgren, Steve Gutreuter, Megan O'Driscoll, Oliver Stevens, Sumali Bajaj, Rob Ashton, Alexandra Hill, Emma Russell, Rachel Esra, Nicolas Dolan, Yusuf O. Anifowoshe, Mark Woodbridge, Ian Fellows, Robert Glaubius, Emily Haeuser, Taylor Okonek, John Stover, Matthew L. Thomas, Jon Wakefield, Timothy M. Wolock, Jonathan Berry, Tomasz Sabala, Nathan Heard, Stephen Delgado, Andreas Jahn, Thokozani Kalua, Tiwonge Chimpandule, Andrew Auld, Evelyn Kim, Danielle Payne, Leigh F. Johnson, Richard G. FitzJohn, Ian Wanyeki, Mary I. Mahy, and Ray W. Shiraishi. Naomi: a new modelling tool for estimating hiv epidemic indicators at the district level in sub-saharan africa. *Journal of the International AIDS Society*, 24(S5):e25788, 2021. doi: https://doi.org/10.1002/jia2.25788. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jia2.25788`.

[20] Rich FitzJohn, Robert Ashton, Alex Hill, Martin Eden, Wes Hinsley, Emma Russell, and James Thompson. *orderly: Lightweight Reproducible Reporting*, 2022. https://www.vaccineimpact.org/orderly/, https://github.com/vimc/orderly.

[21] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. Multi-instance kernels. In *ICML*, volume 2, page 7, 2002.

[22] Alan E Gelfand, Li Zhu, and Bradley P Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45, 2001.

[23] Andrew Gelman and Thomas C Little. Poststratification into many categories using hierarchical logistic regression. 1997.

[24] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[25] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[26] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[27] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952. doi: https://doi.org/10.1111/j.2517-6161.1952.tb00104.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1952.tb00104.x`.

[28] Christoff Gössl, Dorothee P Auer, and Ludwig Fahrmeir. Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562, 2001.

[29] Robert P Haining. *Spatial data analysis: theory and practice*. Cambridge university press, 2003.

[30] Timothy Hallett, Sarah-Jane Anderson, Cynthia Adobea Asante, Noah Bartlett, Victoria Bendaud, Samir Bhatt, Clara Burgert, Diego Fernando Cuadros, Janet Dzangare, Daniela Fecht, et al. Evaluation of geospatial methods to generate subnational HIV prevalence estimates for local level planning. *AIDS*, 30(9):1467–1474, 2016.

[31] O Hamelijnck, T Damoulas, K Wang, and MA Girolami. Multi-resolution multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Leonhard Held, Birgit Schrödle, and Håvard Rue. Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In *Statistical modelling and regression structures*, pages 91–110. Springer, 2010.

[33] Olatunji Johnson, Peter Diggle, and Emanuele Giorgi. A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. *Statistics in Medicine*, 38(24):4871–4887, 2019.

[34] Julia Kelsall and Jonathan Wakefield. Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701, 2002.

[35] Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[36] Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Brad Bell. TMB: automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*, 2015.

[37] Fabian Krüger, Sebastian Lerch, Thordis Thorarinsdottir, and Tilmann Gneiting. Predictive inference based on markov chain monte carlo output. *International Statistical Review*, 2020.

[38] Ho Chung Leon Law, Dino Sejdinovic, Ewan Cameron, Tim CD Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu. Variational learning on aggregate outputs with gaussian processes. *arXiv preprint arXiv:1805.08463*, 2018.

[39] Andrew B Lawson. *Statistical methods in spatial epidemiology*. John Wiley & Sons, 2013.

[40] Kévin Le Rest, David Pinaud, Pascal Monestiez, Joël Chadoeuf, and Vincent Bretagnolle. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7):811–820, 2014.

[41] Duncan Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89, 2011.

[42] Brian G Leroux, Xingye Lei, and Norman Breslow. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191. Springer, 2000.

[43] Ye Li, Patrick Brown, Dionne C Gesink, and Håvard Rue. Log gaussian cox processes and spatially aggregated disease incidence data. *Statistical methods in medical research*, 21(5):479–507, 2012.

[44] Finn Lindgren, Håvard Rue, and Johan Lindström. 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 (Statistical Methodology)*, 73(4):423–498, 2011.

[45] G Matheron. Forecasting block grade distributions: the transfer functions. In *Advanced Geostatistics in the Mining Industry*, pages 237–251. Springer, 1976.

[46] James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

[47] Anita K Nandi, Tim CD Lucas, Rohan Arambepola, Peter Gething, and Daniel J Weiss. disaggregation: An r package for bayesian spatial disaggregation modelling. *arXiv preprint arXiv:2001.04847*, 2020.

[48] Margaret A Oliver and PJ Gregory. Soil, food security and human health: a review. *European Journal of Soil Science*, 66(2):257–276, 2015.

[49] S Openshaw and P.J. Taylor. A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical applications in the spatial science*, pages 127–144, 1979.

[50] Christopher J Paciorek et al. Spatial models for point and areal data using markov random fields on a fine grid. *Electronic Journal of Statistics*, 7:946–972, 2013.

[51] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL `https://doi.org/10.32614/RJ-2018-009`.

[52] Danny Pfeffermann et al. New Important Developments in Small Area Estimation. *Statistical Science*, 28(1):40–68, 2013.

[53] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

[54] JNK Rao and Isabel Molina. Small area estimation, 2015.

[55] Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4): 1145–1165, 2016.

[56] David R Roberts, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929, 2017.

[57] Håvard Rue. Comment on r-inla discussion group thread. URL `https://groups.google.com/g/r-inla-discussion-group/c/l2fSYlbbJJM/m/8vUCjrO_BAAJ`.

[58] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[59] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[60] Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

[61] Teemu Säilynoja, Paul-Christian Bürkner, and Aki Vehtari. Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *arXiv preprint arXiv:2103.10522*, 2021.

[62] James F Saracco, J Andrew Royle, David F DeSante, and Beth Gardner. Modeling spatial variation in avian survival and residency probabilities. *Ecology*, 91(7):1885–1891, 2010.

[63] Volker J Schmid, Brandon Whitcher, Anwar R Padhani, N Jane Taylor, and Guang-Zhong Yang. Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 25(12):1627–1636, 2006.

[64] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

[65] Michael L Stein. Interpolation of spatial data: some theory for kriging. 1999.

[66] Yusuke Tanaka, Toshiyuki Tanaka, Tomoharu Iwata, Takeshi Kurashima, Maya Okawa, Yasunori Akagi, and Hiroyuki Toda. Spatially aggregated Gaussian processes with multivariate areal outputs. In *Advances in Neural Information Processing Systems*, pages 3005–3015, 2019.

[67] Benjamin Taylor, Tilman Davies, Barry Rowlingson, and Peter Diggle. Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. *Journal of Statistical Software*, 63:1–48, 2015.

[68] Yee Whye Teh, Avishkar Bhoopchand, Peter Diggle, Bryn Elesedy, Bobby He, Michael Hutchinson, Ulrich Paquet, Jonathan Read, Nenad Tomasev, and Sheheryar Zaidi. Efficient bayesian inference of instantaneous re-production numbers at fine spatial scales, with an application to mapping and nowcasting the covid-19 epidemic in british local authorities. *URL https://rss. org. uk/RSS/media/File-library/News/2021/WhyeBhoopchand. pdfhttps://localcovid. info/2*, 2021.

[69] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[70] C Edson Utazi, Julia Thorley, VA Alegana, MJ Ferrari, Kristine Nilsen, Saki Takahashi, CJE Metcalf, Justin Lessler, and AJ Tatem. A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. *Statistical Methods in Medical Research*, 28(10-11):3226–3241, 2019.

[71] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[72] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

[73] J Wakefield and S Morris. Spatial dependence and errors-in-variables in environmental epidemiology. *Bayesian statistics*, 6:657–684, 1999.

[74] Jon Wakefield, Taylor Okonek, and Jon Pedersen. Small area estimation of health outcomes. *arXiv preprint arXiv:2006.10266*, 2020.

[75] Jonathan Wakefield and Hilary Lyons. *Spatial Aggregation and the Ecological Fallacy*, volume 2010, pages 541–558. 03 2010. doi: 10.1201/9781420072884-c30.

[76] Daniel J Weiss, Bonnie Mappin, Ursula Dalrymple, Samir Bhatt, Ewan Cameron, Simon I Hay, and Peter W Gething. Re-examining environmental correlates of plasmodium falciparum malaria endemicity: a data-intensive variable selection approach. *Malaria journal*, 14(1):1–18, 2015.

[77] Katie Wilson and Jon Wakefield. Pointless spatial modeling. *Biostatistics*, 21(2):e17–e32, 09 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy041. URL `https://doi.org/10.1093/biostatistics/kxy041`.

[78] Fariba Yousefi, Michael Thomas Smith, and Mauricio A Álvarez. Multi-task learning for aggregated data using gaussian processes. *arXiv preprint arXiv:1906.09412*, 2019.

[79] Harrison Zhu, Adam Howes, Owen van Eer, Maxime Rischard, Yingzhen Li, Dino Sejdinovic, and Seth Flaxman. Aggregated Gaussian Processes with Multiresolution Earth Observation Covariates. *arXiv e-prints*, art. arXiv:2105.01460, May 2021.