Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1

Understanding models for spatial structure in small-area estimation

Adam Howes[*,†,¶], Jeffrey W. Eaton[‡,†], Seth R. Flaxman[§]

Abstract. Spatial correlation between small-areas is typically accounted for using spatial random effects. However, commonly used models for spatial random effects based upon adjacency relations, like the Besag model, make unrealistic and unsatisfying spatial assumptions for non-grid geometries. Furthermore, they do not take into account that across many settings areal data is generated by aggregating continuous geostatistical 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.

In a simulation study, we used a proper scoring rule to evaluate the performance of different spatial random effect specifications. The simulations were performed 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.

1 Introduction

Spatial random effects are frequently used to model spatial variation in areal data (Haining, 2003; Cramb et al., 2018). The most common class of models used to specify spatial random effects are Gaussian Markov random fields (GMRFs), which combine a Gaussian distribution with Markov conditional independence assumptions between areas (Rue and Held, 2005). Under the Markov assumptions, observations made in areas close together are correlated, and more distant relationships are ignored. Perhaps the simplest GMRF model is that of Besag et al. (1991), where information is borrowed equally from each adjacent area. This model is attractive as it requires minimal additional choices and is widely implemented. As such, it has seen widespread use, including in agriculture (Oliver and Gregory, 2015), ecology (Saracco et al., 2010), epidemiology (Lawson, 2013), image analysis (Schmid et al., 2006), neuroscience (Gössl et al., 2001) and public health (Dwyer-Lindgren et al., 2015). However, for irregular geometries like the administrative divisions of a country, the assumptions made by the Besag model about space are unrealistic. As such, in this work we test the hypothesis that using

[*]Imperial College London, Department of Mathematics
[†]Imperial College London, MRC Centre for Global Infectious Disease Analysis
[‡]Harvard University, Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health
[§]University of Oxford, Department of Computer Science
[¶]Corresponding Author

more realistic assumptions about spatial structure improves the performance of small-area estimation models. In doing so, we 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 (Hallett et al., 2016; Cuadros et al., 2017). 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 the covariates most strongly associated with HIV, such as sexual risk behaviour (Howes et al., 2023) or the prevalence of male circumcision, often face analogous measurement difficulties. Possibly as a result, models including covariates have only been found to modestly improve predictions (Dwyer-Lindgren et al., 2019, 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 (Weiss et al., 2015). These circumstances foreground the importance of models for spatial structure 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. (2017) compared seventeen methods for specifying these weights, but did not find any which outperform the Besag model. 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. (2017). Another approach is to take into account that areal data is typically produced by aggregating higher resolution data. Obtaining inference about a variable at a different resolution to that it was observed at is known in geostatistics as the change-of-support problem (Gelfand et al., 2001), and dates to foundational work on block kriging (Krige, 1951; Matheron, 1976). Kelsall and Wakefield (2002) applied ideas from change-of-support modelling 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 (Li et al., 2012; Diggle et al., 2013; Taylor et al., 2015; Johnson et al., 2019). Inference for aggregated data has recently been advanced by Wilson and Wakefield (2018) who consider the SPDE approach of Lindgren et al. (2011), using `R-INLA`, and an empirical Bayes approach, using `TMB` (Kristensen et al., 2016), and made available as a part of the `disaggregation` package (Nandi et al., 2020).

Spatial heterogeneity commonly reflects a combination of both spatially correlated processes and specific circumstances at a given area. It is usually therefore recommended to use a spatial random effect specification which includes both structured and unstructured components, such as the BYM2 model (Simpson et al., 2017) or earlier convolutions such as the BYM (Besag et al., 1991), Leroux (Leroux et al., 2000) or Dean (Dean et al., 2001) models. Evidence for this recommendation includes the comparison studies of Lee (2011) and Riebler et al. (2016). Any improvement to the Besag model

is likely directly transferable to improving these convolution models, and both are of substantive practical interest. For example, the Naomi HIV small-area estimation model (Eaton et al., 2021) uses both Besag and BYM2 random effects.

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 present a simulated case-study, before moving on 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 geostatistical 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 $\mathcal{S}$ such that $A_i \cap A_j = \emptyset$ and $\cup_{i=1}^n A_i = \mathcal{S}$. Areal data are a type of spatial data where observations $y_i \in \mathcal{Y}$ are associated to areas A_i. Examples of areal data include the colour of a pixel, the minimum wage in a state, and the number of disease cases in a region. Geostatistical data are another type of spatial data where instead observations $y(s) \in \mathcal{Y}$ can be made at any location $s \in \mathcal{S}$ of a spatially-continuous stochastic process. Examples of geostatistical spatial data include the temperature at a monitoring station, the response to a household survey, and the quality of a soil sample.

Areal data may often be conceptualised as arising from aggregation of geostatistical data, such that $y_i = \int_{A_i} y(s)\mathrm{d}s$. Notable exceptions include policies determined at an administrative level, such as the minimum wage or disease control measures, which have a truly discrete spatial structure.

In some situations it may be preferable to work using aggregated areal data rather than the original geostatistical data. For one, the geostatistical data may be unavailabile or inaccessible due to privacy constraints, administrative practicality or limitations in storage capacity. Alternatively, opting to model at the area-level can be advantageous as models for geostatistical data tend to be more complex, require more data, have higher computational burden, and offer less immediate applicablility to area-level policy-making. In addition, in some circumstances, seemingly geostatistical data may truly be areal, such as data observed at polling stations or health facilities which individuals from the surrounding area travel to. Ultimately, the decision to model the underlying data generating mechanism as either areal or geostatistical is often a matter of practical considerations and 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,

known as "Tobler's first law of geography" (Tobler, 1970). This property both presents a challenge and an opportunity. Each observation provides less information than it would have had the samples been independent, making it more difficult to estimate global parameters. On the other hand, spatial correlation can be used to improve local parameter estimates, indexed by a particular spatial location, especially 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 (Pfeffermann et al., 2013). In a spatial setting, this is often in small geographic areas. However, 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. For example, in the context of multilevel regression and post-stratification (Gelman and Little, 1997), small-areas are generated by the intersection of demographic variables like age, gender and race, alongside geographic variables. Because of the cost of gathering samples, a survey may be designed to give reliable estimates at at an aggregate 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. See Wakefield et al. (2020) for a recent review of both design-based and model-based approaches to small-area estimation of health-indicators, including both area and unit-level analysis.

2.3 Bayesian model-based approach and latent Gaussian models

Bayesian hierarchical models are an attractive framework for small-area estimation. Let $[n]$ denote $1, \ldots, n$. Areal data $\mathbf{y} = (y_i)_{i \in [n]}$ may be modelled using a three-layer structure (Berliner, 1996; Cressie and Wikle, 2015; Rao and Molina, 2015) given by

$$\text{(Observations)} \qquad \mathbf{y} \sim p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_1), \tag{2.1}$$

$$\text{(Latent field)} \qquad \mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}_2), \tag{2.2}$$

$$\text{(Parameters)} \qquad (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{2.3}$$

where $\mathbf{x} = (x_i)_{i \in [N]}$ is a N-dimensional latent field, and $\boldsymbol{\theta} = (\theta_i)_{i \in [m]}$ are the m-dimensional hyperparameters, with $m < N$. Small-area estimation models are usually descriptive rather than mechanistic, so a natural choice is to model the latent field with a flexible and computationally efficient Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q})$ with mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$. Three-layer models with Gaussian latent fields have been collectively studied as latent Gaussian models (LGMs) (Rue et al., 2009) and comprise a wide array of popular models, including generalised linear mixed models.

2.4 Spatial random effects

The observations in an LGM are conditionally independent and identically distributed given linear predictors $\boldsymbol{\eta} = (\eta)_{i\in[n]}$ such that

$$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_1) = p(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}_1) = \prod_{i=1}^{n} p(y_i \mid \eta_i, \boldsymbol{\theta}_1).$$

Each linear predictor η_i is constructed from the Gaussian latent field such that

$$\eta_i = \beta_0 + \sum_{l=1}^{p} \beta_j z_{ji} + \sum_{k=1}^{r} f_k(u_{ki}),$$

where the intercept β_0, linear effects β_j of the covariates z_{ji}, and functions $f_k(\cdot)$ of the covariates u_{ki} are each Gaussian and may be collected into $\mathbf{x}$. The conditional mean $\mathbb{E}(y_i \mid \eta_i)$ is given by $g(\eta_i)$ where $g : \mathbb{R} \to \mathcal{Y}$ is an inverse link function

Spatial random effects $\phi(\cdot)$ are one Gaussian function which may be included in the model, whose role is to capture the spatial structure between areas not already accounted for by $\beta_0 + \sum_{l=1}^{p} \beta_j z_{ji}$. Should no spatial structure remain, then independent and identically distributed (IID) random effects $\phi_i \sim \mathcal{N}(0, \tau_\phi^{-1})$ may be appropriate, where τ_ϕ is a shared precision. Exploratory data analysis techniques such as visual inspection or Moran's I coefficient (Cliff and Ord, 1981) may be used to determine if there remains any spatial auto-correlation in the data. Specifying $\boldsymbol{\phi} = (\phi_i)_{i\in[n]}$ amounts to specifying the entries of a precision or covariance matrix, as we discuss in the following sections.

3 Spatial random effects defined using adjacency

3.1 Besag

Spatial structure can be encoded using a symmetric relation between areas. Let $i \sim j$ if the areas A_i and A_j are adjacent or neighbouring. Adjacency is often defined by a shared border, though other choices are also possible (Paciorek et al., 2013). The Besag model (Besag et al., 1991) is an improper conditional auto-regressive (ICAR) model where the full conditional distribution of the ith spatial random effect is given by

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim \mathcal{N}\left(\frac{1}{n_{\delta i}} \sum_{j:j\sim i} \phi_j, \frac{1}{n_{\delta i}\tau_\phi}\right), \tag{3.1}$$

where δi is the set of neighbours of A_i with cardinality $n_{\delta i} = |\delta i|$ and $\boldsymbol{\phi}_{-i}$ 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 (Rue and Held, 2005, Chapter 2)

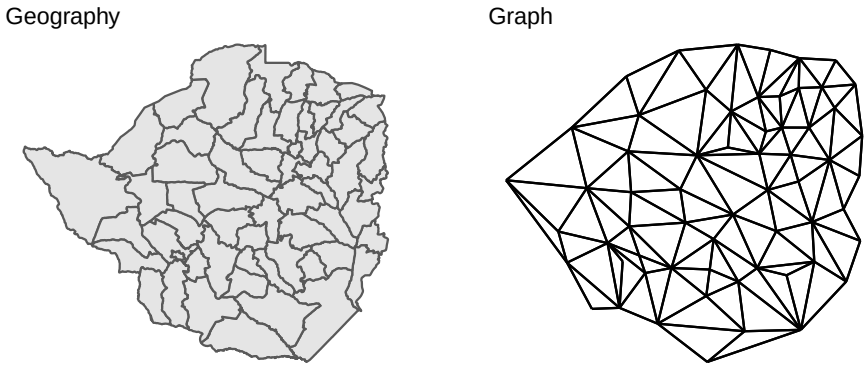

Figure 1: A map of the districts of Zimbabwe together with the corresponding adjacency graph structure $\mathcal{G}$.

the set of full conditionals of the Besag model are equivalent to the Gaussian Markov random field (GMRF) given by

$$\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \tau_\phi^{-1}\mathbf{R}^-), \tag{3.2}$$

where $\mathbf{R}^-$ is the generalised inverse of the rank-deficient structure matrix $\mathbf{R}$, so-called because it defines the structure of the precision matrix, with entries

$$R_{ij} = \begin{cases} n_{\delta i}, & i = j \\ -1, & i \sim j \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

The Markov property arises due to the conditional independence structure $p(\phi_i \,|\, \boldsymbol{\phi}_{-i}) = p(\phi_i \,|\, \boldsymbol{\phi}_{\delta i})$ whereby each area only depends on its neighbours. This is reflected in the sparsity of $\mathbf{R}$ whereby $\phi_i \perp \phi_j \,|\, \boldsymbol{\phi}_{-ij}$ if and only if $R_{ij} = 0$. The structure matrix $\mathbf{R}$ 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. In Section 3.2 we will discuss the appropriateness of using the graph rather than the complete geometry.

Rewriting Equation 3.2, the probability density function of $\boldsymbol{\phi}$ is

$$p(\boldsymbol{\phi}) \propto \exp\left(-\frac{\tau_\phi}{2}\boldsymbol{\phi}^\top \mathbf{R}\boldsymbol{\phi}\right) \propto \exp\left(-\frac{\tau_\phi}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right). \tag{3.4}$$

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 $p(\boldsymbol{\phi}) = p(\boldsymbol{\phi} + c\mathbf{1})$, leading to an improper uniform distribution on the average of the ϕ_i. If $\mathcal{G}$ is connected, in that by traversing the edges,

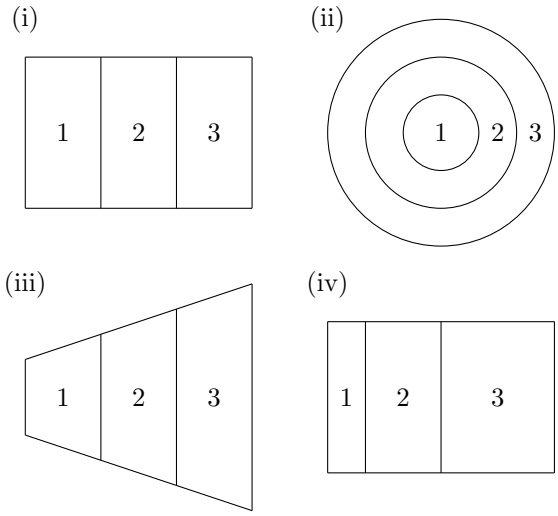

Figure 2: Each of these four geometries have the same adjacency graph.

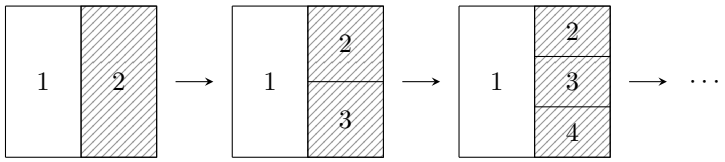

Figure 3: A sequence of geometries where the number of neighbours of area one grows by one at each iteration.

any vertex can be reached from any other vertex, then there is only one impropriety in the model and $\text{rank}(\mathbf{R}) = n - 1$, 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 $\mathbf{R}$ has rank $n - n_c$ and the density is invariant to the addition of a constant to each of the connected components $p(\boldsymbol{\phi}_I) = p(\boldsymbol{\phi}_I + c\mathbf{1})$ 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. We highlight a number of concerns about the model's applicability to this broader setting. This discussion is closely linked to the modifiable areal unit problem (Openshaw and Taylor, 1979) whereby statistical conclusions change as a result of seemingly arbitrary changes in data aggregation, and the challenge of ecological inference and the ecological fallacy (Wakefield and Lyons, 2010), which applies particularly to inference about individuals

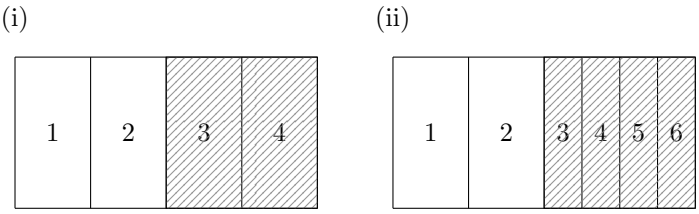

Figure 4: Each of the shaded areas are split into two moving from geometry (i) to (ii).

and groups.

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, but it prompts consideration as to whether the class of geometries with the same adjacency graph is sufficiently similar to merit identical models.[1] Intuitively 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.

Mean structure

In the Besag model is 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, and the proximity of centers of mass. This may be remedied using more general weighted ICAR models, as we outline in Section 3.3.

Variance structure

In Equation 3.1 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 $\text{Var}(\phi_i) \to 0$. This is illustrated by Figure 3 where the area on the right is repeatedly divided such that its number of neighbours increases. This property is a consequence of averaging the conditional mean over a greater number of

[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.

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 seems inappropriate that $\text{Var}(\phi_1)$ should tend to zero simply 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 areas is not expected to correspond to a greater amount of information.

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^2_ϕ 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^2_ϕ 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 $\boldsymbol{\kappa} = (\kappa_i)_{i\in[n]}$ such that $\phi_i \,|\, \boldsymbol{\phi}_{-i} \sim \mathcal{N}(\sum_{j:j\sim i} b_{ij}\phi_j, (\kappa_i\tau_\phi)^{-1})$. The structure matrix $\mathbf{R}$ corresponding to the above full conditionals is given by $\mathbf{R} = \mathbf{D}_\kappa(\mathbf{I} - \mathbf{B})$, where $\mathbf{B}$ 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 $\mathbf{D}_\kappa$ is given by $\text{diag}(\kappa_1, \ldots, \kappa_n)$. 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 $\mathbf{W} = \mathbf{D}_\kappa\mathbf{B}$ such that $\mathbf{R} = \mathbf{D}_\kappa - \mathbf{W}$. 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 $\mathbf{W}$ is provided by Duncan et al. (2017). Much of the work in this area focuses on the case where the geometry is a lattice.

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. (1991), accounts for this in a natural way by decomposing the spatial random effect $\boldsymbol{\phi} = \mathbf{v} + \mathbf{u}$ into a sum of an unstructured IID component $\mathbf{v}$ and a spatially structured Besag component $\mathbf{u}$, each of which with their own respective precision parameters τ_v and τ_u. The resulting distribution is

$$\boldsymbol{\phi} \sim \mathcal{N}(0, \tau_v^{-1}\mathbf{I} + \tau_u^{-1}\mathbf{R}^-). \tag{3.5}$$

Including both $\mathbf{v}$ and $\mathbf{u}$ 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 $\mathbf{Q}$ 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 $\mathbf{Q}$ and $\mathbf{I}$ have the same scale. Additionally, supposing we have a prior belief that the over-dispersion is primarily IID and $\mathbf{v}$ 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. (2000) and Dean et al. (2001) which tackle this the identifiability problem, but do not scale the spatially structured noise, Simpson et al. (2017) propose a reparameterisation $(\tau_v, \tau_u) \mapsto (\tau_\phi, \pi)$ of the BYM model known as the BYM2 model and given by

$$\boldsymbol{\phi} = \frac{1}{\tau_\phi}\left(\sqrt{1-\pi}\,\mathbf{v} + \sqrt{\pi}\,\mathbf{u}^\star\right), \tag{3.6}$$

where τ_ϕ is the marginal precision of $\boldsymbol{\phi}$, $\pi \in [0, 1]$ gives the proportion of the marginal variance explained by each component, and $\mathbf{u}^\star$ is a scaled version of u with precision matrix given by the scaled structure matrix $\mathbf{R}^\star$ (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 (Rue) 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.

4 Spatial random effects defined using kernels

4.1 Areal kernels

In the previous section we 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 which gives a measure of similarity between two areas $K : \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to \mathbb{R}$, where $\mathcal{P}$ denotes the power set such that $\mathcal{P}(\mathcal{S})$ is the space of subsets of the study region. If K is positive semi-definite, then we define areal kernel spatial random effects by

$$\phi \sim \mathcal{N}\left(0, \frac{1}{\tau_\phi}\mathbf{K}\right), \tag{4.1}$$

where the $n \times n$ Gram matrix $\mathbf{K}$ with entries $K_{ij} = K(A_i, A_j)$ is a valid covariance matrix. 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 consider a simple method to construct K from k: namely averaging the kernel k computed on some collection of points from within each area.

4.2 Centroid kernel

The simplest approach is to use a single point. A natural choice is the centroid $c_i \in A_i$, given by the arithmetic mean of the latitude and longitude, which we might hope to be

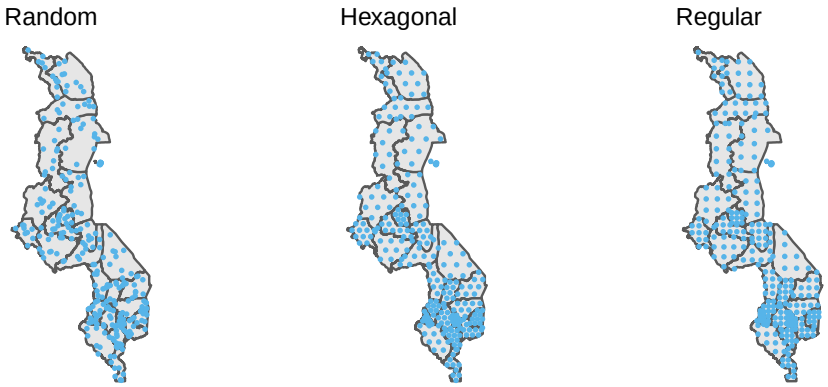

Figure 5: Three ways to choose nodes with $L_i = 10$ for all i in Malawi, as implemented by `sf::st_sample`.

representative of the area. This results in the centroid kernel

$$K(A_i, A_j) = k(c_i, c_j). \tag{4.2}$$

The centroid kernel has been used in environmental epidemiology (Wakefield and Morris, 1999) and to model the reproduction number of COVID-19 (Teh et al., 2021). In a model comparison study Best et al. (2005, 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 oversmooth 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 looking 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

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} k(s, s') \mathrm{d}s \mathrm{d}s'. \tag{4.3}$$

This covariance structure is equivalent to that obtained by aggregating a spatially continuous Gaussian process with kernel k over the areal partition, studied in the machine learning literature under the name aggregated Gaussian processes (Law et al., 2018; Tanaka et al., 2019; Yousefi et al., 2019; Hamelijnck et al., 2019). 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.

Accounting for heterogeneity

If available, additional information accounting for heterogeneity over A_i may be incorporated into the integrated kernel. 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

$$K(A_i, A_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} w_i(s) w_j(s') k(s, s') \mathrm{d}s \mathrm{d}s', \tag{4.4}$$

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.

Computation

Most of the time we do not expect to be able to calculate Equation 4.4 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

$$K(A_i, A_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{m=1}^{L_j} w_i\left(s_l^{(i)}\right) w_j\left(s_m^{(j)}\right) k\left(s_l^{(i)}, s_m^{(j)}\right). \tag{4.5}$$

Equivalently, samples drawn from W_i may be used without weighting by $w_i(s)$. Nodes may also be selected by some deterministic process to give a numerical quadrature estimate of the kernel. These approaches require $\mathcal{O}(\sum_{i=1}^{n} \sum_{j=1}^{n} L_i L_j)$ 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 nodes it is feasible to use. Kelsall and Wakefield (2002) make inference more feasible by using a discrete hyperparameter prior to reduce the number of Gram matrix constructions and inversions required.

Mismatch to data generating process

Aggregation via the integrated kernel occurs at the level of the latent field rather than at the level of the data. 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.

Connection to log-Gaussian Cox processes

The log-Gaussian Cox Process framework (Diggle et al., 2013) arrives naturally 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

$$y_i \,|\, x(s) \sim \text{Poisson}\left(\int_{s \in A_i} x(s)\mathrm{d}s\right). \tag{4.6}$$

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. (2019) obtain Equation 4.4 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. The ith discrete spatial random effect is then $\phi_i = \int_{A_i} w_i(s)\phi(s)\mathrm{d}s$, with covariance structure

$$\text{Cov}\left(\int_{A_i} w_i(s)\phi(s)\mathrm{d}s, \int_{A_j} w_j(s')\phi(s')\mathrm{d}s'\right) = \int_{A_i}\int_{A_j} w_i(s)w_j(s')k(s, s')\mathrm{d}s\mathrm{d}s', \tag{4.7}$$

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

Connection to disaggregation regression

Disaggregation regression, also known as downscaling or interpolation, is another closely related approach. Rather than focusing on the aggregate nature of areal observations as primarily a route towards better area-level estimates, disaggregation regression aims to produce high-resolution or point-level estimates from areal observations (Utazi et al., 2019; Nandi et al., 2020).

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 setting in mind. The R (R Core Team, 2021) code used is available from `github.com/athowes/areal-comparison`. We used `orderly` (FitzJohn et al., 2022) for reproducible research, `ggplot2` for data visualisation (Wickham, 2016) and `rticles` (Allaire et al., 2022a) for reporting via `rmarkdown` (Allaire et al., 2022b).

5.1 Synthetic data-sets

To study robustness to spatial misspecification we simulated synthetic data-sets from three known data generating processes, using spatial random effects $\boldsymbol{\phi}$ generated according to IID, Besag and integrated kernel (IK) simulation models (Table 1). We then generated synthetic data $\mathbf{y} = (y_i)_{i \in [n]}$ 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 linear

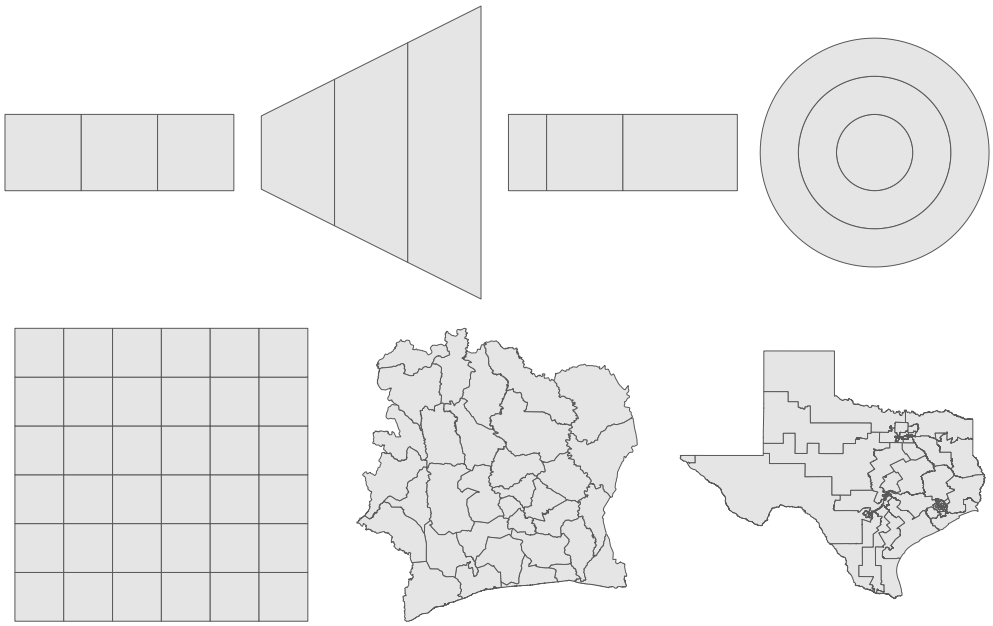

Figure 6: The seven simulation geometries that we considered.

	Simulation model	Additional details
`S1`	IID	$\boldsymbol{\phi} \sim \mathcal{N}(0, \mathbf{I}_n)$
`S2`	Besag	$\boldsymbol{\phi} \sim \mathcal{N}(0, (\mathbf{R}^\star)^-)$ as described in Section 3.1 where $\mathbf{R}^\star$ is the scaled structure matrix
`S3`	IK	$\boldsymbol{\phi} \sim \mathcal{N}(0, \mathbf{K}^\star)$ as described in Section 4.3 where K_{ij} is calculated via Equation 4.5 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 $\boldsymbol{\phi}$.

	Inferential model	Additional details
`I1`	Constant	No spatial random effects
`I2`	IID	$\boldsymbol{\phi} \sim \mathcal{N}(0, \tau_\phi^{-1}\mathbf{I}_n)$
`I3`	Besag	$\boldsymbol{\phi} \sim \mathcal{N}(0, (\tau_\phi \mathbf{R}^\star)^-)$ following additional recommendations of Freni-Sterrantino et al. (2018)
`I4`	BYM2	$\boldsymbol{\phi} = \tau_\phi^{-1}\left(\sqrt{1-\pi}\,\mathbf{v} + \sqrt{\pi}\,\mathbf{u}^\star\right)$ with a Beta$(0.5, 0.5)$ prior on $\pi \in [0,1]$
`I5`	FCK	$\boldsymbol{\phi} \sim \mathcal{N}(0, \tau_\phi^{-1}\mathbf{K})$ with $K_{ij} = k(c_i, c_j)$ where the length-scale l is fixed
`I6`	CK	As `I5`, with length-scale prior $l \sim$ Inv-Gamma(a, b)
`I7`	FIK	$\boldsymbol{\phi} \sim \mathcal{N}(0, \tau_\phi^{-1}\mathbf{K})$ 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	As `I7`, with length-scale prior $l \sim$ Inv-Gamma(a, b)

Table 2: All inferential models are implemented as a part of the R package `arealutils` available from `github.com/athowes/arealutils`. The AGHQ algorithm was used for approximate Bayesian inference of all inferential models.

predictors $\eta_i \in \mathbb{R}$ via

$$\log\left(\frac{\rho_i}{1-\rho_i}\right) = \eta_i = \beta_0 + \phi_i, \quad i \in [n]. \tag{5.1}$$

The sample sizes were fixed as $m_i = 25$ for all $i \in [n]$, the intercept parameter as $\beta_0 = -2$ and the spatial random effect precision parameter as $\tau_\phi = 1$. We used seven different geometries including the four vignette geometries sharing an adjacency graph (Figure 2), 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). The more 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).

Priors

We placed a half-Gaussian prior on the standard deviation (Gelman et al., 2006) 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. 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.

Kernels

We used the Matérn (Stein, 1999) kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ given by

$$k(s, s') = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|s - s'|}{l} \right)^{\nu} B_\nu \left(\frac{\sqrt{2\nu}|s - s'|}{l} \right), \tag{5.2}$$

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 5.2 to be

$$k(s, s') = \left(1 + \sqrt{3}|s - s'|/l\right) \exp\left(-\sqrt{3}|s - s'|/l\right). \tag{5.3}$$

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 (Best et al., 1999). In models `I6` and `I8`, we used a length-scale prior $l \sim \text{Inv-Gamma}(a, b)$, 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 (Betancourt, 2017).

We set $L_i = 10$ using a hexagonal spacing structure, as with the central panel of Figure 6.[2] We are limited to this relatively small number by the computational costs of Model `I8`. For Model `I6` the Gram matrix is only computed once and it is feasible to use a much larger number of nodes, but we chose not to for comparability.

5.3 Inference

We used adaptive Gauss-Hermite quadrature [AGHQ; Stringer et al. (2022)] with $k = 3$ quadrature points per hyperparameter dimension to perform approximate Bayesian inference. Each model is implemented using a Template Model Builder C++ template for the log-posterior via the `TMB` package (Kristensen et al., 2016) We are then able to conduct inference using both AGHQ via the `aghq` package (Stringer, 2021) and the No-U-Turn Sampling (NUTS) Hamiltonian Monte Carlo (HMC) algorithm using Stan (Carpenter et al., 2017) via the `tmbstan` package (Monnahan and Kristensen, 2018). In Appendix we demonstrate that the inference results from AGHQ are comparable to those obtained running NUTS over a longer time. We also compare to empirical Bayes, and where possible integrated nested Laplace approximation via `R-INLA` (Martins et al., 2013).

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

[2] Note that `sf::st_sample` with `type = "hexagonal"` does not guarantee exactly the specified number of samples are returned (Pebesma, 2018).

calibration of the model's estimates.

Continuous ranked probability score

The continuous ranked probability score [CRPS; Matheson and Winkler (1976)] generalises the Brier score (Brier, 1950) to distributional forecasts. Let ρ_i have posterior marginal $f(\rho_i) = p(\rho_i \mid y)$ and ω_i be the true value. Writing F as the cumulative distribution function corresponding to f then

$$\text{CRPS}(f, \omega_i) = \int_{-\infty}^{\infty} (F(\rho_i) - \mathbb{1}\{\rho_i \geq \omega_i\})^2 \text{d}\rho_i = \int_0^1 (F(\rho_i) - \mathbb{1}\{\rho_i \geq \omega_i\})^2 \text{d}\rho_i, \quad (5.4)$$

where $\mathbb{1}$ denotes the indicator function and the second equality follows from $0 \leq \rho_i \leq 1$. $\text{CRPS}(f, \omega_i)$ if and only if $f(\rho_i) = \delta_{\omega_i}$. The CRPS is a strictly proper scoring rule (Gneiting and Raftery, 2007) and can be evaluated directly using samples $\{\rho_i^s\}_{s=1}^S$ as

$$\text{CRPS}(f, \omega_i) \approx \frac{1}{S} \sum_{s=1}^{S} |\rho_i^s - \omega_i| - \frac{1}{2S^2} \sum_{s=1}^{S} \sum_{l=1}^{S} |\rho_i^s - \rho_i^l|. \quad (5.5)$$

Posterior predictive check for coverage

For calibrated models, over repeated simulations, $F(\omega_i) = q_i \sim \mathcal{U}[0, 1]$ such that at any given nominal coverage $1 - \alpha$, the proportion of quantile-based credible intervals containing the true value ω_i is also $1 - \alpha$. We checked uniformity using probability integral transform (PIT) histograms (Dawid, 1984) and empirical cumulative distribution function (ECDF) difference plots (Aldor-Noiman et al., 2013; Säilynoja et al., 2022).

5.5 Results

See Appendix.

6 HIV prevalence study

We compared model performance in estimating HIV prevalence $\rho_i \in [0, 1]$ 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 ($n = 60$). As before, we varied the spatial random effect specification. The R code for this study is available from https://github.com/athowes/areal-comparison.

6.1 Household survey data

In each country we used data from the most recent publicly available Population Health Impact Assessment (PHIA) survey. 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. We used sampling weights $w_{ij} = 1/\pi_{ij}$ to account for the survey design by adjusting the raw data in each district, obtaining the Kish effective sample size (Kish, 1965) $m_i^\star = (\sum_k w_{ij})^2 / \sum_k w_{ij}^2$ 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 make a slight alteration to the inferential models from Table 2. As the effective number of cases $y_i^\star \in \mathbb{R}$ and effective sample size $m_i^\star \in \mathbb{R}$ may not be integers, we use a generalised binomial distribution $y_i^\star \sim \text{xBin}(m_i^\star, \rho_i)$. The working likelihood under this model for $m_i^\star \geq y_i^\star$ is given by

$$p(y_i^\star \mid m_i^\star, \rho_i) = \frac{\Gamma(m_i^\star + 1)}{\Gamma(y_i^\star + 1)\Gamma(m_i^\star - y_i^\star + 1)} \rho_i^{y_i^\star} (1 - \rho_i)^{(m_i^\star - y_i^\star)}. \tag{6.1}$$

6.3 Cross-validation

We assessed each model using two approaches: (1) a standard leave-one-out cross-validation (LOO-CV), and (2) a spatial leave-one-out cross-validation (SLOO-CV).

Leave-one-out cross-validation

In the ith LOO-CV fold, the model was fit using the data $\mathbf{y}_{-i}$ and assessed according to its prediction on y_i. We assessed forecasting performance using the root mean square error (RMSE), mean absolute error (MAE), and CRPS, at the level of the data y_i rather than at the level of the prevalence ρ_i.

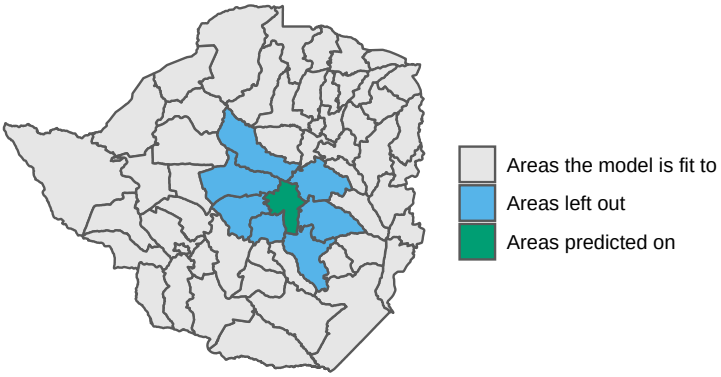

Figure 7: The ith fold in a SLOO-CV in which the model is fitted on the grey areas $A_{-(i,\delta i)}$ with the blue areas $A_{\delta i}$ held-out, to be assessed in predicting the green area A_i.

Spatial leave-one-out cross-validation

In the presence of structural dependencies, LOO-CV tends to underestimate predictive error (Le Rest et al., 2014; Roberts et al., 2017). To counteract this effect, during fold i we left out the block $(i, \delta i)$, increasing the extent to which the training and validation sets are conditionally independent, and predicted on y_i (Figure 7).

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

Areal kernels can be thought of as kernels on sets (Gärtner et al., 2002). Though it may be possible to learn more sophisticated methods for combining kernels using data (Gönen and Alpaydın, 2011), we do believe that this is feasible in the small-area estimation setting, and instead focused on constructing areal kernels which represent our prior beliefs about spatial processes. In more data rich settings it may be possible to learn the nodes used within the integrated kernel (Campbell and Broderick, 2019).

Funding

AH was 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.

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.

Supplementary Material

Appendix. All further materials for this study.

Acknowledgments

AH thanks Robert Ashton for help with R package development, and Harrison Zhu, Theo Rashid, Tim Wolock, Samir Bhatt, Tim Lucas, Elizaveta Semenova, Marta Blangiardo and Oliver Ratmann for helpful conversations. Yidan Xu undertook the Mary Lister McCammon Summer Research Fellowship 2019 on related work.

References

Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W., and Stine, R. A. (2013). "The power to see: A new graphical test of normality." *The American Statistician*, 67(4): 249–260. 17

Allaire, J., Xie, Y., Dervieux, C., R Foundation, Wickham, H., Journal of Statistical Software, Vaidyanathan, R., Association for Computing Machinery, Boettiger, C., Elsevier, Broman, K., Mueller, K., Quast, B., Pruim, R., Marwick, B., Wickham, C., Keyes, O., Yu, M., Emaasit, D., Onkelinx, T., Gasparini, A., Desautels, M.-A., Leutnant, D., MDPI, Taylor and Francis, Öğreden, O., Hance, D., Nüst, D., Uvesten, P., Campitelli, E., Muschelli, J., Hayes, A., Kamvar, Z. N., Ross, N., Cannoodt, R., Luguern, D., Kaplan, D. M., Kreutzer, S., Wang, S., Hesselberth, J., and Hyndman, R. (2022a). *rticles: Article Formats for R Markdown*. R package version 0.23.6. URL https://github.com/rstudio/rticles 13

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2022b). *rmarkdown: Dynamic Documents for R*. R package version 2.14. URL https://github.com/rstudio/rmarkdown 13

Berliner, L. M. (1996). "Hierarchical Bayesian Time Series Models." In *Maximum Entropy and Bayesian Methods*, 15–22. Springer. 4

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43(1): 1–20. 1, 2, 5, 9

Best, N., Arnold, N., Thomas, A., Waller, L., and Conlon, E. (1999). "Bayesian models for spatially correlated disease and exposure data." In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, volume 6, 131. Oxford University Press. 16

Best, N., Richardson, S., and Thomson, A. (2005). "A comparison of Bayesian spatial models for disease mapping." *Statistical Methods in Medical Research*, 14(1): 35–59. 11

Betancourt, M. (2017). "Robust Gaussian processes in Stan."

URL https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html 16

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly weather review*, 78(1): 1–3. 17

Campbell, T. and Broderick, T. (2019). "Automated scalable Bayesian inference via Hilbert coresets." *The Journal of Machine Learning Research*, 20(1): 551–588. 20

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76(1). 16

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*. Taylor & Francis. 5

Cramb, S., Duncan, E., Baade, P., and Mengersen, K. (2018). "Investigation of Bayesian spatial models." 1

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons. 4

Cuadros, D. F., Li, J., Branscum, A. J., Akullian, A., Jia, P., Mziray, E. N., and Tanser, F. (2017). "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. 2

Dawid, A. P. (1984). "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. 17

Dean, C., Ugarte, M., and Militino, A. (2001). "Detecting interaction between random region and fixed age effects in disease mapping." *Biometrics*, 57(1): 197–202. 2, 10

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. (2013). "Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm." *Statistical Science*, 28(4): 542–563. 2, 12

Duncan, E. W., White, N. M., and Mengersen, K. (2017). "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. 2, 9

Dwyer-Lindgren, L., Cork, M. A., Sligar, A., Steuben, K. M., Wilson, K. F., Provost, N. R., Mayala, B. K., VanderHeide, J. D., Collison, M. L., Hall, J. B., et al. (2019). "Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017." *Nature*, 570(7760): 189–193. 2

Dwyer-Lindgren, L., Flaxman, A. D., Ng, M., Hansen, G. M., Murray, C. J., and Mokdad, A. H. (2015). "Drinking patterns in US counties from 2002 to 2012." *American Journal of Public Health*, 105(6): 1120–1127. 1

Eaton, J. W., Dwyer-Lindgren, L., Gutreuter, S., O'Driscoll, M., Stevens, O., Bajaj, S., Ashton, R., Hill, A., Russell, E., Esra, R., Dolan, N., Anifowoshe, Y. O., Wood-

bridge, M., Fellows, I., Glaubius, R., Haeuser, E., Okonek, T., Stover, J., Thomas, M. L., Wakefield, J., Wolock, T. M., Berry, J., Sabala, T., Heard, N., Delgado, S., Jahn, A., Kalua, T., Chimpandule, T., Auld, A., Kim, E., Payne, D., Johnson, L. F., FitzJohn, R. G., Wanyeki, I., Mahy, M. I., and Shiraishi, R. W. (2021). "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. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jia2.25788 3

FitzJohn, R., Ashton, R., Hill, A., Eden, M., Hinsley, W., Russell, E., and Thompson, J. (2022). *orderly: Lightweight Reproducible Reporting*. https://www.vaccineimpact.org/orderly/, https://github.com/vimc/orderly. 13

Freni-Sterrantino, A., Ventrucci, M., and Rue, H. (2018). "A note on intrinsic conditional autoregressive models for disconnected graphs." *Spatial and spatio-temporal epidemiology*, 26: 25–34. 15

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. J. (2002). "Multi-instance kernels." In *ICML*, volume 2, 7. 20

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). "On the change of support problem for spatio-temporal data." *Biostatistics*, 2(1): 31–45. 2

Gelman, A. and Little, T. C. (1997). "Poststratification into many categories using hierarchical logistic regression." 4

Gelman, A. et al. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis*, 1(3): 515–534. 15

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. 17

Gönen, M. and Alpaydın, E. (2011). "Multiple kernel learning algorithms." *The Journal of Machine Learning Research*, 12: 2211–2268. 20

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). "Bayesian spatiotemporal inference in functional magnetic resonance imaging." *Biometrics*, 57(2): 554–562. 1

Haining, R. P. (2003). *Spatial data analysis: theory and practice*. Cambridge university press. 1

Hallett, T., Anderson, S.-J., Asante, C. A., Bartlett, N., Bendaud, V., Bhatt, S., Burgert, C., Cuadros, D. F., Dzangare, J., Fecht, D., et al. (2016). "Evaluation of geospatial methods to generate subnational HIV prevalence estimates for local level planning." *AIDS*, 30(9): 1467–1474. 2

Hamelijnck, O., Damoulas, T., Wang, K., and Girolami, M. (2019). "Multi-resolution multi-task Gaussian processes." *Advances in Neural Information Processing Systems*, 32. 11

Howes, A., Risher, K. A., Nguyen, V. K., Stevens, O., Jia, K. M., Wolock, T. M., Esra, R. T., Zembe, L., Wanyeki, I., Mahy, M., Benedikt, C., Flaxman, S. R., and Eaton,

J. W. (2023). "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa." *PLOS Global Public Health*, 3(4): 1–14.
URL https://doi.org/10.1371/journal.pgph.0001731 2

Johnson, O., Diggle, P., and Giorgi, E. (2019). "A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data." *Statistics in Medicine*, 38(24): 4871–4887. 2, 13

Kelsall, J. and Wakefield, J. (2002). "Modeling spatial variation in disease risk: a geostatistical approach." *Journal of the American Statistical Association*, 97(459): 692–701. 2, 12

Kish, L. (1965). "Survey sampling." 18

Krige, D. G. (1951). "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. 2

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., Bell, B. M., et al. (2016). "TMB: Automatic Differentiation and Laplace Approximation." *Journal of Statistical Software*, 70(i05). 2, 16

Law, H. C., Sejdinovic, D., Cameron, E., Lucas, T., Flaxman, S., Battle, K., and Fukumizu, K. (2018). "Variational learning on aggregate outputs with Gaussian processes." *Advances in neural information processing systems*, 31. 11

Lawson, A. B. (2013). *Statistical methods in spatial epidemiology*. John Wiley & Sons. 1

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V. (2014). "Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation." *Global ecology and biogeography*, 23(7): 811–820. 19

Lee, D. (2011). "A comparison of conditional autoregressive models used in Bayesian disease mapping." *Spatial and Spatio-temporal Epidemiology*, 2(2): 79–89. 2

Leroux, B. G., Lei, X., and Breslow, N. (2000). "Estimation of disease rates in small areas: a new mixed model for spatial dependence." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 179–191. Springer. 2, 10

Li, Y., Brown, P., Gesink, D. C., and Rue, H. (2012). "Log Gaussian Cox processes and spatially aggregated disease incidence data." *Statistical methods in medical research*, 21(5): 479–507. 2

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. 2

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). "Bayesian computing with INLA: new features." *Computational Statistics & Data Analysis*, 67: 68–83. 16

Matheron, G. (1976). "Forecasting block grade distributions: the transfer functions." In *Advanced Geostatistics in the Mining Industry*, 237–251. Springer. 2

Matheson, J. E. and Winkler, R. L. (1976). "Scoring rules for continuous probability distributions." *Management science*, 22(10): 1087–1096. 17

Monnahan, C. C. and Kristensen, K. (2018). "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages." *PloS one*, 13(5): e0197954. 16

Nandi, A. K., Lucas, T. C., Arambepola, R., Gething, P., and Weiss, D. J. (2020). "disaggregation: An R Package for Bayesian Spatial Disaggregation Modelling." *arXiv preprint arXiv:2001.04847*. 2, 13

Oliver, M. A. and Gregory, P. (2015). "Soil, food security and human health: a review." *European Journal of Soil Science*, 66(2): 257–276. 1

Openshaw, S. and Taylor, P. (1979). "A million or so correlation coefficients, three experiments on the modifiable areal unit problem." *Statistical applications in the spatial science*, 127–144. 7

Paciorek, C. J. et al. (2013). "Spatial models for point and areal data using Markov random fields on a fine grid." *Electronic Journal of Statistics*, 7: 946–972. 5

Pebesma, E. (2018). "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal*, 10(1): 439–446.
URL https://doi.org/10.32614/RJ-2018-009 16

Pfeffermann, D. et al. (2013). "New Important Developments in Small Area Estimation." *Statistical Science*, 28(1): 40–68. 4

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/ 13

Rao, J. and Molina, I. (2015). "Small Area Estimation." 4

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). "An intuitive Bayesian spatial model for disease mapping that accounts for scaling." *Statistical methods in medical research*, 25(4): 1145–1165. 2

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, 40(8): 913–929. 19

Rue, H. (????). "Comment on R-INLA Discussion Group thread."
URL https://groups.google.com/g/r-inla-discussion-group/c/l2fSYlbbJJM/m/8vUCjrO_BAAJ 10

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press. 1, 5

Rue, H., Martino, S., and Chopin, N. (2009). "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. 4

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2022). "Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison." *Statistics and Computing*, 32(2): 32. 17

Saracco, J. F., Royle, J. A., DeSante, D. F., and Gardner, B. (2010). "Modeling spatial variation in avian survival and residency probabilities." *Ecology*, 91(7): 1885–1891. 1

Schmid, V. J., Whitcher, B., Padhani, A. R., Taylor, N. J., and Yang, G.-Z. (2006). "Bayesian methods for pharmacokinetic models in dynamic contrast-enhanced magnetic resonance imaging." *IEEE Transactions on Medical Imaging*, 25(12): 1627–1636. 1

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. 2, 10

Stein, M. L. (1999). "Interpolation of spatial data: some theory for kriging." 16

Stringer, A. (2021). "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package." *arXiv preprint arXiv:2101.04468*. 16

Stringer, A., Brown, P., and Stafford, J. (2022). "Fast, scalable approximations to posterior distributions in extended latent Gaussian models." *Journal of Computational and Graphical Statistics*, 1–15. 16

Tanaka, Y., Tanaka, T., Iwata, T., Kurashima, T., Okawa, M., Akagi, Y., and Toda, H. (2019). "Spatially aggregated Gaussian processes with multivariate areal outputs." In *Advances in Neural Information Processing Systems*, 3005–3015. 11

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). "Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R." *Journal of Statistical Software*, 63: 1–48. 2

Teh, Y. W., Bhoopchand, A., Diggle, P., Elesedy, B., He, B., Hutchinson, M., Paquet, U., Read, J., Tomasev, N., and Zaidi, S. (2021). "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*. 11

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." *Economic geography*, 46(sup1): 234–240. 4

Utazi, C. E., Thorley, J., Alegana, V., Ferrari, M., Nilsen, K., Takahashi, S., Metcalf, C., Lessler, J., and Tatem, A. (2019). "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. 13

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. 2

Wakefield, J. and Lyons, H. (2010). *Spatial Aggregation and the Ecological Fallacy*, volume 2010, 541–558. 7

Wakefield, J. and Morris, S. (1999). "Spatial dependence and errors-in-variables in environmental epidemiology." *Bayesian statistics*, 6: 657–684. 11

Wakefield, J., Okonek, T., and Pedersen, J. (2020). "Small Area Estimation of Health Outcomes." *arXiv preprint arXiv:2006.10266*. 4

Weiss, D. J., Mappin, B., Dalrymple, U., Bhatt, S., Cameron, E., Hay, S. I., and Gething, P. W. (2015). "Re-examining environmental correlates of Plasmodium falciparum malaria endemicity: a data-intensive variable selection approach." *Malaria journal*, 14(1): 1–18. 2

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL https://ggplot2.tidyverse.org 13

Wilson, K. and Wakefield, J. (2018). "Pointless spatial modeling." *Biostatistics*, 21(2): e17–e32.
URL https://doi.org/10.1093/biostatistics/kxy041 2

Yousefi, F., Smith, M. T., and Alvarez, M. (2019). "Multi-task learning for aggregated data using Gaussian processes." *Advances in Neural Information Processing Systems*, 32. 11