Spatial models for point and areal data using Markov random fields on a fine grid

Christopher J. Paciorek

Department of Statistics
University of California, Berkeley
Berkeley, CA 94720 USA
e-mail: paciorek@stat.berkeley.edu

Abstract: I consider the use of Markov random fields (MRFs) on a fine grid to represent latent spatial processes when modeling point-level and areal data, including situations with spatial misalignment. Point observations are related to the grid cell in which they reside, while areal observations are related to the (approximate) integral over the latent process within the area of interest. I review several approaches to specifying the neighborhood structure for constructing the MRF precision matrix, presenting results comparing these MRF representations analytically, in simulations, and in two examples. The results provide practical guidance for choosing a spatial process representation and highlight the importance of this choice. In particular, the results demonstrate that, and explain why, standard CAR models can behave strangely for point-level data. They show that various neighborhood weighting approaches based on higher-order neighbors that have been suggested for MRF models do not produce smooth fields, which raises doubts about their utility. Finally, they indicate that an MRF that approximates a thin plate spline compares favorably to standard CAR models and to kriging under many circumstances.

Keywords and phrases: Conditional autoregressive models, Gaussian processes, spatial smoothing, thin plate splines.

Received May 2012.

1. Introduction

Markov random field (MRF) models (also called conditional autoregressive (CAR) models) are the dominant approach to analyzing areally-aggregated spatial data, such as disease counts in administrative units. In contrast, point-referenced data are generally modeled using Gaussian process (GP) models that posit a continuous latent underlying spatial surface, of which the popular approach of kriging for spatial prediction is one variant.

GP models are computationally challenging for large datasets because of manipulations involving large covariance matrices, and there has been a large body of recent work attempting to reduce the computational burden through reduced rank approximations (Kammann and Wand, 2003; Banerjee et al., 2008; Sang and Huang, 2012), inducing sparseness in the covariance (covariance tapering)

(Furrer et al., 2006; Kaufman et al., 2008; Sang and Huang, 2012), and approximate likelihood approaches (Stein et al., 2004), among others. In contrast, MRF models work with the precision matrix directly, so calculation of the likelihood is computationally simple. In MCMC implementations, one can exploit the Markovian structure to sample the value of the field for each area sequentially or exploit the sparsity of the precision matrix when sampling the field values jointly (Rue and Held, 2005).

Given the computational attractiveness of the MRF approach, it is appealing to consider its use for point-referenced data as well. One issue lies in how to define the neighborhood structure for a set of point-referenced observations, but a second more fundamental issue lies in the fact that the model structure changes when one changes the number of sites under consideration. The solution that I highlight here is to relate the point-referenced observations to an underlying regularly-gridded surface modeled as an MRF.

Considering an underlying gridded surface for areally-aggregated data is also appealing. Kelsall and Wakefield (2002) argue for the use of a smooth underlying surface to model areal data, with each areal observation related to the average of the surface over the defined area. They found approximations to the integrals involved and worked with an underlying GP representation; related work includes Fuentes and Raftery (2005) and Hund et al. (2012). Here I relate areal observations to an underlying MRF on a fine grid, approximating the necessary integrals as a simple weighted average of the MRF values for the grid cells that overlap each area. Such an approach also allows one to model multiple areally-aggregated datasets with different boundary structures. In summary, the approach of using an MRF on a fine grid provides an aggregation-consistent model for point or areal data (or a mix of both) and is a strategy also suggested in Besag and Mondal (2005).

The most common form of MRF represents the spatial dependence structure such that areas that share a boundary are considered neighbors, with an area conditionally independent (given its neighbors) of any non-neighboring areas, a so-called first-order neighborhood structure. Fig. 1 shows the results of using this MRF structure on a fine regular grid to model point-referenced data. The fitted smooth surface is not visually appealing, in contrast to an MRF that approximates a thin plate spline (Rue and Held, 2005) and to kriging. The results are consistent with Besag and Mondal (2005), who show that the intrinsic (i.e., improper) first-order MRF on a two-dimensional regular grid produces spatial fields whose distribution approaches two-dimensional Brownian motion (the de Wijs process) asymptotically as the grid resolution increases. Given this continuous but non-differentiable representation of the underlying surface, the local heterogeneity of the surface estimate in Fig. 1 is not surprising. However, note that Besag and Mondal (2005) and Besag in his comments on Diggle et al. (2010) argue that the de Wijs process is preferable to GPs in the Matérn class.

A common alternative to this standard nearest-neighbor structure is to extend the neighborhood structure beyond bordering areas (Pettitt et al., 2002; Hrafnkelsson and Cressie, 2003; Song et al., 2008). Such higher-order neighborhood structure might be expected to produce more smooth process repre-
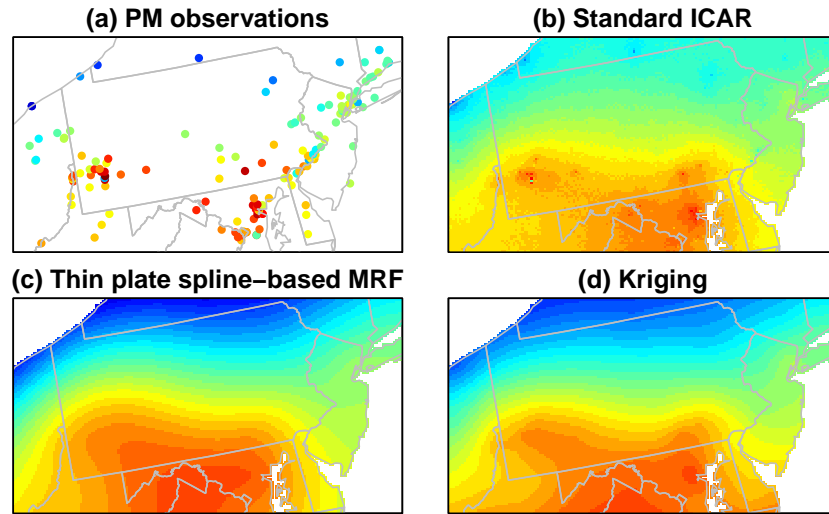
FIG 1. *Example fits to particulate matter (PM) air pollution point observations (a) using the standard first-order MRF on a fine regular grid (b), an MRF approximation to a thin plate spline, again represented on a fine regular grid (c), and using kriging with an exponential covariance (d).*

sentations, but I show that straightforward higher-order neighborhoods do not achieve this. An alternative that I highlight here is an MRF approximation to a thin plate spline (TPS) that involves only nearby grid cells as neighbors (Rue and Held, 2005; Yue and Speckman, 2010). Finally, Lindgren et al. (2011) have recently developed a powerful theory and methodology for approximating GPs in the Matérn class that will likely be widely used. The thin plate spline approximation is a limiting case of the Lindgren et al. (2011) representation, and my results shed light on the distinctions between standard first-order MRF models, the thin plate spline approximation, and GP representations.

In this paper, I present a general model for both areal and point-referenced data that deals simply with spatial misalignment by using an MRF on a fine regular grid (Section 2.1). In the literature, first-order MRF representations are widely-used for areal data, generally with little consideration of the properties of the latent process or resulting suitability for a given application. In light of this and of the lack of smoothness of the popular first-order MRF seen in Fig. 1, I compare several existing approaches for the MRF neighborhood structure (Section 2.2) using both analytic calculations and simulations (Section 3) in the context of the general model presented here. The results suggest that the thin plate spline approximation often performs well and help to explain the strange behavior of the first-order MRF. Importantly for situations in which a smooth process representation is desired, higher-order neighborhood structures can have unappealing properties when they are not constructed as approximations to particular spatial surface representations. Finally, MRF models perform

as well as kriging under a variety of circumstances for both point and areal data. I close by discussing computation (Section 4) and presenting two examples, one with point data and one with areal data (Section 5). Online supplementary material (Paciorek, 2013) contains the R code for all analyses and figures as well as the data for the pollution data only, as the breast cancer data are not publicly available.

2. Spatial model for point and areal data

2.1. Model structure

Here I present a basic model for exponential family data. Let $\mu_i = E(Y_i|\boldsymbol{X}_i, \boldsymbol{g})$ be related via a link function, $h(\cdot)$, to a linear predictor:

$$h(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{K}_i \boldsymbol{g}, \tag{1}$$

where $\boldsymbol{K}_i$ is the ith row of a mapping matrix, $\boldsymbol{K}$, discussed further below. I represent the unknown, latent spatial process, $g(\cdot)$, as a piecewise constant surface on a fine rectangular grid,

$$\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, (\kappa \boldsymbol{Q})^-), \tag{2}$$

where $\boldsymbol{Q}$ is an MRF precision matrix and κ a precision parameter, recognizing the potential singularity of $\boldsymbol{Q}$ by using the generalized inverse notation. Computational issues related to the singularity are discussed in Section 4.

For a point-referenced datum, $\boldsymbol{K}_i$ will be a sparse vector with a single 1 that matches the location of the observation to the grid cell in which it falls. Note that one may include covariates in $\boldsymbol{X}$ that can help to account for within grid cell heterogeneity. For an areally-aggregated datum, I consider the relevant functional of the surface to be the average value of the underlying surface over the area A_i: $g^*(A_i) \equiv \int_{A_i} g(\boldsymbol{s}) d\boldsymbol{s}$. Using the piecewise representation, this can be approximated as $g^*(A_i) \approx \sum_{j \in A_i} w_j g_j$ where $j \in A_i$ indexes the grid cells that the A_i overlaps and w_j is the proportion of A_i that falls in the jth grid cell. Hence the non-zero elements of $\boldsymbol{K}_i$ contain the proportions, $\{w_j\}$, as the appropriate elements. If desired, one could interpolate between the values of $\boldsymbol{g}$ at the grid cell centroids for a more accurate approximation.

2.2. Potential MRF models

Here I present the MRF models that I consider for $\boldsymbol{g}$ and their corresponding precision matrices, $\boldsymbol{Q}$. I consider only intrinsic models with singular precision matrices. These specify improper priors with respect to one or more linear combinations of the process values, as the eigenvectors of $\boldsymbol{Q}$ with zero eigenvalues have infinite prior variance (Banerjee et al., 2003). My focus on intrinsic models is motivated by noting that proper first-order MRF models tend not to allow high correlations between neighbors unless the precision matrices are close to

singular (Banerjee et al., 2003; Wall, 2004). Furthermore, intrinsic models represent the conditional mean for the process in a given area as a weighted average of the process values of the neighboring areas with weights summing one, while proper models have weights summing to less than one.

1. Traditional intrinsic CAR model (ICAR): The most commonly-used MRF model is a simple first-order model that treats any two areas that share a border as neighbors. The corresponding precision matrix has diagonal elements, Q_{ii}, equal to the number of neighbors for the ith area, while $Q_{ij} = -1$ (the negative of a weight of one) when areas i and j are neighbors and $Q_{ij} = 0$ when they are not. On a grid, the simplest version of this model treats the four nearest grid cells as neighbors (i.e., cardinal neighbors). Besag and Mondal (2005) show that the model is asymptotically equivalent to two-dimensional Brownian motion, and Lindgren et al. (2011) show that this model approximates a GP in the Matérn class (3) with the spatial range parameter $\rho \to \infty$ and differentiability parameter $\nu \to 0$.

2. Extended neighborhood model (DICAR and HICAR): One might generalize the first-order Markovian structure of the ICAR model to allow for higher-order dependence by considering areas that do not share a border but are close in some sense to be neighbors. At its simplest, this simply introduces additional values of -1 off the diagonal of $\boldsymbol{Q}$ as in Song et al. (2008). I will call this model the higher-order ICAR (HICAR). A more nuanced version would have the weight for a pair of areas depend on the distance between the two areas (usually declining with distance), such as Euclidean distance or the number of intervening cells between the two areas (Pettitt et al., 2002; Hrafnkelsson and Cressie, 2003). Then $Q_{ij} = -\delta(i,j)$ where $\delta(\cdot, \cdot)$ is the chosen weight or distance function. I will term this model the distance-based ICAR (DICAR) and implement the model using the function in Hrafnkelsson and Cressie (2003), $\delta(i,j) = -d_{ij}^{\log .05/ \log r}$, where d_{ij} is the distance between area centroids and r is the (user-chosen) maximum distance at which the weight is non-zero.

3. Thin plate spline approximation (TPS-MRF): Rue and Held (2005, Sec. 3.4.2) present a second-order model where the weights on neighbors of different orders are constructed so the model approximates a thin plate spline. The approach considers a discretized approximation of the penalty function used in deriving the thin plate spline solution. This penalty function is an integral of second partial derivatives of the unknown surface, so the discretization produces a prior precision matrix based on second order difference operators. In the discretized approach, the nearest cardinal neighbors have a weight of 8 ($Q_{ij} = -8$), the nearest diagonal neighbors a weight of -2 ($Q_{ij} = 2$) and the second nearest cardinal neighbors a weight of -1 ($Q_{ij} = 1$). Note the presence of negative weights, unlike in most MRF models with higher order neighborhood structures. Paciorek and Liu (2012, App. C) describe the derivation of the full $\boldsymbol{Q}$ matrix, including boundary effects, in detail. In one dimension, this model corresponds

to the widely-used second-order auto-regressive model (an IID model on second differences) (Breslow and Clayton, 1993). Lindgren et al. (2011) show that this model approximates a GP in the Matérn class (3) with $\nu = 1$ and the spatial range parameter $\rho \to \infty$.

In this work I compare MRF models to GPs in the Matérn class, where I parameterize the Matérn correlation function as

$$R(d) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}d}{\rho} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{2\sqrt{\nu}d}{\rho} \right), \tag{3}$$

where d is Euclidean distance, ρ is the spatial range parameter, and $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind, whose order is the smoothness (differentiability) parameter, $\nu > 0$. $\nu = 0.5$ gives the exponential covariance. Note that the Lindgren et al. (2011) approach provides a computational strategy for fitting an approximate GP model via a MRF representation but in the material that follows I consider the explicit GP rather than an MRF approximation to it.

3. Comparing MRF Structures

This section compares the different MRF models using a variety of tools, starting with an assessment of the implied eigenstructure and smoothing kernels of the models (Section 3.1). The comparison sheds light on the smoothing behavior of the ICAR model and shows that the extended neighborhood models have unappealing properties. In Section 3.2 I compare predictive performance of the models, both under an oracle setting with the optimal choice of smoothing parameter and in simulations in which all parameters are estimated.

3.1. Covariance and smoothing properties

3.1.1. Eigenstructure

To better understand the spatial dependence structures implied by the various MRF precision matrices, I quantify the magnitude of the variability for different scales (frequencies) of spatial variability by considering the eigendecomposition of $\boldsymbol{Q}$. Given the model $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^-)$, we can consider the eigendecomposition, $\boldsymbol{Q}^- = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^-\boldsymbol{\Gamma}^{\top}$. To generate realizations of $\boldsymbol{g}$, we have $\boldsymbol{g} = \boldsymbol{\Gamma}\boldsymbol{u}$ for $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^-)$. Thus the inverse eigenvalues quantify the magnitude of variability associated with patterns or modes of variability represented by the eigenvectors.

Empirical exploration indicates that the inverse eigenvalues decline as the frequency of variability represented in the eigenvectors increases. Therefore to visualize how different MRF models weight variability at different frequencies, I plot the ordered inverse eigenvalues. Because the matrix $\boldsymbol{Q}$ is multiplied by a scalar precision, only the relative magnitudes are of interest, so I scale the inverse eigenvalues such that the 100th largest inverse eigenvalue for each precision matrix is taken to be equal to one. While the eigenvectors for the various
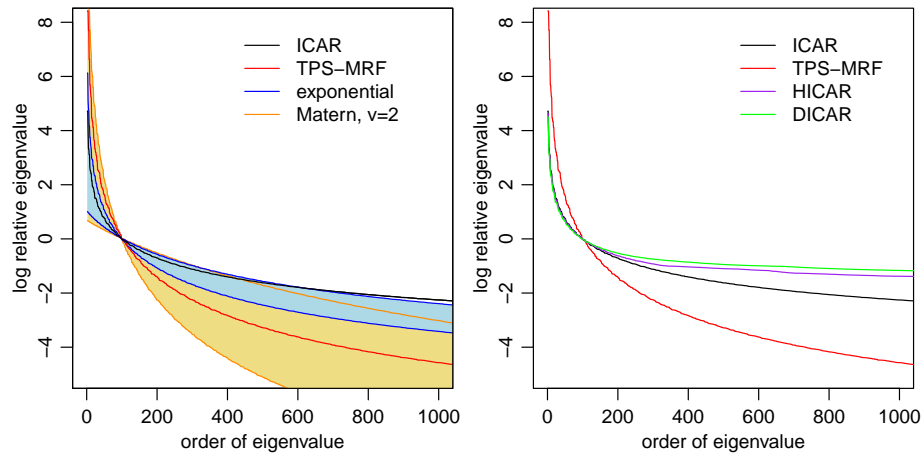
FIG 2. *(a) Log inverse eigenvalues (relative to the 100th largest inverse eigenvalue) for the ICAR and TPS-MRF models, in comparison with the log inverse eigenvalues for the Matérn covariance (with $\nu = 2$) and exponential covariance models. The yellow-orange shading, bounded by orange lines, indicates the range of eigenvalue curves obtained for the Matérn as the range parameter, ρ, varies from the size of the full domain in one dimension to 1/25 of the domain in one dimension. Similarly, the blue shading, bounded by dark blue lines, is for the exponential covariance as the range varies from the full domain to 1/25 of the domain. (b) Log inverse eigenvalues (relative to the 100th largest inverse eigenvalue) for the ICAR and TPS-MRF models, both repeated from (a); the HICAR model with grid cells up to and including three units in distance considered neighbors; and the DICAR model with non-zero weights that decay with distance up to and including five units in distance. Only the first 1000 inverse eigenvalues are plotted, as by the 1000th, the features represented in the eigenvectors occur within groups of 5-10 grid cells, near the limit of scales that could be resolved in a gridded representation.*

precision matrices are not the same, empirical exploration shows that they are quite similar and represent very similar spatial scales of variation for a given position in the ordering of the eigenvector/value pairs. Furthermore, projection onto a common set of eigenvectors (from the ICAR model) confirms that the slight differences in eigenvectors between models do not impact the results shown next.

In the comparisons, I compare the TPS-MRF model to a GP with Matérn correlation with $\nu = 2$, rather than $\nu = 1$ as would be natural given the Lindgren et al. (2011) relationship. I do this because thin plate splines represent smooth functions and the Matérn has M mean square derivatives when $\nu > M$ (Stein, 1999, p. 32). I compare the ICAR model to a GP with an exponential correlation (Matérn with $\nu = 0.5$) despite the Lindgren et al. (2011) result relating the ICAR model to a Matérn covariance with $\nu \to 0$ because the Matérn covariance is only valid for $\nu > 0$.

Fig. 2 plots the size of the ordered inverse eigenvalues for different MRF precision matrices and in comparison to GP models for a regular spatial grid of dimension 75×75. The TPS-MRF puts more weight on the lower frequency eigen-

vectors and less weight on the higher frequency eigenvectors than the ICAR. This is not surprising given the relationship of the ICAR model to Brownian motion and the smoothness of splines. The TPS-MRF eigenvalue curve lies within the set of eigenvalue curves (for varying values of the range parameter, ρ) from the Matérn model with $\nu = 2$. At low frequencies, the ICAR eigenvalue curve lies within the set of eigenvalue curves from the exponential model, while the ICAR model puts more weight on high frequency eigenvectors than the exponential model, consistent with the ICAR approximating a Matérn covariance with $\nu \to 0$ (Lindgren et al., 2011).

Particularly interesting is the behavior of the higher-order and distance-based ICAR variations, which behave similarly to the ICAR model (Fig. 2b). At low frequency, the curves coincide, while the HICAR and DICAR models put more weight on the higher frequencies than the ICAR model. This indicates that one cannot use these approaches to represent surfaces smoother than those represented by the ICAR model. In fact, based on this analysis, it is unclear why one would use these representations, given that the motivation for their use has been to provide more smoothness than the ICAR model.

3.1.2. Equivalent kernels

To understand the smoothing behavior of the various models, I next consider their equivalent kernels, which quantify the local averaging that the models do to make predictions. Under a normal likelihood and ignoring covariates for simplicity, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{Kg}, \tau^2 \boldsymbol{I})$. When there are observations at all the grid cells, the smoothing matrix, $\boldsymbol{S}$, in $\hat{\boldsymbol{g}} = \boldsymbol{Sy}$ can be expressed as

$$\boldsymbol{S} = \frac{1}{\tau^2} \left(\kappa \boldsymbol{Q} + \frac{1}{\tau^2} \boldsymbol{I} \right)^{-1} = (\lambda \boldsymbol{Q} + \boldsymbol{I})^{-1},$$

where $\lambda \equiv \tau^2 \kappa$. For the GP models, we have $\hat{\boldsymbol{g}} = \hat{\mu} \boldsymbol{1} + \sigma^2 \boldsymbol{R}_\theta (\sigma^2 \boldsymbol{R}_\theta + \tau^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \hat{\mu} \boldsymbol{1})$ so

$$\boldsymbol{S} = \sigma^2 \boldsymbol{R}_\theta (\sigma^2 \boldsymbol{R}_\theta + \tau^2 \boldsymbol{I})^{-1} = (\lambda \boldsymbol{R}_\theta^{-1} + \boldsymbol{I})^{-1},$$

where τ^2 is the error variance, σ^2 is the marginal variance of the GP, and $\boldsymbol{R}_\theta$ is the correlation matrix, a function of parameter(s), θ. $\lambda \equiv \tau^2/\sigma^2$ can be thought of as a smoothing parameter, as in the MRF model. Note that the sum of the weights is not one when expressing $\boldsymbol{S}$ as above, because it ignores $\hat{\mu}$, which is also linear in the observations. I ignore this component of the prediction as it involves adding and subtracting a constant that does not vary by location.

Fig. 3 shows the various kernels, plotted as a function of one of the cardinal directions, with the other direction held fixed, for two values of the smoothing parameter, λ. Fig. 3a,d shows that the ICAR kernel puts less weight near the focal cell and more further from the focal cell, compared to the TPS-MRF, but with a spike at the focal cell. This behavior helps to explain the bulls-eye effect and the greater shrinkage towards an overall mean in the gaps between
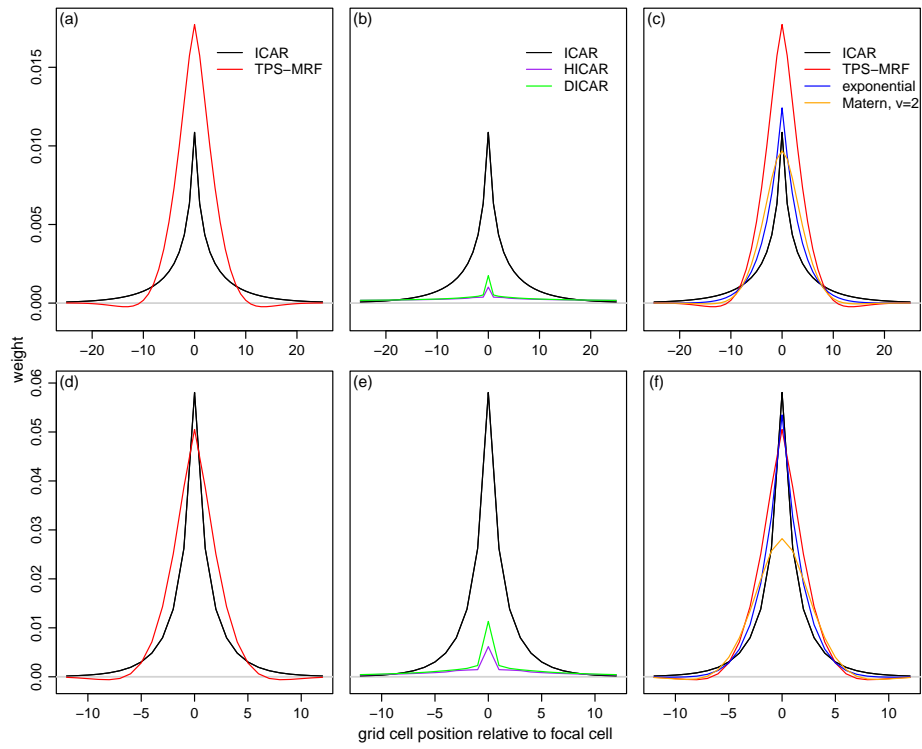
FIG 3. *Equivalent kernel cross-sections in one of the cardinal directions for various models with (a-c) $\lambda = \exp(4)$. which produces less localized weighting and (d-f) $\lambda = \exp(2)$, which produces more localized weighting. In (b) and (e), the HICAR model has grid cells up to and including three units in distance considered neighbors, and the DICAR model has decaying non-zero weights up to and including five units in distance. In (c) and (f), the Matérn (with $\nu = 2$) and exponential covariance models have range, ρ, set to one-tenth of the domain size in one dimension.*

observations seen for the ICAR model in Fig. 1. The TPS-MRF kernel shows some small-magnitude negative weights, which is consistent with the negative weights in the equivalent kernels for spline smoothing and Gaussian process smoothing (Silverman, 1984; Sollich and Williams, 2005). In Fig. 3b,e, we see that the HICAR (with neighbors within three units) and DICAR (with non-zero weights within five units) models place very little weight near the focal observation and spread their weight very widely. The result is that their kernels are even more extreme than the ICAR kernel in making predictions that heavily weight observations far from the focal location. Results are similar but not as extreme for HICAR and DICAR models with smaller neighborhoods. As with the eigenvector analysis, this suggests the HICAR and DICAR models have little practical appeal. Fig. 3c,f compares the ICAR and TPS-MRF to the equivalent kernels for GP models with ρ set to one-tenth of the size of the domain in one dimension. The ICAR model shows some similarity to the exponential-based
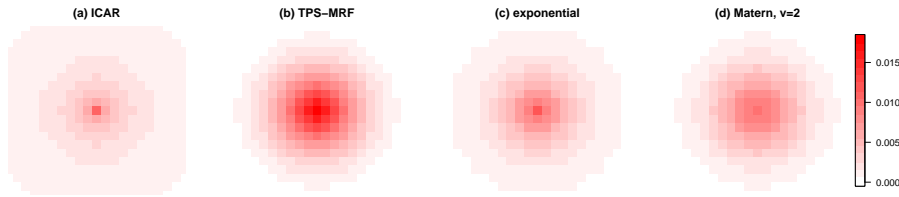
FIG 4. *Equivalent kernels in two dimensions for $\lambda = \exp(4)$: (a) ICAR, (b) TPS-MRF, (c) exponential covariance with range, ρ, equal to one-tenth of the domain in one dimension, and (d) Matérn covariance with $\nu = 2$ and that same range.*

GP in terms of tail behavior and the spike at the focal cell. The TPS-MRF equivalent kernels are rather different than the Matérn-based GP model with $\nu = 2$, but more similar in terms of tail behavior than when compared to the exponential-based model.

Fig. 4 reinforces these points, showing image plots of the equivalent kernels in two dimensions for the ICAR and TPS-MRF models, as well as the Matérn and exponential covariance models. The ICAR kernel puts little weight near the focal cell but spreads positive weight further from the focal cell than the other approaches. The exponential is qualitatively similar, but less extreme than the ICAR. The TPS-MRF and Matérn models are somewhat similar, but the TPS-MRF puts more weight near the focal cells.

3.2. Assessment of predictive ability

Here I compare predictive performance of the ICAR and TPS-MRF models with GP models, via analytic calculation of squared error (using an oracle estimator of the smoothing parameter) and in simulations in which all parameters are estimated.

3.2.1. Data-generating scenarios

I consider the simple generative model for $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$:

$$
\begin{aligned}
\boldsymbol{Y} &\sim \mathcal{N}(\boldsymbol{g}, \tau^2 \boldsymbol{I}) \\
\boldsymbol{g} &\sim \mathcal{N}(\boldsymbol{X\beta}, \boldsymbol{C}),
\end{aligned}
\tag{4}
$$

with $\boldsymbol{C} = \sigma^2 \boldsymbol{R}_\theta$, where $\boldsymbol{R}_\theta$ is a correlation matrix based on the Matérn correlation function with $\nu \in \{0.5, 2\}$ ($\nu = 0.5$ is equivalent to the exponential correlation function). Here $\boldsymbol{X}$ represents spatial basis functions, including an overall mean and possibly linear and higher-order functions of the spatial coordinates, as in ordinary kriging and universal kriging. Using the unit square as the domain, I consider five scenarios: $n = 100$ with locations uniformly sampled, $n = 100$ with observations sampled according to a Poisson cluster process

(PCP) with parent intensity of 25 and standard normal kernels with standard deviation of 0.05, $n = 1000$ with uniform and PCP sampling, and $n = 100$ areal observations on a coarse 10×10 grid overlaid on a regular 100×100 grid. For each scenario, I use a full factorial design with respect to $\nu \in \{0.5, 2\}$, $\rho \in \{0.005, 0.02, 0.08, 0.32, 1.28, 2.56\}$, and $\tau^2 \in \{0.05^2, 0.15^2, 0.45^2, 1.35^2\}$. For $n = 100$ and $n = 1000$, I use 10 replicates to average over the random sampling of locations; the resulting Monte Carlo standard errors are small relative to the point estimates of the reported sum of squared errors (SSE) values. Data are simulated over the unit square in a continuous fashion, while predictions are considered at the grid cell centroids. For simplicity, I simulate with $\sigma^2 = 1$ and $\boldsymbol{\beta} = \mathbf{0}$.

3.2.2. Model fitting

I consider fitting the data using either the ICAR or TPS-MRF models on a regular 100×100 grid:

$$
\begin{aligned}
\boldsymbol{Y} &\sim \mathcal{N}(\boldsymbol{K}\boldsymbol{g}, \tau^2 \boldsymbol{I}), \\
\boldsymbol{g} &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, (\kappa \boldsymbol{Q})^-)
\end{aligned}
$$

where $\boldsymbol{K}$ is a mapping matrix that relates observation locations/areas to grid cells. $\boldsymbol{X}\boldsymbol{\beta}$ represents the overall mean in the ICAR model and the mean plus linear functions of the spatial coordinates in the TPS-MRF model. One can express the best prediction for $\boldsymbol{g}$ as

$$
\begin{aligned}
\hat{\boldsymbol{g}} &= E(\boldsymbol{g}|\boldsymbol{Y}, \cdot) = (\boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{Q})^{-1}(\boldsymbol{K}^\top \boldsymbol{Y} + \lambda \boldsymbol{Q}\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{Q})^{-1}(\boldsymbol{K}^\top + \lambda \boldsymbol{Q}\boldsymbol{X}((\boldsymbol{K}\boldsymbol{X})^\top \boldsymbol{\Sigma}^{*-1}\boldsymbol{K}\boldsymbol{X})^{-1}(\boldsymbol{K}\boldsymbol{X})^\top \boldsymbol{\Sigma}^{*-1})\boldsymbol{Y} \\
&= (\boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{Q})^{-1}\boldsymbol{K}^\top \boldsymbol{Y} \\
&\equiv \boldsymbol{S}_\lambda \boldsymbol{Y}. \tag{5}
\end{aligned}
$$

where $\boldsymbol{\Sigma}^* \equiv \boldsymbol{I} + \boldsymbol{K}(\lambda \boldsymbol{Q})^- \boldsymbol{K}^\top$, $\lambda \equiv \tau^2 \kappa$, and the terms involving $\boldsymbol{X}$ drop out, with estimation of $\boldsymbol{\beta}$ absorbed into $\boldsymbol{g}$, thereby avoiding non-identifiability.

The expected SSE, averaging over randomness in observations and randomness in latent spatial process realizations, can be expressed in terms of a single parameter, λ, that needs to be estimated, plus the parameters of the data-generating model:

$$
\begin{aligned}
\text{SSE}(\lambda) &= E_g E_Y((\hat{\boldsymbol{g}} - \boldsymbol{g})^\top(\hat{\boldsymbol{g}} - \boldsymbol{g})) \\
&= E_g E_Y(\boldsymbol{Y}^\top \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda \boldsymbol{Y} - 2\boldsymbol{g}^\top \boldsymbol{S}_\lambda \boldsymbol{Y} + \boldsymbol{g}^\top \boldsymbol{g}) \\
&= E_g(\tau^2 \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda - 2\boldsymbol{g}^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{g} + \boldsymbol{g}^\top \boldsymbol{g} + \boldsymbol{g}^\top \boldsymbol{K}^\top \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{g}) \\
&= \tau^2 \text{tr}(\boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda) - 2\text{tr}(\boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{C}) + \text{tr}(\boldsymbol{C}) + \text{tr}(\boldsymbol{K}^\top \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{C}) \\
&\quad -2(\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{K}^\top \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} \\
&= \tau^2 \text{tr}(\boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda) - 2\text{tr}(\boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{C}) + \text{tr}(\boldsymbol{C}) + \text{tr}(\boldsymbol{K}^\top \boldsymbol{S}_\lambda^\top \boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{C}). \tag{6}
\end{aligned}
$$

In the derivation, I make use of the fact that $\boldsymbol{S}_\lambda \boldsymbol{K}\boldsymbol{X} = \boldsymbol{X}$ (because $\boldsymbol{Q}\boldsymbol{X} = \mathbf{0}$ when the columns of $\boldsymbol{X}$ are in the space spanned by the eigenvectors of $\boldsymbol{Q}$).

To simplify the presentation of results, I consider an oracle result based on optimizing the SSE (6) across all possible values of λ for each of the ICAR and TPS-MRF models, thereby using the best overall λ for a given generative setting. In this, I assume that τ^2 and σ^2 are known.

Analogous calculations for SSE can be done when fitting the model using the GP approach (thereby fitting with the same model used to generate observations). For the GP approach, I consider predictions at the grid centroids but take the data locations to be the actual locations. As a result the expected SSE can be expressed in similar form to (6), but with S involving R_θ^{-1} in place of Q, $\lambda \equiv \frac{\tau^2}{\sigma^2}$, and a slight bit of additional complexity to account for the correlation between the process values at the data locations and those at the grid cell centroids. Since the GP is the generating model, it provides a baseline for comparison with the MRF models, so I also assume that $\lambda = \frac{\tau^2}{\sigma^2}$ is known, as are $\boldsymbol{\theta} = \{\rho, \nu\}$ and $\boldsymbol{\beta} = \mathbf{0}$.

The oracle assessment just described presumes good choices of the penalty parameter for each representation and does not assess the effect of hyperparameter estimation on performance. Therefore, I also present basic simulation results under the data-generating scenarios described above, with the exception that for the areal scenario, I do not fit the GP model because of computational constraints. I use 100 simulations; the resulting Monte Carlo standard errors are small relative to the reported SSE results. The unknowns are $\boldsymbol{g}$, τ^2, $\boldsymbol{\beta}$ (for the GP model), and $\lambda = \kappa\tau^2$ (for the MRF) or $\lambda = \tau^2/\sigma^2$ (for a GP model). Integrating over $\boldsymbol{g}$ to obtain a marginal likelihood in terms of τ^2 and λ and profiling over τ^2 (and $\boldsymbol{\beta}$ for the GP) gives a likelihood that can be numerically maximized with respect to λ. One can then use (5) to estimate $\boldsymbol{g}$ (with the analogous quantity for the GP case). In fitting the GP model, I use the true ν.

3.2.3. Results for point observations

For $n = 100$ and $n = 1000$, respectively, Figs. 5-6 show the ratio of the SSE using the ICAR to that using the TPS-MRF as well as the ratio of the SSE using the true GP model to that using the TPS-MRF. We see that with uniformly distributed locations, in general the TPS-MRF either matches the SSE of the ICAR or improves upon it. The TPS-MRF strongly outperforms the ICAR when $\nu = 2$, the range is moderate to large, and the noise variance is not too large. The one case in which the ICAR generally beats the TPS-MRF is for $\rho = 0.08$ for $n = 100$ and $\rho = 0.02$ for $n = 1000$, particularly for smaller values of τ^2 and $\nu = 2$, but note that in these cases the absolute SSE (red lines) is close to the SSE of an intercept-only null model for both models. For $\nu = 0.5$, which produces locally heterogeneous surfaces for which we would expect the ICAR model to perform well, we see that the two MRF models perform fairly similarly. Clustering appears to decrease the ratio of SSE of the ICAR relative to the TPS-MRF. This is likely a result of the fact that the TPS-MRF approximates a spline, with no constraint on the spline fit in any large gaps with no observations. Unlike a Gaussian process, a spline does not revert to an overall estimate of the
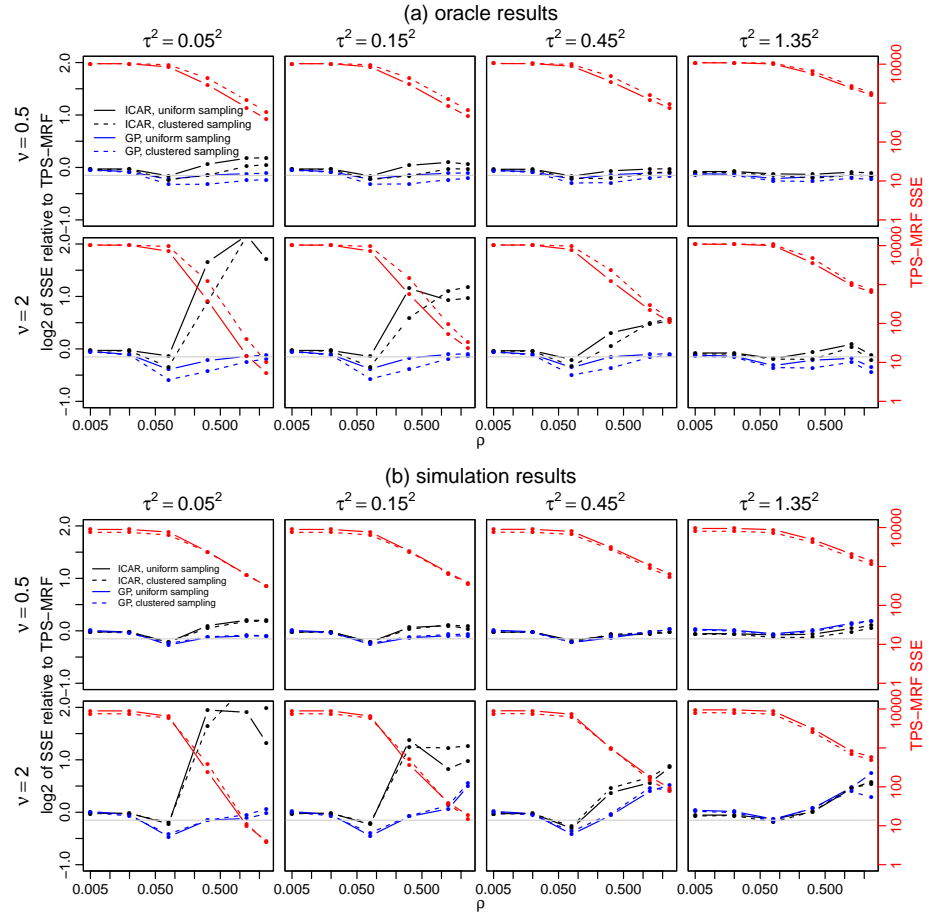
FIG 5. *Predictive performance for $n = 100$ point observations: (a) oracle results and (b) simulation results. Plots show the log (base 2) of the ratio of SSE for the ICAR (black) and GP (blue) models relative to the TPS-MRF and absolute SSE for the TPS-MRF for reference (in red, with axis labels on the right side). In each subplot, ν varies with the row and τ^2 with the column. The horizontal grey line corresponds to the SSE being 90% of the SSE of the TPS-MRF as an informal cutoff below which other models perform substantially better. The SSE in (b) is computed for locations within the convex hull of the observations in a given simulation.*

mean when it is far from any observations, which may lead to poor interpolation. Comparing the TPS-MRF to the GP, the GP generally performs better, which is not surprising given that it is the generative model and is fit based on the true hyperparameter values, but particularly for uniformly-sampled locations, the TPS-MRF is competitive. For larger values of both ρ and τ^2, with $\nu = 2$, we see that the TPS-MRF actually outperforms the GP in the simulations, despite the GP being the generative model, suggesting that for smooth surfaces there is a
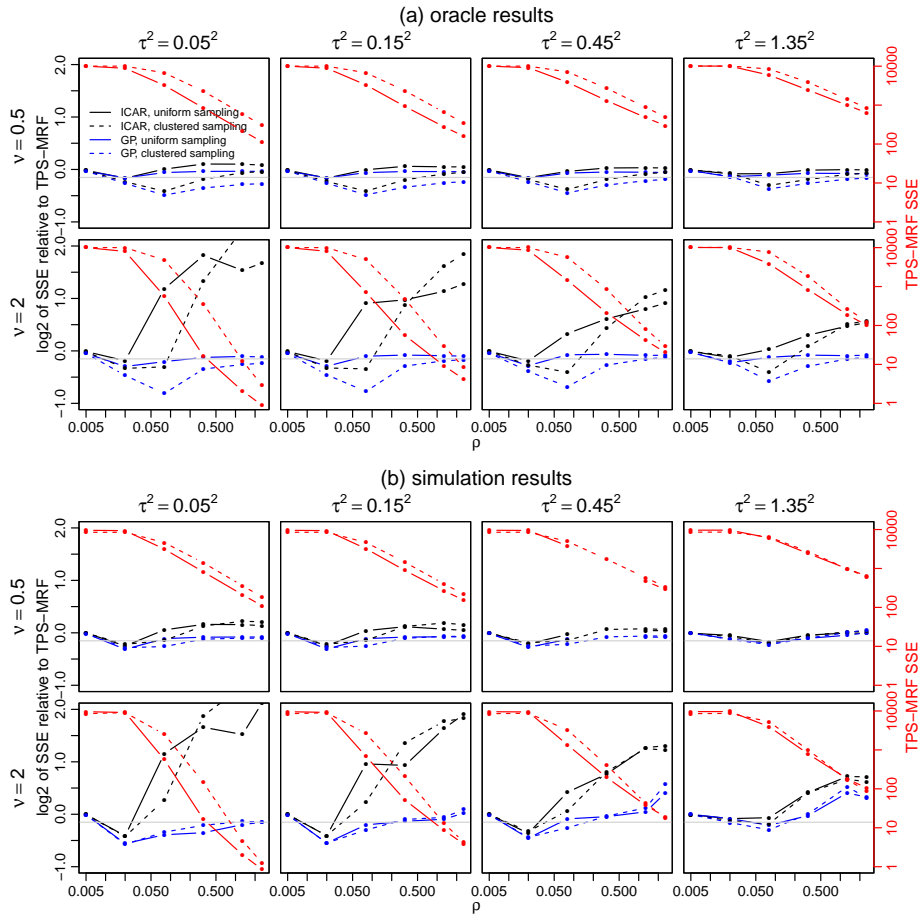
FIG 6. *Predictive performance for* $n = 1000$ *point observations: (a) oracle results and (b) simulation results. Details are as in Fig. 5.*

cost to having to estimate ρ (in particular the GP tends to underestimate ρ and therefore undersmooth), as indicated by the fact that using a GP with known parameter values for these same simulations results in the GP outperforming the TPS-MRF (not shown).

I now interpret these results further in the context of some example fits. For small values of the range, the oracle fit is a constant surface, and both MRF models do this and perform similarly, with SSE approximately equal to $E_g \boldsymbol{g}^\top \boldsymbol{g}$, the SSE of $\hat{\boldsymbol{g}} = \boldsymbol{0}$; see the example fits in Fig. 7, first column. Neither model has any predictive ability, and SSE is essentially squared bias. For slightly larger range values, the ICAR model does modestly better (example fits in Fig. 7, second column), with the TPS-MRF oversmoothing (and experiencing some boundary issues) and the more local behavior and mean reversion of the ICAR
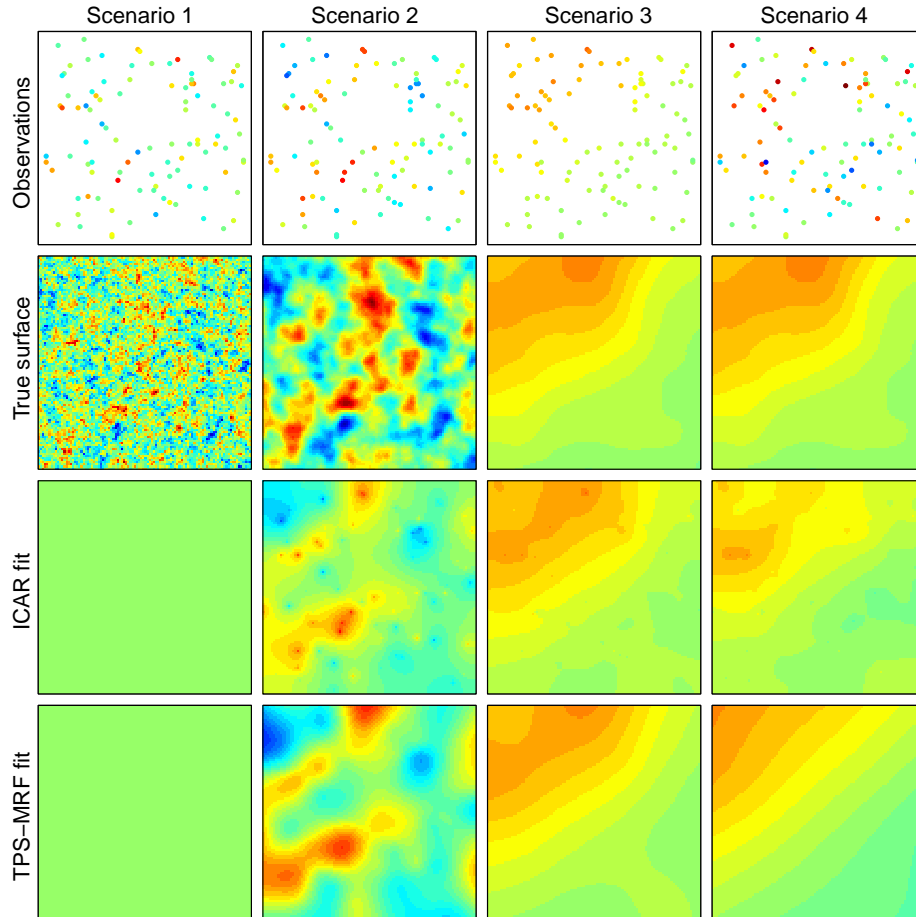
FIG 7. *Example ICAR and TPS-MRF fits (rows) for four generative scenarios (columns) with uniform sampling of locations: (1) $\nu = 0.5$, $\rho = 0.02$, $\tau^2 = 0.15^2$, (2) $\nu = 2$, $\rho = 0.08$, $\tau^2 = 0.15^2$, (3) $\nu = 2$, $\rho = 1.28$, $\tau^2 = 0.15^2$, (4) $\nu = 2$, $\rho = 1.28$, $\tau^2 = 1.35^2$.*

model (recall the results in Section 3.1.2) being beneficial. Then, as the range increases the TPS improves upon the ICAR, most notably when the true surface is smooth ($\nu = 2$) (Fig. 7, third column), with the ICAR showing the bulls-eyes seen in Fig. 1. If the noise variance is also very large, then the advantage of the TPS when $\nu = 2$ moderates, with the TPS oversmoothing (Fig. 7, fourth column). When $\nu = 0.5$, performance of the ICAR and TPS-MRF are comparable because the TPS oversmooths, but follows the larger-scale patterns better than the ICAR, while the ICAR more closely follows the local variability in the vicinity of the observations (not shown).

In terms of uncertainty characterization, in the simulations both the ICAR and TPS-MRF give the nominal 95% coverage for prediction intervals (not shown). However, coverage for the true function, $\boldsymbol{g}$, is quite low in some sit-

uations (not shown). In particular, coverage is well below the nominal coverage levels for both the ICAR and TPS-MRF for $\rho \leq 0.02$, the settings in which both models estimate a constant surface. The TPS-MRF also shows undercoverage for larger values of ρ when $\nu = 0.5$. The good predictive coverage but poor function coverage of the TPS-MRF occurs because the model assumes a smooth underlying surface and attributes some of the true variability in the function surface to the error component, thereby giving standard errors for the function values that are too small. Banerjee et al. (2010) noted a similar problem for reduced rank kriging, with an inflation in the estimated nugget variance. This suggests that one interpret the TPS-MRF uncertainty as relating to the larger-scale spatial variability and accept that one is not able to characterize uncertainty about finer-scale variation.

3.2.4. Results for areal observations

Finally, for areal data, the TPS-MRF model outperforms the ICAR model in some cases, primarily for larger values of ρ, while the ICAR outperforms the TPS-MRF at $\rho = 0.08$ in the simulations but not the oracle results (Fig. 8). The relative advantages and disadvantages of the two models are more pronounced in the simulations, suggesting the difficulty of parameter estimation in the areal context. The primary setting in which the TPS-MRF performs poorly is at $\rho = 0.08$, particularly when the noise level is low. In these settings, at the aggregated resolution of the observations the true surface is fairly heterogeneous between neighboring areas and the ICAR follows the data more closely, while the TPS-MRF oversmooths. However, note that in these settings, the SSE is large, indicating that neither model is able to fit the data well.

Coverage of the TPS-MRF prediction intervals is at the nominal 95% level (with the exception of one setting), but for large values of ρ and smaller values of τ^2, the ICAR model often follows the data exactly, estimating $\tau^2 = 0$ and giving very low coverage (not shown). Considering coverage for the true g, results are similar to the situation for point observations for both models, with nominal coverage maintained only for larger values of ρ, with the ICAR model showing somewhat more settings that maintain nominal coverage (not shown).

4. Computational considerations

4.1. Normal data

Consider an MRF model for g, $g \sim \mathcal{N}_m(\mathbf{0}, (\kappa \mathbf{Q})^-)$, where the zero mean is justified by (5). If the observations are normally distributed, $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X\beta} + \mathbf{Kg}, \tau^2 \mathbf{I})$, we have conjugacy and can integrate g out of the model to obtain a marginal likelihood with which to do maximization or MCMC on the hyperparameter space. Note that one could also use the INLA methodology to quickly approximate the posterior without MCMC (Rue et al., 2009), but here I explore
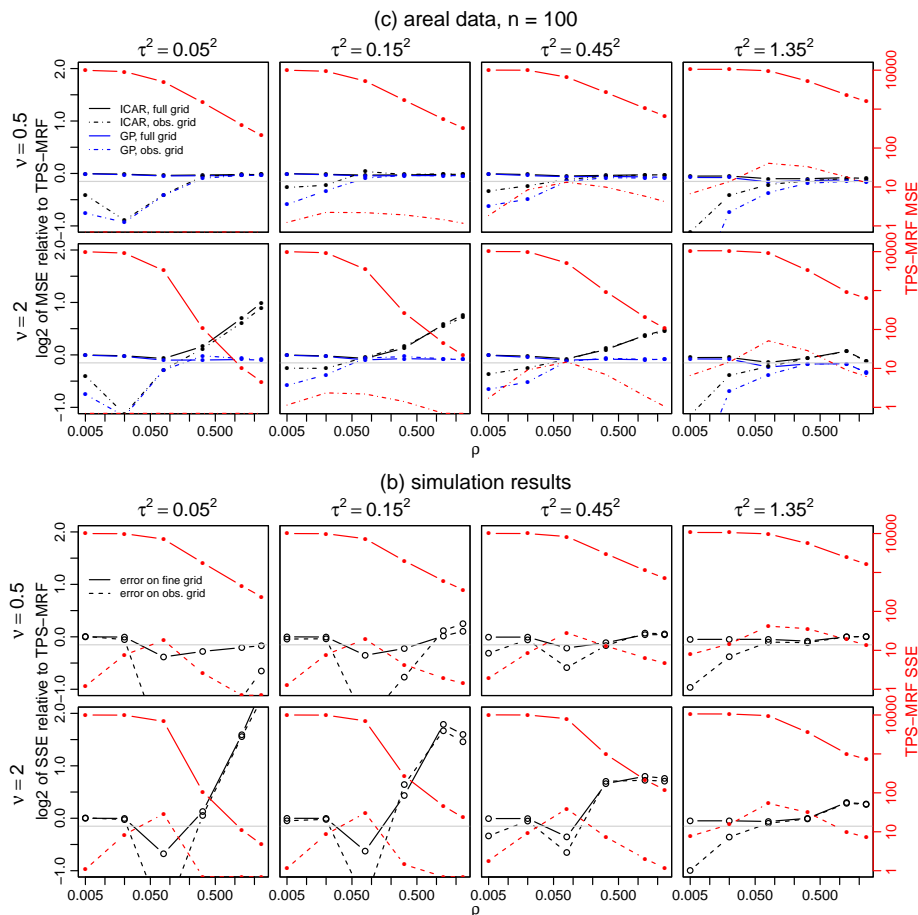
FIG 8. *Predictive performance for areal data: (a) oracle results and (b) simulation results. Plots show the log (base 2) of the ratio of SSE for the ICAR (black) and (for point observations only) GP (blue) models relative to the TPS-MRF and absolute SSE for the TPS-MRF for reference (in red, with axis labels on the right side). Solid lines show SSE for latent process at the cell centroids of the fine grid, while dashed lines show SSE for the values of the latent process averaged within each grid cell of the coarse grid. Additional details are as in Fig. 5.*

the computations needed for maximum likelihood and for 'exact' inference via MCMC.

The marginal precision for $\boldsymbol{Y}$ can be expressed, based on the Sherman-Morrison-Woodbury formula, as

$$\boldsymbol{\Sigma}_\lambda^{-1} = \frac{1}{\tau^2} \left(\boldsymbol{I} - \boldsymbol{K}(\lambda\boldsymbol{Q} + \boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top \right).$$

For maximization, we can express the maxima for $\boldsymbol{\beta}$ and τ^2 as functions of $\boldsymbol{\Sigma}_\lambda^{-1}$ and therefore of λ,

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{\Sigma}_\lambda^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}_\lambda^{-1} \boldsymbol{Y}$$

$$\hat{\tau}_\lambda^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)^\top \boldsymbol{\Sigma}_\lambda^{-1} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)}{n - c},$$

where c is the number of zero eigenvalues of $\boldsymbol{Q}$. Both of these quantities can be calculated efficiently based on a sparse Cholesky decomposition of $\lambda \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{K}$ in the expression for $\boldsymbol{\Sigma}_\lambda^{-1}$, because $\boldsymbol{K}^\top \boldsymbol{K}$ will generally be sparse. If all the data are point locations, $\boldsymbol{K}^\top \boldsymbol{K}$ is diagonal, with the diagonal entries counting the number of observations falling in each grid cell. If all the data are areal observations, only off-diagonal elements of $\boldsymbol{K}^\top \boldsymbol{K}$ corresponding to pairs of grid cells that are overlapped by a common areal observation are non-zero.

The marginal profile likelihood as a function of λ alone is proportional to

$$\frac{\lambda^{(m-c)/2}}{(\hat{\tau}_\lambda^2)^{(n-c)/2} |\lambda \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{K}|^{1/2}}$$

where the determinant can be calculated efficiently based on the already-computed sparse Cholesky decomposition.

MCMC calculations rely on similar quantities that can be computed efficiently. To draw from the posterior of $\boldsymbol{g}$ off-line, given κ, τ^2, and $\boldsymbol{\beta}$, we have

$$\boldsymbol{g} \sim \mathcal{N}((\boldsymbol{K}^\top \boldsymbol{K} + \kappa \tau^2 \boldsymbol{Q})^{-1} \boldsymbol{K}^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tau^2 (\boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{Q})^{-1}),$$

which can be done efficiently based on the same Cholesky decomposition as above.

The computational limitation in the proposed MRF approach is the ability to work with a sparse matrix $\boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{Q}$ whose size scales with the number of grid cells. Thus the approach is limited only by the resolution at which one wishes to do prediction rather than by the sample size, which is similar to the computational constraint in reduced rank kriging, where the computational cost scales with the number of knots. For small n and large m, it may be more computationally efficient to represent the data covariance as

$$\boldsymbol{\Sigma} = \tau^2 \left(\boldsymbol{I} + \frac{1}{\lambda} \boldsymbol{K} \boldsymbol{Q}^- \boldsymbol{K}^\top \right)$$

and precompute the n by n matrix $\boldsymbol{K} \boldsymbol{Q}^- \boldsymbol{K}^\top$, while in the iterations of an MCMC or optimization computing the Cholesky of the dense matrix $\boldsymbol{\Sigma}$. Note that here we need to use the generalized inverse, thereby assigning zero variance to the eigenvectors of $\boldsymbol{Q}$ corresponding to zero eigenvalues and hence necessitating inclusion of the relevant terms in the mean of $\boldsymbol{g}$.

4.2. Non-normal data

The general model (1-2) is a GLMM, where $\boldsymbol{K}$ and $\boldsymbol{g}$ play the roles of the random effects design matrix and random effects, respectively. In this case the

covariance of the random effects has spatial structure and is specified in terms of a precision matrix. Both likelihood-based and Bayesian inference for GLMMs is computationally challenging because inference involves a high-dimensional integral with respect to the random effects that cannot be expressed in closed form.

While the PQL approach (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) is a standard approach for fitting GLMMs, existing implementations such as the glmmPQL() function in the MASS package in R do not include MRF specifications in two dimensions and are not written to take advantage of sparse precision matrices. In contrast the INLA methodology has been implemented specifically for MRF models.

4.2.1. Integrated Nested Laplace Approximation (INLA)

(Rue et al., 2009) present an approach to fitting GLMMs based on nested Laplace approximations involving both the hyperparameters and the latent process values. The result is estimation of the marginal posterior densities of the random effects and the hyperparameters. Note that this accounts for uncertainty in hyperparameters, in contrast to the maximization done with the PQL approach. The INLA R package (http://www.r-inla.org) can make use of sparsity in both Q and K and is therefore very computationally efficient. For analyses using likelihood and prior models that are implemented in INLA and for which one needs only marginal posteriors (or posteriors of linear combinations), INLA is a promising option.

4.2.2. MCMC

While MCMC is a standard approach for Bayesian GLMMs, convergence and mixing are often troublesome (Christensen and Waagepetersen, 2002; Christensen et al., 2006), because of the high-dimensionality of the random effects, the dependence between random effects (particularly when these represent spatial or temporal structure), and cross-level dependence between random effects and their hyperparameters ($\{\tau^2, \kappa\}$ in this work) (Rue and Held, 2005; Rue et al., 2009). The sparse matrix calculations possible with MRF models improve computational efficiency but do not directly address convergence and mixing issues.

Gamerman (1997) describes the use of a weighted least squares proposal for the fixed and random effects, also suggested in Rue and Held (2005, pp. 167-169) to deal with the dependence amongst the random effects, and Rue and Held (2005) suggest combining this with a joint update of the hyperparameter and the random effects to address the cross-level dependence. For high-dimensional random effects vectors such as proposed here, using subblocks of g may also be helpful. To simulate draws of g from the posterior, one might also use INLA to estimate the hyperparameters and then conditionally draw samples of the fixed and random effects using MCMC.

One disadvantage of the TPS-MRF compared to the ICAR may be that the additional smoothness makes it more difficult to accept MCMC proposals for g because the value of the process in one grid cell is more strongly constrained by the values in the other grid cells. The Brownian motion-like behavior in the ICAR (Besag and Mondal, 2005) may help to decouple values of g for different grid cells, similar to the strategy of adding a small amount of white noise to the latent process (e.g., Wikle, 2002), which Paciorek (2007) showed could improve MCMC mixing in a GP context.

5. Examples

5.1. Point-level pollution modeling

This example is based on the work of Paciorek et al. (2009), who modeled spatio-temporal variation in air pollution for the purpose of predicting concentrations for use as the exposure values in a health analysis. The data are average fine particulate matter over 2001-2002 at 339 monitoring stations in the northeast U.S., from the US EPA's Air Quality System database. For this analysis, I averaged the 24 monthly average values and included only locations with at least 22 months of data. I compare the use of the MRF approach for modeling point-level data (based on a 4 km resolution grid) with an additive model built on a reduced rank thin plate spline (using the gam() function from the mgcv package in R) and to universal kriging with both an exponential covariance and a Matérn covariance with $\nu = 2$. For the MRF models, I fitted a model of the form (1-2) with a linear link and normal likelihood (with independent, homoscedastic errors) using the computational approach described in Section 4.1. I used the same set of covariates (log of the distance to nearest road in two road size classes, percent urban land use in a local buffer, log of elevation, and log of the estimated fine PM emissions within a 10 km buffer) as in Paciorek et al. (2009). I included the covariates as linear terms for simplicity and because Paciorek et al. (2009) found only a minor improvement when considering an additive nonparametric structure for the covariates. I used ten-fold cross-validation to assess hold-out error for 219 stations, with 120 stations in boundary states forming a spatial buffer and always used in the training set, to compare the models.

The models all gave very similar prediction results, with the square root of the mean squared prediction errors being 1.30 for the gam() function, 1.34 and 1.29 for universal kriging with the Matérn ($\nu = 2$) and exponential covariances respectively, 1.32 for the TPS-MRF and 1.26 for the ICAR. These correspond to cross-validation R^2 values of about 0.8 (0.787-0.811). The slight apparent advantage for kriging with an exponential covariance and for the ICAR model suggests the presence of fine-scale variability that the other models smooth over. Fig. 9 shows the estimates of residual spatial variability, not including the effect of the spatially-varying covariates, illustrating that the ICAR model fits local effects around the observations but also seems to generally follow the large-scale pattern seen in the TPS-MRF. All the models showed good hold-out prediction coverage, with the MRF models having larger standard errors
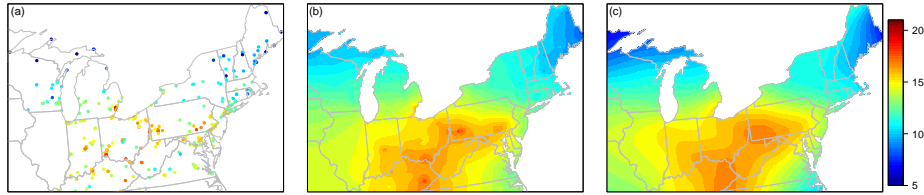
FIG 9. *PM observations (a) and fitted residual spatial surfaces (i.e., not including the effect of covariates) for the ICAR (b) and TPS-MRF (c) models.*

and therefore some overcoverage. The slight performance edge for the ICAR model here occurs in a setting in which it appears there is variation at multiple scales. In light of the scale dependence of the oracle and simulation results, which showed smaller values of ρ and $\nu = 0.5$ tending to favor the ICAR over the TPS-MRF (Section 3.2.3), this suggests the importance of the fine-scale variability to the predictive results here. The advantage for the ICAR may also relate to the fact that the observations are clustered (in metropolitan areas), a setting that the predictive results suggest works in favor of the ICAR model.

Paciorek and Liu (2012) considered more complicated models of fine PM over a variety of spatial and temporal domains using the MRF approach outlined in this paper, including combining point-level monitoring data with areal data from remote sensing and fitting a spatio-temporal extension of the model proposed here.

5.2. Area-level disease mapping

This example is based on the work of Krieger et al. (2006) and Hund et al. (2012), who analyzed variation in breast cancer incidence in Los Angeles County, California, US. Their analyses focused on the relationship between census tract-level poverty and breast cancer incidence and changes over time in that relationship, while accounting for residual spatial variation. Here I modeled breast cancer incidence for data from 1998-2002 for white non-Hispanic women.

The data are counts of cancer incidence in the 5-year period. Following Krieger et al. (2006) and Hund et al. (2012), I fitted Poisson models with a log link, using the log of the expected counts as an offset term,

$$\log \mu_i = \log E_i + \beta_0 + \boldsymbol{\beta}^\top \mathbf{pov}_i + \boldsymbol{K}_i^\top \boldsymbol{g},$$

where μ_i and E_i are the Poisson mean and the expected number of cases in the ith census tract (CT). The expected numbers were calculated based on internal age standardization, described in Hund et al. (2012) and based on CT population (multiplied by five, which assumes constant population over the 5-year period) from the 2000 US Census. CT poverty was a five-level categorical variable with $\mathbf{pov}_i$ being a vector of four indicator variables determining the poverty category of the ith CT. The categorical poverty variable is defined as

follows: (1) $< 5\%$ of residents living below the poverty line and more than 10% of households having high income (at least four times the US median household income), (2) $< 5\%$ of residents living below the poverty line and less than 10% of households having high income, (3) 5.0-9.99% of residents living below the poverty line, (4) 10.0-19.99% of residents living below the poverty line, and (5) at least 20% of residents living below the poverty line, which was used as the baseline category.

I used the INLA package in R to fit the ICAR and TPS-MRF models defined on a fine grid with 99×101 cells of size 1.25^2 km^2 and compared the results to a standard ICAR model based on the neighborhood structure of the irregular census tracts, in all cases using the default hyperparameter priors specified in INLA. The estimated coefficients for the categorical poverty variable were very similar to those in Hund et al. (2012), with lower poverty CTs showing higher breast cancer incidence. Fig. 10 shows the log of the raw incidence rate ratio (observed counts divided by expected number), compared to the estimated log-incidence rate ratios, $\hat{\boldsymbol{g}}$, for the three models. Values of zero indicate no departure from the expected number based on the population and age distribution in the CT. The census-tract-based ICAR and fine grid-based ICAR models show much more spatial variability, while the TPS-MRF smooths quite a bit. DIC and the summed log conditional predictive ordinate (CPO) values, $\sum_i \log P(Y_i | \boldsymbol{Y}_{-i})$ (provided by the INLA package) suggest that the census-tract based ICAR (DIC of 8438, logCPO of -4227) outperforms the grid-based ICAR (8451, -4233) and TPS-MRF models (8476, -4241). The advantage of the ICAR over the TPS-MRF is consistent with the simulation results for areal data in which the ICAR performed better for the moderate value of ρ (Section 3.2.4). Finally, using the Lindgren et al. (2011) approach to approximate a Matérn-based GP on the same fine grid gave results in between the grid-based MRF models, with a DIC of 8462 for $\nu = 1$ (8467 for $\nu = 2$) and logCPO of -4237 (-4239 for $\nu = 2$).

6. Discussion

I have presented a straightforward modeling approach for both areal and point data that relates observations to an underlying smooth spatial surface, represented as an MRF. One important result is that the analytic comparison of various MRF structures indicates that higher-order neighborhood structures that do not have the weighting structure of the TPS-MRF do not produce smoother processes than the standard ICAR model. This suggests that such models are not appealing for spatial modeling. Given these results it would be useful to investigate the smoothing properties of models that empirically choose neighborhood structures (White and Ghosh, 2009; Zhu et al., 2010).

Based on the analytic and simulation assessment of predictive performance, the TPS-MRF outperforms the ICAR model in many scenarios, in particular with smoother surfaces (both in terms of the spatial range and differentiability), as would be expected given that the TPS-MRF approximates a thin plate spline. In both examples, however, the ICAR appeared to outperform the TPS-MRF, perhaps because the true surface was relatively wiggly in those contexts.
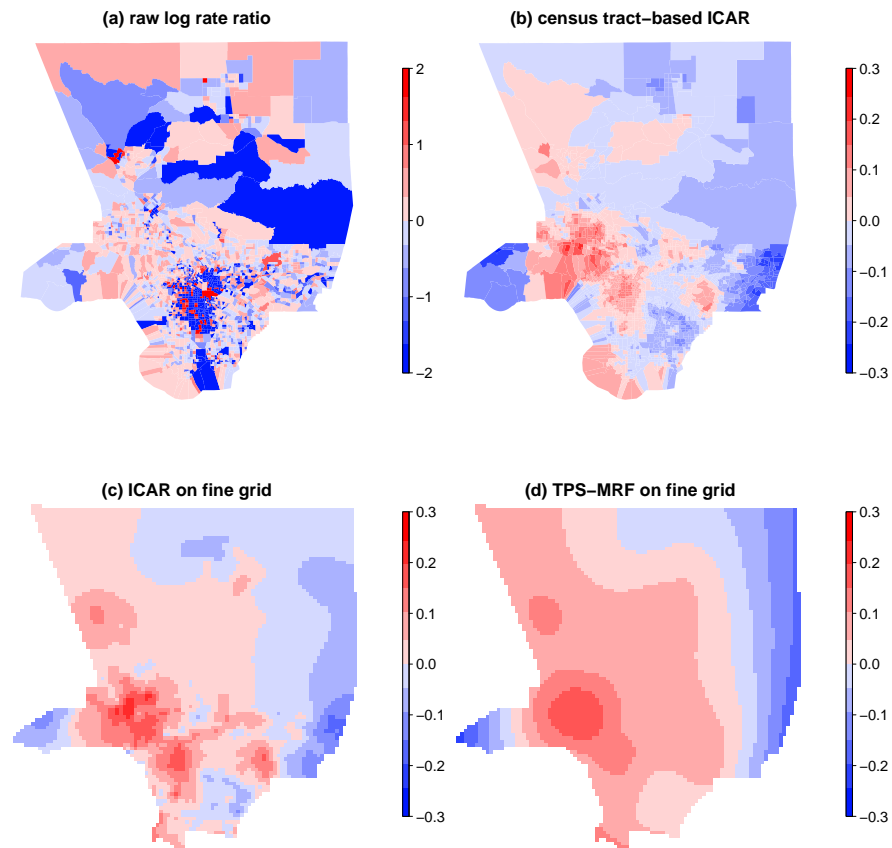
FIG 10. (a) Raw log breast cancer incidence rate ratios by census tract in Los Angeles County and estimated log incidence rate ratios for (b) a standard ICAR model based on the census tract neighborhood structure, (c) the ICAR model on a fine grid, and (d) the TPS-MRF model on a fine grid. Note that the scale in (a) is different than the other panels and that values with magnitude greater than two are censored. Also note that the census tracts extend into the Pacific Ocean in the southwestern portion of panels (a) and (b), while panels (c) and (d) show only the land-based grid cells.

One open question is how the models would perform in a setting with variation at multiple resolutions; these results suggest that the TPS-MRF better represents the large scale while the ICAR model better captures fine-scale variation. How these balance in terms of overall error would likely depend on the relative magnitude of the variation at the different scales.

Spline models can do poorly in situations with large spatial gaps (where large is relative to the spatial range of dependence in the process being modeled) and on the boundary of the domain, as the estimation of basis coefficients is poorly constrained by the data and influenced by data at the extremes of the support

of the basis functions. The TPS-MRF model, by virtue of approximating a thin plate spline, can have this unappealing behavior. In contrast a GP model, being a stationary model, gives predictions that revert to the overall mean for prediction locations far (relative to the estimated spatial range) from observations. In many large datasets, for which the computational efficiency of MRF models is appealing, including deterministic model output and remote sensing observations, gaps are not present or tend to be small, so the issue of extrapolation into large gaps may not be a concern. Furthermore, my ad hoc experience suggests that gaps are less of a problem in two dimensions than in one dimension, although the TPS-MRF can have problems at the boundaries of the domain.

An appealing alternative to the MRF models presented here is the MRF construction of Lindgren et al. (2011), which approximates a Gaussian process with Matérn covariance for integer values of the Matérn smoothness (differentiability) parameter. I expect that much future work with MRFs will involve this construction because of the added flexibility of a representation that includes the GP range parameter. However, I note that the simulations suggest that the TPS-MRF in some cases outperforms an exact GP representation, and in the LA example, I found that the ICAR models outperformed the Lindgren et al. (2011) model. Lindgren et al. (2011) focus on a triangulation rather than a rectangular grid, which has computational advantages when dealing with irregular domains.

I have highlighted the advantages of using a smooth underlying surface for areal data. These include the ability to deal with data aggregation in a consistent manner and with spatial misalignment. Furthermore, in many situations, the area boundaries are essentially arbitrary relative to the process being measured, so the resulting neighborhood structure is arbitrary as well. Rather, it is appealing to imagine an smooth underlying surface, with the areal units merely a measurement artefact that is represented in the measurement model through the mapping matrix, K. In some cases, administrative units might actually have a direct effect on the outcome, in which case more traditional MRF models based on a single random effect per area and standard neighborhood structures may be more appealing, although independent random effects may be appealing in some cases.

Spatio-temporal modeling situations are of course very common. Paciorek and Liu (2012) describe a spatio-temporal extension of the spatial models described here that allows for autoregressive structure in time. In contrast to many spatio-temporal models, the approach has a spatial mean that is shared across time points, which Stein and Fang (1997) emphasize is important for allowing one to properly characterize uncertainty when aggregating over time periods.

Acknowledgments

Supplementary Material

Supplementary data and R code
(doi: [10.1214/13-EJS791SUPP](#); .zip). The supplementary materials contain the data for the PM example as well as R code for all analyses and figures.

References

BANERJEE, S., FINLEY, A.O., WALDMANN, P., AND ERICSSON, T., Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, 105(490):506–521, 2010. [MR2724841](#)

BANERJEE, S., GELFAND, A.E., FINLEY, A.O., AND SANG, H., Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70(4):825–848, 2008. [MR2523906](#)

BANERJEE, S., GELFAND, A.E., AND SIRMANS, C.F., Directional rates of change under spatial process models. *Journal of the American Statistical Association*, 98(464):946–954, 2003. [MR2041483](#)

BESAG, J. AND MONDAL, D., First-order intrinsic autoregressions and the de Wijs process. *Biometrika*, 92(4):909–920, 2005. [MR2234194](#)

BRESLOW, N.E. AND CLAYTON, D.G., Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

CHRISTENSEN, O.F., ROBERTS, G.O., AND SKÖLD, M., Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:1–17, 2006. [MR2269360](#)

CHRISTENSEN, OLE F. AND WAAGEPETERSEN, RASMUS, Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58:280–286, 2002. [MR1908167](#)

DIGGLE, P.J., MENEZES, R., AND SU, T.-L., Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C*, 59:191–232, 2010. [MR2744471](#)

FUENTES, M. AND RAFTERY, A.E., Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45, 2005. [MR2129199](#)

FURRER, R., GENTON, M.G., AND NYCHKA, D., Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:502–523, 2006. [MR2291261](#)

GAMERMAN, D., Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7:57–68, 1997.

HRAFNKELSSON, B. AND CRESSIE, N., Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics*, 10(2):179–200, 2003. [MR1982483](#)

HUND, L., CHEN, J.T., KRIEGER, N., AND COULL, B.A., A geostatistical approach to large-scale disease mapping with temporal misalignment. *Biometrics*, 68:849–858, 2012.

KAMMANN, E.E. AND WAND, M.P., Geoadditive models. *Journal of the Royal Statistical Society, Series C*, 52:1–18, 2003. MR1963210

KAUFMAN, C., SCHERVISH, M., AND NYCHKA, D., Covariance tapering for likelihood-based estimation in large spatial datasets. *Journal of the American Statistical Association*, 103:1556–1569, 2008. MR2504203

KELSALL, J. AND WAKEFIELD, J., Modeling spatial variation in disease risk. *Journal of the American Statistical Association*, 97(459):692–701, 2002. MR1941405

KRIEGER, N., CHEN, J.T., WATERMAN, P.D., REHKOPF, D.H., YIN, R., AND COULL, B.A., Race/ethnicity and changing US socioeconomic gradients in breast cancer incidence: California and Massachusetts, 1978–2002 (United States). *Cancer Causes and Control*, 17(2):217–226, 2006.

LINDGREN, F., RUE, H., AND LINDSTRÖM, J., An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, 73:423–498, 2011. MR2853727

PACIOREK, C.J., Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software*, 19:2, 2007.

PACIOREK, C.J., Supplement to "Spatial models for point and areal data using Markov random fields on a fine grid.", 2013. DOI: 10.1214/13-EJS791SUPP.

PACIOREK, C.J. AND LIU, Y., Assessment and statistical modeling of the relationship between remotely-sensed aerosol optical depth and $PM_{2.5}$. Technical Report 167, Health Effects Institute Research Report (peer-reviewed), 2012.

PACIOREK, C.J., YANOSKY, J.D., PUETT, R.C., LADEN, F., AND SUH, H.H., Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Annals of Applied Statistics*, 3:369–396, 2009. MR2668712

PETTITT, A.N., WEIR, I.S., AND HART, A.G., A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing*, 12(4):353–367, 2002. MR1951708

RUE, H. AND HELD, L., *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, Boca Raton, 2005. MR2130347

RUE, H., MARTINO, S., AND CHOPIN, N., Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71(2):319–392, 2009. MR2649602

SANG, H. AND HUANG, J.Z., A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 74:111–132, 2012. MR2885842

SILVERMAN, B.W., Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, 12(3):898–916, 1984. MR0751281

SOLLICH, P. AND WILLIAMS, C.K.I., Using the equivalent kernel to understand Gaussian process regression. In *Advances in Neural Information Processing Systems 17*, pages 1313–1320. MIT Press, 2005.

SONG, H.R., FUENTES, M., AND GHOSH, S., A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *Journal of Multivariate Analysis*, 99(8):1681–1697, 2008. MR2526171

STEIN, M.L., *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, N.Y., 1999. MR1697409

STEIN, M.L., CHI, Z., AND WELTY, L.J., Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 66(2):275–296, 2004. MR2062376

STEIN, M.L. AND FANG, D., Discussion of ozone exposure and population density in Harris County, Texas, by R.J. Carroll et al. *Journal of the American Statistical Association*, 92:408–411, 1997.

WALL, M.M., A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 2004. MR2038824

WHITE, G. AND GHOSH, S.K., A stochastic neighborhood conditional autoregressive model for spatial data. *Computational Statistics and Data Analysis*, 53(8):3033–3046, 2009. MR2667608

WIKLE, C.K., Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. In A.B. Lawson and D.G.T. Denison, editors, *Spatial Cluster Modelling*, pages 199–209. Chapman & Hall, 2002. MR2015038

WOLFINGER, R. AND O'CONNELL, M., Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.

YUE, Y. AND SPECKMAN, P.L., Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 19:96–116, 2010. MR2654402

ZHU, J., HUANG, H.C., AND REYES, P.E., On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society, Series B*, 72(3):389–402, 2010. MR2758118