Computational Statistics and Data Analysis 53 (2009) 2873–2884

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Improving the performance of predictive process modeling for large datasets

Andrew O. Finley [a,b,*], Huiyan Sang [c,1], Sudipto Banerjee [d,2], Alan E. Gelfand [c,3]

[a] Department of Forestry at the Michigan State University, East Lansing, MI 48824-1222, United States
[b] Department of Geography at the Michigan State University, East Lansing, MI 48824-1222, United States
[c] Department of Statistical Science and Nicholas School, Duke University, Durham, NC 27708-0251, United States
[d] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Available online 16 September 2008

ABSTRACT

Advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged collection of large spatial datasets in many fields and has generated considerable interest in statistical modeling for location-referenced spatial data. The setting where the number of locations yielding observations is too large to fit the desired hierarchical spatial random effects models using Markov chain Monte Carlo methods is considered. This problem is exacerbated in spatial-temporal and multivariate settings where many observations occur at each location. The recently proposed *predictive process*, motivated by kriging ideas, aims to maintain the richness of desired hierarchical spatial modeling specifications in the presence of large datasets. A shortcoming of the original formulation of the predictive process is that it induces a positive bias in the non-spatial error term of the models. A modified predictive process is proposed to address this problem. The predictive process approach is knot-based leading to questions regarding knot design. An algorithm is designed to achieve approximately optimal spatial placement of knots. Detailed illustrations of the modified predictive process using multivariate spatial regression with both a simulated and a real dataset are offered.

1. Introduction

Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged collection of large spatial datasets in many fields and has generated considerable interest in statistical modeling for location-referenced spatial data. This trend is particularly apparent in large-scale natural resource inventories and environmental monitoring initiatives. The data collected in these settings are commonly multivariate, with several spatial variables observed at each location. Given these data, researchers and resource analysts are typically interested in modeling how variables are associated both within and across locations. Specific interest lies in obtaining full inference for model parameters and subsequent predictions along with estimates of associated uncertainty.

* Corresponding author. Tel.: +1 517 432 7219; fax: +1 517 432 1143.
E-mail addresses: finleya@msu.edu (A.O. Finley), huiyan@stat.duke.edu (H. Sang), sudiptob@biostat.umn.edu (S. Banerjee), alan@stat.duke.edu (A.E. Gelfand).
1 Tel.: +1 919 684 8840.
2 Tel.: +1 612 624 0624.
3 Tel.: +1 919 668 5229.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.008

Here, we focus upon the setting where the number of locations yielding observations is too large for fitting desired hierarchical spatial random effects models using Markov chain Monte Carlo methods. That is, such fitting involves matrix decompositions whose complexity increases as $O(n^3)$ in the number of locations, n, at every iteration of the MCMC algorithm, hence the infeasibility or "big n" problem for large datasets. This computational burden is exacerbated in multivariate settings with several spatially dependent response variables as well as when spatial data are collected over time.

In an effort to maintain the richness of desired hierarchical spatial modeling specifications in the presence of large datasets, Banerjee et al. (2008), proposed a class of models based upon the idea of a spatial predictive process (motivated by kriging ideas). The *predictive process* projects the original process onto a subspace generated by realizations of the original process at a specified set of locations (or knots). The approach is in the same spirit as process model approaches using basis functions and kernel convolutions, that is, specifications which attempt to facilitate computation through lower-dimensional process representations. A shortcoming of the original formulation of the predictive process is that it induces a positive bias in the non-spatial error term of the models. Further, Banerjee et al. (2008), identified several open questions regarding the spatial design for placement of knots.

The contribution of this paper is to address both of these issues. In particular, we extend the univariate modified predictive process offered in Finley et al. (in press), by detailing the multivariate modified predictive process that effectively partitions and removes the bias in the non-spatial error terms. Further, we offer an algorithm that places a specified number of knots such that spatially averaged prediction variance is minimized, noting that a predictive process with smaller predictive variance might be viewed as better approximation to the parent process.

The remainder of this manuscript evolves as follows. Section 2 reviews the multivariate predictive process, introduces our proposed bias reducing modification, and describes the Bayesian implementation of the proposed multivariate models. Section 3 illustrates the proposed methods with a simulated dataset and a dataset that couples forest inventory data from the USDA Forest Service Bartlett Experimental Forest with imagery from the Landsat sensor and other variables to map predicted forest biomass by tree species. Section 4 outlines our proposed improved knot design algorithm and provides a simulation study. Finally, Section 5 concludes with a brief discussion including future work.

2. Predictive process models

Geostatistical settings typically assume, at locations $\mathbf{s} \in D \subseteq \Re^2$, a response variable $Y(\mathbf{s})$ along with a $p \times 1$ vector of spatially-referenced predictors $\mathbf{x}(\mathbf{s})$ which are associated through a spatial regression model such as,

$$Y(\mathbf{s}) = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1}$$

where $w(\mathbf{s})$ is a zero-centered Gaussian Process (GP) with covariance function $C(\mathbf{s}, \mathbf{s}')$ and $\epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2)$ is an independent process modeling measurement error or micro-scale variation (see, e.g., Cressie (1993)). In applications, we often specify $C(\mathbf{s}, \mathbf{s}') = \sigma^2\rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ where $\rho(\cdot; \boldsymbol{\theta})$ is a correlation function and $\boldsymbol{\theta}$ includes decay and smoothness parameters, yielding a constant process variance. The likelihood for n observations $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^{\mathrm{T}}$ from (1) is $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$, with $\Sigma_{\mathbf{Y}} = C(\boldsymbol{\theta}) + \tau^2 I_n$, where $X = [\mathbf{x}^{\mathrm{T}}(\mathbf{s}_i)]_{i=1}^n$ is a matrix of regressors and $C(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$. Evidently, both estimation and prediction require evaluating the Gaussian likelihood, hence, evaluating the $n \times n$ matrix $\Sigma_{\mathbf{y}}^{-1}$. While explicit inversion to compute the quadratic form in the likelihood is replaced with faster linear solvers, likelihood evaluation remains expensive for big n, even more so with repeated evaluation as needed in MCMC algorithms.

Recently, Banerjee et al. (2008) proposed a class of models based upon a *predictive process* that operates on a specified lower-dimensional subspace by projecting the original or *parent* process. The lower-dimensional subspace is chosen by the user by selecting a set of "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$, which may or may not form a subset of the entire collection of observed locations $\mathcal{S}$. The predictive process $\tilde{w}(\mathbf{s})$ is defined as the "kriging" interpolator

$$\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*] = \mathbf{c}^{\mathrm{T}}(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \tag{2}$$

where $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$ comprises the parent process realization over the knots in $\mathcal{S}^*$, $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$ is the corresponding $m \times m$ covariance matrix, and $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$.

The predictive process $\tilde{w}(\mathbf{s}) \sim GP(0, \tilde{C}(\cdot))$ defined in (2) has non-stationary covariance function,

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^{\mathrm{T}}(\mathbf{s}; \boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}', \boldsymbol{\theta}), \tag{3}$$

and is completely specified by the parent covariance function. Realizations associated with $\mathbf{Y}$ are given by $\tilde{\mathbf{w}} = [\tilde{w}(\mathbf{s}_i)]_{i=1}^n \sim MVN(\mathbf{0}, c^{\mathrm{T}}(\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})c(\boldsymbol{\theta}))$, where $c^{\mathrm{T}}(\boldsymbol{\theta})$ is the $n \times m$ matrix whose ith row is given by $\mathbf{c}^{\mathrm{T}}(\mathbf{s}_i; \boldsymbol{\theta})$. The attractive theoretical properties of the predictive process including its role as an optimal approximator have been discussed in Banerjee et al. (2008).

The predictive process in (2) immediately extends to multivariate Gaussian process settings. For a $q \times 1$ multivariate Gaussian parent process, $\mathbf{w}(\mathbf{s})$, the corresponding predictive process is

$$\tilde{\mathbf{w}}(\mathbf{s}) = \mathrm{Cov}(\mathbf{w}(\mathbf{s}), \mathbf{w}^*)\mathrm{Var}^{-1}(\mathbf{w}^*)\mathbf{w}^* = \mathcal{C}^{\mathrm{T}}(\mathbf{s}; \boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \tag{4}$$

where $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}') = \mathrm{Cov}(\mathbf{w}(\mathbf{s}),\mathbf{w}(\mathbf{s}')) = [\mathrm{Cov}(w_l(\mathbf{s}),w_m(\mathbf{s}'))]_{l,m=1}^{q}$ is the *cross-covariance* matrix (see, e.g., Banerjee et al. (2004)), $\mathcal{C}^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}_1^*;\boldsymbol{\theta}),\ldots,\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}_m^*;\boldsymbol{\theta})]$ is $q \times mq$ and $\mathcal{C}^*(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i^*,\mathbf{s}_j^*;\boldsymbol{\theta})]_{i,j=1}^{m}$ is the $mq \times mq$ dispersion matrix of $\mathbf{w}^* = [\mathbf{w}(\mathbf{s}_i^*)]_{i=1}^{m}$. Eq. (4) shows $\tilde{\mathbf{w}}(\mathbf{s})$ is a zero mean $q \times 1$ multivariate predictive process with cross-covariance matrix given by $\Gamma_{\tilde{\mathbf{w}}}(\mathbf{s},\mathbf{s}') = \mathcal{C}^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\mathbf{s}';\boldsymbol{\theta})$. This is especially important for the applications we consider, where each location $\mathbf{s}$ yields observations on q dependent variables given by a $q \times 1$ vector $\mathbf{Y}(\mathbf{s}) = [Y_l(\mathbf{s})]_{l=1}^{q}$. For each $Y_l(\mathbf{s})$, we also observe a $p_l \times 1$ vector of regressors $\mathbf{x}_l(\mathbf{s})$. Thus, for each location we have q univariate spatial regression equations which can be combined into the following multivariate regression model:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \tag{5}$$

where $\mathbf{X}^{\mathrm{T}}(\mathbf{s})$ is a $q \times p$ matrix ($p = \sum_{l=1}^{q} p_l$) having a block-diagonal structure with its lth diagonal being the $1 \times p_l$ vector $\mathbf{x}_l^{\mathrm{T}}(\mathbf{s})$ and $\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} MVN(\mathbf{0},\Psi)$. Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p)^{\mathrm{T}}$ is a $p \times 1$ vector of regression coefficients with $\boldsymbol{\beta}_l$ being the $p_l \times 1$ vector of regression coefficients corresponding to $\mathbf{x}_l^{\mathrm{T}}(\mathbf{s})$. Likelihood evaluation from (5) that involves $nq \times nq$ matrices can be reduced to $mq \times mq$ matrices by simply replacing $\mathbf{w}(\mathbf{s})$ in (5) by $\tilde{\mathbf{w}}(\mathbf{s})$.

Further computational gains in computing $\mathcal{C}^{*-1}(\boldsymbol{\theta})$ can be achieved by adopting "coregionalization" methods (Wackernagel, 2003; Gelfand et al., 2004; Banerjee et al., 2008) that model $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}') = A(\mathbf{s})\mathrm{Diag}[\rho_l(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})]_{l=1}^{q}A^{\mathrm{T}}(\mathbf{s}')$, where each $\rho_l(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})$ is a univariate correlation function satisfying $\rho_l(\mathbf{s},\mathbf{s}';\boldsymbol{\theta}) \to 1$ as $\mathbf{s} \to \mathbf{s}'$. Note that $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}) = A(\mathbf{s})A^{\mathrm{T}}(\mathbf{s})$, hence $A(\mathbf{s}) = \Gamma_{\mathbf{w}}^{1/2}(\mathbf{s},\mathbf{s})$ can be taken as any square-root of $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s})$. Often we assume $A(\mathbf{s}) = A$ and assign an inverse-Wishart prior on AA^{T} with A a computationally efficient square-root (e.g., Cholesky or spectral). It now easily follows that $\mathcal{C}^*(\boldsymbol{\theta}) = (I_m \otimes A)\Sigma^*(\boldsymbol{\theta})(I_m \otimes A^T)$, where $\Sigma^*(\boldsymbol{\theta})$ is an $mq \times mq$ matrix partitioned into $q \times q$ blocks, whose (i,j)th block is the diagonal matrix $\mathrm{Diag}[\rho_l(\mathbf{s}_i^*,\mathbf{s}_j^*;\boldsymbol{\theta})]_{l=1}^{q}$. This yields a sparse structure and can be computed efficiently using specialized sparse matrix algorithms. Alternatively, we can write Σ^* as an orthogonally transformed matrix of an $m \times m$ block-diagonal matrix, $P^{\mathrm{T}}[\oplus_{l=1}^{q}[\rho_l(\mathbf{s}_i^*,\mathbf{s}_j^*;\theta_l)]_{i,j=1}^{m}]P$, where $\oplus$ is the block-diagonal operator and P is a permutation (hence orthogonal) matrix. Since $P^{-1} = P^{\mathrm{T}}$, we need to invert $qm \times m$ symmetric correlation matrices rather than a single $qm \times qm$ matrix. Constructing the $nq \times mq$ matrix $\tilde{\Sigma}(\boldsymbol{\theta}) = [\mathrm{Diag}[\rho_l(\mathbf{s}_i,\mathbf{s}_j^*;\boldsymbol{\theta})]_{l=1}^{q}]_{i,j=1}^{n,m}$, we further have

$$\mathrm{Var}(\tilde{\mathbf{w}}) = \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) = (I_n \otimes A)\tilde{\Sigma}(\boldsymbol{\theta})\Sigma^{*-1}(\boldsymbol{\theta})\tilde{\Sigma}^{\mathrm{T}}(\boldsymbol{\theta})(I_m \otimes A^{\mathrm{T}}), \tag{6}$$

where the Kronecker structures and sparse matrices render easier computations.

2.1. Modified predictive process and its implementation

The predictive process systematically underestimates the variance of the parent process $w(\mathbf{s})$ at any location $\mathbf{s}$. This follows immediately since we have $\mathrm{var}(\tilde{w}(\mathbf{s})) = \mathbf{c}^{\mathrm{T}}(\mathbf{s},\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s},\boldsymbol{\theta})$, $\mathrm{var}(w(\mathbf{s})) = C(\mathbf{s},\mathbf{s})$ and that $0 \le \mathrm{var}(w(\mathbf{s})|\mathbf{w}^*) = C(\mathbf{s},\mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s},\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s},\boldsymbol{\theta})$. In practical implementations, this often reveals itself by overestimating the nugget variance in predictive process versions of models such as (1), where the estimated τ^2 roughly captures the $\tau^2 + E(C(\mathbf{s},\mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}))$. (Here, $E(C(\mathbf{s},\mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}))$ denotes the averaged bias underestimation over the observed locations.) Indeed, Banerjee et al. (2008) observed that while predictive process models employing a few hundred knots excelled in estimating most parameters in several complex high-dimensional models for datasets involving thousands of data points, reducing this upward bias in τ^2 was especially problematic.

To remedy this problem, we propose a *modified predictive process*, defined as $\ddot{w}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s})$, where $\tilde{\epsilon}(\mathbf{s}) \overset{indep}{\sim} N(0, C(\mathbf{s},\mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s},\boldsymbol{\theta})C^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s},\boldsymbol{\theta}))$ is a process of independent variables but with spatially adaptive variances. It is now easy to see that $\mathrm{var}(\ddot{w}(\mathbf{s})) = C(\mathbf{s},\mathbf{s}) = \mathrm{var}(w(\mathbf{s}))$, as desired. Furthermore, $E[\ddot{w}(\mathbf{s}) \mid \mathbf{w}^*] = \tilde{w}(\mathbf{s})$ which ensures that $\ddot{w}(\mathbf{s})$ inherits the attractive properties of $\tilde{w}(\mathbf{s})$ (Banerjee et al., 2008). The adjustment for the multivariate predictive process is analogous: following (6), we have $\ddot{\mathbf{w}}(\mathbf{s}) = \tilde{\mathbf{w}}(\mathbf{s}) + \tilde{\boldsymbol{\epsilon}}(\mathbf{s})$, where $\tilde{\boldsymbol{\epsilon}}(\mathbf{s}) \sim MVN(\mathbf{0}, \Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}) - \mathcal{C}^{\mathrm{T}}(\mathbf{s},\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\mathbf{s},\boldsymbol{\theta}))$.

For estimating the modified predictive process model corresponding to (5), we have the data likelihood from the set $\mathcal{S} = \{\mathbf{s}_1,\ldots,\mathbf{s}_n\}$ as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^* + \tilde{\boldsymbol{\epsilon}} + \boldsymbol{\epsilon}; \qquad \tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0},\Sigma_{\tilde{\epsilon}}), \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, I_q \otimes \Psi), \tag{7}$$

where $\mathbf{Y} = [\mathbf{Y}(\mathbf{s}_i)]_{i=1}^{n}$ is the $nq \times 1$ response vector, $\mathbf{X} = [\mathbf{X}^{\mathrm{T}}(\mathbf{s}_i)]_{i=1}^{n}$ is the $nq \times p$ matrix of regressors, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients and $\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_j^*;\boldsymbol{\theta})]_{i,j=1}^{n,m}$ is $nq \times mq$. In addition, $\Sigma_{\tilde{\epsilon}} = \mathrm{Diag}[\Gamma_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_i) - \mathcal{C}^{\mathrm{T}}(\mathbf{s}_i,\boldsymbol{\theta})\mathcal{C}^{*-1}\mathcal{C}(\mathbf{s}_i,\boldsymbol{\theta})]_{i=1}^{n}$. Given priors, model fitting employs a Gibbs sampler with Metropolis–Hastings steps using the marginalized likelihood $MVN(\mathbf{X}\boldsymbol{\beta}, \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta}))$, where $\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta}) = \mathrm{Diag}[\Psi + \Gamma_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_i) - \mathcal{C}^{\mathrm{T}}(\mathbf{s}_i,\boldsymbol{\theta})\mathcal{C}^{*-1}\mathcal{C}(\mathbf{s}_i,\boldsymbol{\theta})]_{i=1}^{n}$. Computing the marginalized likelihood for the predictive process likelihood now requires the inverse and determinant of $\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})$. The inverse is computed using the Sherman–Woodbury–Morrison formula, $\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta}) - \Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta})\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\left[\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta})\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\right]^{-1}\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\boldsymbol{\theta})$, requiring $mq \times mq$ inversions instead of $nq \times nq$ inversions, while the determinant is computed as $|\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})||\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}(\boldsymbol{\theta})\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})|/|\mathcal{C}^*(\boldsymbol{\theta})|$. In particular, with coregionalized models $\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})$ can be expressed as in (6), while

$$\Sigma_{\tilde{\epsilon}}(\boldsymbol{\theta}) = \mathrm{Diag}[AA^{\mathrm{T}} - (\mathbf{1}_m^{\mathrm{T}} \otimes A)[\oplus_{j=1}^{m}[\oplus_{k=1}^{m}\rho_k(\mathbf{s}_i,\mathbf{s}_j^*;\boldsymbol{\theta})]]\Sigma^{*-1}(\boldsymbol{\theta})[\oplus_{j=1}^{m}[\oplus_{k=1}^{m}\rho_k(\mathbf{s}_i,\mathbf{s}_j^*;\boldsymbol{\theta})]](\mathbf{1}_m \otimes A^{\mathrm{T}})].$$

To complete the hierarchical specifications, customarily we set $\boldsymbol{\beta} \sim MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$, while Ψ could be assigned an inverse-Wishart prior. More commonly, independence of pure error for the different responses at each site is adopted, yielding a diagonal $\Psi = \text{Diag}(\tau_i^2)_{i=1}^q$ with $\tau_i^2 \sim IG(a_i, b_i)$. Also we model AA^{T} with an inverse-Wishart prior. Assigning priors to parameters within $\boldsymbol{\theta}$ will again depend upon the choice of correlation function. A particularly flexible choice is the Matérn correlation function, which allows control of spatial range and smoothness (see, e.g., Stein (1999)) and is given by

$$\rho(\mathbf{s}, \mathbf{s}'; \phi, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\|\mathbf{s} - \mathbf{s}'\|\phi)^{\nu}\mathcal{K}_{\nu}(\|\mathbf{s} - \mathbf{s}'\|; \phi); \quad \phi > 0,\ \nu > 0. \tag{8}$$

In (8), $\mathcal{K}_{\nu}$ is a modified Bessel function of the third kind with order ν and $\|\mathbf{s}-\mathbf{s}'\|$ is the Euclidean distance between the sites $\mathbf{s}$ and $\mathbf{s}'$. ϕ controls the decay in spatial correlation and ν is interpreted as a smoothness parameter with higher values yielding smoother process realizations. The spatial decay parameters are generally weakly identifiable and, reasonably informative priors are needed for satisfactory MCMC behavior. Priors for the decay parameters are set relative to the size of D, e.g., prior means that imply the spatial ranges to be a chosen fraction of the maximum distance. The smoothness parameter ν is typically assigned a prior support of (0, 2) as the data can rarely inform about smoothness of higher orders.

We obtain L samples, say $\{\Omega^{(l)}\}_{l=1}^L$, from $p(\Omega \mid Data) \propto p(\boldsymbol{\beta})p(A)p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi)$, where $\Omega = (\boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi)$. Sampling proceeds by first updating $\boldsymbol{\beta}$ from an $MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$ distribution with $\Sigma_{\boldsymbol{\beta}|\cdot} = [\Sigma_{\boldsymbol{\beta}}^{-1} + (\mathbf{X}^{\mathrm{T}}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon})^{-1}\mathbf{X}]^{-1}$ and mean $\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot}\mathbf{X}^{\mathrm{T}}(\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta}) + \Sigma_{\tilde{\epsilon}+\epsilon})^{-1}\mathbf{Y}$. The remaining parameters are updated using Metropolis steps, possibly with block-updates (e.g. all the parameters in Ψ in one block and those in A in another). Typically, random walk Metropolis with (multivariate) normal proposals is adopted; since all parameters with positive support are converted to their logarithms, some Jacobian computation is needed. For instance, while we assign an inverted Wishart prior to AA^{T}, in the Metropolis update we update A, which requires transforming the prior by the Jacobian $2^k\prod_{i=1}^k a_{ii}^{k-i+1}$. Uniform priors on the spatial decay parameters will require a Hastings step due to the asymmetry in the priors.

Once the posterior samples from $P(\Omega \mid Data)$, $\{\Omega^{(l)}\}_{l=1}^L$, have been obtained, posterior samples from $P(\mathbf{w}^* \mid Data)$ are drawn by sampling $\mathbf{w}^{*(l)}$ for each $\Omega^{(l)}$ from $P(\mathbf{w}^* \mid \Omega^{(l)}, Data)$. This composition sampling is routine because $P(\mathbf{w}^* \mid \Omega, Data)$ is Gaussian; in fact, from (7) we have this distribution as

$$MVN[(\mathcal{C}^{*-1}(\boldsymbol{\theta}) + \mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta}))^{-1}\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$
$$(\mathcal{C}^{*-1}(\boldsymbol{\theta}) + \mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\Sigma_{\tilde{\epsilon}+\epsilon}^{-1}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta}))^{-1}].$$

In some instances (e.g., prediction) we desire to recover $\tilde{\boldsymbol{\epsilon}}$, in which case we again use composition sampling to draw $\tilde{\boldsymbol{\epsilon}}^{(l)}$ from the distribution

$$MVN[(\Sigma_{\tilde{\epsilon}}^{-1} + (I_n \otimes \Psi^{-1}))^{-1}(I_n \otimes \Psi^{-1})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^*), (\Sigma_{\tilde{\epsilon}}^{-1} + (I_n \otimes \Psi^{-1}))^{-1}].$$

Once $\mathbf{w}^*$ and $\tilde{\boldsymbol{\epsilon}}$ are recovered, prediction is carried out by drawing $\mathbf{Y}^{(l)}(\mathbf{s}_0)$, for each $l = 1, \ldots, L$ from a $q \times 1$ multivariate normal distribution with mean $\mathbf{X}^{\mathrm{T}}(\mathbf{s}_0)\boldsymbol{\beta}^{(l)} + \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta}^{(l)})\mathcal{C}^{*-1}(\boldsymbol{\theta}^{(l)})\mathbf{w}^{*(l)} + \tilde{\boldsymbol{\epsilon}}^{(l)}$ and variance $\Psi^{(l)}$.

3. Illustrations

We present two simulated data examples followed by an analysis of forest biomass data from a USDA Forest Service experimental forest. Our modified predictive process implementations were written in C++, leveraging threaded and processor optimized BLAS, sparse BLAS, and LAPACK routines for the required matrix computations. The most demanding model (involving 6000 spatial effects) took approximately 5 h to deliver its entire inferential output involving three chains of 25,000 MCMC iterations on two Quad-Core 3.0 GHz Intel Xeon processors with 32.0 GB of RAM running Fedora Linux. Convergence diagnostics and other posterior summarizations were implemented within the R statistical environment (http://cran.r-project.us.org) employing the CODA package.

3.1. Simulated illustrations

We start this section with an example that demonstrates the bias introduced when using the unmodified predictive process; then, a second example of a computationally demanding analysis of a large multivariate dataset that would require 6000 dimensional matrix computations.

For the first example, we generate 2000 locations within a [0, 100] × [0, 100] square and then generate the dependent variable from model (1) with an intercept as the regressor, an exponential covariance function with range parameter $\phi = 0.06$ (i.e., such that the spatial correlation is ∼0.05 at 50 distance units), scale $\sigma^2 = 1$ for the spatial process, and with nugget variance $\tau^2 = 1$. We then fit the predictive process and modified predictive process models using a holding-out set of randomly selected sites, along with a separate set of regular lattices for the knots ($m = 49$, 144 and 900). Table 1 shows the posterior estimates and the root mean square prediction error (RMSPE) based on the prediction for the hold-out dataset. The overestimation of τ^2 by the unmodified predictive process is apparent and we also see how the modified predictive process is able to adjust for the τ^2. Not surprisingly, the RMSPE is essentially the same under either process model.

For the second example, we simulated a response vector $\mathbf{Y}(\mathbf{s})$ of length $q = 6$ for each of 1000 irregularly scattered locations over a [0, 1000] × [0, 1000] square domain using (5) and associated parameter values given in Table 2. Spatial

Table 1
Parameter estimates for the predictive process and modified predictive process models in the univariate simulation

	μ	σ^2	τ^2	RMSPE
True	1	1	1	
$m = 49$				
Predictive process	1.365 (0.292, 2.610)	1.367 (0.652, 2.371)	1.177 (1.067, 1.230)	1.2059
Modified process	1.363 (0.511, 2.392)	1.042 (0.522, 1.915)	0.936 (0.679, 1,140)	1.2048
$m = 144$				
Predictive process	1.363 (0.524, 2.324)	1.387 (0.764, 2.442)	1.095 (0.959, 1.244)	1.1739
Modified process	1.332 (0.501, 2.240)	1.141 (0.643, 1.784)	0.932 (0.764, 1.223)	1.1718
$m = 900$				
Predictive process	1.306 (0.235, 2.545)	1.121 (0.853, 1.581)	0.993 (0.851, 1.155)	1.1685
Modified process	1.307 (0.230, 2.632)	1.045 (0.763, 1.493)	0.984 (0.872, 1.210)	1.1679

Table 2
Simulated data generated with these parameter values and model (5)

Parameter	Value	Parameter	Value	Parameter	Value
$\Gamma_{w;1,1}$	50	$\Psi_{1,1}$	25	β_0	1
$\Gamma_{w;1,2}$	25	$\Psi_{1,2}$	0	β_1	1
$\Gamma_{w;1,3}$	25	$\Psi_{1,3}$	0	β_2	1
$\Gamma_{w;1,4}$	−25	$\Psi_{1,4}$	0	β_3	1
$\Gamma_{w;1,5}$	0	$\Psi_{1,5}$	0	β_4	1
$\Gamma_{w;1,6}$	0	$\Psi_{1,6}$	0	β_5	1
$\Gamma_{w;2,2}$	50	$\Psi_{2,2}$	50	ϕ_{w_1}	0.004
$\Gamma_{w;2,3}$	25	$\Psi_{2,3}$	0	ϕ_{w_2}	0.004
$\Gamma_{w;2,4}$	−25	$\Psi_{2,4}$	0	ϕ_{w_3}	0.004
$\Gamma_{w;2,5}$	0	$\Psi_{2,5}$	0	ϕ_{w_4}	0.004
$\Gamma_{w;2,6}$	0	$\Psi_{2,6}$	0	ϕ_{w_5}	0.015
$\Gamma_{w;3,3}$	50	$\Psi_{3,3}$	25	ϕ_{w_6}	0.015
$\Gamma_{w;3,4}$	−25	$\Psi_{3,4}$	0	ν_{w_1}	0.5
$\Gamma_{w;3,5}$	0	$\Psi_{3,5}$	0	ν_{w_2}	0.5
$\Gamma_{w;3,6}$	0	$\Psi_{3,6}$	0	ν_{w_3}	0.5
$\Gamma_{w;4,4}$	50	$\Psi_{4,4}$	50	ν_{w_4}	0.5
$\Gamma_{w;4,5}$	0	$\Psi_{4,5}$	0	ν_{w_5}	0.5
$\Gamma_{w;4,6}$	0	$\Psi_{4,6}$	0	ν_{w_6}	0.5
$\Gamma_{w;5,5}$	50	$\Psi_{5,5}$	25	Range_{w_1}	750
$\Gamma_{w;5,6}$	45	$\Psi_{5,6}$	0	Range_{w_2}	750
$\Gamma_{w;6,6}$	50	$\Psi_{6,6}$	50	Range_{w_3}	750
				Range_{w_4}	750
				Range_{w_5}	200
				Range_{w_6}	200

association was assumed to follow the Matérn correlation function (8), with response-specific range, ϕ, and smoothness, ν, parameters indexed with subscript 1, ..., 6 in Table 2. The simulated locations and interpolated surfaces of the resulting univariate responses are displayed in Fig. 1. Given these data, we considered sub-models of (5) including the non-spatial (i.e., (5) without $\mathbf{w}(\mathbf{s})$) and spatial non-separable (i.e., coregionalized) models with several knot intensities. For the spatial models, Ψ and Γ_w are considered full $q \times q$ cross-covariance matrices. The iterative inversion of the 6000 dimension matrix (i.e., $q \times n = 6 \times 1000$) makes fitting the full spatial models computationally challenging. Therefore, the candidate spatial models employ the modified predictive process and consider three knot intensities of 64, 100 and 225. Knots were located on a uniform grid over the domain. We judge the performance of these models based on the ability to recover the *true* parameter values, prediction of a hold-out set of 1000 locations, and visual similarity between the predicted and true response surfaces.

Prior distributions are assigned to model parameters to complete the Bayesian specification. As is customary, a flat prior was assigned to each intercept parameter $\boldsymbol{\beta}$. The cross-covariance matrices Ψ and Γ_w each receives an inverse-Wishart, $IW(df, S)$, with the degrees of freedom set to $q + 1$. For Ψ and Γ_w, the scale matrix, S, was constructed with zeros on the off-diagonal elements and diagonal elements taken as the nugget and partial-sill values, respectively, from univariate semi-variograms fit to the residuals of the non-spatial multivariate model. For each response variable, the Matérn correlation function decay parameter ϕ follows a $U(0.003, 3)$ which, when $\nu = 0.5$, corresponds to about 1 to 1000 distance units for the effective spatial range (i.e., $-\log(0.05)/\phi$ is the distance at which the correlation drops to 0.05). As previously noted in Section 2, ν is typically poorly identified by the data and therefore we fix it at 0.5.

For each model, we ran three initially over-dispersed chains for 25,000 iterations. Convergence diagnostics revealed 5000 iterations to be sufficient for initial burn-in and so the remaining 20,000 samples from each chain were used for posterior inference. The non-separable model with 255 knots required the most computing resources, taking approximately 5 h to complete the MCMC sampling; the non-spatial and separable models required substantially less time to collect the specified samples.

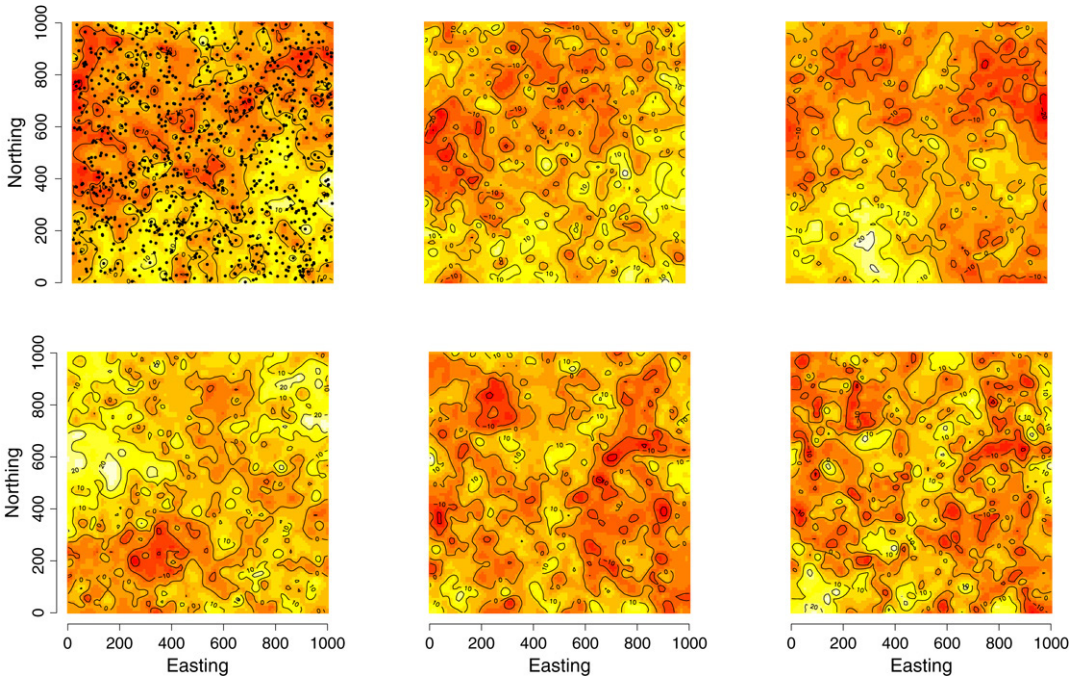

Fig. 1. Interpolated surfaces of the simulated multivariate response values over 1000 sites. Response variables ordered 1–3, top row and 4–6, bottom row. Site locations overlaid on top left panel.

Table 3
Simulated data parameter estimates for the 64 knot non-separable modified predictive process model

Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)
$\Gamma_{w;1,1}$	41.09 (24.81, 51.56)	$\Psi_{1,1}$	23.56 (20.83, 27.09)	β_0	−2.19 (−7.52, 3.07)
$\Gamma_{w;1,2}$	21.94 (18.71, 25.34)	$\Psi_{1,2}$	−0.22 (−1.01, 0.36)	β_1	−0.54 (−5.85, 4.87)
$\Gamma_{w;1,3}$	22.42 (16.80, 25.31)	$\Psi_{1,3}$	0.01 (−0.77, 0.59)	β_2	−2.54 (−8.67, 2.60)
$\Gamma_{w;1,4}$	−22.98 (−25.97, −17.70)	$\Psi_{1,4}$	−0.18 (−0.75, 0.51)	β_3	5.09 (0.34, 10.20)
$\Gamma_{w;1,5}$	−0.12 (−1.92, 0.76)	$\Psi_{1,5}$	0.39 (−0.34, 1.73)	β_4	0.95 (−1.97, 3.79)
$\Gamma_{w;1,6}$	0.31 (−1.74, 2.17)	$\Psi_{1,6}$	−0.42 (−1.04, 0.13)	β_5	1.01 (−1.71, 3.52)
$\Gamma_{w;2,2}$	45.19 (35.86, 64.79)	$\Psi_{2,2}$	50.83 (45.22, 56.46)	ϕ_{w_1}	0.003 (0.003, 0.004)
$\Gamma_{w;2,3}$	24.06 (22.02, 28.40)	$\Psi_{2,3}$	0.03 (−0.64, 1.35)	ϕ_{w_2}	0.004 (0.003, 0.006)
$\Gamma_{w;2,4}$	−24.27 (−27.75, −22.59)	$\Psi_{2,4}$	0.81 (−0.06, 1.62)	ϕ_{w_3}	0.004 (0.003, 0.007)
$\Gamma_{w;2,5}$	−0.30 (−2.06, 0.86)	$\Psi_{2,5}$	0.31 (−1.32, 1.97)	ϕ_{w_4}	0.007 (0.004, 0.011)
$\Gamma_{w;2,6}$	−0.04 (−1.01, 0.97)	$\Psi_{2,6}$	−0.22 (−1.07, 1.27)	ϕ_{w_5}	**0.010 (0.008, 0.012)**
$\Gamma_{w;3,3}$	53.92 (42.78, 74.91)	$\Psi_{3,3}$	**21.12 (18.19, 24.15)**	ϕ_{w_6}	0.015 (0.005, 0.028)
$\Gamma_{w;3,4}$	−24.88 (−28.52, −21.90)	$\Psi_{3,4}$	0.25 (−0.39, 1.07)	Range_{w_1}	911.85 (717.70, 996.68)
$\Gamma_{w;3,5}$	−0.45 (−3.07, 1.21)	$\Psi_{3,5}$	0.07 (−0.55, 1.40)	Range_{w_2}	738.92 (487.80, 974.03)
$\Gamma_{w;3,6}$	0.46 (−0.58, 1.45)	$\Psi_{3,6}$	−0.10 (−1.79, 0.33)	Range_{w_3}	672.65 (412.65, 964.63)
$\Gamma_{w;4,4}$	55.71 (46.88, 75.40)	$\Psi_{4,4}$	43.94 (34.50, 52.04)	Range_{w_4}	439.24 (272.73, 826.45)
$\Gamma_{w;4,5}$	0.24 (−1.18, 2.93)	$\Psi_{4,5}$	0.34 (−0.21, 1.82)	Range_{w_5}	**301.20 (244.50, 372.67)**
$\Gamma_{w;4,6}$	−0.52 (−1.42, 1.19)	$\Psi_{4,6}$	0.00 (−0.98, 2.68)	Range_{w_6}	204.36 (107.68, 551.47)
$\Gamma_{w;5,5}$	58.53 (49.95, 74.85)	$\Psi_{5,5}$	20.30 (16.23, 25.94)		
$\Gamma_{w;5,6}$	49.10 (44.55, 56.00)	$\Psi_{5,6}$	−0.16 (−1.60, 0.47)		
$\Gamma_{w;6,6}$	49.35 (45.98, 53.66)	$\Psi_{6,6}$	54.39 (48.18, 62.49)		

Bold values identify those 95% credible intervals that do not include the *true* parameter values given in Table 2.

For the three knot intensities, there was negligible difference among the parameter estimates. Further, in only two instances, at the 64 knot intensity, did the estimated 95% credible interval not cover the true parameter's value. Table 3 presents the parameter estimates for the 64 knot grid. At the 100 knot intensity and greater, all 95% credible intervals cover the true parameter values and there is a marginal tightening of the spatial range parameters. We now turn our attention to prediction of the hold-out set. The empirical coverage of 95% prediction interval for the three knot intensities 64, 100, and 255 were 91%, 93%, and 96%, respectively. There was no perceptible tightening of the prediction intervals as knot intensity increased; however, increasing the knot intensity allowed estimates of $\tilde{\mathbf{w}}$ to better approximate the local trends in the residual spatial surface. Fig. 2 offers an interpolated surface for the median of the posterior predictive distribution from the 100 knot model. These prediction surfaces closely approximate the true response surfaces in Fig. 1.

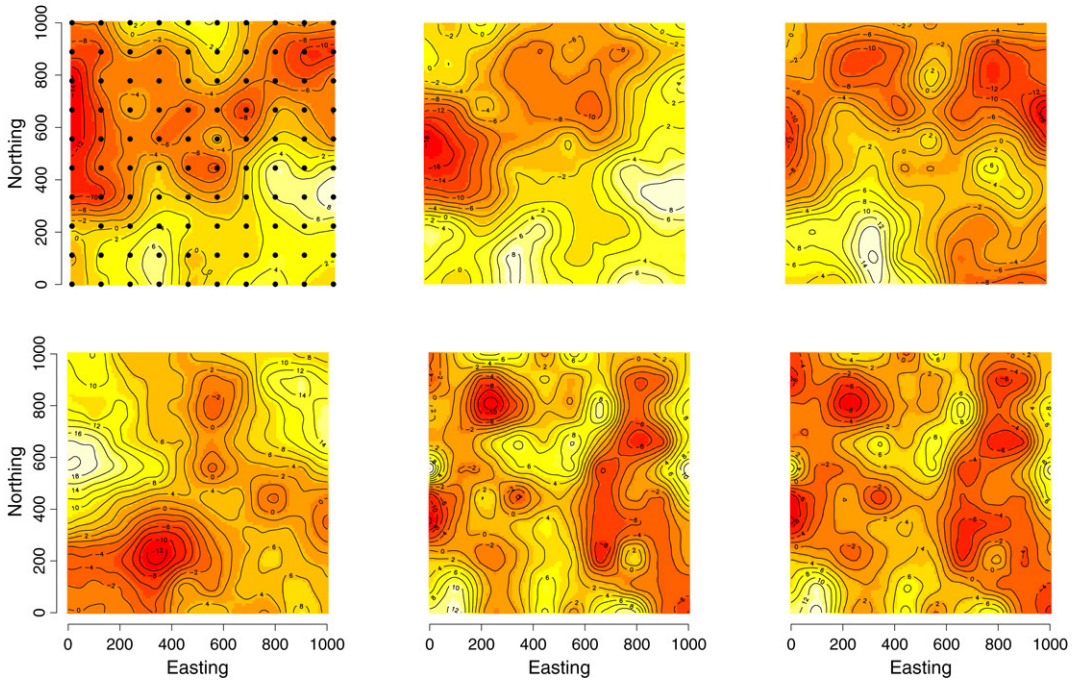

Fig. 2. Interpolated surfaces of the median predicted multivariate response values over a grid of 1000 hold-out sites. The order of panels corresponds to Fig. 1. Predictions based on the 100 knot locations overlaid on top left panel.

3.2. Forest biomass prediction and mapping

Spatial modeling of forest biomass and other variables related to measurements of current carbon stocks and flux have recently attracted much attention for quantifying the current and future ecological and economic viability of forest landscapes. Interest often lies in detecting how biomass changes across the landscape (as a continuous surface) by forest tree species. We consider point-referenced biomass (log-transformed) data observed at 437 forest inventory plots across the USDA Forest Service Bartlett Experimental Forest (BEF) in Bartlett, New Hampshire. Each location yields measurements of metric tons of above-ground biomass per hectare for American beech (BE), eastern hemlock (EH), red maple (RM), sugar maple (SM), and yellow birch (YB) and five covariates: TC1, TC2, and TC3 tasseled cap components (see Huang et al. (2002)) derived from a spring date of mid-resolution Landsat 7 ETM+ satellite imagery from the National Land Cover Database (www.mrlc.gov/mrlc2k_nlcd.asp), and; elevation (ELEV) and slope (SLOPE) derived from a digital elevation model data (see http://seamless.usgs.govfor metadata). Fig. 3 offers interpolated surfaces of the response variables. Covariates were measured on a 30×30 m pixel grid and are available for every location across the BEF. Interest lies in producing pixel-level prediction of biomass by species across large geographic areas. Because data layers such as these serve as input variables to subsequent forest carbon estimation models, it is crucial that each layer also provides a pixel-level measure of uncertainty in prediction. Following our discussion in Section 2.1, we see that basing prediction on a predictive process could substantially reduce the time necessary to estimate the posterior predictive distributions over a large array of pixels. A similar analysis was conducted by Finley et al. (2008); however, due to computational limitations they were only able to fit models using half of the available data and pixel-level prediction was still infeasible.

Here we considered sub-models of (5) including the non-spatial and spatial non-separable models with the modified predictive process and three knot intensities of 51, 126, and 206. For all models Ψ and Γ_w are considered full $q \times q$ cross-covariance matrices where $q = 5$. Predictive process knots were located on a uniform grid within the BEF. We judge the performance of these models based on prediction of a hold-out set of 37 inventory plots, and visual similarity between the predicted and observed response surfaces.

We assigned a flat prior to each of the 30 β parameters (i.e., $p = \sum_{l=1}^{5} p_l = 30$ with each p_l including an intercept, TC1, TC2, TC3, ELEV, and SLOPE). The cross-covariance matrices Ψ and Γ_w each receives an inverse-Wishart, $IW(df, S)$, with the degrees of freedom set to $q + 1 = 6$. Again, diagonal elements in the IW hyperprior scale matrix for Ψ and Γ_w were taken from univariate semi-variograms fit to the residuals of the non-spatial multivariate model. The decay parameter ϕ in the Matérn correlation function spatial follows a $U(0.002, 0.06)$ which corresponds to an effective spatial range between 50 and 1500 m. Again, the smoothness parameter, ν, was fixed at 0.5, which reduces (8) to the common Exponential correlation function. For each model, we ran three initially over-dispersed chains for 35,000 iterations. Unlike in the simulation analysis, substantial effort was required to select tuning values that achieved acceptable Metropolis acceptance rates. Ultimately, we resorted to univariate updates of elements in $\Psi^{1/2}$ and $\Gamma_w^{1/2}$ to gain the control necessary to maintain an acceptance of

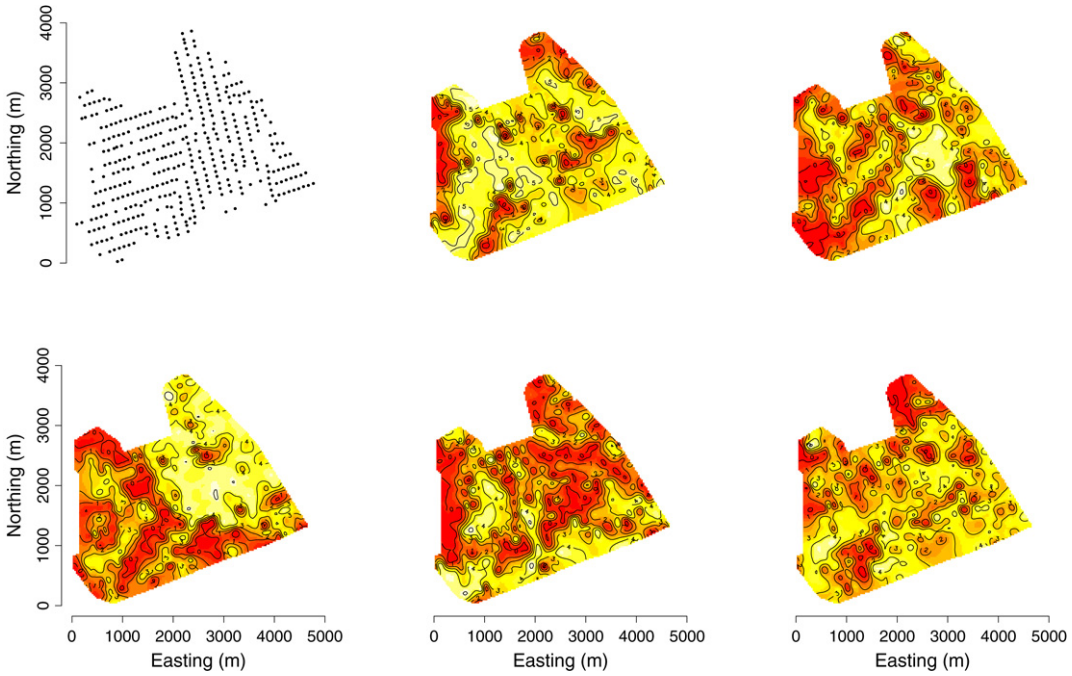

Fig. 3. Interpolation surfaces of log-transformed metric tons of biomass per hectare by species measured on forest inventory plots across the BEF. Response variables ordered BE, EH, top row and RM, SM, YB, bottom row. The set of 437 forest inventory plots is represented as points in the top left panel.

Table 4
BEF biomass parameter estimates for the 126 knot modified predictive process model

Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)
$\Gamma_{w;1,1}$	1.97 (1.93, 2.02)	$\Psi_{1,1}$	1.95 (1.92, 1.98)	ϕ_{w_1}	0.0056 (0.0033, 0.01)
$\Gamma_{w;1,2}$	0.0044 (−0.0029, 0.019)	$\Psi_{1,2}$	−0.01 (−0.031, −0.0002)	ϕ_{w_2}	0.0048 (0.0037, 0.0144)
$\Gamma_{w;1,3}$	−0.014 (−0.034, −0.004)	$\Psi_{1,3}$	−0.0069 (−0.018, 0.001)	ϕ_{w_3}	0.0028 (0.0021, 0.0053)
$\Gamma_{w;1,4}$	0.011 (−0.0004, 0.027)	$\Psi_{1,4}$	0.01 (−0.0026, 0.019)	ϕ_{w_4}	0.0051 (0.0035, 0.0085)
$\Gamma_{w;1,5}$	0.012 (0.0009, 0.018)	$\Psi_{1,5}$	−0.0048 (−0.022, 0.013)	ϕ_{w_4}	0.0059 (0.0032, 0.0102)
$\Gamma_{w;2,2}$	1.96 (1.89, 2.00)	$\Psi_{2,2}$	1.92 (1.88, 1.97)	Range_{w_1}	536.75 (296.06, 903.66)
$\Gamma_{w;2,3}$	0.017 (0.0043, 0.032)	$\Psi_{2,3}$	0.0081 (−0.0001, 0.015)	Range_{w_2}	624.72 (208.76, 806.32)
$\Gamma_{w;2,4}$	0.0032 (−0.01, 0.013)	$\Psi_{2,4}$	−0.0048 (−0.012, 0.0019)	Range_{w_3}	1085.68 (563.5, 1453.63)
$\Gamma_{w;2,5}$	0.0031 (−0.0058, 0.041)	$\Psi_{2,5}$	0.011 (0.0042, 0.038)	Range_{w_4}	586.02 (350.93, 846.24)
$\Gamma_{w;3,3}$	1.98 (1.9, 2.01)	$\Psi_{3,3}$	1.97 (1.95, 1.98)	Range_{w_5}	506.06 (293.25, 934.23)
$\Gamma_{w;3,4}$	−0.0058 (−0.015, 0.012)	$\Psi_{3,4}$	−0.013 (−0.045, −0.0002)		
$\Gamma_{w;3,5}$	0.016 (−0.0017, 0.029)	$\Psi_{3,5}$	0.0018 (−0.0089, 0.016)		
$\Gamma_{w;4,4}$	2.03 (1.99, 2.065)	$\Psi_{4,4}$	1.94 (1.90, 1.98)		
$\Gamma_{w;4,5}$	0.0064 (−0.0091, 0.016)	$\Psi_{4,5}$	0.0044 (−0.003, 0.012)		
$\Gamma_{w;5,5}$	1.91 (1.84, 2.026)	$\Psi_{5,5}$	1.96 (1.93, 1.98)		

Subscripts 1–6 correspond to BE, EH, RM, SM, and YB species.

approximately 20%. Convergence diagnostics revealed 5000 iterations to be sufficient for initial burn-in and so the remaining 30,000 samples from each chain were used for posterior inference. The 206 knot model required approximately 2 h to complete the MCMC sampling with the 106 and 51 knot models requiring substantially less time to collect the specified number of samples.

For the three knot intensities, there was negligible difference among the $\boldsymbol{\beta}$ parameter estimates. The estimated diagonal elements of Ψ and Γ_w for the three models were also nearly identical. Further, all of the 95% credible intervals for the off-diagonal elements in Ψ and Γ_w overlapped between the 126 and 206 knot models; however, the 206 knot model had several more significant off-diagonal elements (i.e., indicated by a credible interval that does not include zero). For the 51 knot model, off-diagonal elements of Γ_w were generally closer to zero and the corresponding elements in Ψ were significantly different from zero, suggesting that the coarseness of this knot grid could not capture the covariation among the residual spatial processes.

Table 4 presents the parameter estimates of Γ_w, Ψ, and $\boldsymbol{\phi}$ for the 126 knot model. For brevity we have omitted $\boldsymbol{\beta}$ estimates but note that 15 were significant at the 0.05 level. Significant off-diagonal elements $\Gamma_{w;2,3}$ and $\Gamma_{w;1,5}$ in Table 4 correspond the spatial correlations between BE and YB and between EH and RM. These associations can also be seen in the interpolated surface of $\tilde{\mathbf{w}}$ depicted in Fig. 4, where surface patterns are similar between BE and YB and between EH and RM.

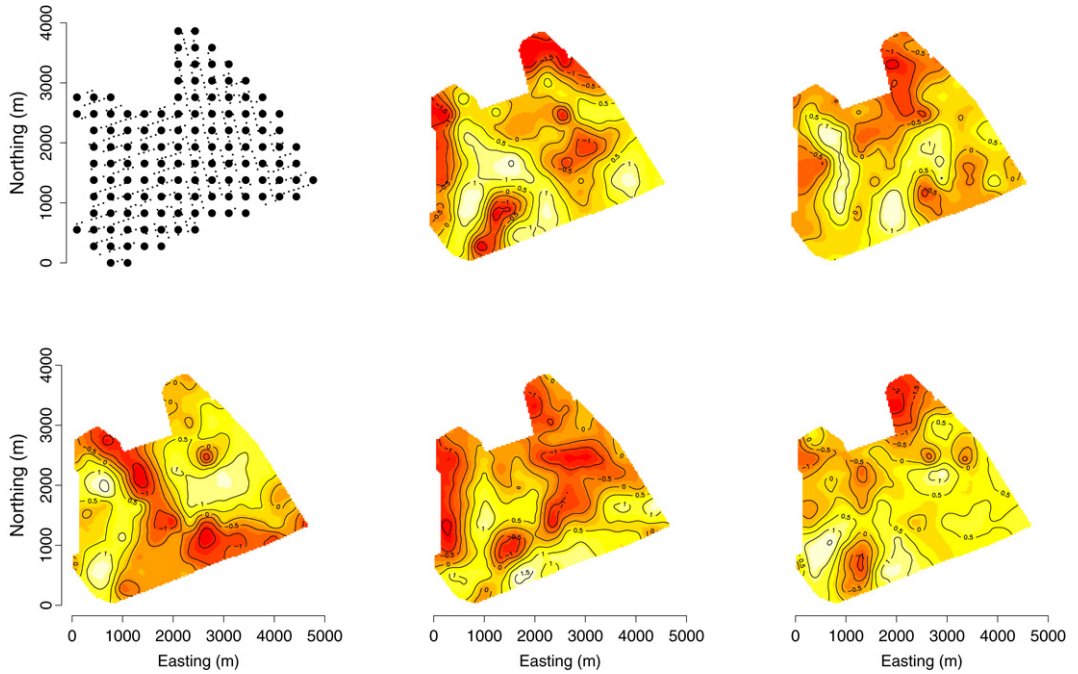

Fig. 4. Interpolated surfaces of the 126 knot model's median $\tilde{\mathbf{w}}$ at each inventory plot. Top left panel shows forest inventory plots (small points) under the 126 knots (large points). The order of response variables in the subsequent panels corresponds to Fig. 3.

Turning to prediction, it appears that the covariates and spatial proximity of observed inventory plots explain a significant portion of the variation in the response variables, perhaps leading to overfitting. We note that for our 37 hold-out plots the 95% prediction intervals are quite broad yielding a 100% empirical coverage for all three knot intensities. Finally, comparing the surface of pixel-level prediction for 1000 randomly selected pixels (Fig. 5) to the observed (Fig. 3) we see that the model can capture landscape-level variation in biomass and spatial patterns in biomass by species.

As described in Finley et al. (2008), the goal of these types of modeling exercises, that couple remotely sensed covariates and georeferenced forest inventory, is to enable fine resolution prediction of forest attributes (e.g., biomass) at the landscape scale. Ideally, the remotely sensed covariates would explain all of the variation in the response variable; however, this is rarely the case and we are often left with substantial spatial dependence in the residuals, as seen here. As noted above, the computational burden of the full multivariate geostatistical model forced Finley et al. (2008) to use only half the available forest inventory plot data. Although, we considered only a subset of the covariates used in Finley et al. (2008) and worked with a log-transformed response variable, we see several common trends in the residual spatial process (e.g., significance among several cross-covariance terms). Ultimately, the predictive process model makes this analysis and subsequent pixel-level prediction trivial for even a common single processor workstation.

4. Optimal knot design

4.1. A brief review of spatial design

As with any knot-based method, selection of knots is a challenging problem with choice in two dimensions more difficult than in one. Suppose for the moment that m is given. We are essentially dealing with a problem analogous to a spatial design problem, with the difference being that we already have samples at n locations. There is a rich literature in spatial design which is summarized in, e.g., the recent paper of Xia et al. (2006). One approach would be the so-called space-filling knot selection following the design ideas of Nychka and Saltzman (1998). Such designs are based upon geometric criteria, measures of how well a given set of points covers the study region, independent of the assumed covariance function. Instead, a number of authors have investigated the problem of optimal spatial sampling design assuming a particular spatial model. Model-based design often involves the minimization of a prediction-driven design criterion which depends on the particular prediction objectives. See, for example, McBratney and Webster (1981), Ritter (1996), and Zhu (2002). A recent work by Zhu and Stein (2005) and Zimmerman (2006) considers designs to achieve good prediction and accounts for covariance parameter estimation uncertainty using the likelihood. Diggle and Lophaven (2006) discuss a Bayesian design criterion which minimizes the spatially averaged prediction variance. Their Bayesian design approach naturally combines the goal of efficient spatial prediction while allowing for uncertainty in the values of model parameters. Application-specific numerical methods are often used to find optimal solutions. For example, Zhu and Stein (2005) implement the optimization

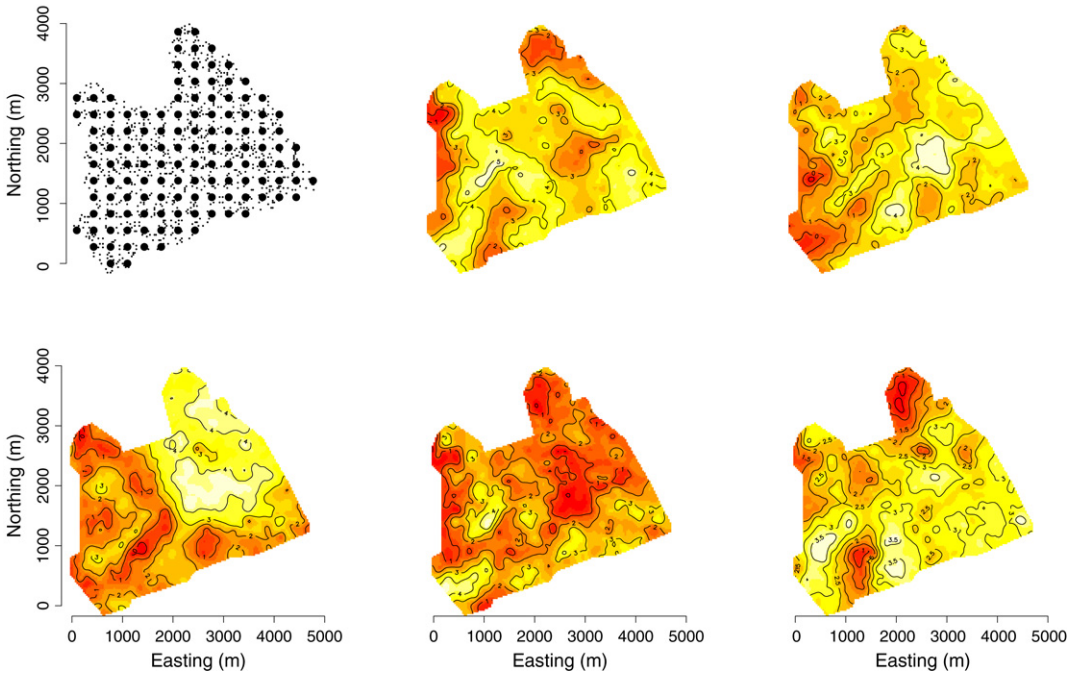

Fig. 5. Interpolated surfaces of the 126 knot model's median predicted response value over a random subset of 1000 pixels in the BEF. Top left panel shows the subset of prediction pixels (small points) under the 126 knots (large points). The order of response variables in the subsequent panels corresponds to Fig. 3.

using a simulated annealing algorithm. Xia et al. (2006) consider algorithms such as sequential selection, block selection and stochastic search.

4.2. Proposed approach

For a given set of observations, our goal is to construct a knot selection strategy such that the induced predictive process is a better approximation to the parent process. For a selected set of knots, $\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*]$ is considered as an approximation to the parent process. Given $\boldsymbol{\theta}$, the associated predictive variance of $w(\mathbf{s})$ conditional on the predictive process $\mathbf{w}^*$ on $\mathcal{S}^*$ can be written as $V_{\boldsymbol{\theta}}(\mathbf{s}, \mathcal{S}^*) = \text{VAR}[w(\mathbf{s})|\mathbf{w}(\cdot), \mathcal{S}^*, \boldsymbol{\theta}] = \boldsymbol{C}(\mathbf{s}, \mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s}, \boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{C}^{*-1}\mathbf{c}(\mathbf{s}, \boldsymbol{\theta})$, which measures how well we approximate $w(\mathbf{s})$ by the predictive process $\tilde{w}(\mathbf{s})$.

One possible criterion in knot selection is then defined as a function of $V_{\boldsymbol{\theta}}(\mathbf{s}, \mathcal{S}^*)$. One commonly used criterion is:

$$V_{\boldsymbol{\theta}}(\mathcal{S}^*) = \int_A V_{\boldsymbol{\theta}}(\mathbf{s}, \mathcal{S}^*)g(\mathbf{s})\mathrm{d}\mathbf{s} = \int_A \text{VAR}[w(\mathbf{s})|\mathbf{w}(\cdot), \mathcal{S}^*, \boldsymbol{\theta}]g(\mathbf{s})\mathrm{d}\mathbf{s},$$

where $g(\mathbf{s})$ is the weight assigned to location $\mathbf{s}$. In this paper we only consider the simple case for which $g(\mathbf{s}) \equiv 1$. $V_{\boldsymbol{\theta}}(\mathcal{S}^*)$ is referred to as spatially averaged predictive variance. In our case, we compute the spatially averaged prediction variance over all the observed locations, i.e.,

$$V_{\boldsymbol{\theta}}(\mathcal{S}^*) = \frac{\sum_{i=1}^{n} \text{VAR}[w(\mathbf{s}_i)|\mathbf{w}(\cdot), \mathcal{S}^*, \boldsymbol{\theta}]}{n}.$$

We ultimately reduce the problem of knot performance to the minimization of a design criterion which is the function $V_{\boldsymbol{\theta}}(\mathcal{S}^*)$.

It can be proved that: (1) $V_{\boldsymbol{\theta}}(\{\mathcal{S}^*, \mathbf{s}_0\}) - V_{\boldsymbol{\theta}}(\mathcal{S}^*) < 0$ for a new site $\mathbf{s}_0$, (2) $V_{\boldsymbol{\theta}}(\{\mathcal{S}^*, \mathbf{s}_0\}) - V_{\boldsymbol{\theta}}(\mathcal{S}^*) \longrightarrow 0$ when $\| \mathbf{s}_0 - \mathbf{s}_i^* \| \longrightarrow 0$, where $\mathbf{s}_i^*$ is any point of the knots, (3) $V_{\boldsymbol{\theta}}(\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}) = 0$, where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ are the original observed locations. The variance covariance matrix under the parent process model in Section 2 is $\Sigma_{\mathbf{Y}} = \boldsymbol{C} + \tau^2\boldsymbol{I}$, and the variance covariance matrix from the corresponding predictive process is given by $\Sigma_{pred} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{C}^{*-1}\boldsymbol{c} + \tau^2\boldsymbol{I}$. The Frobenius norm between $\Sigma_{\mathbf{Y}}$ and Σ_{pred} is $\| \Sigma_{\mathbf{Y}} - \Sigma_{pred} \| \equiv \text{tr}(\boldsymbol{C} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{C}^{*-1}\boldsymbol{c})^2$. Since $\boldsymbol{C} - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{C}^{*-1}\boldsymbol{c}$ is positive definite, the Frobenius norm $\| \Sigma_{\mathbf{Y}} - \Sigma_{pred} \| \equiv \sum \lambda_i^2$, where λ_i is the ith eigenvalue of $\Sigma_{\mathbf{Y}} - \Sigma_{pred}$. Also, the averaged predictive variance $V = \text{tr}(\Sigma_{\mathbf{Y}} - \Sigma_{pred})/n = \sum \lambda_i/n$.

In practice, the values of covariance parameters have to be estimated under the assumed model. An option is to obtain the parameter estimations by using a subset of original data or fitting the predictive process model based on a regular lattice

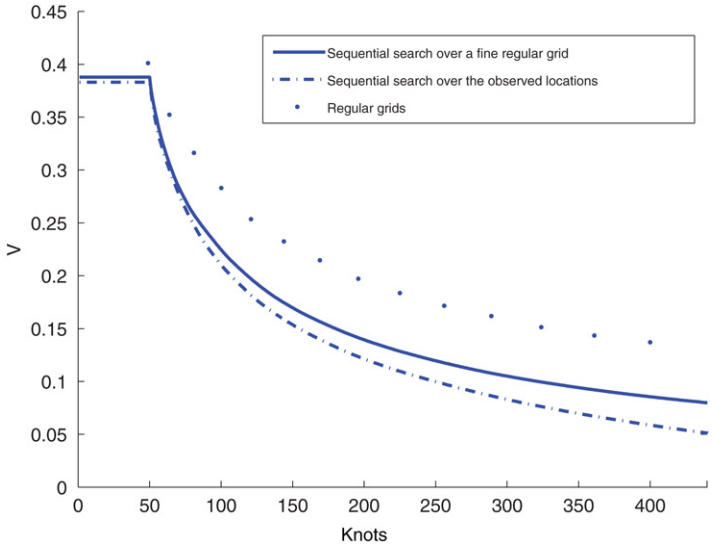

Fig. 6. Averaged prediction variance (V) versus number of knots (m). Solid dots denote results for regular grids; dash-dotted line denotes results for the sequential search over the observed data locations (starting with 49 randomly chosen sites from the observed locations), and; solid line denotes results for the sequential search over a 60×60 regular grid (starting with a 7×7 regular grid).

of knots. (That is what we do.) Another option is to adopt a Bayesian criterion, which places a prior on $\boldsymbol{\theta}$ and then minimizes $E_{\boldsymbol{\theta}}(V_{\boldsymbol{\theta}}(\mathcal{S}^*))$ (see, Diggle and Lophaven (2006)).

Suppose the values of the parameters and the knot size m are given. We consider the following sequential search algorithm approach to find the approximately optimal design:

- Initialization: specify allowable sampling locations of size N; possible choices include a fine grid, the observed locations or the union of these two sets.
- Specify a set of locations of size n_0 as starting points for knot selection; possible choices include a regular grid, or a subset of the observed locations chosen randomly or deterministically.
- At step $t + 1$,
  - For each sample point $\mathbf{s}_i$ in the allowable sample set, evaluate $V(\{\mathcal{S}^{*(t)}, \mathbf{s}_i\})$.
  - Remove the sample point with maximum decrease in V from the allowable sample set and add it to the knot set.
- Repeat the above procedure until we obtain m points in knot set.

The sequential evaluation of V is achieved using a very efficient algorithm incorporating block-matrix computation. We have successfully implemented the sequential algorithm in a simulation study shown in Section 4.3. We remark that the sequential algorithm does not necessarily achieve the global optimization solution. Alternative computational approaches are available to be used in finding approximately optimal designs such as stochastic search and block selection (see Xia et al. (2006)).

As for the choice of m, the obvious answer is "as large as possible". Evidently, this is governed by computational cost and sensitivity to choice. So, in principle, we will have to implement the analysis over different choices of m and consider run time along with stability of predictive inference; in our case, the value of minimized V under different choices of m.

Finally, we can perform a two-step analysis by combining this knot selection procedure with the modified predictive process in a natural way: (1) choose a set of knots to minimize the averaged predictive variances; (2) then use the modified process in the model fitting.

4.3. A simulation example using the two-step analysis

We generated 1000 data points in a $[0, 100] \times [0, 100]$ square and then generated the dependent variable from model (1) with an intercept $\mu = 1$ as a regressor, an exponential covariance function with range parameter $\phi = 0.06$ (i.e., an effective range of ~50 units), scale $\sigma = 1$ for the spatial process, and with nugget variance $\tau^2 = 1$. We illustrate a comparison among three design strategies, including regular grids, sequential search over all the observed locations and sequential search over a fine regular lattice. In Fig. 6, we plot the averaged predictive variances under each strategy. Sequential search algorithm is clearly better than choosing a regular grid as knots. For instance, with 180 sites selected, sequential search over the observed locations yielded an averaged predictive variance approximately 0.15. For the regular grids, roughly 150 additional sites are needed to achieve the same level of performance.

5. Summary and future work

Treating the "big N problem" for spatial data is currently an active research area and, with increased data collection and storage capability, will become even more of an issue. With our proposed modification and approximately optimal knot design, predictive process models offer an attractive tool for handling this problem.

Future work will extend these models to handling space–time datasets, where, with high temporal resolution, additional computational challenges exist. A related problem is scalability of spatial models. Spatial modeling is often done on small scales to achieve high resolution or at large scales, sacrificing resolution. Strategies that implement predictive processes offer the possibility of studying high resolution over large regions.

Acknowledgements

The work of the first and third authors was supported in part by NSF-DMS-0706870, that of the third and fourth authors was supported in part by NIH grant 1-R01-CA95995 and that of the second and fourth authors was supported in part by NSF-DEB-05-16198.

References

Banerjee, S., Carlin, B., Gelfand, A., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall.
Banerjee, S., Gelfand, A., Finley, A., Sang, H., 2008. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society, Series B 70, 825–848.
Cressie, N., 1993. Statistics for Spatial Data, 2nd ed. Wiley, New York.
Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scandinavian Journal of Statistics 33 (1), 53–64.
Finley, A., Banerjee, S., Ek, A., McRoberts, R., 2008. Bayesian multivariate process modeling for prediction of forest attributes. Journal of Agricultural, Biological, and Environmental Statistics 13 (1), 1–24.
Finley, A., Banerjee, S., Waldmann, P., Ericsson, T., Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. Biometrics (in press). doi:10.1111/j.1541-0420.2008.01115.x.
Gelfand, A., Schmidt, A., Banerjee, S., Sirmans, C., 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. Test 13 (2), 263–312.
Huang, C., Wylie, B., Homer, C., Yang, L., Zylstre, G., 2002. Derivation of a tasseled cap transformation based on landsat 7 at-satellite reflectance. International Journal of Remote Sensing 8, 1741–1748.
McBratney, A., Webster, R., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. II. Program and examples. Computers and Geosciences 7 (4), 335–365.
Nychka, D., Saltzman, N., 1998. Design of air quality monitoring networks. Case Studies in Environmental Statistics 51–76.
Ritter, K., 1996. Asymptotic optimality of regular sequence designs. The Annals of Statistics 24 (5), 2081–2096.
Stein, M., 1999. Interpolation of Spatial Data. Springer, New York.
Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer.
Xia, G., Miranda, M., Gelfand, A., 2006. Approximately optimal spatial design approaches for environmental health data. Environmetrics 17 (4), 363–385.
Zhu, Z., 2002. Optimal sampling design and parameter estimation of gaussian random fields. Ph.D. Thesis. University of Chicago. Dept. of Statistics.
Zhu, Z., Stein, M., 2005. Spatial sampling design for parameter estimation of the covariance function. Journal of Statistical Planning and Inference 134 (2), 583–603.
Zimmerman, D., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17 (6), 635–652.