Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 4, *pp.* 825–848

Gaussian predictive process models for large spatial data sets

Sudipto Banerjee,

University of Minnesota, Minneapolis, USA

Alan E. Gelfand,

Duke University, Durham, USA

Andrew O. Finley

Michigan State University, East Lansing, USA

and Huiyan Sang

Duke University, Durham, USA

[Received April 2007. Final revision February 2008]

Summary. With scientific data available at geocoded locations, investigators are increasingly turning to spatial process models for carrying out statistical inference. Over the last decade, hierarchical models implemented through Markov chain Monte Carlo methods have become especially popular for spatial modelling, given their flexibility and power to fit models that would be infeasible with classical methods as well as their avoidance of possibly inappropriate asymptotics. However, fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases in cubic order with the number of spatial locations, rendering such models infeasible for large spatial data sets. This computational burden is exacerbated in multivariate settings with several spatially dependent response variables. It is also aggravated when data are collected at frequent time points and spatiotemporal process models are used. With regard to this challenge, our contribution is to work with what we call predictive process models for spatial and spatiotemporal data. Every spatial (or spatiotemporal) process induces a predictive process model (in fact, arbitrarily many of them). The latter models project process realizations of the former to a lower dimensional subspace, thereby reducing the computational burden. Hence, we achieve the flexibility to accommodate non-stationary, non-Gaussian, possibly multivariate, possibly spatiotemporal processes in the context of large data sets. We discuss attractive theoretical properties of these predictive processes. We also provide a computational template encompassing these diverse settings. Finally, we illustrate the approach with simulated and real data sets.

Keywords: Co-regionalization; Gaussian processes; Hierarchical modelling; Kriging; Markov chain Monte Carlo methods; Multivariate spatial processes; Space–time processes

1. Introduction

Recent advances in geographical information systems and global positioning systems enable accurate geocoding of locations where scientific data are collected. This has encouraged the formation of large spatiotemporal data sets in many fields and has generated considerable

Address for correspondence: Sudipto Banerjee, Division of Biostatistics, School of Public Health, University of Minnesota, Mayo Mail Code 303, Minneapolis, MN 55455-0392, USA.
E-mail: sudiptob@biostat.umn.edu

 1369–7412/08/70825

interest in statistical modelling for location-referenced spatial data; see, for example, Cressie (1993), Møller (2003), Banerjee *et al.* (2004) and Schabenberger and Gotway (2004) for a variety of methods and applications. Here, we focus on the setting where the number of locations yielding observations is too large for fitting desired hierarchical spatial random-effects models. Full inference and accurate assessment of uncertainty often require Markov chain Monte Carlo (MCMC) methods (Banerjee *et al.*, 2004). However, 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 data sets. Evidently, the problem is further aggravated when we have a vector of random effects at each location or when we have spatiotemporal random effects.

Approaches to tackle this problem have adopted several different paths. The first seeks approximations for the spatial process by using kernel convolutions, moving averages, low rank splines or basis functions (e.g. Wikle and Cressie (1999), Lin *et al.* (2000), Higdon (2001), Ver Hoef *et al.* (2004), Xia and Gelfand (2006), Kammann and Wand (2003) and Paciorek (2007)). Essentially, the process $w(\mathbf{s})$ is replaced by an approximation $\tilde{w}(\mathbf{s})$ that represents the realizations in a lower dimensional subspace. A second approach seeks, instead, to approximate the likelihood either by working in the spectral domain of the spatial process and avoiding the matrix computations (Stein, 1999; Fuentes, 2007; Paciorek, 2007) or by forming a product of appropriate conditional distributions to approximate the likelihood (e.g. Vecchia (1988), Jones and Zhang (1997) and Stein *et al.* (2004)). A concern is the adequacy of the resultant likelihood approximation. Expertise in tailoring and tuning of a suitable spectral density estimate or a sequence of conditional distributions is required and does not easily adapt to multivariate processes. Also, the spectral density approaches seem best suited to stationary covariance functions. Another approach either replaces the process (random-field) model by a *Markov* random field (Cressie, 1993) or else approximates the random-field model by a Markov random field (Rue and Tjelmeland, 2002; Rue and Held, 2006). This approach is best suited for points on a regular grid. With irregular locations, realignment to a grid or a torus is required, which is done by an algorithm, possibly introducing unquantifiable errors in precision. Adapting these approaches to more complex hierarchical spatial models involving multivariate processes (e.g. Wackernagel (2003) and Gelfand *et al.* (2004)), spatiotemporal processes and spatially varying regressions (Gelfand *et al.*, 2003) and non-stationary covariance structures (Paciorek and Schervish, 2006) is potentially problematic.

We propose a class of models that is based on the idea of a spatial predictive process (motivated from kriging ideas). We project the original process onto a subspace that is generated by realizations of the original process at a specified set of locations. The rudiments of our idea appear in Switzer (1989). Our approach is in the same spirit as process model approaches using basis functions and kernel convolutions, i.e. specifications which attempt to facilitate computation. Our version is directly connected to whatever valid covariance structure we seek and is applicable to any class of distributions that can support a spatial stochastic process. We typically use such modelling to describe an underlying process; one that is never actually observed. The modelling provides structured dependence for random effects, e.g. intercepts or coefficients, at a second stage of specification where the first stage need not be Gaussian.

Our objectives differ from Gaussian process regressions for large data sets in machine learning (see, for example, Wahba (1990), Seeger *et al.* (2003) and Rasmussen and Williams (2006)), where the regression function is viewed as a Gaussian process realization centred on some *mean* function with a conditionally independent Gaussian model for the data given the regression function. Assuming known process and noise parameters and a large *training set*, the Gaussian posterior will exhibit the big n problem. Recently, Cornford *et al.* (2005) turned this

machinery to geostatistics. In all of the above, explicit forms for Bayesian prediction ('kriging' in the last case) and the associated posterior variance are utilized. However, we are interested in full posterior inference associated with spatiotemporal hierarchical models and parameters therein. Full Bayesian inference would most probably employ MCMC methods (Robert and Casella, 2005), an approach that has been dismissed in the machine learning literature (see, for example, Cornford *et al.* (2005)). In this sense our current work is more ambitious in its spatial modelling and inferential objectives.

We illustrate with two simulated examples yielding rather complex spatial random fields and with a challenging spatially adaptive regression model fitted to forest biomass data. Section 2 introduces and discusses the predictive process models as well as relevant specification issues. Section 3 considers extensions to multivariate process models. Section 4 describes Bayesian computation issues and Section 5 presents the two examples. Section 6 concludes with a brief discussion including future work.

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Univariate predictive process modelling

2.1. The customary univariate model

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

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

i.e. the residual comprises a spatial process $w(\mathbf{s})$, capturing spatial association, and an independent process $\varepsilon(\mathbf{s})$, which is often called the *nugget*. The $w(\mathbf{s})$ are spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The customary process specification for $w(\mathbf{s})$ is a mean 0 Gaussian process with covariance function, $C(\mathbf{s}, \mathbf{s}')$, which is denoted $\mathrm{GP}\{0, C(\mathbf{s}, \mathbf{s}')\}$. 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. In any event, $\varepsilon(\mathbf{s}) \sim^{\mathrm{IID}} N(0, \tau^2)$ for every location $\mathbf{s}$. With n observations $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^{\mathrm{T}}$, the data likelihood is given by $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$, 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 $C(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$. Likelihood-based inference proceeds through maximum likelihood or restricted maximum likelihood methods (e.g. Schabenberger and Gotway (2004)).

With hierarchical models, we assign prior distributions to $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2)$ and inference proceeds by sampling from $p(\Omega|\mathbf{Y})$, whereas prediction at an arbitrary site $\mathbf{s}_0$ samples $p\{Y(\mathbf{s}_0)|\mathbf{Y}\}$ one for one with posterior draws of Ω by composition (Banerjee *et al.*, 2004). This is especially convenient for Gaussian likelihoods since $p\{Y(\mathbf{s}_0)|\Omega, \mathbf{Y}\}$ is itself Gaussian. Evidently, both estimation and prediction require evaluating the Gaussian likelihood: hence, evaluating the $n \times n$ matrix $\Sigma_{\mathbf{y}}^{-1}$. Although explicit inversion is replaced with faster linear solvers, likelihood evaluation remains expensive for big n.

2.2. The predictive process model

We consider 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 Gaussian process in model (1) yields $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim$

$\mathrm{MVN}\{\mathbf{0}, C^*(\boldsymbol{\theta})\}$, where $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$ is the corresponding $m \times m$ covariance matrix. The spatial interpolant (that leads to kriging) at a site $\mathbf{s}_0$ is given by $\tilde{w}(\mathbf{s}_0) = E[w(\mathbf{s}_0)|\mathbf{w}^*] = \mathbf{c}^{\mathrm{T}}(\mathbf{s}_0; \boldsymbol{\theta})\, C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$, where $\mathbf{c}(\mathbf{s}_0; \boldsymbol{\theta}) = [C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$. This single-site interpolator, in fact, defines a spatial process $\tilde{w}(\mathbf{s}) \sim \mathrm{GP}\{0, \tilde{C}(\cdot)\}$ with covariance function

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^{\mathrm{T}}(\mathbf{s}; \boldsymbol{\theta})\, C^{*-1}(\boldsymbol{\theta})\, \mathbf{c}(\mathbf{s}', \boldsymbol{\theta}), \tag{2}$$

where $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$. We refer to $\tilde{w}(\mathbf{s})$ as the *predictive process* derived from the *parent process* $w(\mathbf{s})$. The realizations of $\tilde{w}(\mathbf{s})$ are precisely the kriged predictions conditional on a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. The process is completely specified given the covariance function of the parent process and $\mathcal{S}^*$. So, to be precise, we should write $\tilde{w}_{\mathcal{S}^*}(\mathbf{s})$, but we suppress this implicit dependence. From equation (2), this process is non-stationary regardless of whether $w(\mathbf{s})$ is. Furthermore, the joint distribution that is associated with the realizations at any set of locations in D is non-singular if and only if the set has at most m locations.

Replacing $w(\mathbf{s})$ in model (1) with $\tilde{w}(\mathbf{s})$, we obtain the predictive process model

$$Y(\mathbf{s}) = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \varepsilon(\mathbf{s}). \tag{3}$$

Since $\tilde{w}(\mathbf{s}) = \mathbf{c}^{\mathrm{T}}(\mathbf{s})C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$, $\tilde{w}(\mathbf{s})$ is a spatially varying linear transformation of $\mathbf{w}^*$. The dimension reduction is seen immediately. In fitting model (3), the n random effects $\{w(\mathbf{s}_i),\ i = 1, 2, \ldots, n\}$ are replaced with only the m random effects in $\mathbf{w}^*$; we can work with an m-dimensional joint distribution involving only $m \times m$ matrices. Evidently, model (3) is different from model (1). Hence, though we introduce the same set of parameters in both models, they will not be identical in both models.

Knot-based linear combinations such as $\sum_{i=1}^m a_i(\mathbf{s})\, w(\mathbf{s}_i^*)$ resemble other process approximation approaches. For instance, motivated by an integral representation of (certain) stationary processes as a kernel convolution of Brownian motion on $\Re^2$, Higdon (2001) proposed a finite approximation to the parent process of the form $\sum_{i=1}^m a_i(\mathbf{s}; \boldsymbol{\theta})u_i$ where u_is are independent and identically distributed (IID) $N(0, 1)$ and $a_i(\mathbf{s}; \boldsymbol{\theta}) = k(\mathbf{s}, \mathbf{s}_i^*; \boldsymbol{\theta})$ with $k(\cdot; \boldsymbol{\theta})$ being a Gaussian *kernel* function. Evidently, Gaussian kernels only capture Gaussian processes with Gaussian covariance functions (see Paciorek and Schervish (2006)). Xia and Gelfand (2006) suggested extensions to capture more general classes of stationary Gaussian processes by aligning kernels with covariance functions. However, the class of stationary Gaussian process models admitting a kernel representation is limited.

The integral representation permits kernel convolution with spatially varying kernels as in Higdon *et al.* (1999). It can also replace Brownian motion with a stationary Gaussian process on $\Re^2$. Finite approximations allow replacing the white noise realizations u_i with stationary process realizations such as $\mathbf{w}^*$. Then, the original realizations are projected onto an m-dimensional subspace that is generated by the columns of the $n \times m$ matrix $K = [k(\mathbf{s}_i, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^{n,m}$, where $\tilde{\mathbf{w}} = K\mathbf{w}^*$. Alternatively, one could project as $\tilde{\mathbf{w}} = Z\mathbf{u}$, where $\mathbf{u} \sim N(\mathbf{0}, I)$, and set $Z = [\mathbf{c}^{\mathrm{T}}(\mathbf{s}_i; \boldsymbol{\theta})]_{i=1}^n C^{*-1/2}(\boldsymbol{\theta})$ to yield the same joint distribution as the predictive process model (3). This approach has been used in 'low rank kriging' methods (Kammann and Wand, 2003). We note that we do not want to work with the induced process $\tilde{\tilde{w}}(\mathbf{s}) = \mathbf{c}^{\mathrm{T}}(\mathbf{s})\, C^{*-1/2}(\boldsymbol{\theta})\mathbf{w}^*$. It has the induced covariance function $\mathbf{c}^{\mathrm{T}}(\mathbf{s})\, \mathbf{c}(\mathbf{s})$ but sacrifices properties that we seek below in Section 2.3 (e.g. exact interpolation or Kullback–Leibler (KL) minimization). More general low rank spline models were also discussed in Ruppert *et al.* (2003), chapter 13, and Lin *et al.* (2000). Again, we regard the predictive process as a competing model specification with computational advantages, but induced by an underlying full rank process. In fact, these models are, in some sense, *optimal* projections as we clarify in the next subsection. Also, $\tilde{w}(\mathbf{s})$ does not arise as a discretization of an integral representation of a process and we only require a valid covariance function to induce it.

A somewhat similar reduced rank kriging method has been recently proposed by Cressie and Johannesson (2008). Letting $\mathbf{g}(\mathbf{s})$ be a $k \times 1$ vector of specified basis functions on $\Re^2$, the proposed covariance function is $C(\mathbf{s}, \mathbf{s}') = \mathbf{g}(\mathbf{s})^{\mathrm{T}} K\, \mathbf{g}(\mathbf{s}')$ with K an unknown positive definite $k \times k$ matrix that is estimated from the data by using a method-of-moments approach. Such an approach may be challenging for the hierarchical models that we envision here. We shall be providing spatial modelling with random effects at the second stage of the specification. We have no 'data' to provide an empirical covariance function.

Lastly, we can draw a connection to recent work in spatial dynamic factor analysis (see, for example, Lopes *et al.* (2006) and references therein), where K is viewed as an $n \times m$ matrix of factor loadings. Neither K nor $\mathbf{w}^*$ is known but replication over time in the form of a dynamic model is introduced to enable the data to separate them and to infer about them. In our case, the entries in K are 'known' given the covariance function C.

2.3. Properties of the predictive process

First, note that $\tilde{w}(\mathbf{s}_0)$ is an orthogonal projection of $w(\mathbf{s}_0)$ onto a particular linear subspace (e.g. Stein (1999)). Let $\mathcal{H}_{m+1}$ be the Hilbert space that is generated by $w(\mathbf{s}_0)$ and the m random variables in $\mathbf{w}^*$ (with $\mathcal{H}_m$ denoting the space that is generated by the latter); hence, $\mathcal{H}_{m+1}$ comprises all linear combinations of these $m+1$ zero-centred, finite variance random variables along with their mean-square limit points. If we seek the element in $\tilde{w}(\mathbf{s}_0) \in \mathcal{H}_m$ that is closest to $w(\mathbf{s}_0)$ in terms of the inner product norm that is induced by $E[w(\mathbf{s})\, w(\mathbf{s}')]$, we obtain the linear system $E[\{w(\mathbf{s}_0) - \tilde{w}(\mathbf{s}_0)\}\, w(\mathbf{s}_j^*)] = 0$, $j = 1, \ldots, m$, with the unique solution $\tilde{w}(\mathbf{s}_0) = \mathbf{c}^{\mathrm{T}}(\mathbf{s}_0)\, C^{*-1}(\boldsymbol{\theta})\mathbf{w}^*$. Being a conditional expectation, it immediately follows that $\tilde{w}(\mathbf{s}_0)$ minimizes $E[w(\mathbf{s}_0) - f(\mathbf{w}^*)|\mathbf{w}^*]$ over all real-valued functions $f(\mathbf{w}^*)$. In this sense, the predictive process is the best approximation for the parent process.

Also, $\tilde{w}(\mathbf{s}_0)$ *deterministically* interpolates $w(\mathbf{s})$ over $\mathcal{S}^*$. Indeed, if $\mathbf{s}_0 = \mathbf{s}_j^* \in \mathcal{S}^*$ we have

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

since $\mathbf{e}_j^{\mathrm{T}}\, C^*(\boldsymbol{\theta}) = \mathbf{c}^{\mathrm{T}}(\mathbf{s}_j^*, \boldsymbol{\theta})$, where $\mathbf{e}_j$ denotes the vector with 1 in the jth position and 0 elsewhere. So, equation (4) shows that $E[\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*] = w(\mathbf{s}_j^*)$ and $\mathrm{var}\{\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*\} = 0$ (a property of kriging). At the other extreme, suppose that $\mathcal{S}$ and $\mathcal{S}^*$ are disjoint. Then $\mathbf{w}|\mathbf{w}^* \sim \mathrm{MVN}(c^{\mathrm{T}}\, C^{*-1}\mathbf{w}^*, C - c^{\mathrm{T}} C^{*-1} c)$ where c is the $m \times n$ matrix whose columns are the $\mathbf{c}(\mathbf{s}_i)$ and C is the $n \times n$ covariance matrix of $\mathbf{w}$. We can write $\mathbf{w} = \tilde{\mathbf{w}} + (\mathbf{w} - \tilde{\mathbf{w}})$ and the choice of $\mathbf{w}^*$ determines the (conditional) variability in the second term on the right-hand side, i.e. how close $\Sigma_{\mathbf{w}}$ is to $\Sigma_{\tilde{\mathbf{w}}}$. It also reveals that there will be less variability in the predictive process than in the parent process as n variables are determined by $m < n$ random variables.

KL-based justification for $\tilde{w}(\mathbf{s}^*)$ is discussed in Csató (2002) and Seeger *et al.* (2003). The former offers a general theory for KL projections and proposes sequential algorithms for computations. As a simpler and more direct argument, let us assume that $\mathcal{S}^*$ and $\mathcal{S}$ are disjoint and let $\mathbf{w}_a = (\mathbf{w}^*, \mathbf{w})^{\mathrm{T}}$ be the $(m+n) \times 1$ vector of realizations over $\mathcal{S}^* \cup \mathcal{S}$. In model (1), assuming all other model parameters fixed, the posterior distribution for $p(\mathbf{w}_a|\mathbf{Y})$ is proportional to $p(\mathbf{w}_a)\, p(\mathbf{Y}|\mathbf{w})$ since $p(\mathbf{Y}|\mathbf{w}_a) = p(\mathbf{Y}|\mathbf{w})$. The corresponding posterior in model (3) replaces $p(\mathbf{Y}|\mathbf{w})$ with a density $q(\mathbf{Y}|\mathbf{w}^*)$. Letting $\mathcal{Q}$ be the class of all probability densities satisfying $q(\mathbf{Y}|\mathbf{w}_a) = q(\mathbf{Y}|\mathbf{w}^*)$, suppose that we seek the density $q \in \mathcal{Q}$ that minimizes the reverse KL divergence $\mathrm{KL}(q, p) = \int q \log(q/p)$. In Appendix A we argue that $\mathrm{KL}\{q(\mathbf{w}_a|\mathbf{Y}), p(\mathbf{w}_a|\mathbf{Y})\}$ is minimized when $q(\mathbf{Y}|\mathbf{w}^*) \propto \exp(E_{\mathbf{w}|\mathbf{w}^*}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])$. Subsequent calculations from standard multivariate normal theory reveal this to be the Gaussian likelihood corresponding to the predictive process model.

2.4. Selection of knots

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. First, in the spline smoothing literature (and in most of the literature on functional data or regression modelling using basis representations), it is customary to place knots at every data point (e.g. Ramsay and Silverman (2005)). This is not an option for us and raises the question of whether to use a subset of the observed spatial locations or a disjoint set of locations. If we use a subset of the sampled locations, should we draw this set at random? If we do not use a subset then we are dealing with a problem that is analogous to a 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, for example, Xia *et al.* (2006). We need a criterion to decide between a regular grid and placing more knots where we have sampled more. One approach would be a so-called space filling knot selection following the design ideas of Nychka and Saltzman (1998). Such designs are based on geometric criteria, measures of how well a given set of points covers the study region, independent of the assumed covariance function. Stevens and Olsen (2004) showed that spatial balance of design locations is more efficient than simple random sampling. Recently, Diggle and Lophaven (2006) have discussed spatial designs suggesting modification to regular grids. These designs augment the lattice with close pairs or infill. We examine such designs for knot selection in our simulation examples in Section 5.1.1.

To bring in the covariance function, in our setting, since the joint distributions of $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are both multivariate normal, we might think in terms of using KL distance between these distributions. According to choice, the non-symmetrized distance will involve the covariance matrix of one distribution and the inverse covariance matrix of the other. We cannot obtain the inverse covariance matrix for $\mathbf{w}$ and the inverse covariance matrix for $\tilde{\mathbf{w}}$ does not exist since the distribution is singular. Moreover, if we try to work with the KL distance for any set of m knots, we have argued above that in this case $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are the same realization; the KL distance is 0.

In working with kernel convolution approximation, KL distance can be used with any subset of locations because it is not an exact interpolator. In fact, Xia and Gelfand (2005) showed that, working with a regular grid, introduction of a lattice that is larger than the study area is desirable. However, this is justified by the nature of the discrete approximation that is made to an integral representation that is over all of $\Re^2$. Since the predictive process is not driven by a finite approximation to an integral representation, such expansion does not seem warranted here.

A direct assessment of knot performance is comparison of the covariance function of the parent process with that of the predictive process, given in equation (2) and dependent on $\mathcal{S}^*$. To illustrate, 200 locations are uniformly generated over a $[0, 10] \times [0, 10]$ rectangle. The knots consist of a 10×10 equally spaced grid. We employ the Matérn covariance function to make comparisons, setting $\sigma^2 = 1$ without loss of generality. Fixing the range parameter $\phi = 2$, Fig. 1 overlays the covariances of the parent process and those of the predictive processes for four different values of the smoothness parameter ν. (Covariances for 2000 of the roughly 40 000 distance pairs are plotted for the predictive process.) Alternatively, setting $\nu = 0.5$ (the exponential covariance function), we compare covariances by using four values of the range parameter in Fig. 2.

In general, we find that the functions agree better at larger distances and even more so with increasing smoothness and range. There is little additional information in varying m for a given ϕ and ν. What matters is the size of the range relative to the spacing of the grid for the knots. Indeed, the fact that the comparison is weakest at short distances and is even weaker for shorter ranges merely reflects the fact that observations at the knots will not, in this case, provide much

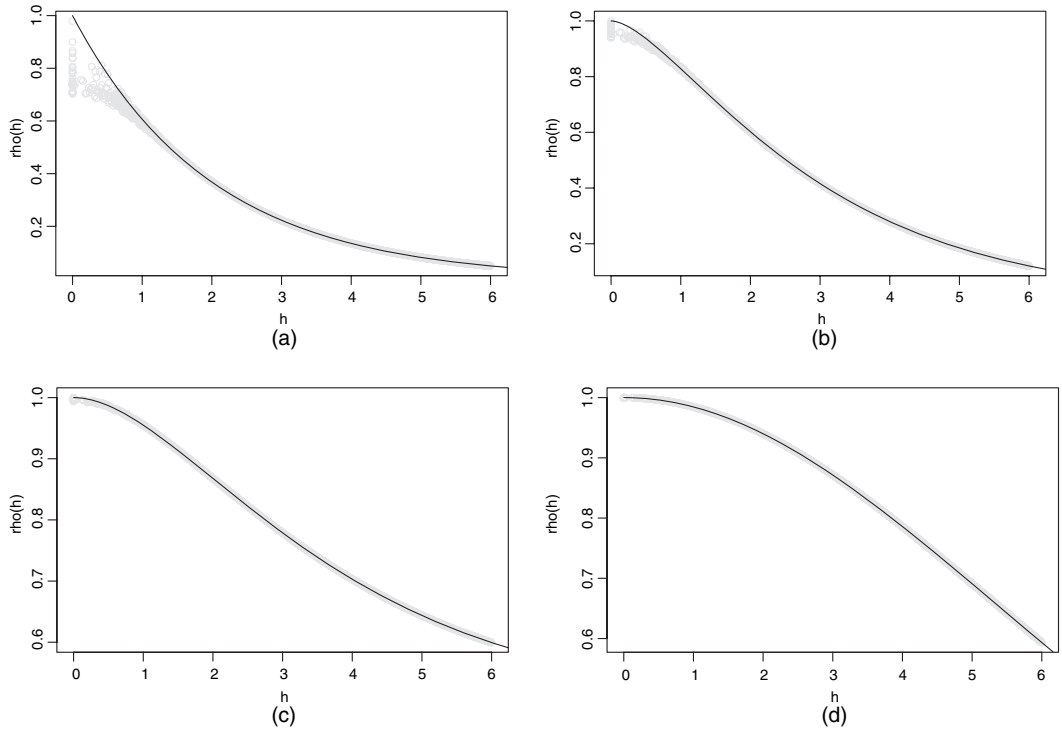

Fig. 1. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance (●): (a) smoothness parameter 0.5; (b) smoothness parameter 1; (c) smoothness parameter 1.5; (d) smoothness parameter 5

information about dependence for pairs of sites that are very close to each other. With concern about fine scale spatial variation or very short spatial range, a rather dense set of knots may be required (see, for example, Stein (2007)). Then, knot selection incorporating varying density over the domain including a 'packed' subset would enable us to inform better about the spatial range and the variance components.

Finally, the choice of m is governed by computational cost and sensitivity to choice. So, in principle, we shall have to implement the analysis over different choices of m. Since we cannot work with the full set of n locations, comparison must be relative and will consider run time (with associated computational stability) along with stability of predictive inference. Indeed, Section 5 below illustrates model performance with various choices of m.

2.5. Non-Gaussian first-stage models

There are two typical non-Gaussian first-stage settings:

(a) binary response at locations modelled by using logit or probit regression and
(b) count data at locations modelled by using Poisson regression.

Diggle *et al.* (1998) unified the use of generalized linear models in spatial data contexts. See also Lin *et al.* (2000), Kammann and Wand (2003) and Banerjee *et al.* (2004). Essentially we replace model (1) with the assumption that $E[Y(\mathbf{s})]$ is linear on a transformed scale, i.e. $\eta(\mathbf{s}) \equiv g\{E[Y(\mathbf{s})]\} = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s})$ where $g(\cdot)$ is a suitable link function. With the Gaussian first stage, we can marginalize over the ws to achieve the covariance matrix $\tau^2 I + \mathbf{c}^{\mathrm{T}} C^{*-1} \mathbf{c}$. Though

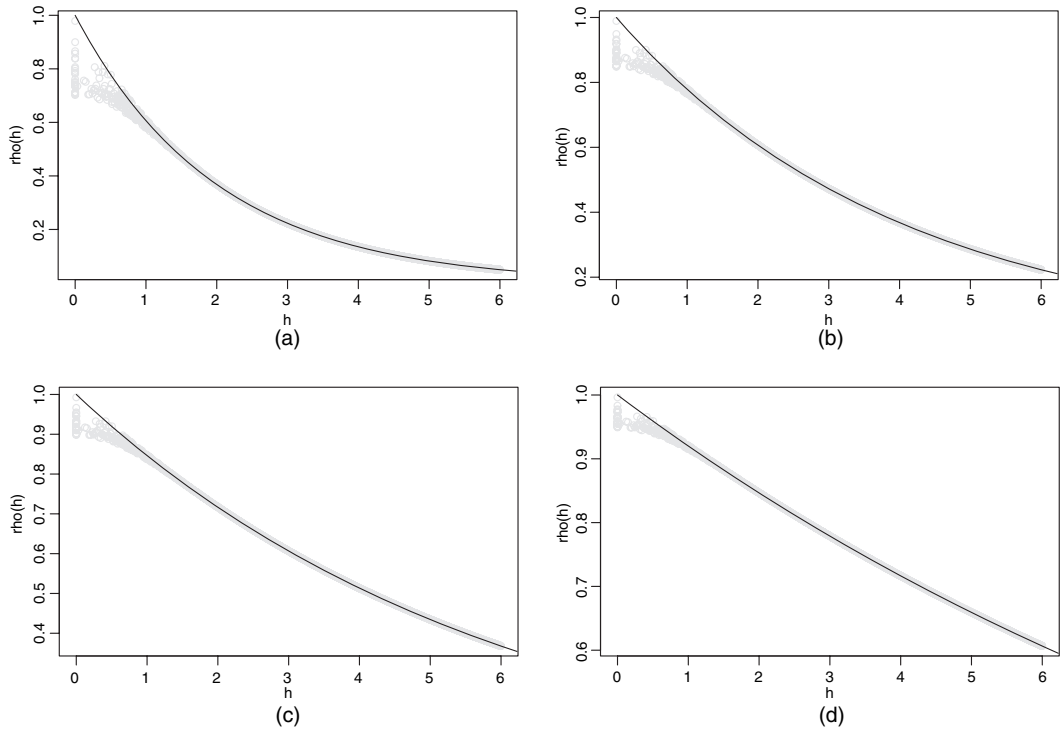

Fig. 2. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance: (a) range parameter 2; (b) range parameter 4; (c) range parameter 6; (d) range parameter 12

this matrix is $n \times n$, using the Sherman–Woodbury result (Harville (1997); also see Section 4), inversion requires only C^{*-1}. With, say, a binary or Poisson first stage, such marginalization is precluded; we must update the ws in running our Gibbs sampler. Using the predictive process, we must update only the $m \times 1$ vector $\mathbf{w}^*$.

A little more clarification may be useful. As described in the previous paragraph, the resulting model would take the form $\Pi_i\, p\{Y(\mathbf{s}_i)|\boldsymbol{\beta}, \mathbf{w}^*, \phi\}\, p(\mathbf{w}^*|\sigma^2, \phi)\, p(\boldsymbol{\beta}, \phi, \sigma^2)$. Though $\mathbf{w}^*$ is only $m \times 1$, updating this vector through its full conditional distribution may be awkward owing to the way that $\mathbf{w}^*$ enters the likelihood. Computations can be simplified by introducing a small amount of pure error to let $\eta(\mathbf{s}) \equiv g\{E[Y(\mathbf{s})]\} = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \varepsilon(\mathbf{s})$ where the $\varepsilon(\mathbf{s}) \sim^{\mathrm{IID}} N(0, \tau^2)$ with τ^2 known and very small. The full model takes the form

$$\prod_i p\{Y(\mathbf{s}_i)|\eta(\mathbf{s}_i)\} \prod_i p\{\eta(\mathbf{s}_i)|\boldsymbol{\beta}, \mathbf{w}^*, \phi\}\, p(\mathbf{w}^*|\phi, \sigma^2)\, p(\boldsymbol{\beta}, \phi, \sigma^2).$$

The full conditional distribution for $\mathbf{w}^*$ is now multivariate normal and the $\eta(\mathbf{s}_i)$ are conditionally independent. General categorical data settings can be treated by using these ideas.

2.6. Spatiotemporal versions

There are various spatiotemporal contexts in which predictive processes can be introduced to render computation feasible. We illustrate three of them here. First, we generalize model (1) to

$$Y(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)\boldsymbol{\beta} + w(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t) \tag{5}$$

for $\mathbf{s} \in D$ and $t \in [0, T]$. In model (5), the εs are, again, pure error terms, the $\mathbf{x}(\mathbf{s}, t)$ are local space–time covariate vectors and $\boldsymbol{\beta}$ is a coefficient vector, which is here assumed constant over space and time but can be spatially and/or temporally varying (see Section 3). We have replaced the spatial random effects $w(\mathbf{s})$ with space–time random effects $w(\mathbf{s}, t)$ that come from a Gaussian process with covariance function $\operatorname{cov}\{w(\mathbf{s}, t), w(\mathbf{s}', t')\} \equiv C(\mathbf{s}, \mathbf{s}'; t, t')$. There has been recent discussion regarding valid non-separable space–time covariance functions; see, for example, Stein (2005) and discussion therein. Now, we assume data $Y(\mathbf{s}_i, t_i)$, $i = 1, 2, \ldots, n$, where n can be very large because there are many distinct locations, times or both. In any event, the predictive process will be defined analogously to that above: $\tilde{w}(\mathbf{s}, t) = c(\mathbf{s}, t)^{\mathrm{T}} C^{*-1} \mathbf{w}^*$ where now $\mathbf{w}^*$ is an $m \times 1$ vector associated with m knots over $D \times [0, T]$ having covariance matrix C^* and $c(\mathbf{s}, t)$ is the vector of covariances of $w(\mathbf{s}, t)$ with the entries in $\mathbf{w}^*$. The spatiotemporal predictive process model $\tilde{w}(\mathbf{s}, t)$ will enjoy the same properties as $\tilde{w}(s)$. Now, the issue of knot selection arises over $D \times [0, T]$.

Next, suppose that we discretize time to, say, $t = 1, 2, \ldots, T$. Now, we would write the response as $Y_t(\mathbf{s})$ and the random effects as $w_t(\mathbf{s})$. Dynamic evolution of $w_t(\mathbf{s})$ is natural, leading to a spatial dynamic model as discussed in, for example, Gelfand *et al.* (2005). In one scenario the data may arise as a time series of spatial processes, i.e. there is a conceptual time series at each location $\mathbf{s} \in D$. Alternatively, it may arise as cross-sectional data, i.e. there is a set of locations that are associated with each time point and these can differ from time point to time point. In the latter case, we can expect an explosion of locations as time goes on. Use of predictive process modelling, defined through a dynamic sequence of $\mathbf{w}_t^*$s sharing the same knots, enables us to handle this.

Finally, predictive processes offer an alternative to the dimension reduction approach to space–time Kalman filtering that was presented by Wikle and Cressie (1999). With time discretized, they envisioned evolution through a discretized integrodifferential equation with spatially structured noise, i.e.

$$w_t(\mathbf{s}) = \int h_{\mathbf{s}}(\mathbf{u})\, w_{t-1}(\mathbf{u})\, \mathrm{d}\mathbf{u} + \eta_t(\mathbf{s})$$

with $h_{\mathbf{s}}$ a location interaction function and η a *spatially coloured* error process. $w_t(\mathbf{s})$ is decomposed as $\Sigma_{k=1}^{K} \phi_k(\mathbf{s}) a_{kt}$ where the $\phi_k(\mathbf{s})$s are deterministic orthonormal basis functions and the as are mean 0 time series. Then, each $h_{\mathbf{s}}$ has a basis representation in the ϕs, i.e. $h_{\mathbf{s}}(\mathbf{u}) = \Sigma_{l=1}^{\infty} b_l(\mathbf{s})\, \phi_l(\mathbf{u})$ where the bs are unknown. A dynamic model for the $k \times 1$ vector a_t driven by a linear transformation of the spatial noise process $\eta(\mathbf{s})$ results. Instead of the above decomposition for $w_t(\mathbf{s})$, we would introduce a predictive process model using $\mathbf{w}_t^*$. We replace the projection onto an arbitrary basis with a projection based on a desired covariance specification.

3. Multivariate predictive process modelling

The multivariate predictive process extends the preceding concepts to multivariate Gaussian processes. A $k \times 1$ multivariate Gaussian process, which is written as $\mathbf{w}(\mathbf{s}) \sim \mathrm{MVGP}_k\{\mathbf{0}, \Gamma_{\mathbf{w}}(\cdot)\}$ with $\mathbf{w}(\mathbf{s}) = [w_l(\mathbf{s})]_{l=1}^{k}$, is specified by its *cross-covariance* function $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}')$ which is defined for any pair of locations as the $k \times k$ matrix $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = \operatorname{cov}\{\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}')\} = [\operatorname{cov}\{w_l(\mathbf{s}), w_m(\mathbf{s}')\}]_{l,m=1}^{k}$. When $\mathbf{s} = \mathbf{s}'$, $\Gamma_{\mathbf{w}}(\mathbf{s}, \mathbf{s})$ is precisely the dispersion matrix for the elements of $\mathbf{w}(\mathbf{s})$. For any integer n and any collection of sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$, we write the multivariate realizations as a $kn \times 1$ vector $\mathbf{w} = [\mathbf{w}(\mathbf{s}_i)]_{i=1}^{n}$ with $\mathbf{w} \sim \mathrm{MVN}(\mathbf{0}, \Sigma_{\mathbf{w}})$, where $\Sigma_{\mathbf{w}} = [\Gamma_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1}^{n}$. A *valid* cross-covariance function ensures that $\Sigma_{\mathbf{w}}$ is positive definite. In practice, cross-covariances will involve spatial parameters $\boldsymbol{\theta}$ and we shall write $\Gamma(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$.

Analogous to the univariate setting, we again consider a set of knots $\mathcal{S}^*$ and denote by $\mathbf{w}^*$ the realizations of $\mathbf{w}(\mathbf{s})$ over $\mathcal{S}^*$. Then the multivariate predictive process is defined as

$$\tilde{\mathbf{w}}(\mathbf{s}) = \text{cov}\{\mathbf{w}(\mathbf{s}), \mathbf{w}^*\}\,\text{var}^{-1}(\mathbf{w}^*)\mathbf{w}^* = \mathcal{C}^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\,\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^*, \tag{6}$$

where $\mathcal{C}^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta}) = (\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}_1^*;\boldsymbol{\theta}),\ldots,\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}_m^*;\boldsymbol{\theta}))$ is $k \times mk$ and $\mathcal{C}^*(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i^*,\mathbf{s}_j^*;\boldsymbol{\theta})]_{i,j=1}^m$ is the $mk \times mk$ dispersion matrix of $\mathbf{w}^*$. Equation (6) shows that $\tilde{\mathbf{w}}(\mathbf{s})$ is a zero-mean multivariate predictive process ($k \times 1$) with cross-covariance matrix given by $\Gamma_{\tilde{\mathbf{w}}}(\mathbf{s},\mathbf{s}') = \mathcal{C}^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\,\mathcal{C}^{*-1}(\boldsymbol{\theta})\,\mathcal{C}(\mathbf{s}';\boldsymbol{\theta})$. $\tilde{\mathbf{w}}(\mathbf{s})$ has properties that are analogous to its univariate counterpart. Its realizations over a finite set with more than m locations have singular joint distributions. A multivariate analogue to equation (4) is immediate, showing that the multivariate predictive process is an interpolator with $\tilde{\mathbf{w}}(\mathbf{s}_j^*) = \mathbf{w}(\mathbf{s}_j^*)$ for any $\mathbf{s}_j^* \in \mathcal{S}^*$. Also, the optimality property in terms of the KL metric carries through.

Specification of the process requires only $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})$ along with $\mathcal{S}^*$. Here, we adopt a spatially adaptive version of the 'linear model of co-regionalization' (Wackernagel, 2006; Gelfand *et al.*, 2004). We model the parent process to be a (possibly) space varying linear transformation $\mathbf{w}(\mathbf{s}) = A(\mathbf{s})\,\mathbf{v}(\mathbf{s})$, where $\mathbf{v}(\mathbf{s}) = [v_i(\mathbf{s})]_{i=1}^k$ and each $v_i(\mathbf{s})$ is an independent Gaussian process with unit variance and correlation function $\rho_i(\mathbf{s},\mathbf{s}')$. Thus, $\mathbf{v}(\mathbf{s}) \sim \text{MVGP}\{\mathbf{0}, \Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}')\}$ has a diagonal cross-covariance $\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}') = \text{diag}\{[\rho_i(\mathbf{s},\mathbf{s}')]_{i=1}^k\}$ and yields a valid non-stationary cross-covariance $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}') = A(\mathbf{s})\,\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}')\,A^{\mathrm{T}}(\mathbf{s}')$ for $\mathbf{w}(\mathbf{s})$. In general, $\mathbf{w}(\mathbf{s})$ is non-stationary even when $\mathbf{v}(\mathbf{s})$ is. When $A(\mathbf{s}) = A$ is constant, $\mathbf{w}(\mathbf{s})$ inherits stationarity from $\mathbf{v}(\mathbf{s})$; $\Gamma_{\mathbf{w}}(\mathbf{s}-\mathbf{s}') = A\Gamma_{\mathbf{v}}(\mathbf{s}-\mathbf{s}')A^{\mathrm{T}}$. Regardless, as in the one-dimensional case, the induced multivariate predictive process is non-stationary.

Since $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s}) = A(\mathbf{s})\,A^{\mathrm{T}}(\mathbf{s})$, without loss of generality we can assume that $A(\mathbf{s}) = \Gamma_{\mathbf{w}}^{1/2}(\mathbf{s},\mathbf{s})$ is a lower triangular square root; the one-to-one correspondence between the elements of $A(\mathbf{s})$ and $\Gamma_{\mathbf{w}}(\mathbf{s},\mathbf{s})$ is well known (see, for example, Harville (1997), page 229). Thus, $A(\mathbf{s})$ determines the association between the elements of $\mathbf{w}(\mathbf{s})$ within $\mathbf{s}$. Choices for modelling $A(\mathbf{s})$ include an inverse spatial–Wishart process for $A(\mathbf{s})\,A^{\mathrm{T}}(\mathbf{s})$ (Gelfand *et al.*, 2004) or elementwise modelling with Gaussian and log-Gaussian processes. When stationarity is assumed (so $A(\mathbf{s}) = A$) we could either assign a prior, e.g. inverse Wishart, to AA^{T} or could further parameterize it in terms of eigenvalues and the Givens angles which are themselves assigned hyperpriors (Daniels and Kass, 1999).

We point out that this approach may accrue further benefits in computing $\mathcal{C}^{*-1}$. Letting $\mathcal{A}^* = \oplus_{i=1}^m A(\mathbf{s}_i^*)$ and $\Sigma_{\mathbf{v}^*} = [\Gamma_{\mathbf{v}}(\mathbf{s}_i^*,\mathbf{s}_j^*)]_{i,j=1}^m$, we have $\mathcal{C}^* = \mathcal{A}^*\Sigma_{\mathbf{v}^*}\mathcal{A}^{*\mathrm{T}}$. Since $\mathcal{A}^{*-1} = \oplus_{i=1}^m A^{-1}(\mathbf{s}_i^*)$, this requires m $k \times k$ triangular inversions. Turning to $\Sigma_{\mathbf{v}^*}$, the diagonal $\Gamma_{\mathbf{v}}$ delivers a sparse structure. Permuting the elements of $\mathbf{v}^*$ to stack the realizations of the $v_i(\mathbf{s})$s over $\mathcal{S}^*$ yields $\Sigma_{\mathbf{v}^*} = P^{\mathrm{T}}\{\oplus_{i=1}^k H_i^*(\boldsymbol{\theta}_i)\}P$, where $H_i^*(\boldsymbol{\theta}_i) = [\rho_i(\mathbf{s}_j^*,\mathbf{s}_{j'}^*;\theta_i)]_{j,j'=1}^m$. Since $P^{-1} = P^{\mathrm{T}}$, the computational complexity resides in the inversion of k $m \times m$ symmetric correlation matrices, $H_i^{*-1}(\theta_i)$, rather than a $km \times km$ matrix.

Now, suppose that 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. They can be combined into a multivariate regression model that is written as

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + \mathbf{Z}^{\mathrm{T}}(\mathbf{s})\,\mathbf{w}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s}), \tag{7}$$

where $\mathbf{X}^{\mathrm{T}}(\mathbf{s})$ is a $q \times p$ matrix ($p = \Sigma_{l=1}^q p_l$) having a block diagonal structure with its lth diagonal being the $1 \times p_l$ vector $\mathbf{x}_l^{\mathrm{T}}(\mathbf{s})$. 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 $\mathbf{x}_l^{\mathrm{T}}(\mathbf{s})$.

The spatial effects $\mathbf{w}(\mathbf{s})$ form a $k \times 1$ coefficient vector of the $q \times k$ design matrix $\mathbf{Z}^{\mathrm{T}}(\mathbf{s})$, where $\mathbf{w}(\mathbf{s}) \sim \mathrm{MVGP}\{\mathbf{0}, \Gamma_{\mathbf{w}}(\cdot,\cdot)\}$. The $q \times 1$ vector $\varepsilon(\mathbf{s})$ follows an $\mathrm{MVN}(\mathbf{0}, \Psi)$ distribution modelling the measurement error effect with dispersion matrix Ψ. Model (7) acts as a general framework admitting several spatial models. For instance, letting $k = q$ and $\mathbf{Z}^{\mathrm{T}}(\mathbf{s}) = I_q$ leads to the multivariate analogue of model (1) where $\mathbf{w}(\mathbf{s})$ acts as a *spatially varying intercept*. However, we could envision all coefficients to be spatially varying and set $k = p$ with $\mathbf{Z}^{\mathrm{T}}(\mathbf{s}) = \mathbf{X}^{\mathrm{T}}(\mathbf{s})$. This yields *multivariate spatially varying coefficients*, which are multivariate analogues of those discussed in Gelfand *et al.* (2003). Banerjee *et al.* (2004), chapter 7, discussed model-based spatial interpolation using model (7) when there are locations where one or several of a multivariate datum are missing. The predictive process versions would simply replace $\mathbf{w}(\mathbf{s})$ with $\tilde{\mathbf{w}}(\mathbf{s})$ in model (7).

4. Bayesian implementation and computational issues

For fitting the predictive process model corresponding to model (7), we form the data equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbb{Z}^{\mathrm{T}}\, \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\, \mathcal{C}^{*-1}(\boldsymbol{\theta})\mathbf{w}^{*} + \varepsilon, \qquad \varepsilon \sim N(\mathbf{0}, I_q \otimes \Psi), \tag{8}$$

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, $\mathbb{Z}^{\mathrm{T}} = \oplus_{i=1}^{n} \mathbf{Z}^{\mathrm{T}}(\mathbf{s}_i)$ is $nq \times nk$, $\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta}) = [\Gamma_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_j^{*}; \boldsymbol{\theta})]_{i,j=1}^{n,m}$ is $nk \times mk$ and $\mathcal{C}^{*}(\boldsymbol{\theta})$ and $\mathbf{w}^{*}$ are as described in Section 3. Given priors, model fitting employs a Gibbs sampler with Metropolis steps using the marginalized likelihood $\mathrm{MVN}\{\mathbf{X}\boldsymbol{\beta}, \mathbb{Z}^{\mathrm{T}}\, \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\, \mathcal{C}^{*-1}(\boldsymbol{\theta})\, \mathcal{C}(\boldsymbol{\theta})\mathbb{Z} + I_n \otimes \Psi\}$, after integrating out $\mathbf{w}^{*}$.

To complete hierarchical specifications, customarily we set $\boldsymbol{\beta} \sim \mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$, whereas Ψ could be assigned an inverse Wishart prior although, more commonly, independence of pure error for the different responses at each site is adopted, yielding a diagonal $\Psi = \mathrm{diag}\{(\tau_i^2)_{i=1}^{q}\}$ with $\tau_i^2 \sim \mathrm{IG}(a_i, b_i)$. Also, $\mathcal{A}^{*}$ is unknown and needs to be stochastically specified. For A constant, $\mathcal{A}^{*} = I_n \otimes A$ and we model AA^{T} with an inverse Wishart prior. Under stationarity, $\Sigma_{\mathbf{v}} = [\Gamma_{\mathbf{v}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^{n}$ and we need to assign priors on $\boldsymbol{\theta} = \{\phi_k, \nu_k\}_{k=1}^{q}$. This will again depend on the choice of correlation functions. In general, the spatial decay parameters are weakly identifiable and prior specifications become somewhat delicate. Reasonably informative priors are needed for satisfactory MCMC behaviour. Priors for the decay parameters are set relatively to the size of D, e.g. prior means that imply the spatial ranges to be a chosen fraction of the maximum distance. For the Matérn correlation function, 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 $p(\Omega|\mathrm{data}) \propto p(\boldsymbol{\beta})\; p(A)\; p(\boldsymbol{\theta})\; p(\mathbf{Y}|\boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi)$, where $\Omega = (\boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi)$. Sampling proceeds by first updating $\boldsymbol{\beta}$ from an $\mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$ distribution with

$$\Sigma_{\boldsymbol{\beta}|\cdot} = \{\Sigma_{\boldsymbol{\beta}}^{-1} + \mathbf{X}^{\mathrm{T}}(\mathbb{Z}^{\mathrm{T}}\, \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\, \mathcal{C}^{*-1}(\boldsymbol{\theta})\, \mathcal{C}(\boldsymbol{\theta})\mathbb{Z} + I_N \otimes \Psi)^{-1}\mathbf{X}\}^{-1}$$

and mean

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot}\mathbf{X}^{\mathrm{T}}(\mathbb{Z}^{\mathrm{T}}\, \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\, \mathcal{C}^{*-1}(\boldsymbol{\theta})\, \mathcal{C}(\boldsymbol{\theta})\mathbb{Z} + I_N \otimes \Psi)^{-1}\mathbf{Y}.$$

The remaining parameters are updated by using Metropolis steps, possibly using block updates (e.g. all the parameters in Ψ in one block and those in A in another). Typically, random-walk Metropolis steps with (multivariate) normal proposals are adopted; since all parameters with positive support are converted to their logarithms, some Jacobian computation is needed. For instance, although we assign an inverted Wishart prior to AA^{T}, in the Metropolis update we update A, which requires transforming the prior by the Jacobian $2^{k}\, \Pi_{i=1}^{k}\, a_{ii}^{k-i+1}$.

This scheme requires the determinant and inverse of $(\mathbb{Z}^{\mathrm{T}}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\mathcal{C}^{*-1}(\boldsymbol{\theta})\mathcal{C}(\boldsymbol{\theta})\mathbb{Z} + I_n \otimes \Psi)$, which is $nq \times nq$. These tasks are accomplished efficiently in terms of only $mk \times mk$ matrices by using Sherman–Woodbury–Morrison-type computations (e.g. Harville (1997)) that evaluate the determinant as $|\Psi|^n|\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\mathbb{Z}(I_n \otimes \Psi^{-1})\mathbb{Z}^{\mathrm{T}}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})|/|\mathcal{C}^*(\boldsymbol{\theta})|$ and the inverse as

$$I_n \otimes \Psi^{-1} - (I_n \otimes \Psi^{-1})\mathbb{Z}^{\mathrm{T}}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\{\mathcal{C}^*(\boldsymbol{\theta}) + \mathcal{C}(\boldsymbol{\theta})\mathbb{Z}(I_n \otimes \Psi^{-1})\mathbb{Z}^{\mathrm{T}}\mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta})\}^{-1}\mathcal{C}(\boldsymbol{\theta})\mathbb{Z}(I_n \otimes \Psi^{-1}).$$

Note that, in updating Ω by using the marginal model, the $\mathbf{w}^*$s are not sampled. However, with first-stage Gaussian models the posterior samples of $\mathbf{w}$ can be recovered by sampling from

$$p(\mathbf{w}^*|\text{data}) \propto \int p(\mathbf{w}^*|\Omega, \text{data})\ p(\Omega|\text{data})\ \mathrm{d}\Omega$$

by using composition since we have posterior samples from $p(\Omega|\text{data})$ and the first distribution under the integral is a multivariate normal distribution. The sampling is one for one with $\Omega^{(l)}$, yielding $\mathbf{w}^{*(l)}$ and hence samples $\tilde{\mathbf{w}}^{(l)} = \mathcal{C}^{\mathrm{T}}(\boldsymbol{\theta}^{(l)})\mathcal{C}^{*-1}(\boldsymbol{\theta}^{(l)})\mathbf{w}^{*(l)}$ as well. For predicting $\mathbf{Y}(\mathbf{s}_0)$ at a new location $\mathbf{s}_0$, the $\mathbf{w}^{*(l)}$s produce $\tilde{\mathbf{w}}^{(l)}(\mathbf{s}_0)$. Then, $\mathbf{Y}^{(l)}(\mathbf{s}_0)$ is sampled by using model (7) with $\Omega^{(l)}$ and $\tilde{\mathbf{w}}^{(l)}(\mathbf{s}_0)$.

The Sherman–Woodbury–Morrison matrix identities have been used in other low rank kriging approaches with marginalized likelihoods, e.g. Cressie and Johannesson (2008). With the likelihood in equation (8) we could avoid marginalizing over the $(m \times 1)$-dimensional vector $\mathbf{w}^*$ and instead update by using its full conditional distribution (multivariate normal when the first-stage likelihood is Gaussian). Of course, generally, the marginalized sampler achieves faster convergence. However, for models whose first-stage likelihood is non-Gaussian (as we discussed in Section 2.5), marginalization is not feasible, we must update the $\mathbf{w}^*$, and the Sherman–Woodbury–Morrison formula plays no role.

5. Illustrations

Our predictive process implementations were written in C++, leveraging processor-optimized BLAS, sparse BLAS and LAPACK routines (`www.netlib.org`) for the required matrix computations. The most demanding model (involving 28 500 spatial effects) took approximately 46 h to deliver its entire inferential output involving 15 000 MCMC iterations on a single 3.06-GHz Intel Xeon processor with 4.0 Gbytes of random-access memory running Debian LINUX. Convergence diagnostics and other posterior summarizations were implemented within the R statistical environment (`http://cran.r-project.us.org`) employing the `CODA` package.

5.1. Simulation studies

We present two simulation examples. The first example (Section 5.1.1) involves 3000 locations and an anisotropic random field where we estimate the parent model itself to offer comparisons with the predictive process models, whereas the second (Section 5.1.2) is a much larger example with 15 000 locations and a more challenging non-stationary spatial structure that precludes estimation of the parent model with the computational specifications above.

5.1.1. Simulation example 1

Here, we simulated the response $Y(\mathbf{s})$ by using model (1) from 3000 irregularly scattered locations over a 1000×1000 domain. In this case we can fit model (1) without resorting to the predictive process; comparison with various choices of m and knot design can be made. The regression

Table 1. Parameter credible intervals, 50% (2.5% 97.5%), and predictive validation for the predictive process models by using a regular grid of knots†

Parameter	*True value*	*Results for the following numbers of knots:*			
		144	*256*	*529*	*3000*
β	1.0	0.94 (0.56, 1.35)	0.73 (0.34, 1.16)	0.77 (0.34, 1.21)	0.72 (0.43, 1.01)
ψ	45.0°	*36.45* (*34.66*, *38.14*)	42.09 (37.62, 45.80)	43.83 (40.93, 46.77)	44.47 (43.18, 45.74)
λ_1	300.0	*390.4* (*330.1*, *399.6*)	279.0 (258.6, 311.0)	323.1 (289.9, 349.0)	302.6 (275.5, 330.2)
λ_2	50.0	*62.42* (*52.7*, *71.99*)	*79.41* (*59.77*, *103.40*)	61.47 (41.50, 84.30)	47.45 (40.03, 55.13)
σ^2	1.0	1.14 (0.87, 1.49)	1.02 (0.80, 1.54)	1.31 (0.83, 1.52)	0.95 (0.87, 1.05)
τ^2	0.2	*0.56* (*0.53*, *0.59*)	*0.45* (*0.42*, *0.49*)	*0.26* (*0.21*, *0.29*)	0.16 (0.13, 0.22)
Prediction	95%	91%	92%	93%	95%

†Entries in italics indicate where the true value is missed. The last column shows results for the parent model.

component included only an intercept, whereas the spatial process $w(\mathbf{s})$ was generated by using a stationary *anisotropic* Matérn covariance function given by

$$C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \{\sigma^2 / 2^{\nu-1}\Gamma(\nu)\}[2\sqrt{\{\nu\, d(\mathbf{s}, \mathbf{s}')\}}]^{\nu}\, \kappa_{\nu}[2\sqrt{\{\nu\, d(\mathbf{s}, \mathbf{s}')\}}],$$

where $d(\mathbf{s}, \mathbf{s}') = (\mathbf{s} - \mathbf{s}')^{\mathrm{T}}\Sigma^{-1}(\mathbf{s} - \mathbf{s}')$. We further parameterize $\Sigma = G(\psi)\Lambda\, G^{\mathrm{T}}(\psi)$ where $G(\psi)$ is a rotation matrix with angle ψ and Λ is the diagonal matrix with positive diagonal elements λ^2. The vector $\boldsymbol{\theta} = (\nu, \psi, \Lambda)$ denotes the spatial parameters: ν controls the smoothness, whereas the rate of spatial decay is controlled by the λ^2s.

Parameter values generating the simulated process are given in the second column in Table 1. Fig. 3(a) clearly shows the dominant 45.0° orientation of the process. We assign a flat prior to the intercept β, a $U(0, \pi/2)$ prior for the rotation parameter ψ and a $U(10, 400)$ prior for the λs. This support on the λs corresponds to about 30–1200 distance units for the effective spatial ranges along those axes (i.e. approximately 3λ is the distance at which the correlation drops to 0.05). The remaining process parameters σ^2 and τ^2 followed IG(2, 1) and IG(2, 0.2) priors respectively. We kept $\nu = 0.5$ as fixed for the analysis in this subsection.

We carried out several simulation experiments with varying knot sizes and configurations. In addition to regular lattices or grids, we also explored two different knot configurations which were described by Diggle and Lophaven (2006) in spatial design contexts. The first, which is called the *lattice plus close pairs* configuration, considers a regular $k \times k$ lattice of knots but then intensifies this grid by randomly choosing m' of these lattice points and then placing an additional knot close to each of them—say within a circle having the lattice point as centre and a radius that is some fraction of the spacing on the lattice. The second configuration, which is called the *lattice plus infill* design, also starts with knots on a regular $k \times k$ lattice but now intensifies the grid by placing a more finely spaced lattice within m' randomly chosen cells of the original lattice.

Here we present illustrations with the above designs with knot sizes of 144, 256 and 529. For the uniform grid these were arranged on a square lattice with knots spaced at 91.0, 66.8 and 45.5 units respectively. For the close pair and infill designs we held the number of knots at 144, 256 and 529 (for a fair comparison with the uniform lattice) and adjusted the lattice accordingly. For instance, Fig. 3(b) shows the close pair design with 256 knots by randomly selecting 60 knots from a 14×14 lattice and then adding a knot to each of them to form close pairs. Similarly,

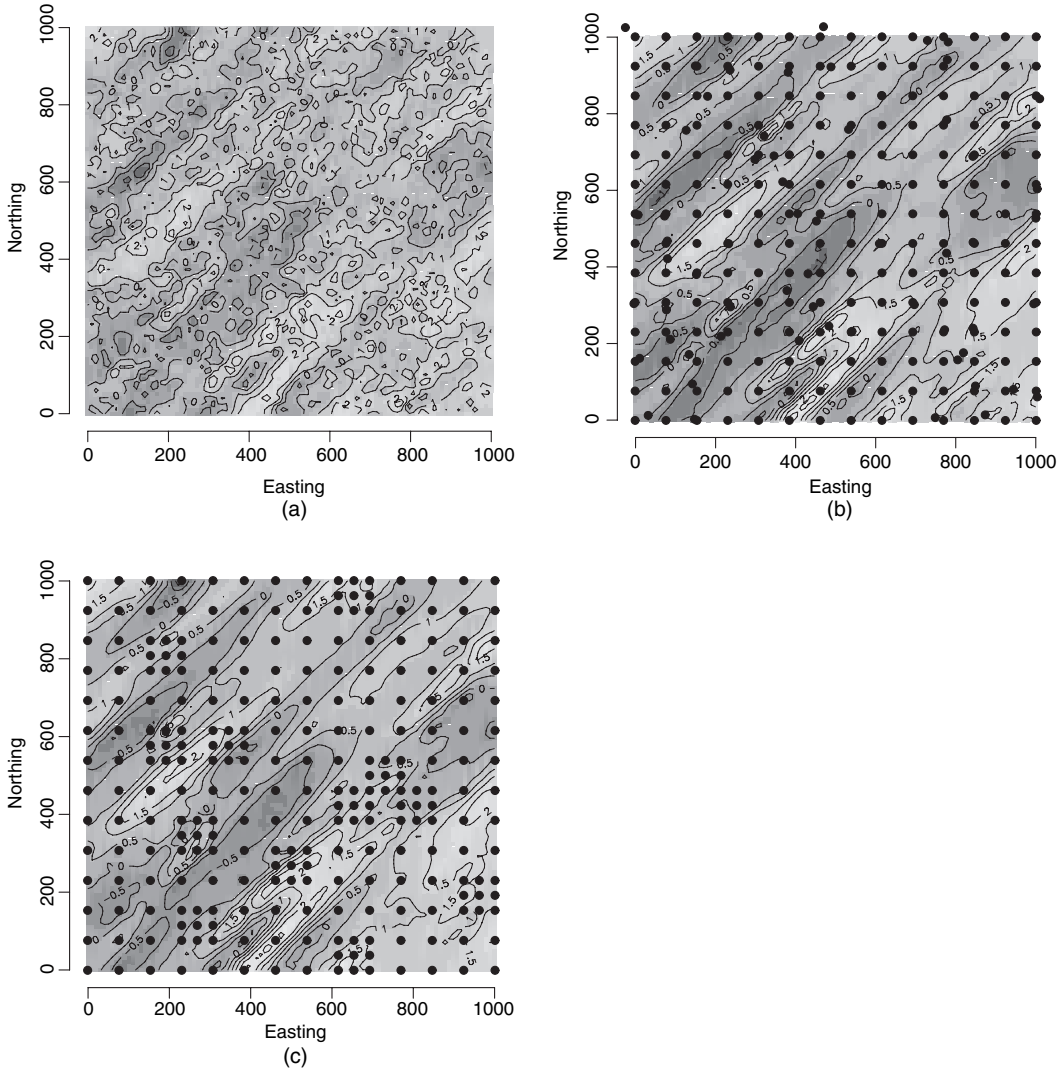

Fig. 3. (a) Simulated stationary anisotropic process generated with 3000 sites by using the parameter values given in Table 1, (b) interpolated (posterior mean) surface for the predictive process model overlaid with 256 knots in the lattice plus close pair configuration and (c) interpolated (posterior mean) surface for the predictive process model overlaid with 256 knots in the lattice plus infill configuration

Fig. 3(c) shows the infill design with 256 knots formed by randomly selecting 12 cells of the original 14×14 lattice and adding a finely spaced sublattice in each of these cells. This results in five additional knots in each of those cells, making the total number of knots $14^2 + 12 \times 5 = 256$.

For each knot configuration, three parallel MCMC chains were run for 5000 iterations each. Convergence diagnostics revealed 1000 iterations to be sufficient for initial burn-in so the remaining 12 000 (4000×3) samples from each chain were used for posterior inference. Table 1 provides parameter estimates for the three knot intensities on a uniform grid along with those from the parent model, i.e. with each of the 3000 locations as a knot, whereas Tables 2 and 3 provide those from the close pair and infill designs respectively. Central processor unit times with the machine specifications that were described earlier were approximately 0.75 h, 1.5 h and 4.25 h for the

Table 2. Parameter credible intervals, 50% (2.5% 97.5%), for the predictive process models with 3000 locations by using the lattice plus close pair design with different knot intensities†

Parameter	*True value*	*Results for the following numbers of knots:*		
		144	*256*	*529*
β	1.0	1.07 (0.77, 1.40)	0.63 (0.26, 1.01)	0.72 (0.35, 1.10)
ψ	45.0°	*40.58* (*38.65*, *42.59*)	44.77 (42.68, 46.74)	43.76 (42.35, 45.98)
λ_1	300.0	*386.62* (*344.01*, *399.69*)	291.29 (267.57, 386.78)	330.00 (295.33, 358.85)
λ_2	50.0	49.24 (43.86, 54.58)	53.40 (46.20, 60.72)	51.08 (43.98, 60.07)
σ^2	1.0	1.34 (1.0, 1.70)	1.42 (0.89, 1.65)	1.39 (0.91, 1.66)
τ^2	0.2	*0.55* (*0.52*, *0.58*)	*0.45* (*0.43*, *0.48*)	*0.24* (*0.22*, *0.29*)
Prediction	95%	91%	92%	92%

†Entries in italics indicate where the true value is missed. The last row provides the empirical coverage of 95% prediction intervals for a set of 100 hold-out locations.

Table 3. Parameter credible intervals, 50% (2.5% 97.5%), for the predictive process models with 3000 locations by using the lattice plus infill design†

Parameter	*True value*	*Results for the following numbers of knots:*		
		144	*256*	*529*
β	1.0	1.18 (0.76, 1.66)	0.77 (0.39, 1.17)	0.71 (0.39, 1.02)
ψ	45.0°	*42.12* (*40.70*, *43.21*)	45.55 (44.46, 46.75)	43.45 (41.89, 45.13)
λ_1	300.0	*392.34* (*343.23*, *399.74*)	316.73 (270.11, 368.57)	345.85 (286.32, 369.79)
λ_2	50.0	58.26 (45.79, 66.46)	56.05 (47.97, 64.27)	47.31 (39.64, 55.38)
σ^2	1.0	1.72 (0.98, 2.66)	1.35 (0.92, 1.76)	1.14 (0.94, 1.37)
τ^2	0.2	*0.57* (*0.54*, *0.60*)	*0.48* (*0.45*, *0.50*)	*0.25* (*0.22*, *0.29*)
Prediction	95%	92%	92%	93%

†Entries in italics indicate where the true value is missed. The last row provides the empirical coverage of 95% prediction intervals for a set of 100 hold-out locations.

144, 256 and 529 knot models respectively, whereas for the parent model it was approximately 18 h.

All the tables reveal the improvements in estimation with increasing number of knots, irrespectively of the design. In all three tables we find substantial overlaps in the credible intervals of the predictive process models with those from the original model. Although 144 knots are adequate for capturing the regression term β, higher knot densities are required for capturing the anisotropic field parameters and the nugget variance τ^2. The nugget variance, in particular, is a difficult parameter to estimate here with much of the variability being dominated by σ^2, yet we see a substantial improvement in moving from 256 knots to 529 knots. Tables 1–3 suggest that estimation is more sensitive to the number of knots than to the underlying design, although the close pair designs appear to improve estimation of the shorter ranges as seen for λ_2 with 256 knots. Predictions, however, are much more robust as is seen from the last row of Tables 1–3. These show the empirical coverage of 95% prediction intervals based on a hold-out set of 100

locations. The coverage, although expected to be lower given that there is less uncertainty in the predictive process than in the parent process (Section 2.3), is only slightly so.

5.1.2. *Simulation example 2*

We now present a more complex illustration with 15000 locations (a fivefold increase from the preceding example) and a more complex non-stationary random field, which renders evaluation of the full model computationally infeasible. We now divide the domain into three subregions and generate $Y(\mathbf{s})$ from these 15000 locations over a 1000×1000 domain by using model (1) and assign a *different* intercept to each of the three regions. Fig. 4(a) shows the domain and

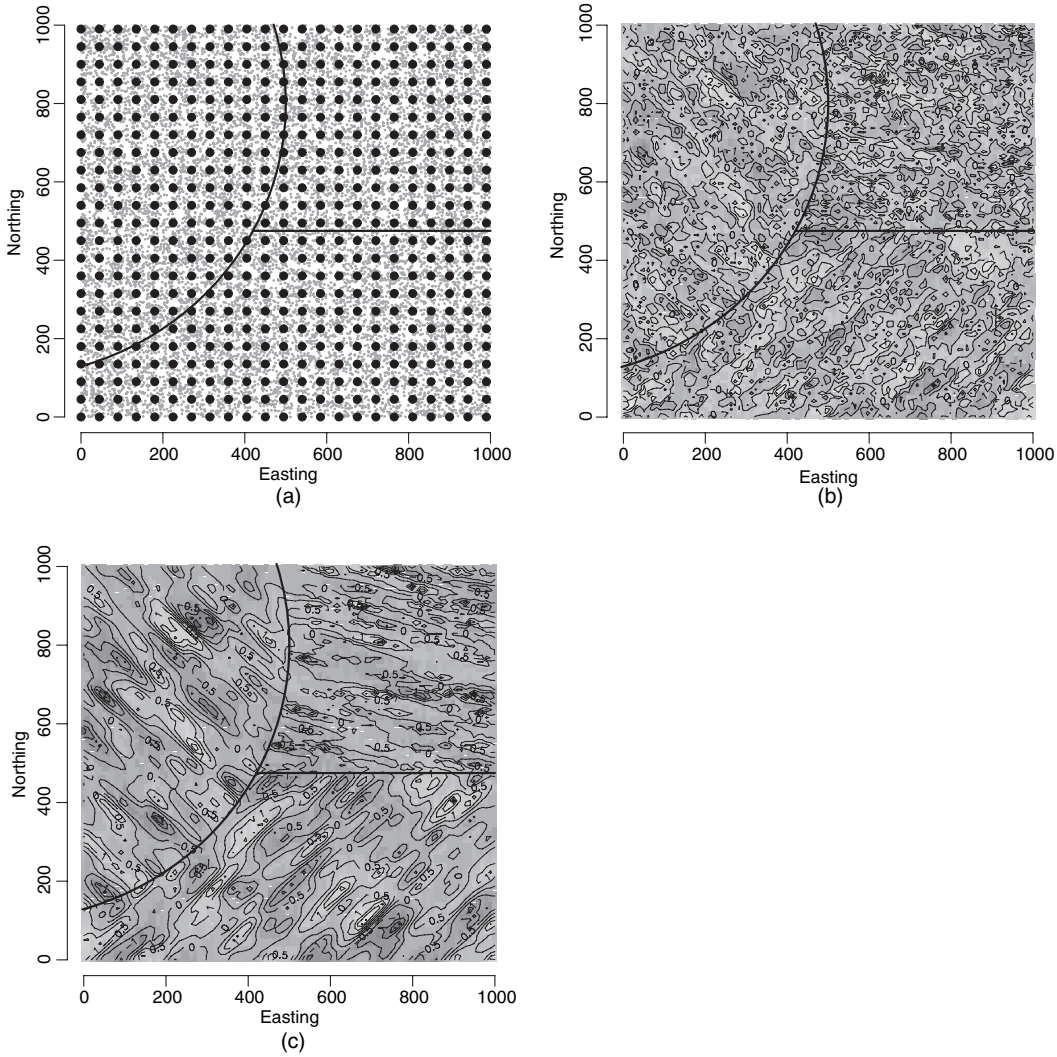

Fig. 4. (a) Three-region domain with 15000 simulated sites (·) beneath the 529 knots (●) that were used to estimate the parent process, (b) interpolated (mean) surface for the ordinary least squares residuals and (c) interpolated (posterior mean) surface for the spatial residuals from the predictive process model with 529 knots

sampling locations along with 529 overlaid knots. We extend the covariance function in the preceding example to a non-stationary version (Paciorek and Schervish, 2006)

$$C(\mathbf{s},\mathbf{s}';\boldsymbol{\theta}) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} |\Sigma_{D(\mathbf{s})}|^{1/4} |\Sigma_{D(\mathbf{s}')}|^{1/4} \left| \frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}')}}{2} \right|^{-1/2} [2\sqrt{\{\nu\, d(\mathbf{s},\mathbf{s}')\}}]^{\nu} \kappa_{\nu}[2\sqrt{\{\nu\, d(\mathbf{s},\mathbf{s}')\}}],$$

where

$$d(\mathbf{s},\mathbf{s}') = (\mathbf{s}-\mathbf{s}')^{\mathrm{T}} \left(\frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}')}}{2} \right)^{-1} (\mathbf{s}-\mathbf{s}')$$

with $\Sigma_{D(\mathbf{s})}$ now varying over space and $D(\mathbf{s})$ indicating the subregion (1, 2 or 3) where $\mathbf{s}$ belongs. We again parameterize $\Sigma_{D(\mathbf{s})} = G(\psi_{D(\mathbf{s})})\Lambda G^{\mathrm{T}}(\psi_{D(\mathbf{s})})$ but now acknowledge the rotation angle to depend on the location. The second column in Table 4 presents the parameter values that were used to generate the data. The region-specific parameters are labelled 1–3 clockwise starting with the rounded region in the north-west.

We assign flat priors to the three intercepts, a $U(0, \pi/2)$ prior for each of the three region-specific rotation parameters and $U(1.7, 300)$ prior for the λs (corresponding to about 5–900 distance units as spatial range). The remaining process parameters ν, σ^2 and τ^2 followed $U(0, 2)$, IG(2, 1) and IG(2, 0.2) priors respectively. We again employed uniform grids as well as the close pair and lattice-plus-infill grids with 144, 256 and 529 knots; central processor unit times here were approximately 6 h, 15 h and 33 h respectively. The inference did not vary significantly for these designs, so we present the results for the uniform grid only. For each model, we ran three initially overdispersed chains for 3000 iterations. Convergence diagnostics revealed that 1000 iterations

Table 4. Parameter credible intervals, 50% (2.5% 97.5%), for three knot intensities with 15000 locations†

Parameter	*True value*	*Results for the following numbers of knots:*		
		144	*256*	*529*
β_1	50.0	50.00 (49.91, 50.07)	49.99 (49.86, 50.12)	49.96 (49.82, 50.10)
θ_1	45.0°	*35.90* (*31.93*, *39.46*)	*31.52* (*28.02*, *44.81*)	50.80 (41.73, 59.23)
$\lambda_{1,1}$	16.69	16.87 (16.51, 17.03)	16.91 (16.62, 17.66)	16.86 (16.67, 17.02)
$\lambda_{1,2}$	66.7	66.70 (66.56, 66.92)	66.65 (66.48, 66.79)	66.63 (66.45, 66.83)
β_2	10.0	*10.05* (*10.00*, *10.08*)	10.0 (9.95, 10.04)	10.05 (9.99, 10.11)
θ_2	75.0°	*70.97* (*67.95*, *73.15*)	*77.59* (*75.62*, *79.03*)	72.07 (70.11, 75.84)
$\lambda_{2,1}$	5.0	*5.87* (*5.19*, *5.94*)	5.53 (4.96, 6.02)	5.10 (4.86, 5.55)
$\lambda_{2,2}$	50.0	49.83 (49.65, 50.40)	49.82 (49.51, 50.29)	50.17 (49.88, 50.54)
β_3	25.0	*24.88* (*24.80*, *24.96*)	*24.84* (*24.73*, *24.95*)	25.04 (24.90, 25.18)
θ_3	45.0°	*57.12* (*54.71*, *59.55*)	*36.70* (*33.51*, *42.59*)	38.41 (33.63, 56.98)
$\lambda_{3,1}$	66.7	66.63 (66.43, 67.23)	66.69 (66.47, 66.85)	66.73 (66.63, 66.97)
$\lambda_{3,2}$	16.69	16.674 (16.53, 16.83)	16.80 (16.65, 17.40)	16.77 (16.55, 17.36)
ν	0.5	*0.26* (*0.25*, *0.37*)	0.43 (0.26, 0.59)	0.56 (0.45, 0.67)
σ^2	1.0	*2.46* (*1.98*, *3.13*)	*1.99* (*1.00*, *3.27*)	1.66 (0.97, 2.09)
τ^2	0.2	*1.03* (*1.01*, *1.06*)	*0.94* (*0.88*, *0.96*)	*0.53* (*0.24*, *0.86*)
Predictive validation	95%	90%	90%	92%

†Entries in italics indicate where the interval misses the true value. The last row provides the empirical coverage of 95% prediction intervals for a set of 100 hold-out locations.

were sufficient for initial burn-in so the remaining 2000 samples from each chain were used for posterior inference.

Table 4 presents the parameter estimates. The overall picture is quite similar to that in Section 5.1.1 with increasing knot density leading to improved estimation, especially for the spatial and nugget variances. Although 144 knots, with their larger separation between knots, provided an inadequate approximation to the underlying non-stationary structure, the models with 256 and 529 knots performed much better, especially the latter. The last row of Table 4 shows the empirical coverage of 95% prediction intervals based on a hold-out set of 100 locations. The coverage, although expected to be lower given that there is less uncertainty in the predictive process than in the parent process (Section 2.3), is only slightly lower. Fig. 4(b) is an image plot of ordinary least squares residuals, and Fig. 4(c) is the spatial residual surface from the predictive process model with 529 knots. These images were constructed by using the same interpolation and contouring algorithm (the MBA package in R). They reveal the smoothing in Fig. 4(c) that is brought about by the predictive process; Fig. 4(c) also makes the region-specific anisotropy more apparent. Indeed, note that regions 1 and 3 have the same rotation angle ($\theta_1 = \theta_2 = 45^\circ$) but reciprocal range parameters (i.e. $\lambda_{1,1} = \lambda_{3,2} = 16.69$ and $\lambda_{1,2} = \lambda_{3,1} = 66.7$), causing the opposite orientations of the contours, whereas in region 2 the shorter spatial range ($\lambda_{2,1}$) yields more concentrated contours along the 75° axis.

5.2. Spatially varying regression example

Spatial modelling of forest biomass and other variables that are 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) and how homogeneous it is across the region. We consider point-referenced biomass (log-transformed) data observed at 9500 locations that were obtained from the US Department of Agriculture Forest Service 'Forest inventory and analysis' programme. Each location yields measurements on biomass from trees in that location and two regressors: the cross-sectional area of all stems above 1.37 m from the ground (basal area) and the number of tree stems (stem density) at that location. Often spatial interpolation of biomass is sought at locations by using either typical values of basal area and stem density measurements or, where available, values from historic data sources. Fig. 5 shows the domain and sampling locations with 144 knots overlaid.

Spatial regression models with only a spatially varying intercept are often found inadequate in explaining biomass. Instead, we opt for spatially varying regression coefficients for the intercept as well as the two regressors (Gelfand *et al.*, 2003). More specifically, we use model (7) with $q = 1$, $k = p = 3$ and $\mathbf{Z}^{\mathrm{T}}(\mathbf{s}) = \mathbf{X}^{\mathrm{T}}(\mathbf{s})$, resulting in $Y(\mathbf{s}) = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\,\tilde{\boldsymbol{\beta}}(\mathbf{s}) + \varepsilon(\mathbf{s})$, where $\tilde{\boldsymbol{\beta}}(\mathbf{s}) = \boldsymbol{\beta} + \mathbf{w}(\mathbf{s})$ are spatially varying regression coefficients. Though we have a univariate response for the model, the modelling for the dependent coefficient surfaces introduces a multivariate spatial Gaussian process $\mathbf{w}(\mathbf{s})$. The power of hierarchical modelling is revealed here; we can learn about this multivariate process without ever seeing any observations from it. We cast this into equation (8) as described in Section 4 with $n = 9$ and $n = 500$, $q = 1$, $k = p = 3$ (yielding 28 500 spatial effects) and $\mathbb{Z} = \oplus_{i=1}^{n} \mathbf{x}^{\mathrm{T}}(\mathbf{s}_i)$. The advantages of such models have been detailed in, for example, Gelfand *et al.* (2003), who recognized the computational infeasibility of these models with large spatial data sets and resorted to separable covariance structures that restrict the same spatial decay to all the coefficients. Predictive process models enable us to move beyond separability and to employ the more general co-regionalization structures that were discussed in Section 3.

The parameters that we estimate are $\Omega = (\boldsymbol{\beta}, A, \boldsymbol{\theta}, \Psi)$, where $\boldsymbol{\beta}$ is 3×1, Ψ is $I_n \otimes \tau^2$ and A is

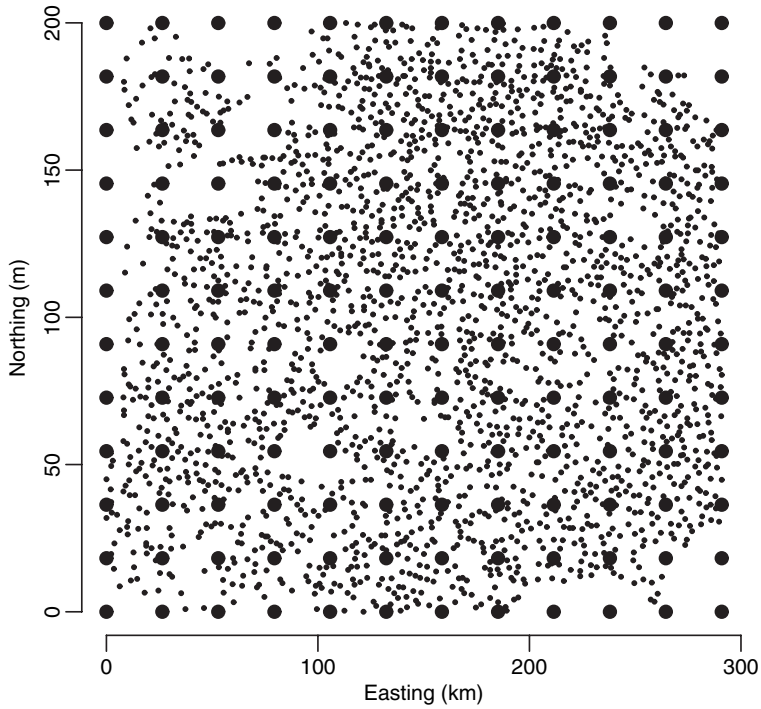

Fig. 5. Spatial distribution of forest inventory sites: the 9500 georeferenced forest inventory sites (•) overlain with 144 knots (●)

a 3×3 lower triangular matrix. The parameters in the $\boldsymbol{\beta}$-vector comprise the model intercept β_0, the basal area β_1 and tree stem density β_2. We assume the Matérn correlation function which adds three ϕs and three νs. A flat prior was assigned to $\boldsymbol{\beta}$. We used the nugget value from an empirical semivariogram calculated by using ordinary least squares residuals to centre the IG(2,0.03) prior that was assigned to the pure error term, τ^2. $\Gamma(\mathbf{0}) = AA^{\mathrm{T}}$ was assigned an IW$\{4, \mathrm{diag}(0.04)\}$ prior, where, again, semivariograms of the response, scaled by the regressors, were used to guide the magnitude of the scale hyperparameters. Following the discussion in Section 4, we assume that $\nu \sim U(0, 2)$ and $\phi \sim U(3 \times 10^{-4}, 0.003)$, which, if $\nu = 0.5$, describes an effective spatial range interval of 1–100 km.

Predictive process models with 64, 144 and 256 knots were each run with three parallel chains for 5000 iterations (central processor unit times 10 h, 18 h and 46 h respectively). They revealed fairly rapid convergence, benefiting from marginalized likelihoods, with a burn-in of 1000 iterations. The remaining 12 000 samples (4000×3) were retained for posterior analysis. (With only 36 knots the distance between adjacent knots (40 km) seemed to exceed the effective spatial ranges that were supported by the data and led to unreliable convergence of process parameters.) Table 5 provides the posterior inference for model parameters corresponding to the various knot densities. The regression in the data is seen to be quite strong with basal area having a significant negative effect and stem density having a significant positive effect on the response. We see a hint of negative association between two of the pairs of coefficient processes and positive association in the other. A pronounced nugget effect (τ^2) is seen. The spatial decay parameters for the various slope parameters are quite similar, indicating, perhaps, that a separable covariance model would be adequate. Some shrinkage is seen in the smoothness parameter. These

Table 5. Inference summary, 50% (2.5%, 97.5%), for spatially varying coefficient models based on three knot densities

Parameter	*Results for the following numbers of knots:*		
	64	*144*	*256*
β_0	−0.117 (−0.168, 0.034)	−0.115 (−0.163, 0.022)	−0.113 (−0.163, 0.027)
β_1	−0.200 (−0.242, −0.029)	−0.200 (−0.243, −0.049)	−0.200 (−0.246, −0.040)
β_2	1.266 (1.213, 1.428)	1.260 (1.211, 1.412)	1.262 (1.204, 1.438)
Γ_{00}	0.006 (0.005, 0.010)	0.005 (0.004, 0.007)	0.005 (0.004, 0.007)
Γ_{11}	0.004 (0.002, 0.019)	0.005 (0.002, 0.017)	0.004 (0.002, 0.016)
Γ_{22}	0.006 (0.004, 0.018)	0.008 (0.006, 0.021)	0.005 (0.004, 0.016)
$\Gamma_{1,0}/\sqrt{(\Gamma_{0,0}\Gamma_{1,1})}$	−0.449 (−0.740, 0.847)	−0.576 (−0.768, 0.880)	−0.732 (−0.783, 0.817)
$\Gamma_{2,0}/\sqrt{(\Gamma_{0,0}\Gamma_{2,2})}$	0.132 (−0.729, 0.964)	0.345 (−0.591, 0.971)	0.673 (−0.014, 0.971)
$\Gamma_{2,1}/\sqrt{(\Gamma_{1,1}\Gamma_{2,2})}$	−0.843 (−0.944, 0.980)	−0.842 (−0.939, 0.972)	−0.849 (−0.946, 0.973)
τ^2	0.041 (0.041, 0.044)	0.040 (0.040, 0.043)	0.041 (0.040, 0.043)
ϕ_{β_0}	8×10^{-5} (7×10^{-5}, 0.00009)	7×10^{-5} (7×10^{-5}, 0.00008)	7×10^{-5} (6×10^{-5}, 0.00010)
ϕ_{β_1}	7×10^{-5} (6×10^{-5}, 0.00011)	7×10^{-5} (7×10^{-5}, 0.00012)	7×10^{-5} (6×10^{-5}, 0.00011)
ϕ_{β_2}	9×10^{-5} (7×10^{-5}, 0.00011)	1×10^{-4} (8×10^{-5}, 0.00013)	9×10^{-5} (7×10^{-5}, 0.00012)
ν_{β_0}	0.426 (0.387, 0.489)	0.471 (0.366, 0.596)	0.391 (0.287, 0.569)
ν_{β_1}	0.437 (0.407, 0.534)	0.408 (0.391, 0.487)	0.383 (0.330, 0.511)
ν_{β_2}	0.471 (0.422, 0.547)	0.443 (0.372, 0.548)	0.432 (0.397, 0.502)
Range_{β_0} (km)	36.405 (30.817, 46.381)	38.373 (31.946, 46.661)	35.238 (28.695, 45.151)
Range_{β_1} (km)	42.139 (25.965, 58.568)	37.811 (22.765, 46.444)	37.931 (26.622, 48.112)
Range_{β_2} (km)	32.825 (26.040, 57.579)	29.241 (19.920, 43.809)	31.311 (22.697, 60.225)

estimates are generally robust across the different knot densities; further increasing the number of knots delivers little gain in estimation.

Fig. 6 shows the image contour plots of the residual coefficient processes (i.e. the $\tilde{w}(\mathbf{s})$s corresponding to each coefficient) as well as that for biomass (using typical values for basal area and stem density). The intercept process seems to be absorbing much of the spatial variation; the two covariate processes are smoother. The predicted biomass surface offers a spatially smoothed version, adjusted for covariates, compared with what is obtained by interpolating the raw response data (which are not shown) and helps to articulate better the zones of higher biomass.

6. Summarizing remarks and future work

We have addressed the problem of fitting desired hierarchical spatial modelling specifications to large data sets. To do so, we propose simply to replace the parent spatial process specification by its induced predictive process specification. One need not digress from the modelling objectives to think about choices of basis functions, or kernels or alignment algorithms for the locations. The resulting class of models essentially falls under the generalized linear mixed model framework (as in equation (8)).

As in existing low rank kriging approaches, knot selection is required and as we demonstrated in Section 5.1 some sensitivity to the number of knots is expected. Although for most applications a reasonable grid of knots should lead to robust inference, with fewer knots the separation between them increases and estimating random fields with fine scale spatial dependence becomes difficult. Indeed, learning about fine scale spatial dependence is always a challenge (see, for example, Cressie (1993), page 114).

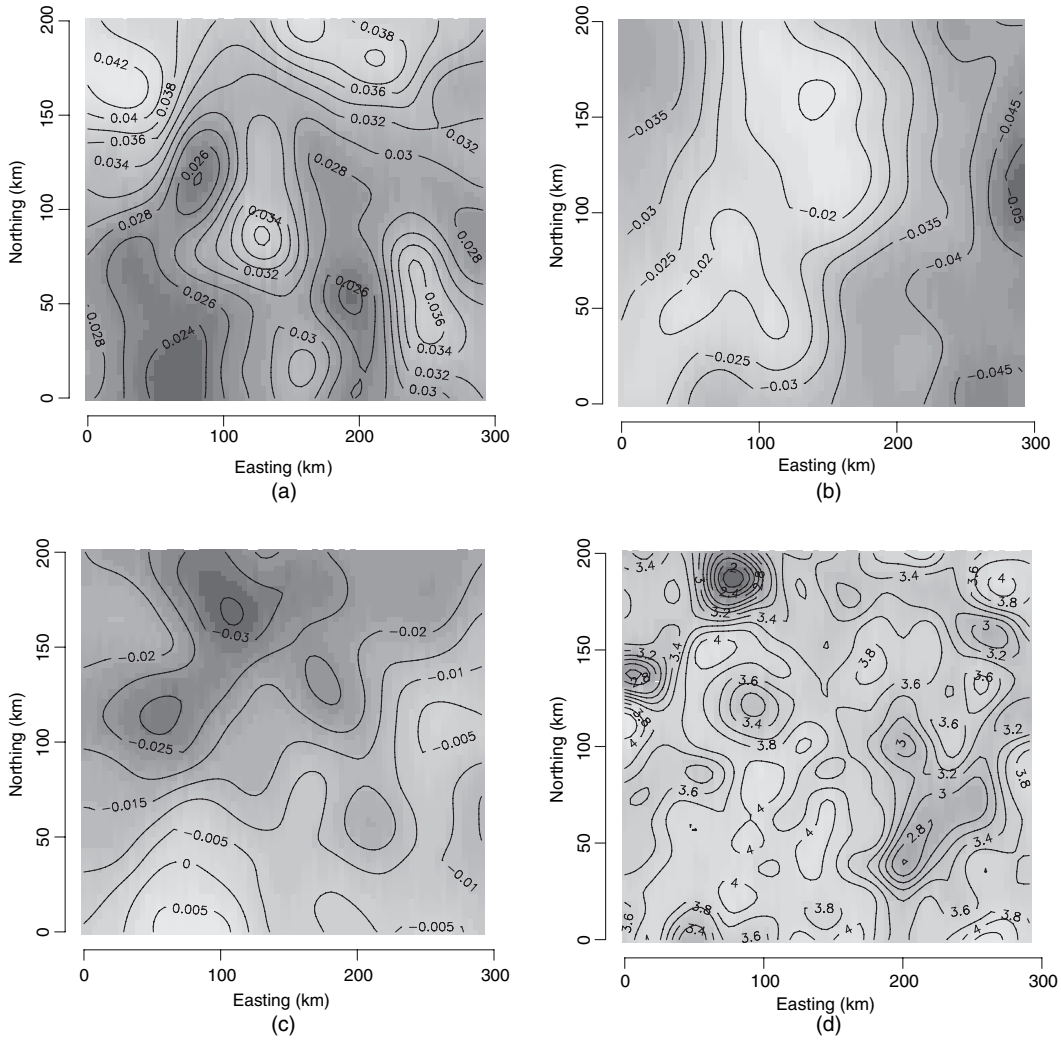

Fig. 6. Posterior (mean) estimates of spatial surfaces from the spatially varying coefficients model: (a) interpolated surface for the space varying intercept parameter; (b) interpolated surface for the space varying basal area parameter; (c) interpolated surface for the space varying stem density parameter; (d) interpolated surface for predicted (log-)biomass

Our examples in Section 5.1 showed that even with fairly complex underlying spatial structures the predictive process model could effectively capture most of the spatial parameters with 529 knots (irrespectively of whether the total number of locations was 3000 or 15000). A further challenge which was noted in our simulated examples was the situation where the variance components ratio ($\sigma^2/\tau^2 = 5.0$) is large so that estimation of τ^2 becomes difficult. One possible remedy is reparameterizing (σ^2, τ^2) in terms of their ratio and the larger variance component (see, for example, Diggle and Ribeiro (2007)). Another option to explore is to modify the predictive process as $\dot{w}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\varepsilon}(\mathbf{s})$, where $\tilde{\varepsilon}(\mathbf{s})$ is an independent Gaussian process with variance $C(\mathbf{s}, \mathbf{s}) - \mathbf{c}^{\mathrm{T}}(\mathbf{s}; \boldsymbol{\theta}) C^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}; \boldsymbol{\theta})$.

Finally, the goal has been dimension reduction to facilitate likelihood evaluation, to facilitate simulation-based model fitting. Although we have employed MCMC methods to fit these

models, faster alternatives that avoid MCMC sampling can also be employed (see, for example, Rue *et al.* (2007)). In fact, the predictive process approach within a full MCMC implementation is perhaps limited to the order of 10^4 observations on modest single-processor machines (see Section 5); strategies that are empirical Bayesian in flavour, combining deterministic and simulation aspects, are likely to be the future for fitting very large space–time data sets. Indeed, it is quite common to find that spatial data sets, especially in scientific studies of large-scale global phenomena, contain far more locations than the illustrations here. For instance, Cressie and Johannesson (2008) worked with data of the order of hundreds of thousands. It is also quite common to find space–time data sets with a very large number of distinct time points, possibly with different time points observed at different locations (e.g. real estate transactions). With multiple processors, substantial gains in computing efficiency can be realized through parallel processing of matrix operations (see, for example, Heroux *et al.* (2006)). We intend to investigate extensively the potential of predictive process models in such settings. More immediately, we intend to migrate our lower level C++ code to the existing `spBayes` (`http://cran.r-project.org`) package in the R environment to facilitate accessibility to predictive process models.

Acknowledgements

The work of the first author was supported in part by National Science Foundation grant DMS-0706870, that of the first and second authors was supported in part by National Institutes of Health grant 1-R01-CA95995, that of the third author was partly supported by National Institutes of Health grant 2-R01-ES07750 and that of the last author was supported in part by National Science Foundation grant DEB05-16198.

Appendix A

Consider the set-up in Section 2.3. Letting $\mathcal{Q}$ be the class of all probability densities satisfying the conditional independence relation $q(\mathbf{Y}|\mathbf{w}_a) = q(\mathbf{Y}|\mathbf{w}^*)$, we want to find the $q \in \mathcal{Q}$ that minimizes the relative entropy or KL divergence. The above conditional independence restriction for q implies that

$$q(\mathbf{w}_a|\mathbf{Y}) = \frac{p(\mathbf{w}^*, \mathbf{w})\, q(\mathbf{Y}|\mathbf{w}^*)}{q(\mathbf{Y})} = p(\mathbf{w}|\mathbf{w}^*)\, q(\mathbf{w}^*|\mathbf{Y}).$$

Using this, we can simplify the KL metric as follows:

$$\begin{aligned}
\mathrm{KL}\{q(\mathbf{w}_a|\mathbf{Y}), p(\mathbf{w}_a|\mathbf{Y})\} &= \int q(\mathbf{w}_a|\mathbf{Y}) \log\left\{\frac{q(\mathbf{w}_a|\mathbf{Y})}{p(\mathbf{w}_a|\mathbf{Y})}\right\} \mathrm{d}\mathbf{w}_a \\
&= \int q(\mathbf{w}_a|\mathbf{Y}) \log\left\{\frac{q(\mathbf{w}^*|\mathbf{Y})}{p(\mathbf{w}^*)p(\mathbf{Y}|\mathbf{w}_a)}\right\} \mathrm{d}\mathbf{w}_a + \log\{p(\mathbf{Y})\} \\
&= \int q(\mathbf{w}^*|\mathbf{Y}) \log\left\{\frac{q(\mathbf{w}^*|\mathbf{Y})}{p(\mathbf{w}^*)}\right\} \mathrm{d}\mathbf{w}^* \\
&\quad - \int q(\mathbf{w}^*|\mathbf{Y}) \left[\int p(\mathbf{w}|\mathbf{w}^*) \log\{p(\mathbf{Y}|\mathbf{w}_a)\}\, \mathrm{d}\mathbf{w}\right] \mathrm{d}\mathbf{w}^* + \log\{p(\mathbf{Y})\} \\
&= \int q(\mathbf{w}^*|\mathbf{Y}) \log\left\{\frac{q(\mathbf{w}^*|\mathbf{Y})}{p(\mathbf{w}^*)\exp(E_{\mathbf{w}|\mathbf{w}^*\sim p}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])}\right\} \mathrm{d}\mathbf{w}^* + \log\{p(\mathbf{Y})\} \\
&= E_{\mathbf{w}^*|\mathbf{Y}\sim q}\left[\log\left\{\frac{q(\mathbf{w}^*|\mathbf{Y})}{p(\mathbf{w}^*)\exp(E_{\mathbf{w}|\mathbf{w}^*\sim p}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])}\right\}\right] + \log\{p(\mathbf{Y})\}.
\end{aligned}$$

Since $\log\{p(\mathbf{Y})\}$ is a constant, minimizing the KL divergence is equivalent to minimizing the first term in this expression. But that expectation must itself be non-negative (in fact it can again be looked on as a

KL divergence up to a proportionality constant) so the numerator and denominator must be proportional. The minimizing $q(\cdot)$ density must satisfy $q(\mathbf{w}^*|\mathbf{Y}) \propto p(\mathbf{w}^*)\exp(E_{\mathbf{w}|\mathbf{w}^*\sim p}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])$, which implies that $q(\mathbf{Y}|\mathbf{w}^*) \propto \exp(E_{\mathbf{w}|\mathbf{w}^*\sim p}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])$. The remainder follows from standard multivariate normal theory by noting that $p(\mathbf{Y}|\mathbf{w}_a)$ is an $\mathrm{MVN}(X\beta+\mathbf{w}, \tau^2 I_n)$ distribution and $p(\mathbf{w}^*, \mathbf{w})$ is $\mathrm{MVN}(\mathbf{0}, \Sigma_{\mathbf{w}_a})$ whence $\exp(E_{\mathbf{w}|\mathbf{w}^*}[\log\{p(\mathbf{Y}|\mathbf{w}_a)\}])$ identifies itself as the desired normal density up to a normalizing constant.

References

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall–CRC.

Cornford, D., Csato, L. and Opper, M. (2005) Sequential, Bayesian geostatistics: a principled method for large datasets. *Geogr. Anal.*, **37**, 183–199.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*, 2nd edn. New York: Wiley.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Csato, L. (2002) Gaussian processes—iterative sparse approximation. *PhD Thesis*. Aston University, Birmingham.

Daniels, M. J. and Kass, R. E. (1999) Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Am. Statist. Ass.*, **94**, 1254–1263.

Diggle, P. and Lophaven, S. (2006) Bayesian geostatistical design. *Scand. J. Statist.*, **33**, 53–64.

Diggle, P. J. and Ribeiro, P. J. (2007) *Model-based Geostatistics*. New York: Springer.

Diggle P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.

Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321–331.

Gelfand, A. E., Banerjee, S. and Gamerman, D. (2005) Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, **16**, 465–479.

Gelfand, A. E., Kim, H., Sirmans, C. F. and Banerjee, S. (2003) Spatial modelling with spatially varying coefficient processes. *J. Am. Statist. Ass.*, **98**, 387–396.

Gelfand, A. E., Schmidt, A., Banerjee, S. and Sirmans, C. F. (2004) Nonstationary multivariate process modelling through spatially varying coregionalization (with discussion). *Test*, **13**, 1–50.

Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Heroux, M. A., Padma, R. and Simon, H. D. (eds) (2006) *Parallel Processing for Scientific Computing*. Philadelphia: Society for Industrial and Applied Mathematics.

Higdon, D. (2001) Space and space time modeling using process convolutions. *Technical Report*. Institute of Statistics and Decision Sciences, Duke University, Durham.

Higdon, D., Swall, J. and Kern, J. (1999) Non-Stationary spatial modeling. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 761–768. Oxford: Oxford University Press.

Jones, R. H. and Zhang, Y. (1997) Models for continuous stationary space-time processes. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions* (eds P. J. Diggle, W. G. Warren and R. D. Wolfinger). New York: Springer.

Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000) Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, **28**, 1570–1600.

Lopes, H. F., Salazar, E. and Gamerman, D. (2006) Spatial dynamic factor analysis. *Technical Report*. Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Møller, J. (ed.) (2003) *Spatial Statistics and Computational Methods*. New York: Springer.

Nychka, D. and Saltzman, N. (1998) Design of air-quality monitoring networks. *Lect. Notes Statist.*, **132**, 51–76.

Paciorek, C. J. (2007) Computational techniques for spatial logistic regression with large datasets. *Computnl Statist. Data Anal.*, **51**, 3631–3653.

Paciorek, C. J. and Schervish, M. J. (2006) Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. New York: Springer.

Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Robert, C. P. and Casella, G. (2005) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.

Rue, H. and Held, L. (2006) *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall–CRC.

Rue, H., Martino, S. and Chopin, N. (2007) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Technical Report*. Norwegian University of Science and Technology, Trondheim.

Rue, H. and Tjelmeland, H. (2002) Fitting Gaussian Markov Random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–49.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schabenberger, O. and Gotway, C. A. (2004) *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman and Hall–CRC.

Seeger, M., Williams, C. K. I. and Lawrence, N. (2003) Fast forward selection to speed up sparse Gaussian process regression. In *Proc. 9th Int. Wrkshp Artificial Intelligence and Statistics* (eds C. M. Bishop and B. J. Frey). KeyWest: Society for Artificial Intelligence and Statistics.

Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory of Kriging*. New York: Springer.

Stein, M. L. (2005) Space-time covariance functions. *J. Am. Statist. Ass.*, **100**, 310–321.

Stein, M. L. (2007) Spatial variation of total column ozone on a global scale. *Ann. Appl. Statist.*, **1**, 191–210.

Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.

Stevens, Jr, D. L. and Olsen, A. R. (2004) Spatially balanced sampling of natural resources. *J. Am. Statist. Ass.*, **99**, 262–278.

Switzer, P. (1989) Non-stationary spatial covariances estimated from monitoring data. In *Geostatistics* (ed. M. Armstrong), vol. 1, pp. 127–138. Dordrecht: Kluwer.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.

Ver Hoef, J. M., Cressie, N. A. C. and Barry, R. P. (2004) Flexible spatial models based on the fast Fourier transform (FFT) for cokriging. *J. Computnl Graph. Statist.*, **13**, 265–282.

Wackernagel, H. (2006) *Multivariate Geostatistics: an Introduction with Applications*, 3rd edn. New York: Springer.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wikle, C. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829.

Xia, G. and Gelfand, A. E. (2006) Stationary process approximation for the analysis of large spatial datasets. *Technical Report*. Institute of Statistics and Decision Sciences, Duke University, Durham.

Xia, G., Miranda, M. L. and Gelfand, A. E. (2006) Approximately optimal spatial design approaches for environmental health data. *Environmetrics*, **17**, 363–385.