Spatial Design for knot selection in knot-based dimension reduction models

Alan E. Gelfand, Sudipto Banerjee and Andrew O. Finley

June 28, 2011

Abstract

1 Introduction

This chapter has a different flavor from others in this volume. Here, the problem is not one of spatial design for data collection but, rather, one of spatial design to facilitate fitting of spatial models. That is, we have a post-data collection, pre-data analysis design problem. More precisely, we address the setting where there is need to specify knots in order to fit desired spatial models. That is, we are unable to directly fit the model we would like and so, we have to adopt an approximate model to make computation feasible. Such a situation arises when we work with very large datasets in space (and in time).

We note that, with the advent of inexpensive high speed computing, statisticians today routinely encounter geographically referenced datasets containing multiple variables observed across thousands of locations, perhaps across thousands of time points. There is, by now, a rich literature in probabilistic/statistical modeling for location-referenced spatial data; see, for example, the books by Cressie (1993), Banerjee, Carlin and Gelfand (2004), and Schabenberger and Gotway (2004), Gelfand et al. (2010) and Cressie and Wikle (2011) for broad discussion and applications. The need to model spatially-referenced outcomes, perhaps vector-valued, across large spatial (and possibly temporal) domains is emerging in many fields including geological and environmental sciences, ecological systems, digital terrain cartography, computer experiments, public health, social and economic systems, etc.

The general challenge of the approximation is to express a surface over a spatial domain through the use of knots, often called low rank or reduced rank representations (Wahba, 1990; Rasmussen and Williams, 2006). Here, the problem takes a critical subdivision. Is the surface viewed as “random”, i.e., a realization of a stochastic process? Are there an uncountable number of variables over the
spatial domain, say $D$? Or, is the surface viewed as “fixed”\(^1\), i.e., a parametric function?

In the former case, we write the surface as \( \{ w(s); s \in D \} \) where \( w(s) = \sum_{j=1}^{m} l(s, s_j^*) Z(s_j^*) \). Thus, the surface/process realization is completely determined by the function \( l(\cdot, \cdot) \) and the set of variables, \( \{ Z(s_j^*), j = 1, 2, ..., m \} \). The collection of \( s_j^* \)'s are the knots and their specification, both number and location, is the design issue of this chapter. Evidently, knot selection depends upon both \( l(\cdot, \cdot) \) and the distribution of the \( Z \)'s.

In the latter case, we tend to think in terms of basis representations for the function, in particular, using splines. We might write the surface as \( f(s) = \sum_{j=1}^{m} b_j g_j(s) \) where the \( b_j \) are a set of coefficients and \( \{ g_j(s), j = 1, 2, ..., m \} \) are a spline basis set (typically, but not necessarily orthonormal). The spline functions are defined with regard to a set of knots which, again, we can label as \( \{ s_j^*, j = 1, 2, ..., m \in D \} \). The literature here is enormous and, in fact, is a subset of functional data analysis which is often applied over domains of dimension higher than 2 and is called multivariate function approximation in the applied math community, nonparametric regression in the statistics realm, and neural nets in the machine learning/computer science world. In this literature, with interest in the function explicitly, typically there is a knot are typically taken at the “location” (covariate level) of each data point (see, e.g., Ramsay and Silverman, 2005). In this case, there is no dimension reduction. Alternatively, knot selection is subsumed within the fitting problem. The fitting is algorithmic, e.g., an algorithm for fitting the curve to the observed data. It is usually done through recursive partitioning as in CART and MARS (Friedman, 1991; Breiman et al., 1984; Berk, 2008; Hastie et al., 2009) or as \( L^p \) (usually \( p = 1, 2 \)) optimization, introducing regularization through a roughness penalty (Gu, 2002; Wahba, 1990). Knot selection can also be achieved using Generalized Cross-Validation (GCV), which is a form of regularization that trades off goodness-of-fit against model complexity using a formula to approximate the error that would be determined by leave-one-out validation (see, e.g., Wahba, 1990; Lin et al. 2000; Ruppert et al., 2003). Bayesian analogues of the above include variants of Bayesian adaptive regression models as discussed in, for example, Lang and Brezger (2004), the book by Denison et al. (2004), and in Chipman et al. (1998, 2010).

Returning to the stochastic process context, we find the situation to be much different if we are interested in model-based inference. While some of the foregoing can be applied (and has been applied - refs), when we think in terms of a stochastic process, we think in terms of a covariance function, say \( C(s, s') \) to measure the covariance between \( w(s) \) and \( w(s') \), which captures the pairwise dependence between the variables in \( D \). If we are going to build an approximation to the process, we need to incorporate this dependence structure into the knot design.

First, let us clarify why we need to approximate the process. It arises primarily due to our desire to fit Bayesian hierarchical models (e.g. Gelman et al. 2004; Banerjee et al., 2004) which\(^1\)

---

\(^1\)Here, we mean fixed in the classical sense, i.e., a parametric form. Of course, in the Bayesian view, if the parametric function is unknown, it is taken to be random.
are recognized as versatile inferential tools for capturing the rich dependence structures underlying spatial data and offering full inference regarding the spatial processes. Hierarchical models implemented through Markov chain Monte Carlo (MCMC) methods have become especially popular for spatial model fitting, given their ability to handle models that would be infeasible otherwise. Under such modeling, covariance matrices arise; determinants and inverses for these matrices are required to work with the likelihoods. When the dimension, say \( n \), of these matrices is large, fitting hierarchical spatial models involves expensive matrix factorizations of the order of \( n^3 \). In a fully Bayesian paradigm, where one seeks formal inference on the “variance” parameters, these matrix computations occur in each iteration of the MCMC. As modern data technologies acquire and exploit massive amounts of data, statisticians analyzing spatially referenced datasets confront settings where the number of geo-referenced locations is so large as to make hierarchical modeling infeasible or impractical. The situation is further exacerbated in multivariate settings with several spatially dependent response variables, where the matrix dimensions increase by a factor of the number of spatially dependent variables being modeled. It is also aggravated when data is collected at frequent time points and spatiotemporal process models are used, again increasing dimension by a factor of the number of time points.

So, we see the value of representations of the spatial process in lower-dimensional subspaces, representations that easily generalize to multivariate and/or spatiotemporal processes. As noted above, these are often referred to as low-rank or reduced-rank spatial models and have been explored in different contexts (Wikle and Cressie, 1999; Lin et al., 2000; Higdon 2002; Kamman and Wand, 2003; Ver Hoef et al. 2004; Xia and Gelfand, 2006; Stein, 2007, 2008; Cressie and Johannesson, 2008; Banerjee et al., 2008; Crainiceanu et al., 2008). Many of these methods are variants of the so-called “subset of regressors” methods used in Gaussian process regressions for large data sets in machine learning (e.g. Rasmussen and Williams, 2006). The idea here is to assume, quite reasonably, that the spatial information available from the entire set of observed locations can be summarized in terms of a smaller, but representative, sets of locations, or “knots”. That is, we still use all of the data but we represent the spatial structure through a dimension reduction. In implementing the reduction, we need to design the knots. This issue forms the balance of the chapter.

The format of the chapter is as follows. Section 2 provides a brief review of different approaches to modelling large point-referenced datasets. Section 3 turns the focus to knot-based dimension reduction using low rank models/approximations and the predictive process in particular. Section 4 discusses some basic ideas for designing knots and connects this objective with existing knowledge on spatial designs. A specific strategy for knot-design for predictive process models is outlined in Section 4.2. Section 5 presents some numerical examples using simulated data as well as forestry data. Finally, we conclude the chapter with some afterthoughts and an eye toward future work.
2 Handling large spatial datasets

Modelling large point-referenced spatial datasets have been receiving increased attention. Here, we provide a brief review of this work. Vecchia (1988) proposed approximating the likelihood with a product of local conditional distributions (in the spirit of Markov random fields (ref)) to obtain maximum-likelihood estimates. Stein et al. (2004) adapt this effectively for restricted maximum likelihood estimation, pointing out the need to include some non-local locations in the approximation in order to better learn about the overall behavior of the parametric covariance function that has been assumed. Another possibility is to approximate the likelihood using spectral representations of the spatial process (Fuentes, 2007). These likelihood approximations yield a joint distribution, but not a process, that facilitates spatial interpolation. However, approximation in the spectral domain, e.g., a periodogram estimate of the spectral density can be subtle. Expertise in tailoring and tuning is required and does not easily adapt to multivariate processes. As a result, concern arises regarding the adequacy of the resultant likelihood approximation. Also, the spectral density approaches seem best suited to stationary covariance functions. Yet another approach considers compactly supported correlation functions (Furrer et al., 2006; Kaufman et al., 2009; Du et al., 2009) which yield sparse correlation structures. More efficient computation using sparse solvers can then be employed for kriging and variance estimation, but the tapered structures may limit modeling flexibility. Also, full likelihood-based inference still requires determinant computations that may be problematic. 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, done by an algorithm, possibly introducing unquantifiable errors in precision.

In recent work Rue, Martino and Chopin (2009) propose a promising INLA (Integrated Nested Laplace Approximation) algorithm as an alternative to MCMC that utilizes the sparser matrix structures to deliver fast and accurate posterior approximations. This uses conditional independence to achieve sparse spatial precision matrices that considerably accelerate computations, but relaxing this assumption would significantly detract from the computational benefits of the INLA and the process needs to be approximated by a Gaussian Markov Random Field (GMRF) (Rue and Held, 2006). Furthermore, the method involves a mode-finding exercise for hyper-parameters that may be problematic when the number of hyperparameters is more than 10. Its effectiveness is unclear for multivariate models with different structured random effects and unknown covariance matrices as hyperparameters. Also, its inference is limited; we do not see the entire posterior for a parameter, the entire predictive distribution when we seek to krig to a new location.

Adapting the above approaches to more complex hierarchical spatial models involving multivari-
ate processes (e.g. Wackernagel, 2003; Gelfand et al, 2004), spatiotemporal processes and spatially varying regressions (Gelfand et al., 2003) and nonstationary covariance structures (Paciorek and Schervish, 2006) is potentially problematic. When retaining richness and flexibility of hierarchical models is of primary interest, knot-based low rank models seem to be the preferred option.

3 Dimension reduction approaches

In Section 3.1 we review some of the properties of a dimension reduction representation. Then, in Section 3.2 we turn to the predictive process, a particular reduced rank approach which we have most considered with regard to knot design.

3.1 Basic properties of low rank models

Following the Introduction, we write the reduced rank specification for \( \{w(s); s \in D\} \) as

\[
\tilde{w}(s) = \sum_{j=1}^{m} l(s, s_j^*) Z(s_j^*). \tag{1}
\]

The surface/process realization is completely determined by the function \( l(\cdot, \cdot) \) and the set of variables, \( \{Z(s_j^*), j = 1, 2, ..., m\} \). The collection of \( s_j^* \)'s are the knots. For a collection of locations, with vector \( \tilde{w} = (\tilde{w}(s_1), \tilde{w}(s_2), ..., \tilde{w}(s_n))^T \), we write

\[
\tilde{w} = Lz^*, \tag{2}
\]

where \( L \) is the \( n \times m \) matrix with \((i,j)\)-th element \( l(s_i, s_j^*) \) and \( z^* \) is the \( m \times 1 \) vector with entries \( Z(s_j^*) \) with \( m < n \).

From (2) we see that, despite there being \( n \) \( \tilde{w} \)'s, we will only have to work with \( m \) \( Z \)'s. Since we anticipate \( m << n \), the dimension reduction is evident and, since we will write the model in terms of the \( Z \)'s (with the \( \tilde{w} \)'s being deterministic from the \( Z \)'s, given \( l(\cdot, \cdot) \)), the associated matrices we work with will be \( m \times m \). Evidently, the \( \tilde{w}(s) \) process spans only an \( m \)-dimensional space; we create an uncountable number of variables through a finite number of variables. When \( n > m \), the joint distribution of \( \tilde{w} \) is singular. However, we do create a valid stochastic process. In particular, the covariance function is

\[
\text{cov}(\tilde{w}(s), \tilde{w}(s')) = l(s)^T \Sigma Z \cdot l(s') \tag{3}
\]

where \( l(s) \) is the \( m \times 1 \) vector with entries \( l(s, s_j^*) \). From (3), we see that, even if \( l(\cdot, \cdot) \) is stationary, i.e., of the form \( l(\cdot - \cdot) \), the induced covariance function is not. Also, if the \( Z \)'s are Gaussian, then \( \tilde{w}(s) \) is a Gaussian process.

How do we view the \( Z \)'s? Are they \( w \)'s from a process of interest, whence the \( \tilde{w} \)'s are an approximation to provide computational tractability? Or are they merely a specification to provide...
a spatial model? Also, are the $s^*_j$ a subset of the observed $s$’s or chosen otherwise? We conclude here with a few more words about the first three questions. The last question takes us back to the design problem.

The most prevalent specification for the $Z$’s is i.i.d. normal with mean 0 and variance $\sigma^2$, i.e., $Z(s)$ is a white noise process, whence (3) simplifies to $\sigma^2 I(s)^T I(s')$. This form appears in Barry and ver Hoef (1996) and in a series of papers by Higdon and collaborators (Higdon et al., 1998; Higdon, 2002), the former calling a “moving average” model, the latter, “kernel convolution.” In particular, a natural choice for $l$ is a kernel function, say $K(s-s')$ which puts more weight on $s'$ near $s$. The kernel would have parameters (which induces a parametric covariance function) and might be spatially varying (Higdon, 2002; Paciorek and Schervish, 2006). The reduced rank form can be viewed as a discretization of a process specification of the form $\tilde{w}(s) = \int_{R^2} K(s-s')Z(s')ds'$ (see Xia and Gelfand, 2006, for discussion regarding this discrete approximation). Gaussian kernels are frequently used though they lead to Gaussian covariance functions which, typically, yield process realizations too smooth to be satisfactory in practice (see Stein, 1999; Paciorek and Schervish, 2006). Moreover, the scope of processes that can be obtained through kernel convolution is limited; for instance the widely used exponential covariance function does not arise from kernel convolution.

A different approach to specification for the $Z$’s is to endow them with a stochastic process model with a selected covariance function. Again, from (3), this will impart a covariance function to the $\tilde{w}$’s. Reversing the perspective, if we have a particular covariance function that we wish for the $\tilde{w}(s)$, what covariance function shall we choose for the $Z$’s? We argue in the next subsection that the predictive process provides an optimal choice in some sense. Moreover, with regard to knot selection, we have most experience with the predictive process.

### 3.2 Predictive process models: A brief review

Detailed descriptions of hierarchical Gaussian predictive process models are given in Banerjee et al. (2008), Finley et al. (2009), and Eidsvik et al. (2010). Here, we offer a brief review.

For an $n \times 1$ vector of observed outcomes, $\mathbf{y} = (y(s_1), y(s_2), \ldots, y(s_n))^T$ with a first stage conditionally independent Gaussian specification and associated priors, we consider the Bayesian hierarchical model

$$p(\theta, \beta, \mathbf{w} \mid \mathbf{y}) \propto p(\theta) \times N(\beta \mid \mathbf{\mu}_\beta, \Sigma_\beta) \times N(\mathbf{w} \mid \mathbf{0}, C(\theta_1)) \times \prod_{i=1}^n N(y(s_i) \mid \mathbf{x}(s_i)^T \beta + w(s_i), \tau^2),$$

where $\theta = \{\theta_1, \tau^2\}$. The parameter $\tau^2$ is called the nugget and captures unstructured noise that may arise in the form of measurement error or micro-scale variability. Customarily, either a flat or a multivariate Gaussian prior is assigned to $\beta$. Zhang (2004) demonstrated, rather remarkably, that the process parameters $\theta_1$ were not consistently (in the classical sense) estimable for a rather general
class of covariance functions. This has consequence in the Bayesian framework since it implies that
the impact of the prior on inference is not obliterated with increasing sample size. Hence, informative
priors will be needed for some parameters in order to identify all of the process unknowns.

Fitting (4) customarily proceeds using Markov chain Monte Carlo (MCMC) methods (e.g. Robert
and Casella, 2004). With Gaussian likelihoods, we can integrate out the spatial effects $w$. This
marginalization results in $N \left( y \mid X\beta, C(\theta_1) + \tau^2 I_n \right)$. In any case, as noted above, MCMC fitting
involves $n \times n$ matrix decompositions of cubic order in the number of locations, which become
exorbitant for large $n$. Evidently, multivariate and spatial-temporal settings aggravate the situation.

Reduced-rank spatial process models use a fixed set of “knots” $S^* = \{s_1^*, \ldots, s_m^*\}$ with $m << n$, which, as suggested above, need not, be a subset of $S$. An optimal projection of the process
$w(s)$ at location $s$, based upon its realization over $S^*$, is given by the “kriging equation” $\hat{w}(s) = E\{w(s) \mid w^*\}$, where $w^* = (w(s_1^*), w(s_2^*), \ldots, w(s_m^*))^T$. We refer to $\hat{w}(s)$ as the \textit{predictive process}
derived from the \textit{parent process} $w(s)$.

For a zero-centered Gaussian process with covariance function $C(s_1, s_2; \theta_1)$ as the parent, the
predictive process is $\hat{w}(s) = E\{w(s) \mid w^*\} = c(s, \theta_1)^T C^*(\theta_1)^{-1} w^*$, where $c(s; \theta_1)^T$ is the $1 \times m$
vector whose $j$-th element is $C(s, s_j^*; \theta_1)$ and $C^*(\theta_1)$ is the $m \times m$ covariance matrix with elements
$C(s_j^*, s_j^*; \theta_1)$. The product $c(s, \theta_1)^T C^*(\theta_1)^{-1}$ provides the $l(s, s_j^*)$ entries that appear in the general
expression (xxxx) for the reduced rank for in Section 2.1. Since $w^*$ is multivariate normal with zero
mean and $m \times m$ dispersion matrix $C^*(\theta_1)$, $\hat{w}(s)$ is itself a nonstationary Gaussian process arising
from a spatially adaptive linear transformation of the parent process over the set of knots. In the
spirit of Section 2.1, replacing $w(s)$ with $\hat{w}(s)$ in (4), yields its predictive process counterpart
\begin{equation}
p(\theta, \beta, w^* \mid y) \propto p(\theta) \times N(\beta \mid \mu_\beta, \Sigma_\beta) \times N(w^* \mid 0, C^*(\theta_1)) \times \prod_{i=1}^n N(y(s_i) \mid x(s_i)^T \beta + \hat{w}(s_i), \tau^2).
\end{equation}
Computational gains are achieved since matrix computations now involve the $m \times m$ matrix $C^*(\theta_1)$,
where $m$ is chosen to be much smaller than $n$. Unlike some other knot-based approaches, the
predictive process does not introduce additional parameters nor does it involve projecting data onto
a grid. Thus, it avoids identifiability issues or spurious decrease in uncertainty (see, e.g., Banerjee
et al. 2008). Indeed, predictive process models are attractive since they are directly induced by
the parent process (adopting its associated covariance function) without requiring choices of basis
functions or kernels or alignment algorithms for the locations.

4 Some basic knot design ideas

In Section 4.1 we offer a short review of potential knot design approaches that we could use,
settling upon an approach of Diggle and Lophaven (2006) for illustration. we elaborate this approach
and the associated optimality criterion in Section 4.2.

4.1 A brief review of spatial design

Selection of knots is a challenging problem, evidently more difficult in two dimensions 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/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 but 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 design problem, with the difference being that we already have samples at \( n \) locations.

The question to ask is whether and how the fact that we have already collected observations at \( (s_1, s_2, ..., s_n) \) would affect the choice of knots for the MCMC model fitting. We depart from the well-studied pre-posterior analysis (Bayesian design) literature. Cost also takes on a different role. Customarily, it is the cost associated with data collection, e.g., erecting a monitoring site, maintaining a monitoring site, collecting the data from that site. For us it is computational cost in terms of run times for MCMC model fitting. Also, from a Bayesian (and a likelihood perspective, as well), we can not revise our model to include a new prior which is the updated posterior resulting from the data collection and then embark upon knot selection. We need to select the knots in order to fit the model.

There is a rich literature in spatial and spatio-temporal design, i.e., design for data with structured dependence. Indeed, this volume provides the current state of the art, as well as full referencing. In our context, we only explore a very small portion of this work. One approach would be a 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. A variation on this would be a probability based approach which includes widely-used simple random sampling without replacement. In our case, this would amount to placing a uniform distribution over the region \( D \) and choosing \( m \) points at random from this distribution. Either of these strategies will tend to be robust in that they make no population assumptions regarding, for example, mean structure or dependence structure.

Opportunities for more interesting design arise when the problem is cast as one of optimal spatial sampling design, assuming a particular spatial model. Model-based design has followed a regression model path or a random process model path. Under regression modeling with independent data, optimality is defined with regard to efficiency of the estimates of the regression coefficients. An optimization criterion that is a function of the design matrix is specified and then the “best” design optimizes this criterion over all design matrices. See Pukelsheim (1993) or Müller (2001) for details.
This theory is not directly extensible to spatial design but approximately optimal solutions based upon information-theoretic measures have emerged, most notably the recursion in Brimkulov, Krug, and Savanov (1986). (See Fedorov, 1996, in this regard). Xia et al. (2006) consider algorithms such as sequential selection, block selection and stochastic search. Recent work by Zhu and Stein (2005) focuses on designs based upon optimization using the likelihood. They suggest working with the Fisher information as a measure in the form of a ratio of determinants and implement the optimization using a simulated annealing algorithm.

Model-based design, motivated by a random process specification, has been promoted in Le and Zidek (1992) and Zidek, Sun and Le (2000) as well as references therein. The proposal is an entropy-based design where the selection of the next site to be added will be the one with the largest entropy where entropy can be viewed as uncertainty. Under a Gaussian process assumption, the criterion that emerges is the conditional variance of an observation at a new location given the locations already selected. (This conditional variance depends only upon the previously selected locations but not on the data already collected at those locations.) The site with the most uncertainty is the one with the largest conditional variance given the selected sites. Extension to multivariate data at a location converts the criterion to a conditional covariance matrix. This approach has no interest in mean structure. In fact, quoting Zidek, Sun and Le (2000, p.66), “[I]t avoids the need to specify objectives like parameter estimation.”

In the multiparameter case (almost certainly the case of interest in applications), both the information and entropy criteria emerge as matrices. So, to achieve a single number summary for a design, one has to summarize the resulting matrix either through a determinant or a (possibly weighted) trace.

Recently, Diggle and Lophaven (2006) discuss spatial designs suggesting modification to design on regular grids. They explore 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. These designs augment the lattice with close pairs or infill. We examine such designs for knot selection in our simulation example.

4.2 A strategy for selecting knots

For a given set of observations, Finley et al. (2009) proposed a knot selection strategy designed to improve the induced predictive process as an approximation to the parent process. For a selected set of knots, \( \tilde{w}(s) = E[w(s) | w^*] \) is considered as an approximation to the parent process. Given \( \theta_1 \), the associated predictive variance of \( w(s) \) conditional on the predictive process \( w^* \) on \( S^* \) is

\[
V_{\theta_1}(s, S^*) = \text{var}[w(s) | w(\cdot), S^*, \theta_1] = C(s, s; \theta_1) - c(s, \theta_1)^T C^{*-1}(\theta_1) c(s, \theta_1),
\]
which measures how well we approximate \( w(s) \) by the predictive process \( \hat{w}(s) \). This measure is in the spirit of the foregoing work of Zidek and colleagues (Le and Zidek, 1992; Zidek et al. 2000), i.e., measuring knot value in terms of conditional variance. Then, the best knot-selection maximizes conditional variance given the selected knots. Here, we measure the effectiveness of a collection of selected knots through small conditional variance.

In particular, the knot selection criterion is then defined as a function of \( V_{\theta_i}(s, S^*) \). One commonly used criterion is:

\[
V_{\theta_i}(S^*) = \int_D V_{\theta_i}(s, S^*) g(s) ds = \int_D \text{var}[w(s) | w^*, \theta_1] g(s) ds
\]

where \( g(s) \), integrable over \( D \) is the weight assigned to location \( s \) (Zidek et al., 2000; Diggle and Lophaven, 2006). Here, we only consider the simple case for which \( g(s) \equiv 1 \). \( V_{\theta_i}(S^*) \) can be regarded as a spatially averaged predictive variance. The integral in (7) is analytically intractable and discrete approximations such as numerical quadrature or Monte Carlo integration will be required. We compute the spatially averaged prediction variance over all the observed locations,

\[
V_{\theta_i}(S^*) \approx \frac{1}{n} \sum_{i=1}^{n} \text{var}[w(s_i) | w^*, \theta_1]
\]

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

The following facts are easily verified for \( V_{\theta_i}(S^*) \) defined in (7):

- \( V_{\theta_i}(\{S^*, s_0\}) - V_{\theta_i}(S^*) < 0 \) for a new site \( s_0 \);
- \( V_{\theta_i}(\{S^*, s_0\}) - V_{\theta_i}(S^*) \to 0 \) when \( \|s_0 - s_i^*\| \to 0 \), where \( s_i^* \) is any member of \( S^* \), and
- \( V_{\theta_i}(S) = 0 \), where \( S = \{s_1, \ldots, s_n\} \) are the original observed locations.

The variance-covariance matrix under the parent process model is \( \Sigma_Y = C(\theta_1) + \tau^2 I \), while that from the corresponding predictive process is given by \( \tilde{\Sigma}_Y = C(\theta_1)^T C^{* - 1}(\theta_1) C(\theta_1) + \tau^2 I \). The Frobenius norm between \( \Sigma_Y \) and \( \tilde{\Sigma}_Y \) is \( \|\Sigma_Y - \tilde{\Sigma}_Y\|_F \equiv \text{tr}((C(\theta_1) - C(\theta_1)^T C^{* - 1}(\theta_1) C(\theta_1))^2) \). Since \( C(\theta_1) - C(\theta_1)^T C^{* - 1}(\theta_1) C(\theta_1) \) is positive definite, the Frobenius norm \( \|\Sigma_Y - \tilde{\Sigma}_Y\|_F \equiv \sum \lambda_i^2 \), where \( \lambda_i \) is the \( i \)-th eigenvalue of \( \Sigma_Y - \tilde{\Sigma}_Y \). Also, the averaged predictive variance is given by

\[
\bar{V} = \frac{1}{n} \text{tr}(\Sigma_Y - \tilde{\Sigma}_Y) = \frac{1}{n} \sum \lambda_i.
\]

Note that, even after discretization, we can not evaluate \( V_{\theta_i}(S^*) \) since it depends upon the unknown \( \theta_1 \). Available options to accommodate this include obtaining parameter estimates by using a subset of original data or more fully Bayesian strategies that place a prior on \( \theta_1 \) and then minimizes \( E_{\theta_1}(V_{\theta_i}(S^*)) \) (see, Diggle and Lophaven, 2006). In fact, we might naturally use the same prior as we would use to fit the model. Regardless of which of these strategies we adopt, how shall
we proceed to find a good $S^*$? Suppose the values of the parameters and the knot size $m$ are given. The following sequential search algorithm finds an approximately optimal design:

- **Initialization:** As in all cases where the domain is continuous, for implementation of an optimal design, we need to reduce the possible sampling locations to a finite set. Natural choices include a fine grid, the observed set of locations or the union of these two sets.

- **Specify an initial set of locations of size $m_0$ as starting points for knot selection; possible choices include a coarse grid, or a subset of the observed locations, chosen randomly or deterministically.**

- **At step $t + 1$,**
  - For each sample point $s_i$ in the allowable sample set, evaluate $V\{S^{(t)}, s_i\}$.
  - Remove the sample point with maximum decrease in $\bar{V}^*$ from the allowable sample set and add it to the knot set.

- **Repeat the above procedure until we obtain $m$ knots.**

The sequential evaluation of $\bar{V}$ is achieved using a very efficient routine incorporating block matrix computation. Using this, we have successfully implemented the sequential algorithm in a simulation study shown in Section 5.2. We remark that the sequential algorithm does not necessarily achieve the global optimization solution. Alternative computational approaches are available to us in finding approximately optimal designs such as stochastic search and block selection (see Xia et al., 2006).

As to the choice of $m$, the obvious answer is “as large as possible.” Evidently, this is governed by computational cost and sensitivity to choice. So, for the former, we will have to implement the analysis over different choices of $m$ to consider run time. For the latter, we look for stability of predictive inference as $m$ increases. We measure this by the value of minimized $\bar{V}$ under different choices of $m$. Unlike more formal sampling design contexts, our goal here is to achieve “good” knot selection to enable model fitting. We find coarse progression of $m$ to be adequate. 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.

## 5 Illustrations

### 5.1 A simulation example

A basic illustrative model employed in Banerjee et al. (2008) simulated the response $Y(s)$ using

$$Y(s) = \beta_0 + w(s) + \epsilon(s),$$

(9)
where $\beta_0$ is an intercept, while the spatial process $w(s)$ was generated using a stationary anisotropic Matérn covariance function given by

$$C(s_1, s_2; \theta) = \left( \sigma^2 / \Gamma(\nu) 2^{\nu-1} \right) \left( 2\sqrt{d(s_1, s_2)} \right)^\nu \kappa_\nu \left( 2\sqrt{d(s_1, s_2)} \right) \nu^{\nu-1},$$

where $d(s_1, s_2) = (s_1 - s_2)^T \Sigma^{-1} (s_1 - s_2)$. We further parameterize $\Sigma = G(\psi) \Lambda G(\psi)^T$ where $G(\psi)$ is a rotation matrix with angle $\psi$ and $\Lambda$ is a diagonal matrix with positive diagonal elements, say $\lambda$'s. The vector $\theta = (\nu, \psi, \Lambda)$ denotes the spatial parameters; $\nu$ controls the smoothness, while the rate of spatial decay is controlled by the $\lambda$'s.

We generated $Y(s_i)$'s using 3000 irregularly scattered locations over a 1000 x 1000 domain. In this case we can fit (9) without resorting to the predictive process; comparison with various choices of $m$ along with knot design can be made. Parameter values generating the simulated process are given in the second column in Table 1. The left panel in Figure 1 clearly shows the dominant 45.0° orientation of the process. We assign a flat prior to the intercept $\beta$, a $U(0, \pi/2)$ prior for the rotation parameter $\psi$, and $U(10, 400)$ prior for the $\lambda$'s. This support for the $\lambda$'s corresponds to about 30 to 1200 distance units for the effective spatial ranges along those axes (i.e., approximately 3$\lambda$ is the distance at which the correlation drops to 0.05). The remaining process parameters $\sigma^2$ and $\tau^2$ 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 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 center and a radius that is some fraction of the spacing on the lattice. The second configuration, called the lattice plus in-fill 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 using 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, the top-right panel in Figure 1 shows the close-pair design with 256 knots by randomly selecting 60 knots from a $14 \times 14$ lattice and then adding a knot to each of them to form close pairs. Similarly, the lower panel in Figure 1 shows the infill design with 256 knots formed by randomly selecting 12 cells of the original $14 \times 14$ lattice and adding a finely spaced sublattice in each of these cells. This results in 5 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 5,000 iterations each. Convergence diagnostics revealed 1,000 iterations to be sufficient for initial burn-in so the remaining 12,000 (4,000 \times 3) samples 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, while Tables 2 and 3 provide those from the close-pair and infill designs respectively. CPU times with the machine specifications described earlier were approximately 0.75 hours, 1.5 hours and 4.25 hours for the 144, 256 and 529 knot models respectively, while for the parent model it was approximately 18 hours.

All the tables reveal the improvements in estimation with increasing number of knots, irrespective 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. While 144 knots is adequate for capturing the regression term (\(\beta\)), higher knot densities are required for capturing the anisotropic field parameters and the nugget variance (\(\tau^2\)). The latter, in particular, is a difficult parameter to estimate here with much of the variability being dominated by \(\sigma^2\), yet we see a substantial improvement in moving from 256 knots to 529 knots. These tables 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 \(\lambda_2\) with 256 knots. Predictions, on the other hand, are much more robust as is seen from the last row of Tables 1, 2 and 3. These show the empirical coverage of 95% prediction intervals based upon a hold-out set of 100 locations. The coverage is expected to be lower since there is less uncertainty in the predictive process than in the parent process (Section 2.3); however, it is only slightly so.

5.2 A simulation example using the two step analysis

We generated 1,000 data points in a \([0, 100] \times [0, 100]\) square and then generate the dependent variable from model (9), but using an exponential covariance function with range parameter \(\phi = 0.06\) (i.e., an effective range of \(\sim 50\) units), scale \(\sigma = 1\) for the spatial process \(w(s)\), and with nugget variance \(\tau^2 = 1\). We illustrate a comparison among three design strategies: regular grids, sequential search over all the observed locations, and sequential search over a fine regular lattice. In Figure 2, we plot the averaged predictive variances under each strategy. The sequential search algorithm is clearly better than choosing a regular grid as knots. For instance, with 180 sites selected, the sequential search over the observed locations yielded an averaged predictive variance of approximately 0.15. For the regular grids, roughly 150 additional sites are needed to achieve the same level of performance.

5.3 Tree height and diameter data analysis

Tree diameter at breast height (DBH; 1.37m above the forest floor) and total height (TH) are used, in combination with other information, to assess a tree’s economic and ecological value. Measuring a tree’s TH is very time consuming compared to measuring its DBH. Therefore TH is
often measured on a small subset of those trees that receive DBH measurements. Then a statistical
model that relates TH to DBH is used to predict TH for the complement of this subset. Here, we
illustrate how the proposed methods can help improve the predictive performance of such models
for a moderately large dataset.

The relationship between DBH and TH is influenced by many individual and environmental
factors. Individual factors include age, species, and genetics, whereas, environmental factors include
quality of soil, quantity of water and light, and competition for these resources. These factors often
vary spatially across the domain. For example, a cohort of trees of the same species, age, and
parentage will likely depict similar DBH and TH relationship. An example of an environmental
factor influencing tree growth characteristics is soil productivity which can result in parent material
or disturbance history. Given these unobserved covariates, we can expect some level of spatial
dependence in TH even after accounting for DBH.

The analysis was based on 2,391 mature trees measured on a 200 × 200 m portion of an uneven-
aged softwood stand located near Sault Ste. Marie, Ontario (Ek, 1969) illustrated in Figure 3(a).
The major tree species are balsam fir (Abies balsamea) and black spruce (Picea mariana). Location,
DBH, and TH were recorded for each tree. Here, the candidate models use an intercept and single
covariate log(DBH) to explain the variability in the outcome variable log(TH). Candidate models
include a simple non-spatial regression and predictive process models.

Knot locations were chosen following a: 1) geometric space-filling design, referred to as a “cov-
erage design” (Johnson et al., 1990; Nychka and Saltzman, 1998), computed using the spatial.design
function in the fields R package; 2) the approximately optimal design computed using the algorithm
detained in Section 4.2. In the approximately optimal design algorithm, the initial knot set, \( m_0 \),
consisted of a single point in the middle of the domain and the 2,391 observed locations served as
the possible sampling locations set. Similar to the simulation example, we considered a range of knot
intensities.

Based on results from an initial variogram analysis of the non-spatial model’s residuals (Fig-
ure 3(b)) the priors for \( \tau^2 \) and \( \sigma^2 \) followed IG(2 0.01). Assuming an Exponential spatial correlation
function the prior for the spatial decay parameter \( \phi \) followed a \( U(0.012 \ 1) \), which corresponds to
support from 1–275 m. Again, this is a broad range of support given the maximum distance between
any two trees is 276 m. Note, Figure 3(b) suggests an effective spatial range (i.e., the range at which
the spatial correlation drops to 0.05) of ~50. For all models the regression coefficients, \( \beta_0 \) and \( \beta_{DBH} \),
associated with the intercept and covariate log(DBH) each received flat priors.

Candidate models were assessed based on their fit to observed data as well as predictive perfor-
ance. Model fit to the observed data was assessed using independent replicates for each observed
outcome: for each \( s_i \in S \), we draw \( y_{rep}(s_i)^{(l)} \) from \( N \left( x(s_i)^T \beta^{(l)} + \tilde{w}(s_i)^{(l)}, \tau^{2(l)} \right) \), one for one for

14
the posterior samples. Letting $\mu_{\text{rep},i}$ and $\sigma^2_{\text{rep},i}$ be the posterior predictive mean and variance for each $y_{\text{rep}}(s_i)$, we compute $G = \sum_{i=1}^{n} (y(s_i) - \mu_{\text{rep},i})^2$ and $P = \sum_{i=1}^{n} \sigma^2_{\text{rep},i}$. We use $D = G + P$ (e.g., Gelfand and Ghosh, 1998) as a model selection criteria, lower values of $D$ indicate better models.

Further, a holdout set was used to assess each models’ predictive performance by computing the mean squared prediction error (MSPE), $\frac{1}{q} \sum_{i=1}^{q} (y(s_i) - \tilde{y}(s_i))^2$, where $\tilde{y}(s_i)$ is the predicted outcome at the $i$-th holdout location and $q$ is the number of locations in the holdout set. The holdout set consisted of $q=598$ observations (i.e., 25%) selected randomly from the dataset. The remaining 1,793 observations were used to fit the candidate models. The location of the observed and holdout observations are depicted in Figure 3(a), with point and plus symbols respectively.

The median and 95% CI parameter estimates for the non-spatial model are 0.825 (0.769, 0.880), 0.611 (0.591, 0.631), and 0.018 (0.016, 0.019) for $\beta_0$, $\beta_{\text{DBH}}$, and $\tau^2$, respectively. Posterior predictive loss and associated values were $G=31.33$, $P=31.42$ and $D=62.75$. The model’s holdout set MSPE was 6.52. Again, as suggested by the variogram estimated from this model’s residuals (Figure 3(b)) there is relatively strong spatial dependence even after accounting for DBH. This residual dependence encourages the addition of spatial random effects.

Table 4 offers parameter estimates for the predictive process models fit using the two knot designs and at the 25 and 50 knot intensities. All models produces nearly identical estimates of $\beta_0$ and $\beta_{\text{DBH}}$. The approximately optimal design produces consistently smaller $\tau^2$ and larger $\sigma^2$ compared to estimates based on the coverage design. Also, as suggested by the simulation example, the approximately optimal design results in marginally improved model fit and predictive performance. For instance, the 50 knot models produce $D$ of 53.56 versus 52.79 and MSPE of 5.63 versus 5.43 for the coverage and approximately optimal designs, respectively.

Figures 3(c) and (d) were generated by interpolating over the median of each location’s spatial random effect posterior distribution for the 50 knot coverage and approximately optimal designs, respectively. In these figures the estimated knot locations are indicated with black circles. For both knot intensities, the approximately optimal design permits spatial random effects to more closely approximate the residual surface and hence provide improved model fit and prediction over the space filling design.

6 Discussion and future work

I have added three paragraphs at the end but the first four paragraphs here are too long and repeat too much of the earlier material. I will edit them.

Our overall objective has been to investigate the issue of knot selection in fitting spatial models incorporating Gaussian process We have addressed the problem of fitting desired hierarchical spatial modelling specifications to large datasets. To do so, we propose simply replacing 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.

As in existing low-rank kriging approaches, knot selection is required and as we demonstrated in Section 5 some sensitivity to the number of knots is expected. While 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 become difficult. Indeed, learning about fine scale spatial dependence is always a challenge (see, e.g., Cressie, 1993, p.114).

Our examples in Section 5 showed that even with fairly complex underlying spatial structures the predictive process model was able to effectively capture most of the spatial parameters with 529 knots (irrespective of whether the total number of locations was 3,000 or 15,000). A further challenge 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 \( \tau^2 \) becomes difficult. One possible remedy is reparametrizing \((\sigma^2, \tau^2)\) in terms of their ratio and the larger variance component (see, e.g., Diggle and Ribeiro Jr, 2007). Another option to explore is to modify the predictive process as \( \tilde{w}(s) = \hat{w}(s) + \tilde{\epsilon}(s) \), where \( \tilde{\epsilon}(s) \) is an independent Gaussian process with variance \( C(s, s)' \)

Finally, the goal has been dimension reduction to facilitate likelihood evaluation, to facilitate simulation-based model fitting. While we have employed MCMC to fit these models, faster alternatives that avoid MCMC can also be employed (see, e.g., Rue, Martino and Chopin, 2007). In fact, the predictive process approach within a full MCMC implementation is perhaps limited to order \( 10^4 \) observations on modest single processor machines (see Section 5); strategies that are empirical Bayesian in flavor, combining deterministic and simulation aspects, are likely the future for fitting very large space time datasets. Indeed, it is not uncommon to find spatial data sets, especially in scientific studies of large scale global phenomena, to contain far more locations than the illustrations here. For instance, Cressie and Johannesson (2007) work with data in the order of hundreds of thousands. It is also not uncommon to find space-time datasets 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, e.g., Heroux et al, 2006). We intend to extensively investigate the potential of predictive process models in such settings.

Evidently, there is much possible future work; here, we have only just opened the door to possible problems to explore. Indeed, we have raised more questions than we have answered. For instance, we could certainly return to the design approaches discussed in Section 4.1 with regard to knot selection adopting alternative criteria in place of the one adopted in Section 4.2. We could make further comparison with geometric and simple random sampling designs. We could look at other
dimension reduction approaches besides the predictive process including the modified predictive process (ref), kernel convolution, and the newly proposed approach in Sang and Huang (2011).

We note that knot selection carries over directly to the spatio-temporal setting. Conceptually, knot selection in time is more straightforward; we might simply use equally spaced knots along the observed time scale. However, we could argue that such choice ignores possible space-time interaction. Also, we do face the potential for the so-called “curse of dimensionality,” here, for a fixed number of knots, taking the form of sparser coverage of the domain of interest, as we go from two to three dimensions.

Finally, in very recent work (Guhanifyogi et al., 2011), the authors experimented with random knot selection, allowing the number of knots and their locations to vary with iteration of the MCMC model fitting. This seems like a promising path in consequentially reducing the effective number of knots. However, investigation to date has yielded only modest gains in exchange for substantial increase in computing time.

REFERENCES


Table 1: Parameter credible intervals, 50% (2.5% 97.5%) and predictive validation for the predictive process models using a regular grid of knots. Bold entries indicate where the true value is missed. The last column shows results for the parent model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>true</th>
<th>144</th>
<th>256</th>
<th>529</th>
<th>3,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.0</td>
<td>0.94 (0.56, 1.35)</td>
<td>0.73 (0.34, 1.16)</td>
<td>0.77 (0.34, 1.21)</td>
<td>0.72 (0.43, 1.01)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>45.0</td>
<td>36.45 (34.06, 38.14)</td>
<td>42.09 (37.62, 45.80)</td>
<td>43.83 (40.95, 46.77)</td>
<td>44.47 (43.18, 45.74)</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>300.0</td>
<td>390.4 (330.1, 399.6)</td>
<td>279.0 (258.6, 311.0)</td>
<td>323.1 (289.9, 349.0)</td>
<td>302.6 (275.5, 330.2)</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>50.0</td>
<td>62.42 (52.7, 71.99)</td>
<td>79.41 (59.77, 103.40)</td>
<td>61.47 (41.50, 84.30)</td>
<td>47.45 (40.03, 55.13)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0</td>
<td>1.14 (0.87, 1.49)</td>
<td>1.02 (0.80, 1.54)</td>
<td>1.31 (0.83, 1.52)</td>
<td>0.95 (0.87, 1.05)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.2</td>
<td>0.56 (0.53, 0.59)</td>
<td>0.45 (0.42, 0.49)</td>
<td>0.26 (0.21, 0.29)</td>
<td>0.16 (0.13, 0.22)</td>
</tr>
</tbody>
</table>

Table 2: Parameter credible intervals, 50% (2.5% 97.5%) for the predictive process models with 3000 locations using the lattice plus close-pair design with different knot intensities. Bold entries 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>
<thead>
<tr>
<th>Parameter</th>
<th>true</th>
<th>144</th>
<th>256</th>
<th>529</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.0</td>
<td>1.07 (0.77, 1.40)</td>
<td>0.63 (0.26, 1.01)</td>
<td>0.72 (0.33, 1.10)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>45.0</td>
<td>40.58 (38.65, 42.59)</td>
<td>44.77 (42.68, 46.74)</td>
<td>43.76 (42.35, 45.98)</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>300.0</td>
<td>386.62 (344.01, 399.69)</td>
<td>291.29 (267.57, 386.78)</td>
<td>330.00 (295.33, 358.85)</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>50.0</td>
<td>49.24 (43.86, 54.58)</td>
<td>53.40 (46.20, 60.72)</td>
<td>51.08 (43.98, 60.07)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0</td>
<td>1.34 (1.0, 1.70)</td>
<td>1.42 (0.89, 1.65)</td>
<td>1.39 (0.91, 1.66)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.2</td>
<td>0.55 (0.52, 0.58)</td>
<td>0.45 (0.43, 0.48)</td>
<td>0.24 (0.22, 0.29)</td>
</tr>
</tbody>
</table>

Table 3: Parameter credible intervals, 50% (2.5% 97.5%) for the predictive process models with 3000 locations using the lattice plus infill design. Bold entries 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>
<thead>
<tr>
<th>Parameter</th>
<th>true</th>
<th>144</th>
<th>256</th>
<th>529</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.0</td>
<td>1.18 (0.76, 1.66)</td>
<td>0.77 (0.39, 1.17)</td>
<td>0.71 (0.39, 1.02)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>45.0</td>
<td>42.12 (40.70, 43.21)</td>
<td>45.55 (44.46, 46.75)</td>
<td>43.45 (41.89, 45.13)</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>300.0</td>
<td>392.34 (343.23, 399.74)</td>
<td>316.73 (270.11, 368.57)</td>
<td>345.85 (286.32, 369.79)</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>50.0</td>
<td>58.26 (57.89, 66.46)</td>
<td>56.05 (47.97, 64.27)</td>
<td>47.31 (39.64, 55.38)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0</td>
<td>1.72 (0.98, 2.66)</td>
<td>1.35 (0.92, 1.76)</td>
<td>1.14 (0.94, 1.37)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.2</td>
<td>0.57 (0.54, 0.60)</td>
<td>0.48 (0.45, 0.50)</td>
<td>0.25 (0.22, 0.29)</td>
</tr>
</tbody>
</table>

Table 4: Predictive process candidate models’ parameter posterior credible intervals 50 (2.5 97.5), model fit criterion, and mean squared error of prediction (MSPE) for the tree height dataset.

<table>
<thead>
<tr>
<th>Coverage design</th>
<th>Approx. optimal design</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>0.874 (0.805, 0.944)</td>
</tr>
<tr>
<td>$\beta_{DBH}$</td>
<td>0.596 (0.577, 0.616)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.007 (0.004, 0.012)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.012 (0.008, 0.014)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.033 (0.019, 0.053)</td>
</tr>
<tr>
<td>G</td>
<td>27.62</td>
</tr>
<tr>
<td>P</td>
<td>28.21</td>
</tr>
<tr>
<td>D</td>
<td>55.84</td>
</tr>
<tr>
<td>MSPE</td>
<td>5.75</td>
</tr>
</tbody>
</table>
Figure 1: Top-left panel: simulated stationary anisotropic process generated with 3,000 sites using parameter values given in Table 1. Top-right panel: interpolated (posterior mean) surface for predictive process model overlaid with 256 knots in the lattice plus close-pair configuration. Bottom-panel: Interpolated (posterior mean) surface for predictive process model overlaid with 256 knots in the lattice plus infill configuration.
Figure 2: Averaged prediction variance ($V$) versus number of knots ($m$). Solid dots denote results for regular grids; dash-dot 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).
Figure 3: Tree height data set and associated candidate model output: (a) location of trees used to fit the candidate models (●) and those used to assess predictive performance (+); (b) empirical semivariogram of residuals from regressing log(HT) on log(DBH) along with maximum likelihood estimates of nugget $\tau^2$ (lower horizontal line), sill $\tau^2 + \sigma^2$ (upper horizontal line), and effective spatial range (vertical line); (c) 50 knot locations based on a coverage design and estimated spatial random effects surface; (d) 50 knot locations based on the proposed approximately optimal design and estimated spatial random effects surface.