

Multi-Scale Shotgun Stochastic Search for Large Spatial Datasets

Daniel Kirsner* and Bruno Sansó

Department of Statistics, University of California Santa Cruz

Abstract

Large spatial datasets often have fine scale features that only occur in sub-domains of the space, coupled with large scale features at much larger ranges. We develop a multi-scale spatial kernel convolution model where fine scale local features are captured by high resolution knots while lower resolution terms are used to describe large scale features. To achieve parsimony and explicitly identify the sub-domains of the space that exhibit fine scale attributes, we develop a form of shotgun stochastic search coupled with a stochastic process prior that induces structured sparsity that results in spatially varying resolution. In contrast to existing approaches, our approach does not require Markov chain Monte Carlo. In addition, the model does not require the spatially varying maximum resolution to be specified in advance. Our model fitting approach, based on Bayesian model averaging, is computationally feasible on large datasets, as computations for shotgun stochastic search can be performed in parallel, and it is possible to leverage the availability of convenient formulas for updating the coefficients when a single new knot is added. Competitive performance for computations, prediction, and interval estimation is demonstrated using simulation experiments and real data. Supplementary material for this article is available online.

Keywords: Discrete kernel convolution, Gaussian Process, Nonstationary spatial models, Parallel computing

*The authors gratefully acknowledge National Science Foundation award DMS-1513076 for partially funding this research.

1 Introduction

The traditional problem of model-based spatial statistics is to use a collection of spatially referenced observations to produce an estimate of the mean function of the data generating process, together with uncertainty intervals, across the entire domain. It is usually the case that observations are irregularly scattered over a large domain, and increasingly often, there is a need to handle very large amounts of data. Furthermore, it is desirable that models for this kind of data are able to capture behavior that varies due to differences in scales and in locations. For example, to model sea surface temperature in the Mediterranean, a model must be able to account for large scale features like the fact that the sea is warmer near Turkey than near Spain, and small scale features like how tiny islands in Greece can affect the temperature near the island. Gaussian processes provide a flexible framework for modeling this kind of data.

A well established literature has been developed on the idea of using Gaussian processes as the main tool for model based geostatistics (see, for example, Gelfand et al., 2010, for a comprehensive review). However, for n data points, the computation of the likelihood for a Gaussian process requires inversion of an n by n covariance matrix, which is computationally expensive ($O(n^3)$). There are numerous approaches to resolving this issue in a big spatial data context, see Heaton et al. (2018) for a comparative review, and Banerjee (2017) for a review of Bayesian methods. Briefly reviewing some of the popular approaches, we see that sparsity inducing techniques seek to reduce the number of non-zero elements in the covariance matrix of the Gaussian process through compactly supported covariance functions (Furrer et al. (2006), Kaufman et al. (2008)). Alternatively, they build sparse precision matrices using Gaussian Markov random fields (Rue and Held, 2005) or nearest neighbor Gaussian processes (Datta et al., 2016). Dimension reduction is another common approach. These techniques express the underlying spatial process as a sum of J basis functions, where $J \ll n$. Some examples include predictive processes (Banerjee et al., 2008), and discrete process convolutions (Higdon (1998), Stein (2007) Lemos and Sansó (2009) among many others). Discrete process convolutions focus on basis functions that are generated by kernels or radial basis functions usually centered on a grid. Conditional on the data and the parameters, the model reduces to a linear regression with J coefficients, which entails a reduction of the computational complexity to $O(J^2n + J^3)$. However, setting J too small can cause the model to miss the small scale features, but increasing J can make the parameter space unfeasibly large.

Traditional Gaussian process geostatistical models make very strong assumptions regarding the symmetry of the Gaussian field. In particular they assume stationarity, namely that covariance functions depend only on the displacement vector between two points, not their locations. Many approaches have been devel-

oped to account for non-stationarity. Some representative examples are: Schmidt and O’Hagan (2003) that uses the idea of deforming the space to map the original field onto a stationary field; Bornn et al. (2012) that embeds the nonstationary field into a higher dimensional space where the field will exhibit stationarity; Fuentes and Smith (2001) that allows the parameters of the covariance function to change in space, and Paciorek and Schervish (2006) that derives a class of nonstationary covariance functions. By construction, finite basis function representations of Gaussian processes, like discrete process convolutions, are non-stationary, but most models in the literature using such formulations do not attempt to explicitly describe the characteristics of the non-stationarity. Lemos and Sansó (2009); Lemos and Sansó (2012) seek to capture non-stationarity explicitly by considering kernels with spatially varying elliptical shapes.

Multi resolution models layer multiple processes on top of each other at different resolutions to accomplish dimension reduction while accounting for both fine and large scale features in the data. Examples include the approach of Nychka et al. (2015), which uses a Gaussian Markov random field prior on coefficients of basis functions at each resolution, and enforces prior independence between coefficients at different resolutions. The multi-resolution predictive process in Katzfuss (2017) recursively fits a predictive process (Banerjee et al., 2008) at increasing resolutions by refining an original set of knots. Both of these examples allow for nonstationary covariance functions, but enforce the same multiresolution structure across the entire field. A Bayesian approach that partially relaxes this structure was proposed in Guhaniyogi and Sansó (2017). They propose discrete process convolution with a nested set of knots and isotropic basis functions at differing resolutions linked by a shrinkage prior that takes into account the multiresolution structure of the knots. A related model, that provides an extensions to of Katzfuss (2017) approach that allows for spatially varying shrinkage was proposed by Benedetti et al. (2018). These two methods both can characterize non-stationarity though spatially varying shrinkage, which is desirable. However, both methods require MCMC, and require setting in advance a maximum number of resolutions, which makes them susceptible to not having enough knots to model a spatial surface well.

In this paper we consider a multi-resolution model that uses kernel convolutions in an increasingly refined set of nested grids. Rather than spatially varying shrinkage, we obtain spatially varying resolution by assuming a prior on the coefficients that sets some of them to be zero in a manner that forces parts of the nested grid to be empty. The sparsity also allows us to consider arbitrarily high number of resolutions, with no pre-specified bound. Having no upper bound in resolution makes our model extremely flexible, and quite robust to not having enough knots at the first resolution. To explore the space of possible sparse knot configurations, we extend shotgun stochastic search (Hans et al., 2007) to this setting, which

allows our method to take advantage of parallel computing environments. We demonstrate how to use this method to perform prediction, uncertainty quantification, and demonstrate competitive computational performance when compared with other approaches on a variety of spatial fields.

2 Background and Notation

We will begin by exploring the background material necessary for our approach. Let $\{w(s) : s \in D\}$ be the spatial process of interest on the domain $D \in \mathbb{R}^d$, where $d \in \{1, 2\}$. We can construct this Gaussian process in the manner of Higdon (1998). Let $k(s)$ be a kernel function, and $\beta_j, j = 1, \dots, J$ a set of Gaussian random variables corresponding to a set of points in D , s_1, \dots, s_J , usually defined over a regular grid. We focus on the finite dimensional representation of the process, $w(s) = \sum_{j=1}^J K(s - s_j | \phi) \beta_j$. Following Lemos and Sansó (2012) we term this a discrete process convolution (DPC).

These models are subject to the choice of the knot locations s_j , their total number J , the kernel functions K and their associated parameters ϕ . Even with a small number of knots, DPCs are able to capture the long range behavior of a spatial field. But, unless J is taken as a very large number, a DPC can miss short range features. And clearly, using a very large number of knots defeats the dimension reduction purpose of the DPC representation. In addition, it is often the case that some areas of the domain will show substantially more variability than others. To approach this issue, we will introduce the multi-resolution DPC. This embeds multiple DPCs at different resolutions into the same model. We will next define the notation for the structure we will use for the multiple resolutions.

2.1 Domain Partitioning

To define the structure of our multiple resolutions, we will follow the notation of Guhaniyogi and Sansó (2017). Start by partitioning the spatial domain D into $J(1)$ square subregions $D_1 \dots D_{J(1)}$. The centers of these regions define the first resolution of knots. To define resolution 2, each of the square subregions D_i will be partitioned into 2^d square subregions, giving us $J(2) = J(1) * 2^d$ subregions on the second resolution. The 2^d partitions of domain D_i are labeled as D_{i,i_2} where $i_2 \in \{1 \dots 2^d\}$. We can now iteratively define resolution r by partitioning the subregions at resolution $r - 1$ into 2^d square regions, and can index a domain in this region as D_{i_1, \dots, i_r} where $i_1 \in 1 \dots J(1)$ and $i_2 - i_r \in 1 \dots 2^d$. We will refer to the center of domain D_{i_1, \dots, i_r} as a knot s_j^r where $j = \sum_{l=1}^{r-1} ((i_l - 1)(2^d)^{r-l}) + i_r$. Figure 1 displays both one and two dimensional examples of the knot placements.

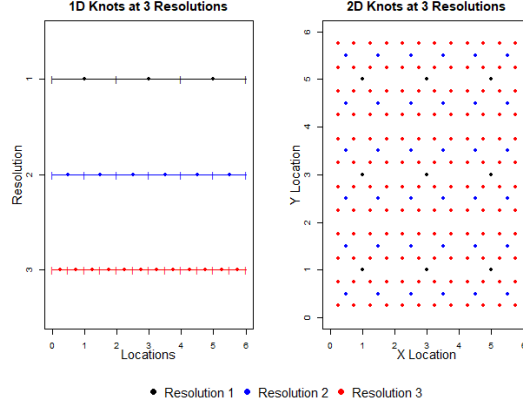


Figure 1: On the left, a plot of how knot locations work at different resolutions, and on the right, the same but in 2 dimensions.

From this definition, we can see that $J(r) = 2^{d(r-1)}$. We can view this partitioning as forming a tree, with the highest nodes at the lowest resolution, and lower nodes representing higher resolutions. 2^d branches come from each node to the nodes at the next level. Motivated by this tree structure, we will define $\text{parent}(D_{i_1, \dots, i_{r-1}, i_r}) = D_{i_1, \dots, i_{r-1}}$ and $\text{children}(D_{i_1, \dots, i_{r-1}}) = \{D_{i_1, \dots, i_{r-1}, i_r} : i_r \in 1, \dots, 2^d\}$. These definitions are also useful to apply to the knots. We define $\text{parent}(s_j^r) = s_{\lfloor \frac{j-1}{2^d} \rfloor + 1}^{r-1}$ and $\text{children}(s_j^{r-1}) = \{s_k^r : k \in 2^p(j-1) + 1, 2^p(j-1) + 2 \dots 2^p(j-1) + 2^p\}$. Lastly, we define the subtree, which is the set of all domains that are ancestors of a particular domain. Formally, $\text{subtree}(D_{i_1, \dots, i_r}) = \{D_{i_1, \dots, i_r, \dots}\}$.

3 A Bayesian multiresolution model

We start with a standard spatial regression model,

$$y(s)_i = \mathbf{x}(\mathbf{s})_i^T \alpha + w(s) + \epsilon(s)_i, \quad \epsilon(s) \sim N(0, \sigma^2),$$

where $\mathbf{x}(\mathbf{s})_i$ is a set of individual level predictors, α are the fixed effect coefficients associated with the predictors, $w(s)$ is the spatial effect, i is the index for replicates at a particular point s , and $\epsilon(s)_i$ is the random noise. Note that the predictors occur on the individual level, not the level of the spatial process. The spatial process is defined by a multiresolution DPC, $w(s) = \sum_{r=1}^{\infty} \sum_{j=1}^{J(r)} K(s, s_j^r, \phi_r, \nu) \beta_j^r$. For computational purposes, we require that K is compactly supported, with range ϕ_r . To facilitate our desire that higher resolution kernels reflect small scale behavior, we will let the kernel width decrease linearly as the resolution increases, i.e. $\phi_r = \tau \|s_j^r - s_{j-1}^r\|$ for some $\tau > 1$. We propose to use a Bezier kernel (Brenning,

2001), which is compactly supported, and has a parameter ν that controls the differentiability. In section 4 we will discuss the sensitivity of this method to the parameters ν and τ .

We will now turn our attention to the coefficients β_j^r . To achieve spatially varying *resolution*, we require sparsity, i.e. $\beta_j^r = 0$ for some r and j , that is structured in a manner such that the number of resolutions varies in space. As an aside, when comparing spatially varying shrinkage with spatially varying resolution, an analogy can be made to shrinkage versus variable selection in the regression context. Although a full review is omitted here, the review by Hahn and Carvalho (2015) covers many of the trade-offs between shrinkage and sparsity in that context.

3.1 A prior that induces spatially varying resolution

Motivated by this analogy, we will adapt a standard variable selection prior on the coefficients of our model (Hans et al., 2007) to this setting in order to induce spatially varying resolution. First, some notation must be introduced. Let $\gamma = [\gamma_1, \gamma_2, \dots]$ be a vector of infinite length with $\text{length}(\gamma_r) = J(r)$. Let the j th entry of γ_r be called $\gamma_{r,j}$, $j \in 1, 2, \dots, J(r)$. We will set $\gamma_{r,j} = 1$ if $\beta_j^r \neq 0$, and we will set our prior on this vector. This is vector of infinite length, so any prior will be better understood as a stochastic process.

For our prior to induce spatially varying resolution, we would like to satisfy three properties. First, every resolution 1 knot must be associated with a nonzero coefficient. Without the entire resolution 1 grid, it is conceivable that parts of our spatial field would not be modeled as not spatially varying, which does not make sense. And second, to allow the resolution to vary spatially, with a different number of resolutions possible at different locations, we only consider configurations that satisfy

$$\beta_j^r \neq 0 \implies \beta_{\lfloor \frac{j-1}{2^d} \rfloor + 1}^{r-1} \neq 0, \text{ which is identical to } \gamma_{r,j} = 1 \implies \gamma_{\lfloor r-1, \frac{j-1}{2^d} \rfloor + 1}.$$

In other words, if a coefficient associated with a knot is nonzero, then the coefficient associated with the parent of the knot must also be nonzero. Throughout this paper, we will interchangeably refer to the set of nonzero coefficients in the model as the knot configuration. Models with these restrictions will produce a field that has locally varying resolution. This is a sparse analogue to a feature of the model in Guhaniyogi and Sansó (2017), where the shrinkage applied to the coefficient of a parent is also applied to its children. Lastly, we would like our prior to not result in infinitely many nonzero coefficients, as these models will not be computationally feasible. Motivated by this, we set $Pr(\gamma_{1,j} = 1) = 1$, and

$$Pr(\gamma_{r,j} = 1 | \gamma_{r-1}) = \pi \times \gamma_{r-1, \lfloor \frac{j-1}{2^d} \rfloor + 1}.$$

This prior follows the properties above. Every resolution one knot is in the model, and if a knot at resolution $r > 1$ is in the model, then its parent must be as well. To understand some of the other features of this prior, we can consider the random variable $X_r = \sum_{j=1}^{J(r)} \gamma_{r,j}$, the number of nonzero β_j^r at resolution r . X_r can be thought of as a branching process (Chung, 2012). The initial state of the process is $X_1 = J(1)$, and the offspring distribution be $\text{Binomial}(2^d, \pi)$. The extinction probability of this process is analogous to the probability of having a finite number of nonzero β_j^r . By the properties of a branching process, the extinction probability is 1 as long as the expected value of the offspring distribution is less than 1. Therefore, if we set π such that $\pi 2^d < 1$, then the extinction probability of this process is 1, and the prior favors a finite number of nonzero coefficients.

To complete the specification of our prior, we must either fix π at some constant less than 1, or assume π to be a random variable and choose a prior for it. Fixing π was shown to be inadequate in the setting of linear model selection in Scott and Berger (2010). Specifically, a fixed value of π results in inadequate correction for multiplicity, which can lead to models that are too large, which in our context translates to overfitting. Scott and Berger (2010) recommend the use of a Beta prior on π , and show that this corrects for multiplicity and results in smaller models in the linear regression context while still preserving a closed form prior model probability, which we will need for our model selection procedure. Following this approach we let $\pi \sim \text{Beta}(a_\pi, b_\pi)$, so that under the prior, $\mathbb{E}(\pi) = a_\pi / (a_\pi + b_\pi)$.

This prior provides several attractive features. As shown in section 4, it is not very sensitive to varying a_π and b_π , and those parameters can be used to control the prior expected number of nonzero coefficients in a way that is easy to interpret. Recalling again the properties of a branching process, $\mathbb{E}\left(\sum_{r=1}^{\infty} \sum_{j=1}^{J(r)} \gamma_{j,r}\right) = J(1)/(1 - 2^d \mathbb{E}(\pi))$, provided that $2^d a_\pi / (a_\pi + b_\pi) < 1$. The prior probability of a particular coefficient being nonzero is decreasing geometrically with resolution, as $\text{Pr}(\gamma_{r,j} = 1) = \mathbb{E}(\pi)^{r-1}$. The prior probability for a particular set of nonzero coefficients γ is

$$p(\gamma) = \frac{B\left(a_\pi + \sum_{r=2}^{\infty} \sum_{j=1}^{J(r)} \gamma_{r,j}, b_\pi + \sum_{r=2}^{\infty} \left[2^d \sum_{j=1}^{J(r-1)} \gamma_{r-1,j} - \sum_{j=1}^{J(r)} \gamma_{r,j}\right]\right)}{B(a_\pi, b_\pi)}.$$

where $B(a, b)$ is a Beta function. For further interpretability of the hyper-parameters we use the alternative parameterization $\theta = a_\pi + b_\pi$ and $\mu = a_\pi / (a_\pi + b_\pi)$.

It is important to notice that in the multiresolution context, the Beta prior on π favors smaller models more as the number of knots increases, which makes the Beta prior favor sparser models. To demonstrate this, we will explore the prior odds in favor of a smaller model under a simple context with 1 dimension of knots. We will compare a prior with fixed $\pi = .5$, and call this p_1 , and a prior with $\pi \sim \text{Beta}(1, 1)$,

and call this p_2 . Under these two priors, we have the same prior expected number of knots, but have very different prior odds in favor of a smaller model. Let m_0 be a model with $J(1)$ first resolution knots, and no additional knots, and m_1 be a model with a single second resolution knot, and the same $J(1)$ first resolution knots. Under the first prior, the prior odds $p_1(m_0)/p_1(m_1) = 1/((1 - \pi)\pi) = 4$, which is constant in $J(1)$. Under the second prior, using the fact that $B(x + 1, y) = B(x + y)x/(x + y)$, the prior odds $p_2(m_0)/p_2(m_1) = (2J(1) + 2)(2J(1) + 3)/(2J(1) + 1)$. This expression indicates that under p_2 , the prior odds in favor of the smaller model are increasing as $J(1)$ increases, which favors the smaller model more strongly for larger models. And even for $J(1) = 1$, the smaller model is favored more under p_2 than p_1 . This has been confirmed by our empirical explorations, which indicate that using a random prior on π in our spatial multiresolution model produces a smaller number of knots than the one that is obtained with a fixed value of π , without compromising goodness of fit.

3.2 Spatially varying resolution and multifractal analysis

The spatially varying resolution induced by the proposed prior causes the resulting surface to have spatially varying regularity. One way to measure this is multifractal analysis (Jaffard et al., 2006), which uses discrete wavelet analysis to estimate changes in the regularity of the signal, as measured by Holder exponents. To explore how spatially varying resolution causes the local signal regularity to change, and how the varying degrees of sparseness controlled by π can affect this, we simulate 100 trajectories from priors corresponding to grid of values for ranging from 0 to 1, in a one dimensional setting. We then perform a multifractal analysis by recording the resulting ranges of Holder exponents. In all our simulations we use $J(1) = 7$. Notice that $\pi = 0$ results in just the 7 knots, and for increasing π , the average number of resolutions and knots will increase. For the $\pi = 1$ example, we truncate the maximum number of resolutions to 10, but as shown in section 3.1, no truncation is necessary for $\pi < .5$. Next, conditional on the knots and locations, a design matrix will be generated from our Bezier kernel, with $\phi_r = 2.5$ and $\nu = 1$. Finally, for each knot s_j^r , we generate $\beta_j^r \sim N(0, 1/r^2)$, which makes the coefficients on average smaller at higher resolutions. A wide range of Holder exponents suggests that the fractal behavior varies substantially in the resulting curve, which means that the roughness of the response curve differs at different points in the domain. Results are displayed in figure 2, where a clear increasing trend is observed in the range of Holder exponents, save for $\pi = 1$, as, in such case, the Holder exponents are virtually unchanged in the space. This makes intuitive sense, since for a dense multiresolution grid, the resolution is not spatially varying.

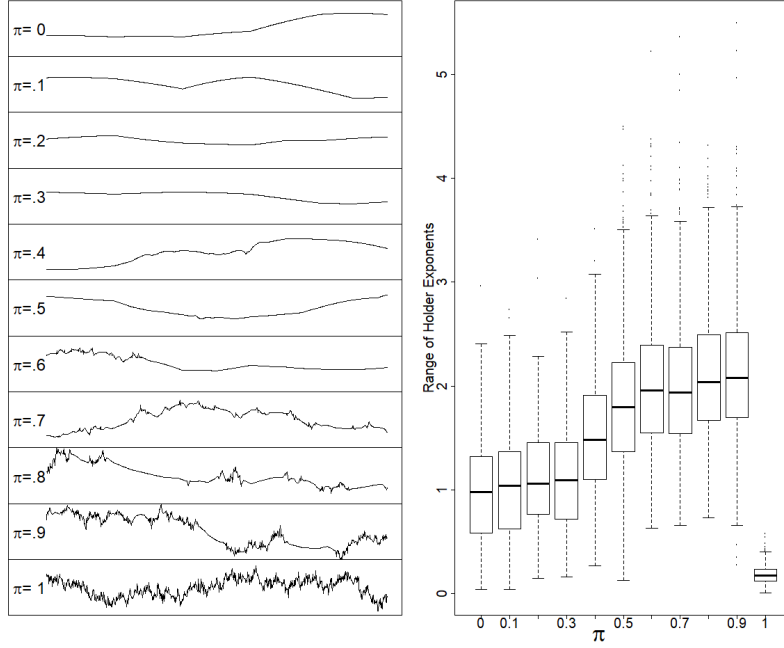


Figure 2: Left panel: randomly selected curves, one for each sparsity level, plotted on the same axes. Smaller values of π correspond to stronger spatial variability of the roughness of the sample paths. For $\pi = 1$ we observe homogeneous local variability across space. Right panel: distribution of Holder exponents as a function of π . An evident increasing pattern is present, except for $\pi = 1$, where Holder exponents typically have a very small range, indicating that the fields with dense multiresolution grids do not exhibit multifractal behavior.

3.3 Prior for the nonzero β_j^r and σ^2

The prior on the nonzero coefficients must be compatible for the spatial structure as well as computationally tractable. Let $\beta = \{\beta_1^1, \dots, \beta_{J(1)}^1, \beta_1^2, \dots, \beta_{J(2)}^2, \dots\}$. Conditional on the vector γ , we let β_γ be a vector of length $\sum_{r=1}^{\infty} \sum_{j=1}^{J(r)} \gamma_j^r$ that contains the nonzero β_j^r . Since γ specifies which β_j^r are zero, $p(\beta, \gamma) = p(\beta|\gamma)p(\gamma) = p(\beta_\gamma|\gamma)p(\gamma)$, so we can focus on specifying a prior on β_γ . In order to define the design matrix, let K_r be an $n \times J(r)$ matrix where the $K_r(i, j) = K(s_i, s_j^r, \phi_r, \nu)$, and $K_{r,\gamma}$ be the $n \times \sum_{i=1}^{J(r)\gamma_{r,j}}$ matrix with columns that correspond to nonzero $\gamma_{r,j}$. Finally, let $K_\gamma = [K_{1,\gamma}, K_{2,\gamma}, \dots]$ be a $\sum_{r=1}^{\infty} \sum_{j=1}^{J(r)} \gamma_j^r \times n$ matrix. This is the design matrix that corresponds to the nonzero β_j^r .

A g-prior (Zellner, 1986) on the coefficients associated with the knots, coupled with a reference prior on $\sigma^2|\gamma$ satisfies our desiderate, and has analytically tractable marginals. Note that putting a reference prior on coefficients common to all models being compared, and a g-prior on the other coefficients is a commonly used approach in the model selection context (Liang et al., 2008). For this multi-resolution model, the g-prior is of the form $p(\beta_\gamma, \sigma^2|\gamma) = p(\beta_\gamma|\sigma^2, \gamma)p(\sigma^2|\gamma)$ where $p(\beta_\gamma|\sigma^2, \gamma) = N(0, g\sigma^2(K_\gamma^T K_\gamma)^{-1})$, with a reference prior on the fixed effects α , the error σ^2 , and $p(\gamma)$ specified in the manner of section 3.1.

Notice that, usually, the argument for using a reference prior on α is made by assuming that the columns of K_γ have mean zero. However, centering this matrix would result in our basis functions no longer being compact. Fortunately, as Li and Clyde (2018) point out, the posterior distributions of the centered and non-centered models would have equivalent posteriors through a change of variables. Therefore, we will not center our design matrix.

An important property of the g-prior is that it induces shrinkage to high resolution knots that is, on average, larger than the one applied to low resolution ones. This behavior is due to the fact that more locations are in the range of kernels at lower resolutions. Therefore, the prior variance for the coefficients associated with the low resolution knots is higher than for the high resolution knots. We demonstrate this with a simple simulation. First, 10,000 locations are generated from a Uniform(0,10) distribution. Then, a number of multiresolution design matrices $K|\tau, \nu$ are formed for the Bezier kernel with a smoothness $\nu = 1$ and a kernel width $\tau = 1.5$, 7 resolution, and 5 knots at 1. This is approximately equally spaced data with a dense grid of knots unlikely to occur in MSSS, but is useful for demonstration purposes. We compute the diagonal of $(K^T K)^{-1}$ and take the average by resolution. The results are displayed in the supplementary material. We observe that the shrinkage is approximately linear on the log scale, save for the jump from resolution 1 to 2, which makes the shrinkage geometric in resolution.

To set the value of g we observe that small values of g result in large shrinkage

of the posterior mean. A popular default choice is $g = n$, which is known as a unit information prior (Kass and Wasserman, 1995), and provides reasonable performance in our context. The marginal likelihood for fixed g is available in closed form. However, Liang et al. (2008) observe that choosing g in this manner produces an information paradox. Loosely speaking, the marginal probability of model should approaches 1 as $r^2 \rightarrow 1$, but in the case of a g-prior with fixed g , this converges to a constant. We can resolve this issue by using the hyper-g prior suggested by the authors, which is of the form $g/(1+g) \sim \text{Beta}(1, a/2 - 1)$. This prior resolves the information paradox for non-null models and still results in a closed expression for the marginal likelihood, though it involves the Gauss hypergeometric ${}_2F_1$ function. Due to instability in the computation of ${}_2F_1$, for moderate to large n , this will require a Laplace approximation.

As a final note, the g-prior is improper if any columns of the design matrix are empty. In the context of this multi-resolution spatial model, this means that the prior does not make sense for a kernel function that has no data points within its range. To account for this, we propose to set $\beta_j^r = 0$ if it is associated with an empty column, regardless of the resolution. This is sensible, as even in a situation with proper priors, the coefficients associated with an empty column could not be learned well by the data.

We have now specified a prior on the model space λ , and the marginal likelihood of the data conditional on λ , so up to a normalizing constant, our posterior model probabilities are

$$p(\gamma|y) \propto \frac{a-2}{\sum_{i=1}^{J(r)} \gamma_{r,j} + a - 2} {}_2F_1 \left(\frac{n-1}{2}, 1, \frac{\sum_{i=1}^{J(r)} \gamma_{r,j} + a}{2}, R_\gamma^2 \right) \times \\ \frac{B \left(a_\pi + \sum_{r=2}^{\infty} \sum_{j=1}^{J(r)} \gamma_{r,j}, b_\pi + \sum_{r=2}^{\infty} \left[2^d \sum_{j=1}^{J(r-1)} \gamma_{r-1,j} - \sum_{j=1}^{J(r)} \gamma_{r,j} \right] \right)}{B(a_\pi, b_\pi)}.$$

Therefore, for a particular γ , we can compute the posterior model probability by forming the design matrix $K\gamma$, estimating β_γ using least squares, calculating R_γ^2 , and using the Laplace approximation to compute the ${}_2F_1$ function. In the next sections, we will discuss how to use these model probabilities to explore the space of possible knot configurations, and how to updated the least squares estimate of β_γ in a computationally efficient manner.

3.4 Extending shotgun stochastic search

Since the priors we have chosen results in closed form marginal model probabilities of particular configurations of knots, we can use shotgun stochastic search (Hans et al., 2007) to explore the space of possible knot configurations in a quick manner

that takes advantage of modern computing architecture, namely multiple core processors. Shotgun stochastic search (SSS) proceeds as follows:

1. Given a current model m_c , a set of the top Q models evaluated, and their respective marginal model probabilities and coefficients, define a neighborhood of possible new models N .
2. Evaluate the marginal probability of each model in N in parallel, and update the top Q models.
3. Choose a new current model from the neighborhood with probabilities proportional to their marginal probabilities.

In order to fit spatial fields with locally varying resolution, we would like to extend SSS, but rather than selecting variables from a finite set, selecting configurations of multiresolution knots arranged in nested grids. Note that this is a countably infinite set, as we are not truncating the number of resolution to consider. To use SSS, we need to define the neighborhood in a manner that is consistent with the prior from section 3.1.

To perform SSS, N is split into three groups, $N = N_- \cup N_o \cup N_+$. N_- is defined as all models of size $p-1$ that contain predictors that are all selected from γ . Moving to a model in this set is termed a *deletion move*. N_+ is defined as all models of size $p+1$ that contain all p predictors from γ and one from κ . Moving to a model in this set is termed an *addition move*. N_o is defined as all models of size p that contain $p-1$ predictors from γ and one from κ . Moving to a model in this set is termed a *replacement move*.

In the multiresolution knot selection context, if m_c is the current model, and $\kappa = \{\kappa_1, \dots, \kappa_p\}$ is the set of knots in mode the restrictions above lead to the following neighborhood definitions. For addition moves, only models that add a single knot that is a child of one of the knots already in m_c will be considered. The potential knots to add S_+ will be defined as $S_+ = \{\text{children}(\kappa_i) \mid i \in \{1, \dots, p\} \setminus \kappa\}$. So N_+ is just all models one knot from S_+ , and every knot in κ . For deletion moves, only knots that have no children will be considered for deletion. Formally, the potential deletion S_- will be defined as $S_- = \{\kappa_i \mid [\text{children}(\kappa_i) \setminus \kappa] = \text{children}(\kappa_i)\}$. Therefore, N_- is just all models with all but one knot in κ , with the knot removed $\kappa_{del} \in S_-$.

It is not very reasonable in our context for N_o to be all possible swap moves. This is because our space of possible variables is quite different in nature to the regression context. In regression, the swap moves are designed to explore spaces with correlated variables. For example, consider two possible predictors x_i and x_j that are highly correlated. If m_c contains x_i , it would be relatively unlikely for an add move to bring x_j into the model. But in the spatial context with compactly

supported kernels, the columns that will have the highest correlations are parents and children, which cannot be swapped due to the restrictions we place on the knot placements. Knots on the same resolution have fairly low correlation as long as the kernel width is not very wide. For example, in a 1D setting, with uniformly distributed locations and one resolution of knots, for a kernel width of 1.5 and a smoothness of 1 (which we suggest as a default in section 4.4), the correlation between adjacent knots is only about .5.

3.5 Computational details

Given these choices, we can now formulate the algorithm for multi-scale shotgun stochastic search (MSSS). Given a current model m_c , and a list of the Q top models,

1. Form $N = N_+ \cup N_-$ as defined above.
2. In parallel, for every $m_p \in N$, evaluate the marginal probability using the expressions above, and update the top Q models.
3. Sample m_{p-} from N_- and m_{p+} from N_+ with probability proportional to the marginal model probabilities. Then sample a new m_p from $\{m_{p+}, m_{p-}\}$ with probabilities proportional to their marginal probabilities. Return to step 1.

We run this algorithm until it reaches a local maximum, i.e. when the Q top models does not change for some number of iterations.

In order to calculate the marginal model probabilities quickly, we provide convenient formulas for updating the regression parameters of a model for all possible add 1 knot and subtract 1 knot models without computing the entire regression from scratch. These will be provided in the supplementary material. From these updated coefficients, we can calculate R^2 in the usual manner and then can evaluate the marginal likelihood. This is significantly faster than calculating the regression from scratch for each model.

3.6 Prediction and interval estimation

To get predictions that account for model uncertainty, we use Bayesian model averaging over the top knot configurations. Let the top Q configurations of knots found be $M = \{m_1, \dots, m_Q\}$ with marginal model probabilities $\{p_1, \dots, p_Q\}$. Correspondingly, consider their R^2 values, $\{R_1^2, \dots, R_Q^2\}$, least squared estimates of the coefficient vectors, $\hat{\beta}_1, \dots, \hat{\beta}_Q$, least squared estimates of the error variance, $\{\hat{\sigma}_1^2, \dots, \hat{\sigma}_Q^2\}$, the covariance matrices of the estimates V_1, \dots, V_Q , and the number of knots, $\{b_1, \dots, b_Q\}$.

For prediction at a point s_{new} in the spatial field, Bayesian model averaging works as follows:

1. For each of the Q knot configurations, calculate the values of the kernel functions at s_{new} , which is analogous to the rows of the design matrix for an observation. Call them $\{k_{new,1}, \dots, k_{new,Q}\}$.
2. Using each of the Q kernel function vectors, calculate the expected value $\mathbb{E}(y(s_{new})|m_i)$ for each $m_i \in M$. For the hyper- g prior, we have that

$$\mathbb{E}(y(s_{new})|m_i) = E\left(\frac{g}{1+g} \middle| m_i\right) k_{new,i}^T \hat{\beta},$$

where

$$E\left(\frac{g}{1+g} \middle| m_i\right) = \hat{s} = \frac{2}{p_i + a_g} \frac{{}_2F_1(.5(n-1), 2, .5(p_i + a_g), R_i^2)}{{}_2F_1(.5(n-1), 1, .5(p_i + a_g), R_i^2)}.$$

3. The Bayesian model averaging estimate is

$$y_{new}^*(s) = \frac{\sum_{i=1}^Q \mathbb{E}(y(s)|m_i) * p_i}{\sum_{i=1}^Q p_i}.$$

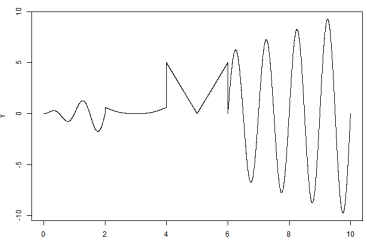
In practice, the largest of the posterior model probabilities is usually much larger than the others, so the averaging step is not always necessary. For intervals, the same averaging procedure can be used, but instead of using the expected value, we use the quantiles of the posterior predictive distribution. Since the posterior predictive distribution under the Hyper- g prior is not analytically available, we use the plug in estimator for the shrinkage factor, \hat{s} , from step 2 above. Then, the posterior predictive distribution is

$$p(y(s_{new})|m_i) \sim T_{n-p}(\mathbb{E}(y(s_{new})|m_i), \hat{s}\sigma_i^2(1 + k_{new,1}(V_i)k_{new,1})).$$

4 Assessing the proposed model

There are several things we would like to explore with respect to the performance of this model in a situation without covariates. Namely, we will assess the predictive accuracy and runtime of the model using a holdout set when changing the values of parameters that affect model size, model fit, and the smoothness of the predicted surface. Specifically, we will vary the prior sparsity parameters a_π and b_π , the resolution 1 size $J(1)$, the kernel width τ , and the kernel smoothness ν . For each of a number of simulated datasets and parameters, we fit an MSSS with a

Table 1: Equation for and plot of the mean function for the nonstationary 1D dataset

Function	Plot
$f(s) = \begin{cases} \sin(2\pi s) + 5 & \text{if } 0 \leq s < 2 \\ \sin(s - 3) ^3 + 5 & \text{if } 2 \leq s < 4 \\ 5 s - 5 + 5 & \text{if } 4 \leq s < 6 \\ \sin(2\pi s)s + 5 & \text{if } 6 \leq s < 10 \end{cases}.$	

10% randomly chosen holdout group, and quantify the predictive accuracy for the different parameter combinations. In addition, we compare the performance and runtime of our model to that of other multiresolution models that we were able to implement. There are many possible competing models (Heaton et al., 2018, see, for example), but here we limit ourselves with models that have a multi-resolution structure. We focus on the model proposed in Nychka et al. (2015), abbreviated as LK, for which the R package `LatticeKrig` (Nychka et al., 2016) is available, and the multiresolution process convolution model of Guhaniyogi and Sansó (2017), referred to as MDCT. To demonstrate how the multiresolution process convolution models behave differently than single resolution models, we will also compare the model with a single resolution DPC with a varying number of kernels with a kernel width of 1.5 times the distance of the grid.

Another natural competitor is the model in Katzfuss (2017). Code for implementing this model on 2d spatial fields is available. However, for the 2d datasets described below, we were unable to obtain sensible results. Mean estimation was possible, but when calculating intervals, the model consistently returned negative variances for some points in our spatial field. Therefore, this model will be excluded from our comparisons.

4.1 The datasets

Our first example consists of a one dimensional piecewise function that is meant to demonstrate the flexibility of our method in tackling highly nonstationary processes, and was used in Guhaniyogi and Sansó (2017). We generated one example with 20,000 observations from the mean curve, and added $N(0,1)$ noise. Plots and details of the function are presented in table 1. The next three simulated datasets consist of 2-dimensional fields. The first two were generated from stationary Gaussian processes with Matern covariance functions using the `RandomFields` package

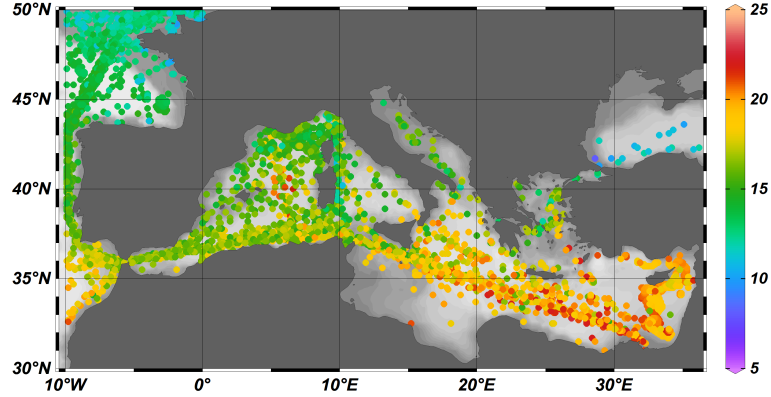


Figure 3: Mediterranean sea surface temperature observational data

(Schlather et al., 2017) on the interval $[0, 10] \times [0, 10]$. For the first of these datasets, the scale parameter was 1 and the smoothness was 2, and for the second, the scale was 1 and the smoothness was $1/2$. The second of these fields is continuous, but non-differentiable. For both, 20,000 observations were sampled from unequally spaced locations, and random noise with variance .1 was added to the generated data. The third dataset was generated from the nonstationary kernel convolution model of Lemos and Sansó (2009) using a 9 by 9 grid of kernels that are rotated differently across the space. This makes for a very smooth, nonstationary field. The same unequally spaced sampling and variance of .1 were used. The unequal spacing is displayed in the supplementary material, and the fields are displayed with the results in table 4.

The last examples corresponds to 12,210 temperature in situ measurements from the Mediterranean Sea during the month of December 2003. These data are obtained from four different types of devices, namely: buckets launched from navigating vessels; readings from the water intake of ship’s engine rooms; moored buoys; and drifting buoys. The result is a set of very unequally spaced, with many observations taken along shipping lanes, and large areas of the ocean scarcely covered by the sampling. In addition, it is known that the complexity of the shapes of the coastlines and the action of the currents, produce a very heterogeneous field of temperatures.

4.2 Parameter settings and competitors

For each of the examples discussed in the previous section we implemented MSSS with an intercept term, and a number of different parameter settings under a fully crossed design, resulting in 243 total runs. For prior sparsity, kernel size, and

kernel smoothness, the parameter settings are listed in table 2. The number of r_1 knots was varied between 10, 15, and 30 in the 1d example, and 42, 132 and 272 in the 2D simulated examples, and 91, 312, and 663 in the SST data example. For each setting of parameters, the top 100 models were stored for creating the prediction and intervals described in 3.6.

θ	μ	τ	ν
1	.1	1.5	1
5	.2	2	2
10	.5	2.5	3

Table 2: The simulation study uses a fully crossed design with these settings for the parameters of the prior and kernel function, with $\mu = \frac{a_\pi}{a_\pi + b_\pi}$ and $\theta = a_\pi + b_\pi$.

To fully leverage the parallel nature of MSSS, the model was implemented in C++ using OpenMP. All data preparations were done in R, and the RCPP package was used to pass information from R to C++. The single resolution DPC was implemented using MCMC in R under independent priors, and run for 10,000 iterations with 3 levels and varying numbers of resolution 1 knots. The MDCT of Guhaniyogi and Sansó (2017) was implemented in R using MCMC in R and run for 10,000 iterations under varying numbers of resolution 1 knots and 3 levels of resolution. The model of Nychka et al. (2015) was implemented using the `LatticeKrig` package in R with varying number of first resolution basis functions and 3 levels. Since we have run hundreds of different configurations of MSSS, in the numerical summaries, we will show the best, worst, and median result for each individual statistic, and in the graphical summaries, we will show plots of the best and worst of the MSSS models measured by the top posterior model probability.

4.3 Results

For the 1D example plots of the estimated mean function under the different models are shown in figure 4, and numerical results are found in table 3. In this setting, MSSS worked better than other models when $J(1)$ was large, and the kernel width τ was small. When a small number of wide, smooth kernels were used, the MSPE increased to as high as 1.1, but the interval coverage was still very close to .9. Given the piecewise nature of this function, this finding is not a surprising. It is also clear from the kernel locations that more knots are added near the discontinuities in the mean function. This is intuitive, and is how MSSS explicitly accounts for the local, high frequency behavior that occurs at those points. The MDCT also performed very well, as long as enough first resolution kernels were used. Despite very good results with respect to the estimation of the

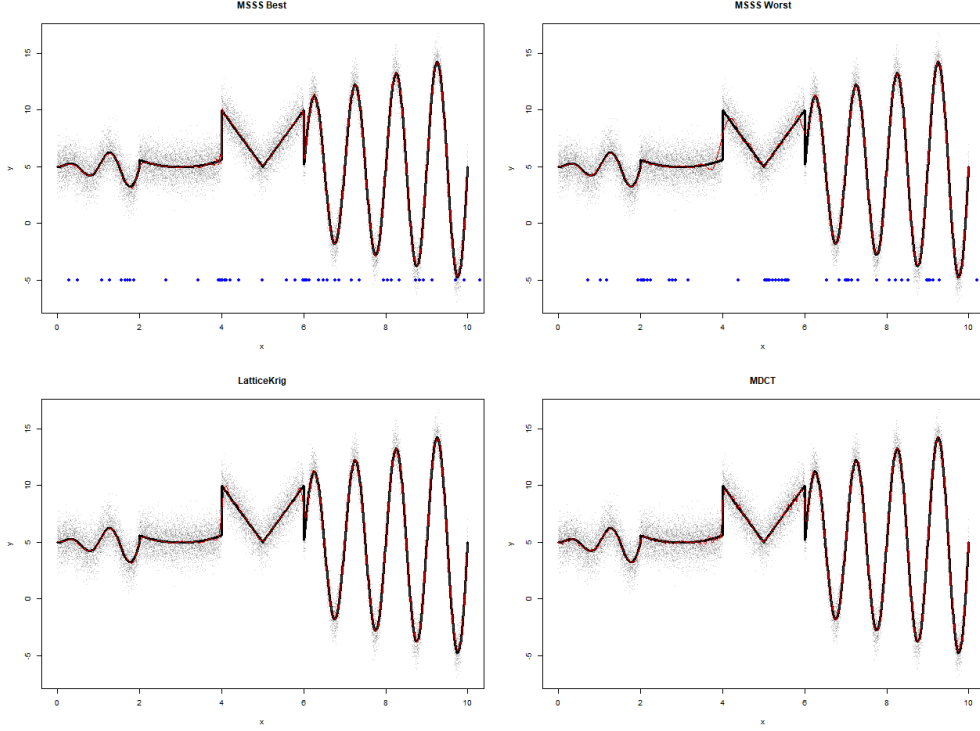


Figure 4: Comparison of predicted mean function, in red, and true mean function, in black, with kernel locations for MSSS in blue and the observed data in gray.

mean function, LatticeKrig had prediction interval coverage that was higher than the intended confidence level. This behavior repeats itself in all of the 2D examples, and reinforces the empirical findings of Heaton et al. (2018), where LatticeKrig demonstrated the same characteristics.

For the 2D simulated examples, predicted surfaces are displayed in table 4, and numerical summaries are presented in the appendix. Every model save for the DPC with the fewest kernels performed very well with the smooth, $\nu = 2$ Matern, returning appropriate looking predicted surfaces, and good numerical results. Every model struggled with the extremely jagged, $\nu = .5$ GP. The MSPE was much higher than the true variance (which was .1) for every model. However, MSSS, the DPCs, and the MDCT all did well in interval coverage. The best of the MSSS models, which had the largest number of initial kernels and the least smooth basis functions, did particularly well both in prediction and interval coverage. All of the models performed well in fitting the simulations from the nonstationary kernel convolutions, with good MSPE for every model except for the DPC with only 42 kernels, and excellent coverage probabilities for all but the LatticeKrig. It is worth mentioning that for this surface, the MSSS with a very large number of R1 knots

Model	MSPE	90% Coverage	Runtime (sec)
MSSS Min	.99	.89	4
MSSS Med	1.02	.90	13
MSSS Max	1.10	.91	76
MDCT 10	1.17	.9	381
MDCT 20	1.02	.9	518
LK 10	1.08	.94	102
LK 20	1.03	.95	105
LK 40	1.00	.95	107
DPC 10	15.30	.84	269
DPC 100	1.17	.89	419
DPC 1000	1.13	.89	911

Table 3: Numerical summaries for the competitor models on the 1D piecewise example. MSSS always provides excellent predictive interval coverage, and gives excellent out of sample fit when either enough R1 kernels are used or the kernels are of the appropriate shape.

sometimes added no knots at all, which is a desirable behavior when one resolution is sufficient.

The spatially varying resolution created by MSSS allows for an additional visualization. We can plot the posterior average number of resolutions active at each point in the space, as seen for the best MSSS fit, in figure 5. Note that since we are using model averaging over the top 100 models, this quantity can be a fraction. This graphic provides information about the regions of the space where there is more fine scale variation. The smooth, stationary GP with $\nu = 2$ requires fewer resolutions than the jagged GP with $\nu = .5$. The stationarity in these datasets is reflected by a similar pattern in resolutions across the space. In other words, there is not a single area in the space where the resolution is much higher than in other places. When MSSS is fit to the nonstationary kernel convolution, the behavior is quite different. The number of resolutions required is different across the space, with just one section of the space requiring 3 or 4, while the vast majority just requires two.

On the SST data, to ensure that out of sample predictions were reasonable on data this unequally spaced, MSSS required a small modification. Kernels were only allowed to enter the model if there was at least one data point within one kernel width from the center of the kernel. Without this restriction, kernel edge effects can cause the out of sample predictions to be unreasonable. When a location is far from the kernel, but still within the compact support, a low value in the design matrix will be compensated for by a high value of the coefficient β associated

Table 4: 2D true surfaces and predicted surfaces. The models are in the columns, and the three datasets are in the rows. Row A is the GP, Matern covariance with $\nu = 2$, row B is the Row A is the GP, Matern covariance with $\nu = .5$, and row C is the nonstationary Kernel Convolution.

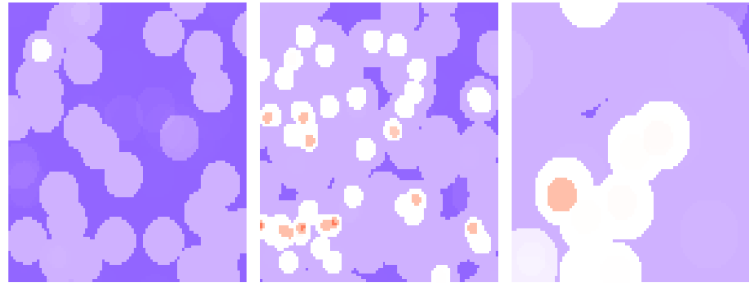
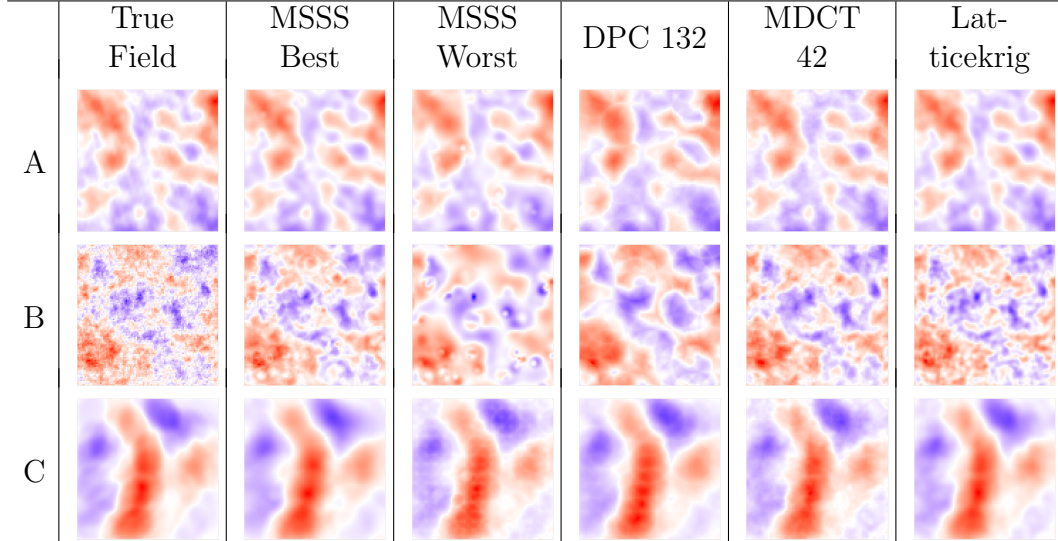


Figure 5: Plots of the maximum resolution active at each point on the surface for the best MSSS model by marginal model probability for each of the three simulated 2D datasets.

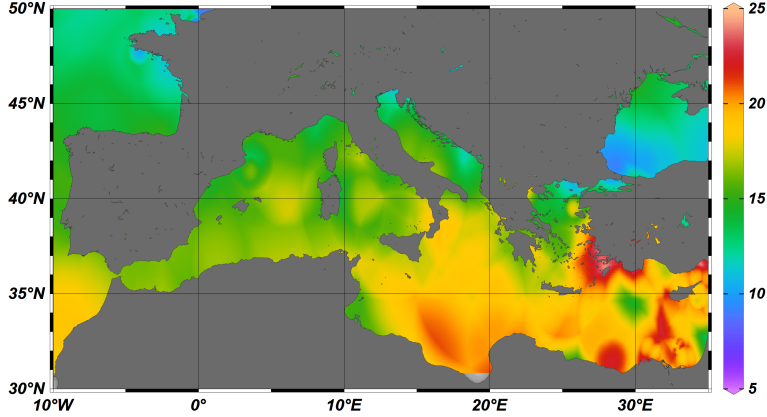


Figure 6: SST Predictions with $\nu = 3$, a kernel width of 2.5, and the additional restrictions discussed above.

with that column. The problem with this behavior is that closer the center of the kernel, the larger coefficient can cause unrealistically large predictions out of sample. As for the other settings, we set $\nu = 3$ and the kernel width to 2.5 since for sea surface temperature, we expect a relatively smooth mean function. The SST estimates are shown in figure 6, and the number of resolutions at each point is shown in figure 7. The plot of the number of resolutions at each point in the Mediterranean identifies regions with temperatures that vary differently than the surrounding areas. Some areas with higher resolutions include the region between Palma and Sardinia, which is warmer than its surroundings, the region adjacent to the Brittany peninsula on the northwest end of the dataset that is colder than its surroundings, and the southeast end of the Mediterranean, which has a large amount of temperature variation in the observed data, with observations varying between 15 and 23 degrees in a very small region. Numerical results are in table 5. MSSS and the MDCT with 42 kernels were the only models with both low MSPE and well calibrated interval coverage. Unlike the GP examples, predictions were substantially better using MSSS when compared to Latticekrig or the MDCT.

4.4 Discussion of default parameters

The results obtained in our data analysis lead to some guidelines for the selection of the parameters of the MSSS. First, the different parameters used in the beta-binomial prior on γ do not change the resulting surface or sparsity substantially, unless very extreme values are used. Therefore, we propose setting $\mu = 1/2^d$ and $\theta = 2$ as a safe default for data of the size that was dealt with here.

The remaining parameters τ , $J(1)$, and ν can be set by maximizing the pre-

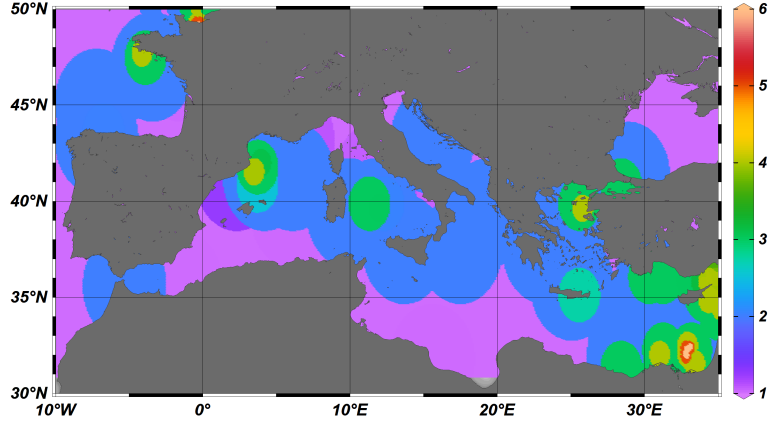


Figure 7: Plot of the maximum resolution active at each point on the surface for the MSSS model on SST in the Mediterranean.

Model	MSPE	90% Coverage	Runtime (sec)
MSSS	.63	.89	235
MDCT 42	.70	.88	1020
MDCT 132	1.58	.86	1505
LK 10	.68	.95	162
LK 20	.67	.94	204
DPC 91	.99	.88	309
DPC 312	.88	.89	433
DPC 1144	1.12	.87	717

Table 5: Numerical summaries for the models on the SST dataset.

dictive distribution over a grid of possible values. For large datasets such strategy can impose a steep computational cost. For the kernel parameters we require that $\tau > 1.5$. This ensures enough kernel overlap to prevent gaps. Beyond this strict restriction, the ability of MSSS to include an unlimited number of resolutions provides some robustness with respect to τ and $J(1)$. This is demonstrated in the simulation study, where the MSPE does not change very much among the different settings. For example, if $J(1)$ is not large enough to fit the data well, MSSS is able to add more kernels at high resolution to compensate for the lack of fit. Some attention must be paid, though, to the smoothness parameter ν , as the shape of the resulting predicted surface can be highly dependent on this parameter. However, specific knowledge of the application can inform the choice of ν . For example, in the SST dataset, it would be unreasonable for a predicted field to vary in too jagged of a manner, so a larger value of ν is preferable.

5 Conclusion

We have proposed a novel method for fitting nonstationary spatial models that achieves spatially varying resolution through a stochastic process prior. By avoiding MCMC, utilizing sparse matrix methods, an add one column regression updating formula, and modern parallel computing, MSSS has competitive computational performance when compared to other spatial methods. We have also shown that MSSS provides competitive out of sample fit and uncertainty quantification on a variety of unequally spaced spatial datasets, both stationary and non-stationary. We have also shown that the spatially varying resolution that this method enforces allows the statistician to simply and explicitly identify regions of non-stationarity in spatial datasets, which can have physical meaning in the context of specific applications.

Supplementary Materials

Online Appendix: Supplementary tables, figures, and detailed formulas for add one and subtract one knot updating in regression.

Data and Code: Zip file containing R and C++ code for the statistical methods and the sea surface temperature dataset from the Met Office.

Acknowledgments

We thank Peter Mueller for a discussion on the add and subtract one knot regression formulas, and John Kennedy at the Met Office, UK for making available the in situ sea surface temperature dataset.

References

- Banerjee, S. (2017, 06). High-dimensional bayesian geostatistics. *Bayesian Anal.* 12(2), 583–614.
- Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(4), 825–848.
- Benedetti, M., V. Berrocal, and N. Narisetty (2018). Identifying regions of inhomogeneities in spatial processes via an m-ra and mixture priors. Technical report, Technical report, Univiersity of Michigan.
- Bornn, L., G. Shaddick, and J. V. Zidek (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association* 107(497), 281–289.
- Brenning, A. (2001). *Geostatistics without stationarity assumptions within geographical information systems*. Citeseer.
- Chung, K. L. (2012). *Elementary probability theory with stochastic processes*. Springer Science & Business Media.
- Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* 111(514), 800–812.
- Fuentes, M. and R. L. Smith (2001). A new class of nonstationary spatial models. Technical report, Technical report, North Carolina State University, Raleigh, NC.
- Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15(3), 502–523.
- Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of spatial statistics*. CRC press.
- Guhaniyogi, R. and B. Sansó (2017). Large Multi-scale Spatial Modeling Using Tree Shrinkage Priors. Technical report, University of California, Santa Cruz.
- Hahn, P. R. and C. M. Carvalho (2015). Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* 110(509), 435–448.

- Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for large p regression. *Journal of the American Statistical Association* 102(478), 507–516.
- Heaton, M. J., A. Datta, A. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, et al. (2018). Methods for analyzing large spatial data: A review and comparison. *arXiv preprint arXiv:1710.05013*.
- Heaton, M. J., A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. (2018). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 1–28.
- Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics* 5(2), 173–190.
- Jaffard, S., B. Lashermes, and P. Abry (2006). Wavelet leaders in multifractal analysis. In *Wavelet analysis and applications*, pp. 201–246. Springer.
- Kass, R. E. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association* 90(431), 928–934.
- Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association* 112(517), 201–214.
- Kaufman, C. G., M. J. Schervish, and D. W. Nychka (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* 103(484), 1545–1555.
- Lemos, R. T. and B. Sansó (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association* 104(485), 5–18.
- Lemos, R. T. and B. Sansó (2012). Conditionally linear models for non-homogeneous spatial random fields. *Statistical Methodology* 9(1), 275 – 284. Special Issue on Astrostatistics + Special Issue on Spatial Statistics.
- Li, Y. and M. A. Clyde (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association* 113(524), 1828–1845.
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association* 103(481), 410–423.

- Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain (2015). A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics* 24(2), 579–599.
- Nychka, D., D. Hammerling, S. Sain, and N. Lenssen (2016). Latticekrig: Multiresolution kriging based on markov random fields. R package version 7.0.
- Paciorek, C. J. and M. J. Schervish (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 17(5), 483–506.
- Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*. CRC press.
- Schlather, M., A. Malinowski, M. Oesting, D. Boecker, K. Strokorb, S. Engelke, J. Martini, F. Ballani, O. Moreva, J. Auel, P. J. Menck, S. Gross, U. Ober, Christoph Berreth, K. Burmeister, J. Manitz, P. Ribeiro, R. Singleton, B. Pfaff, and R Core Team (2017). *RandomFields: Simulation and Analysis of Random Fields*. R package version 3.1.50.
- Schmidt, A. M. and A. O’Hagan (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(3), 743–758.
- Scott, J. G. and J. O. Berger (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 2587–2619.
- Stein, M. L. (2007). Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics*, 191–210.
- Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.