Nonstationary positive definite tapering on the plane

Ethan Anderes* Raphaël Huser† Douglas Nychka‡ Marc Coram§

October 31, 2011

Abstract

A common problem in spatial statistics is to predict a random field f at some spatial location t_0 using observations $f(t_1), \ldots, f(t_n)$ at $t_1, \ldots, t_n \in \mathbb{R}^d$. Recent work by Kaufman et al. (2009) and Furrer et al. (2006) study the use of tapering for reducing the computational burden associated with likelihood-based estimation and prediction in large spatial data sets. Unfortunately, highly irregular observation locations present problems for stationary tapers. In particular, there can exist local neighborhoods with too few observations for sufficient accuracy, while others have too many for computational tractability. In this paper we show how to generate *nonstationary* covariance tapers $T(s,t)$ such that the number of observations in $\{t\colon T(s,t) > 0\}$ is approximately a constant function of s. This ensures that tapering neighborhoods do not have too many points to cause computational problems but simultaneously have enough local points for accurate prediction. We focus specifically on tapering in two dimensions where quasiconformal theory can be used.

Keywords: Covariance tapering, optimization, random fields, kriging.

AMS subject classifications: 62M30, 62P12, 62G07, 65K10

1 Introduction and examples

A common problem in spatial statistics is to predict a random field f at some spatial location t_0 using observations $f(t_1), \ldots, f(t_n)$ at $t_1, \ldots, t_n \in \mathbb{R}^d$. In geostatistics, the field f may represent permeability of soil and the observations from physical measurements in a well-bore. In another example, of particular importance for environmental problems, f represents the concentration of pollutants in a contaminated site and the observations from remote sensing equipment. To generate predictions one often estimates the covariance structure of f then predicts using the

*Statistics Department, University of California at Davis, One Shields Avenue, Davis, California, 95616 (anderes@stat.ucdavis.edu). Supported by NSF grant DMS-1007480.

†Ph. D. candidate, École Polytechnique Fédérale De Lausanne, Switzerland. Supported by the Competence Center Environment and Sustainability (CCES) and the Swiss National Science Foundation (SNF).

‡National Center for Atmospheric Research.

§Department of Health Research and Policy (Biostatistics), Stanford University.

estimated covariance. Kriging is the geostatistical term for optimal unbiased linear interpolation (coined by Matheron (1963)). It has been the predominant prediction tool in spatial statistics since its introduction in the 1960's (see Cressie (1990) for a historical account) and can be shown to be equivalent to spline interpolation under certain assumptions on the covariance structure of f (Wahba, 1990). One of the difficulties with kriging is that it often requires computing the Cholesky decomposition of the estimated covariance matrix of the observations. This nominally requires $\mathcal{O}(n^3)$ operations (n is the number of observations) and is therefore prohibitive for even moderately large n. In this paper we present a technique for tapering the covariance of a two dimensional random field in such a way as to preserve the accuracy of prediction while significantly reducing the computational burden of the Cholesky decomposition.

Recent work by Kaufman et al. (2009) and Furrer et al. (2006) study the use of tapering for reducing the computational burden associated with likelihood-based estimation and prediction in large spatial data sets. Instead of using the covariance matrix Σ directly, the authors use a tapered matrix Σ_T which is obtained by component-wise multiplication of a correlation matrix constructed from a compactly supported stationary autocorrelation function. The ensuing sparsity in Σ_T often makes the Cholesky tractable. Kaufman et al. (2009) shows that by additionally tapering the sample covariance matrix of the observations, the resulting maximum likelihood estimate, of certain covariance parameters, consistently estimates the truth. Other results for estimates based on tapered covariances, including asymptotic distributions for these estimates, can be found in Zhang and Du (2008); Du et al. (2009); Shaby and Ruppert (2011). Furrer et al. (2006) studies tapering from the kriging perspective and demonstrates significant computational improvements. These papers exclusively use stationary tapers which can be problematic when the observation locations are highly irregular. In particular, there can exist local taper neighborhoods with too few observations for sufficient accuracy, while others have too many for computational tractability.

In this paper we show how to generate *nonstationary* covariance tapers $T(s,t)$ for spatial arguments $s,t \in \mathbb{R}^2$ such that the taper neighborhoods $\{t\colon T(s,t) > 0\}$ are bounded and have a pre-specified size. Of particular interest is the construction of nonstationary tapers $T(s,t)$ that satisfy

$$\text{area}\{t\colon T(s,t) > 0\} \approx \frac{c}{\rho(s)} \tag{1}$$

for each $s \in \mathbb{R}^2$ where $\rho(s)$ is the density of the observation locations and $c > 0$ is a constant. The advantage of such a taper is that the expected number of observation locations in $\{t\colon T(s,t) > 0\}$ is a constant function of s. This ensures that tapering neighborhoods do not have too many points to cause computational problems but simultaneously have enough local points for accurate prediction. In this paper we focus specifically on tapering in two dimensions. The reason is our method uses results from the theory of quasiconformal maps on the plane to construct T. It is possible to extend our methods to general dimension but at a considerable loss of simplicity provided by the elegant quasiconformal theory.

There are two main challenges for finding nonstationary tapering covariance functions that satisfy (1). The first is that T must be positive definite. This is difficult to ensure (let alone check) without the use of Bochner's Theorem which is only available for stationary tapers. Secondly, we don't actually measure the density ρ in (1), just a finite number of samples from ρ. This introduces an inherent ill-posedness for solving (1). In Section 2 we outline how to handle these difficulties

using quasiconformal theory and in Section 3 we discuss how tapering can be used for fast kriging prediction. We finish the paper with sections 4 and 5 which contain simulations and a numerical example that demonstrates the advantages of using nonstationary tapers, and some concluding discussion is given in Section 6.

2 Transformed probability measures and tapers

In this section we show how to generate our nonstationary tapers. We start by representing the density of the observation locations, ρ, as a transformed uniform density on a bounded region in $\mathbb{R}^2$ and show how this transformation, denoted φ, can be used to solve (1). In Section 2.1 we construct the estimate of φ using a penalized maximum likelihood approach. Finally, Section 2.2 contains a discussion of the advantages provided by quasiconformal theory.

We start by making the assumption that the observation locations $t_1, \ldots, t_n$ are independent samples from a continuous density $\rho(t)$ on a bounded observation region $\Omega \subset \mathbb{R}^2$. Let φ be a smooth (orientation preserving) bijection from Ω to a compact set $\Omega' \subset \mathbb{R}^2$ such that ρ is the pull-back of the uniform measure on Ω'. In particular the random variable $\varphi^{-1}(U)$ has density ρ where U is uniformly distributed on Ω'. By the transformation formula for densities

$$\rho(s) = \frac{\det D\varphi(s)}{\operatorname{area} \Omega'}$$

almost everywhere in Ω where $D\varphi(s)$ denotes the Jacobian of φ at s. Under mild assumptions on the density ρ, such a map always exists (McCann, 1995). To see how this representation of ρ is useful, let $K(|s-t|)$ be a stationary taper (positive definite on $\mathbb{R}^2$ with compact support) and consider the nonstationary taper obtained by deforming K with φ:

$$T_\varphi(s,t) := K(|\varphi(s) - \varphi(t)|).$$

Notice that the taper neighborhood for T_φ centered at s is the pullback of the circular neighborhood for K centered at $\varphi(s)$. In particular $\{t\colon T_\varphi(s,t) > 0\} = \varphi^{-1}\{v\colon K(|\varphi(s) - v|) > 0\}$. Therefore by changing variables and letting r denote the tapering radius of K we get

$$\operatorname{area}\{t\colon T_\varphi(s,t) > 0\} = \int_{v\colon K(|\varphi(s)-v|)>0} \frac{dy}{\det D\varphi(\varphi^{-1}(y))} = \frac{1}{\operatorname{area}\Omega'} \int_{v\colon |\varphi(s)-v|\le r} \frac{dy}{\rho(\varphi^{-1}(y))}.$$

For a small taper radius r this gives the correct approximation:

$$\operatorname{area}\{t\colon T_\varphi(s,t) > 0\} \approx \frac{\pi r^2}{\rho(s)\operatorname{area}\Omega'}.$$

If the density, ρ, that governs the sampling distribution of the observations were known, estimating φ would be a matter of numerical approximation. What makes the problem more difficult is that we only have access to ρ through a finite number of independent samples from ρ. The following section discusses the technique of penalized maximum likelihood estimation for estimating φ.

2.1 Penalized maximum likelihood estimation of φ

By representing the true, but unknown, monitoring distribution (governed by density ρ) as a deformed uniform distribution on $\Omega' \subset \mathbb{R}^2$ one can essentially treat the map φ as an unknown statistical parameter. The log likelihood function for this problem, considering the observation locations $t_1, \ldots, t_n$ as random draws from ρ, has the form $\ell(\varphi) = \sum_{k=1}^{n} \log\det D\varphi(t_k) + c$ where the constant c does not depend on φ. Maximizing $\ell(\varphi)$ is untenable since the functional $\ell(\varphi)$ is typically unbounded over nonparametric classes of maps. A popular technique for overcoming the ill-posed nature of the inverse problem is to regularize. In our case this is done by introducing a complexity or smoothness penalty $\mathcal{P}(\varphi)$ to the objective function $\ell(\varphi)$. In particular, our estimate of φ is defined by

$$\hat{\varphi} := \arg\max_{\varphi} \left[\sum_{k=1}^{n} \log\det D\varphi(t_k) - \lambda \mathcal{P}(\varphi) \right] \tag{2}$$

where $\lambda > 0$ is a tuning parameter. There is some flexibility in defining the smoothness penalty $\mathcal{P}(\varphi)$ but the main goal is to penalize maps φ which have large amounts of distortion (see Section 2.2 for more details). The estimation of ρ by $\det D\hat{\varphi}/\operatorname{area}\Omega'$ has been recently studied in the case of nonparametric density estimation (Anderes and Coram, 2011). In that paper, the objective is to estimate ρ and the actual map $\hat{\varphi}$ is unimportant. This paper has the alternative objective to estimate $\hat{\varphi}$, for generating the taper $T_{\hat{\varphi}}$, whereas the estimated density $\det D\hat{\varphi}$ is of secondary importance.

Before we discuss the class of maps over which (2) is optimized—in Section 2.2—we briefly mention how the objective function in (2) varies with vector field perturbations of φ. In particular, let $\varphi_\epsilon(t) = \varphi(t) + \epsilon u(\varphi(t)) + o(\epsilon)$ be a perturbation of φ for some vector field u and $\epsilon \in \mathbb{R}$. Then $\dot{\ell}[\varphi](u) = \sum_{k=1}^{n} \operatorname{div} u(\varphi(t_k)) - \lambda \dot{\mathcal{P}}[\varphi](u)$ where $\dot{l}[\varphi](u) := \lim_{\epsilon\to 0} \frac{\ell(\varphi_\epsilon) - \ell(\varphi)}{\epsilon}$ when the limit exists (a similar definition for $\dot{\mathcal{P}}[\varphi](u)$). This formula is used in Anderes and Coram (2011) for developing a gradient ascent algorithm to construct $\hat{\varphi}$.

2.2 Quasiconformal maps

The theory of quasiconformal maps provides a flexible theoretical framework for approximating φ (see Ahlfors (2006) for a classic reference on quasiconformal maps). In two dimensions, an important object is the complex dilatation $\mu\colon \Omega \to \mathbb{D}$ (here Ω is the observation region and $\mathbb{D}$ is the unit disk in the complex plane). The complex dilatation is important for a number of reasons. First, μ characterizes quasiconformal maps up to post composition with a conformal map. This, along with the fact that the only requirement on μ is measurability and $|\mu|_\infty < 1$, will allow us to construct rich families of maps for the transformation φ. Another advantage of the complex dilatation is that there is a constructible relationship between perturbations of μ and the corresponding perturbation of φ. This allows us to design a gradient ascent algorithm for estimating μ. In our algorithm we restrict the class of maps φ to be quasiconformal on all of $\mathbb{R}^2$ with complex dilatations that are twice continuously differentiable and have support in some known compact domain Ω which contains the observation locations. This simplifies a number of technicalities without significantly restricting the flexibility of the class of allowable maps.

To define the complex dilatation we start by writing the quasiconformal map φ in terms of its coordinate functions (φ^1, φ^2), each mapping $\mathbb{R}^2 \to \mathbb{R}$. We then define $\partial_z\varphi$ and $\partial_{\bar{z}}\varphi$ as functions

mapping $\mathbb{R}^2 \to \mathbb{C}$ such that

$$\begin{aligned}\partial_z \varphi &:= \tfrac{1}{2}\left(\partial_x \varphi^1 + \partial_y \varphi^2\right) + \tfrac{i}{2}\left(\partial_x \varphi^2 - \partial_y \varphi^1\right) \\ \partial_{\bar{z}} \varphi &:= \tfrac{1}{2}\left(\partial_x \varphi^1 - \partial_y \varphi^2\right) + \tfrac{i}{2}\left(\partial_x \varphi^2 + \partial_y \varphi^1\right).\end{aligned}$$

The complex dilatation for μ is now defined as

$$\mu_\varphi := \frac{\partial_{\bar{z}} \varphi}{\partial_z \varphi}.$$

The complex dilatation μ can be interpreted as measuring the ellipticity and inclination of the local ellipse which gets mapped to a local circle under the quasiconformal map which it characterizes. In particular, the closer μ_φ to 0 the more circular the local neighborhoods of T_φ are. This is important since by penalizing the departure of μ_φ from 0 (using the smoothness penalty $\mathcal{P}$) one can force T_φ to prefer circular local neighborhoods. For the reminder of the paper we drop the subscript and write μ when the dependency on φ is clear in context. We also let φ^μ denote the unique quasiconformal map with complex dilatation μ which maps $\mathbb{R}^2$ onto $\mathbb{R}^2$ and fixes $0, 1$. All other maps with complex dilatation μ that map onto $\mathbb{R}^2$ are of the form $a\varphi^\mu + b$ $(a, b \in \mathbb{C},\, a \neq 0)$.

Now we discuss the variation relation between the complex dilatation μ and φ^μ. Let μ be a complex dilatation that has support in Ω, $|\mu|_\infty < 1$ and $\mu \in C^2$ (the differentiability assumption is not critical and is mainly to forgo some technicalities). Perturbing μ in direction $\nu \in L_\infty$ (also with compact support in Ω), so that $\mu_\epsilon = \mu + \epsilon\nu$ for sufficiently small $\epsilon > 0$, leads to the ordinary differential equation characterization of φ^{μ_ϵ} (the map associated with μ_ϵ) by

$$\varphi^{\mu_\epsilon} = \varphi^\mu + \epsilon u \circ \varphi^\mu + o(\epsilon) \tag{3}$$

where $o(\epsilon)/\epsilon \to 0$ uniformly on compact subsets as $\epsilon \to 0$. Note that we are using '$\circ$' to denote composition of functions in equation (3). The vector field u depends on both μ and ν and is given by $u(\zeta) = u_0(\zeta) - \zeta u_0(1) + (\zeta - 1)u_0(0)$ where u_0 is the unique vector field such that

$$\partial_{\bar{z}} u_0 = \left(\frac{\nu}{1 - |\mu|^2} \frac{\partial_z \varphi^\mu}{\overline{\partial_z \varphi^\mu}}\right) \circ (\varphi^\mu)^{-1} \tag{4}$$

with boundary condition $u_0(z) = o(1)$ as $|z| \to \infty$ (Theorem 5, page 61 of Ahlfors (2006)). See Anderes and Coram (2011) for a discussion on numerical techniques for solving (4).

One can immediately see the advantage of this characterization. First, the unstructured nature of complex dilatations allow one to use linear basis functions to generate flexible classes of complex dilatations. Secondly, by rewriting the log likelihood ℓ and the smoothness penalty $\mathcal{P}$ as a function of the complex dilatation μ, instead of the associated map φ^μ, the directional derivative of $\ell(\mu) - \lambda\mathcal{P}(\mu)$ in direction $\nu \in L_\infty$ can be written

$$\dot{\ell}[\mu](\nu) = \sum_{k=1}^{n} \operatorname{div} u \circ \varphi^\mu(t_k) - \lambda \dot{\mathcal{P}}[\mu](\nu) \tag{5}$$

where $u(\zeta) = u_0(\zeta) - \zeta u_0(1) + (\zeta - 1)u_0(0)$ and u_0 is the solution to (4). The advantage here is that $\dot{\ell}[\mu](\nu)$ and $\dot{\mathcal{P}}[\mu](\nu)$ are linear over the reals in argument ν so that one can easily design a gradient

ascent algorithm for estimating the linear coefficients in the basis expansion of μ (see Anderes and Coram (2011) for details). In addition a basis expansion of μ can be used for designing natural smoothness penalties $\mathcal{P}(\mu)$. In Anderes and Coram (2011) they use this technique to define the smoothness penalty $\mathcal{P}(\mu)$ which essentially corresponds to $\int \left|\nabla^2 \frac{\mu(z)}{1-|\mu(z)|}\right|^2 dxdy$ where ∇^2 denotes the Laplacian.

3 Kriging with tapered covariances

In this section we describe how the technique of kriging can be used with tapered covariance matrices. We start with a description of kriging, or best linear unbiased estimation, for filtering and interpolating spatial observations; then finish with a discussion of how tapering gives significant computational improvements for producing kriging surfaces.

The basic structure of kriging starts with a linear model for the observational spatial process $\{z(t)\colon t \in \mathbb{R}^d\}$

$$z(t) = \sum_{k=1}^{m} \beta_k \theta_k(t) + f(t) + \epsilon(t) \tag{6}$$

where the functions $\theta_1(t), \dots \theta_m(t)$ are known spatial covariates, $\beta_1, \dots \beta_m$ are unknown coefficients, the centered Gaussian random field $\{f(t)\colon t \in \mathbb{R}^d\}$ is spatially correlated and $\epsilon(t)$ is Gaussian white noise error with variance σ^2. The observation locations of z are denoted $t_1, \dots, t_n$ and the kriging problem is to then estimate $\sum_{k=1}^m \beta_k \theta_k(t_0) + f(t_0)$ at some unobserved location t_0 using a linear combination of the observed quantities $z(t_1), \dots, z(t_n)$.

The vector of observations $\boldsymbol{z} := (z(t_1), \dots, z(t_n))^\dagger$ has the following multivariate Gaussian representation

$$\boldsymbol{z} \sim \mathcal{N}(X\boldsymbol{\beta}, \Sigma)$$

where $\Sigma := \left[\operatorname{cov}(f(t_i), f(t_j)) + \sigma^2 \delta_{i,j}\right]_{i,j=1}^n$, $X = \left[\theta_k(t_j)\right]_{j=1,k=1}^{n,m}$ and $\boldsymbol{\beta} := (\beta_1, \dots, \beta_m)^\dagger$. The kriging estimate is found in two stages: the first stage is to estimate $\boldsymbol{\beta}$ using generalized least squares; the second to smooth and interpolate the residuals. In particular, the estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \{X^\dagger \Sigma^{-1} X\}^{-1} X^\dagger \Sigma^{-1} \boldsymbol{z}. \tag{7}$$

Then by smoothing and interpolating $\boldsymbol{z} - X\hat{\boldsymbol{\beta}}$ one gets the following kriging estimate of $\sum_{k=1}^m \beta_k \theta_k(t_0) + f(t_0)$:

$$\underbrace{\sum_{k=1}^{m} \hat{\beta}_k \theta_k(t_0)}_{\text{estimated trend}} + \underbrace{\Sigma_0 \Sigma^{-1} (\boldsymbol{z} - X\hat{\boldsymbol{\beta}})}_{\text{smoothed, interpolated residuals}} \tag{8}$$

where Σ_0 is the vector $\left[\operatorname{cov}(f(t_0), f(t_j))\right]_{j=1}^n$.

Notice that the estimate (8) requires computation of $X^\dagger \Sigma^{-1} X$, $\Sigma^{-1}\boldsymbol{z}$ and $\Sigma^{-1}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})$. Nominally, the inversion Σ^{-1} takes $\mathcal{O}(n^3)$ operations which is computationally prohibitive for even moderate n. However, if Σ is sparse, the computations can be done quickly. To see this notice one can use a sparse Cholesky routine to obtain a lower triangular matrix L such that $\Sigma = LL^\dagger$. Now to compute $\Sigma^{-1}\boldsymbol{z}$, for example, one simply performs two triangular linear solves: first solving the

triangular system $L\,\boldsymbol{x}_1 = \boldsymbol{z}$ to obtain $\boldsymbol{x}_1$; then a second triangular system $L^\dagger \boldsymbol{x}_2 = \boldsymbol{x}_1$ which gives $\boldsymbol{x}_2 = \Sigma^{-1}\boldsymbol{z}$. Similar triangular systems can be solved to obtain the other two quantities $X^\dagger \Sigma^{-1} X$ and $\Sigma^{-1}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})$. Unfortunately, Σ is rarely sparse itself. This leads to the process of tapering Σ.

Tapering is essentially voluntary attenuation of Σ by multiplying, component-wise, by a sparse tapering correlation matrix. The tapers we consider here, for $\mathbb{R}^2$, are of the form $T_{\hat{\varphi}}(s,t) := K(|\hat{\varphi}(s) - \hat{\varphi}(t)|)$ where K is a stationary compactly-supported autocorrelation function on $\mathbb{R}^2$ and $\hat{\varphi}$ is the estimated transformation discussed in the previous section. Since $T_{\hat{\varphi}}$ is positive definite, the tapered covariance function $C_{\text{tap}}(t,s) = \text{cov}(f(s), f(t)) T_{\hat{\varphi}}(s,t)$ is also positive definite on $\mathbb{R}^2$. Moreover, the authors of Furrer et al. (2006) show that inference based on C_{tap} can still lead to asymptotically optimal kriging estimates under appropriate conditions on the spectral densities of the covariance and tapering functions. This motivates replacing Σ with the tapered matrix

$$\Sigma_{T_{\hat{\varphi}}} := \left[\text{cov}(f(t_i), f(t_j)) T_{\hat{\varphi}}(t_i, t_j) + \sigma^2 \delta_{i,j}\right]_{i,j=1}^{n}.$$

The advantage is that $\Sigma_{T_{\hat{\varphi}}}$ is sparse when the diameter of the compact support of K is small. Now by replacing Σ^{-1} in (8) and (7) with $\Sigma_{T_{\hat{\varphi}}}^{-1}$ results in the tapered kriging estimate.

Remark: In the above derivation of kriging we worked under the assumption that the covariance structure of the random field $\{f(t) : t \in \mathbb{R}^d\}$ is known. A more typical situation when working with real data is that one must estimate the covariance structure of $\{f(t) : t \in \mathbb{R}^d\}$ from observations. To streamline the exposition, however, we choose to treat the covariance structure as known.

4 Simulation Study

In this section we explore the behavior of nonstationary tapering techniques through simulations. We use prediction mean square error (MSE) and computational time as metrics for the success of nonstationary tapers. All of the following simulation times are recorded from the same computer with 2×2.93 GHz 6-Core Intel Xeon processors and 16 GB of ram.

The basic structure of our simulations are as follows: randomly generate irregular monitor locations; simulate a random field at the monitor locations plus a prediction grid; add noise to the monitor location observations; use stationary and nonstationary tapering to predict at the prediction grid; finally, record mean square prediction error and computational times. All the simulations are done in the region $(-1.5, 1.5] \times (-1, 2]$ which contain the randomly generated monitor locations and the prediction locations. The prediction mean square error is determined on a square prediction grid of 49 observation points, which stays the same throughout all simulations (shown as red dots in the left image in Figure 1). To generate the random monitor locations we use a discrete approximation to a Cox point process. In particular, a positive random field is generated, the log of which is a mean zero Gaussian random field, then 10^4 random locations are generated with density given by the normalized positive random field[1]. A realization of the random monitor locations is shown in the left plot of Figure 1 as black dots.

[1]To generate random densities with sufficiently high resolution we use a discrete FFT approximation to an isotropic random field with spectral density given by $8\frac{10^4}{(10^3+|w|^2)^2}$. To avoid the same monitor location sampled multiple times (which becomes possible when using a discrete approximation to the density) each location is perturbed slightly by a small Gaussian vector.

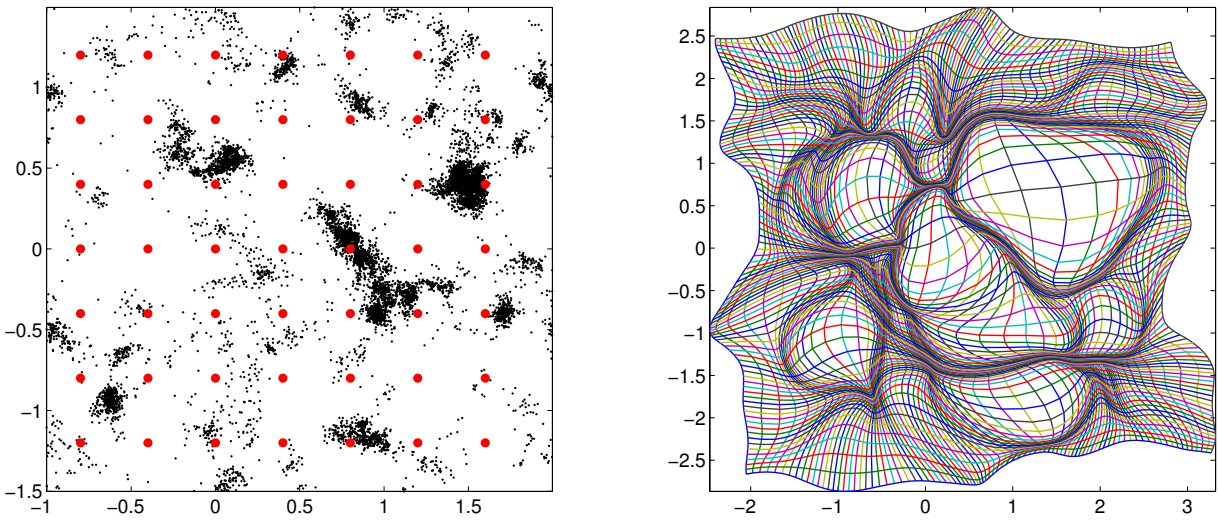

Figure 1: *Left:* Randomly generated monitor locations (black) and the prediction grid used for evaluation (red). *Right:* Plot of the estimated transformation $\hat{\varphi}$ used for nonstationary tapering on the monitor locations shown at left.

On each set of monitor and prediction locations we simulated a mean zero Gaussian Matérn random field[2]. We considered two different settings for the Matérn smoothness parameter, $\nu = 3/2$ and $\nu = 1/2$, but fixed the range parameter to $\rho = 1$ (which is 1/3 of the side length of the observation region) and the variance parameter to $\sigma^2 = 1$. We also added mean zero independent Gaussian noise to the monitor observations with standard deviation 0.1. The Matérn parameters were all considered known when generating the kriging predictions.

We used a fairly aggressive warping estimate $\hat{\varphi}$ for the nonstationary tapers based on the Wendland_2 autocorrelation taper. The estimates were generated using the same truncated Fourier basis and regularization penalty $\mathcal{P}(\varphi)$ found in Section 6 of Anderes and Coram (2011) with tuning coefficient $\lambda = 10$ (an example of $\hat{\varphi}$ is shown in the right plot of Figure 1). The average time for generating $\hat{\varphi}$ was approximately 1 hour. This was the most time consuming step of each simulation. However, it should be noted that the computational complexity for generating $\hat{\varphi}$ grows linearly with the number of observation locations. In contrast, kriging grows cubically in the number of monitor locations. Moreover, if cross-validation is performed for the estimation of covariance parameters then kriging needs to to be computed multiple times. Therefore, for very large data sets we expect the dominant computational bottleneck to be either estimation or prediction and not the warping estimation.

Table 1 shows the simulation results for the Matérn simulations with smoothness parameter set to $\nu = 3/2$. The entries of the table show average MSE over 50 independent simulations with computational time reported in parentheses. The columns designate the use of stationary versus nonstationary tapers for prediction. The rows of the table correspond to different methods for selecting the taper radii. The first row corresponds to selecting the smallest taper radii for which

[2]The random field values on the prediction grid were set aside and used only for prediction evaluation.

$\nu = 3/2$	stationary	nonstationary
min 20 points	0.059 (74.7 sec)	0.051 (73.2 sec)
max 100 points	0.582 (27.6 sec)	0.251 (27.2 sec)

Table 1: Simulation results for the Matérn simulations with smoothness parameter set to $\nu = 3/2$. Entries of the table correspond to average MSE over 50 independent simulations with computational time reported in parentheses. See Section 4 for details.

every prediction neighborhoods contains at least 20 monitor locations. The second row corresponds to selecting the largest taper radii for which no prediction neighborhood contains more than 100 monitor locations. The main feature of the results in Table 1 is the uniform reduction in average MSE when using a nonstationary taper (by as much as a factor of two). The computational time also decreases for nonstationary tapers, although the reduction is marginal. It seems that the computational gains provided by a reduction in the size of the largest tapering neighborhoods is offset by the increase in the size of the smallest tapering neighborhoods. Indeed, this is most likely the explanation for the dramatic decrease in average MSE when using a nonstationary taper: the nonstationary taper fills sparse prediction neighborhoods which have an insufficient number of observations for prediction. We do expect, however, that the computational gains will improve as the number of monitor locations grows to the point where the cubic complexity of the over-dense regions begins to dominate the computational time.

$\nu = 1/2$	stationary	nonstationary
min 20 points	0.135 (73.9 sec)	0.154 (72.8 sec)
max 100 points	0.645 (29.0 sec)	0.326 (29.5 sec)

Table 2: Simulation results for the Matérn simulations with smoothness parameter set to $\nu = 1/2$. Entries of the table correspond to average MSE over 50 independent simulations with computational time reported in parentheses. See Section 4 for details.

Table 2 shows the corresponding simulation results when smoothness parameter set to $\nu = 1/2$. Again, the entries of the table show average MSE over 50 independent simulations with computational time reported in parentheses. The interesting feature of this simulation is that the average MSE goes up in the first row when using a nonstationary taper. We speculate that this is caused by the aggressive nature of the warping which can negatively effect prediction when too extreme. Notice, however, the second row still shows a factor of two improvement over stationary tapers.

5 Numerical Example

We finish with an illustration of our method on summer temperatures monitored at $n = 4408$ observation locations $t_1, \ldots, t_n$ over the United States. The observations are shown in Figure 2. Each dot represents a monitoring location with the corresponding color representing the mean summer daily maximum temperature for June-August, 1990. The goal is to smooth and interpolate the data on a dense grid of $\sim 8.7 \times 10^5$ unobserved locations. Without tapering this is computationally intensive. Moreover, the non-uniformity of the monitoring locations makes stationary tapers

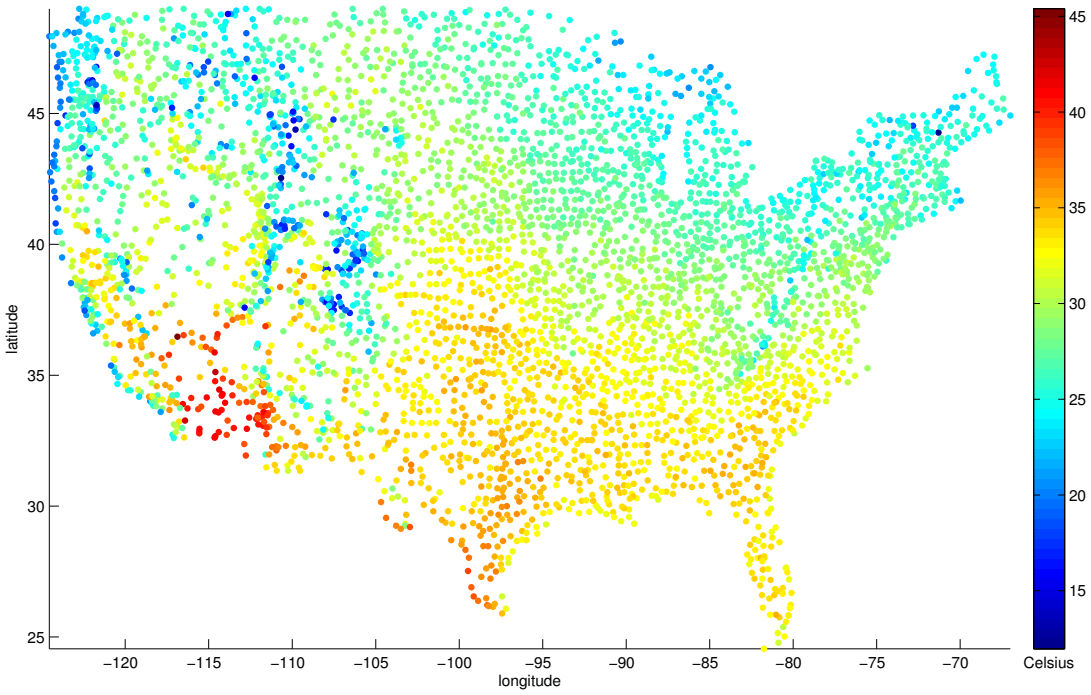

Figure 2: Mean summer daily maximum temperatures for June-August, 1990.

problematic, yielding some neighborhoods with too few predictive points and others with far to many for computational efficiency. This motivates using nonstationary tapers.

We start with the same setup as in Section 3, where $z(t)$ denotes the mean summer daily maximum temperatures for June-August, 1990, at spatial location $t \in \mathbb{R}^2$ (the longitude and latitude coordinates are scaled by a factor of 1/18 to fit in $[-1, 2] \times [-2/3, 2/3]$). Our linear model (6) has 4 spatial covariates: $\theta_1(t) = 1$; $\theta_2(t) =$ *altitude of location* t; $\theta_3(t) =$ *longitude of location* t; $\theta_4(t) =$ *latitude of location* t. We suppose $f(t)$ is a smooth isotropic random field on $\mathbb{R}^2$ with variance τ^2 and autocorrelation function $C(|s-t|) := \mathrm{cov}(f(s), f(t))/\tau^2$. The function C is chosen to belong to the Matérn family which is widely used within the geostatistics community for its flexibility and interpretability (see Stein (1999)). The class of isotropic Matérn autocorrelation functions is defined as

$$C(|h|) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}}{\rho} |h| \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\sqrt{\nu}}{\rho} |h| \right)$$

where $\Gamma(\cdot)$ is the Gamma function and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of second kind of order ν. The parameter ρ controls the range of spatial correlation and ν controls sample path smoothness. In this example we set $\rho = 0.3$, $\nu = 1.5$ and estimate the ratio σ^2/τ^2 by cross validation. The quantity σ^2/τ^2 corresponds to a type of noise-to-signal ratio and plays a role similar to the regularization coefficient for splines (Wahba, 1990). The parameter $\rho = 0.3$ corresponds to a range of roughly half the width of Colorado (in the scaled space) and $\nu = 1.5$ corresponds to

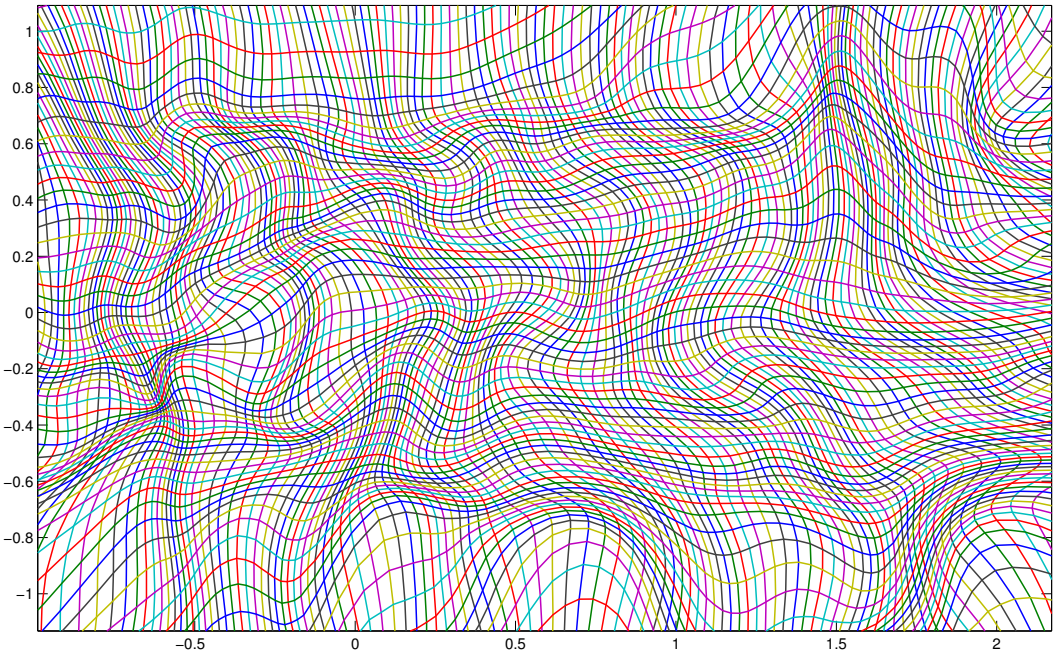

Figure 3: Plot of the estimated transformation $\hat{\varphi}$ using the methodology described in Section 2.1. The map $\hat{\varphi}$ transforms the monitor locations to an approximate uniform density on $[-1, 2] \times [-2/3, 2/3]$. The monitor locations before and after the transformation are shown in Figure 8. We used the same truncated Fourier basis and regularization penalty $\mathcal{P}(\varphi)$ found in Section 6 of Anderes and Coram (2011) (with tuning coefficient $\lambda = 10$).

a mean square differentiable random field. We remark that for modest changes in ν and ρ our kriging results are similar.

Due to the non-uniformity of the monitoring locations we use the nonstationary taper developed in Section 3. Our taper $T_{\hat{\varphi}}$ is defined as $T_{\hat{\varphi}}(s,t) = K(|\hat{\varphi}(s) - \hat{\varphi}(t)|/b)$ for all $s, t \in [-1, 2] \times [-2/3, 2/3]$ (the scaled longitude and latitude space) where K is the Wendland$_2$ autocorrelation function, $\hat{\varphi}$ is the estimated transformation discussed Section 2.1 and b is the tapering bandwidth. The Wendland$_2$ autocorrelation is given in Table 1 of Furrer et al. (2006) (see also Wendland (1995, 1998)) and has 4 continuous derivatives at the origin, compared with $C(|h|)$ which has two. Indeed, a smoother taper than process covariance is a fundamental assumption for most of the asymptotic results found in the literature (see Furrer et al. (2006) for example). The estimated map $\hat{\varphi}$, shown in Figure 3, transforms the approximate density of the monitor locations to the uniform density on $[-1, 2] \times [-2/3, 2/3]$, as discussed in Section 2.1. The transformed monitor locations are shown in the bottom diagram of Figure 8. We used the same truncated Fourier basis and regularization penalty $\mathcal{P}(\varphi)$ found in Section 6 of Anderes and Coram (2011). The tuning parameter λ was set to 10 which was picked by visual inspection of the resulting map. A more systematic approach for choosing λ is possible, using cross validation for example. However, we were still able to produce good results by simple trial and error: looking for a λ which produced a

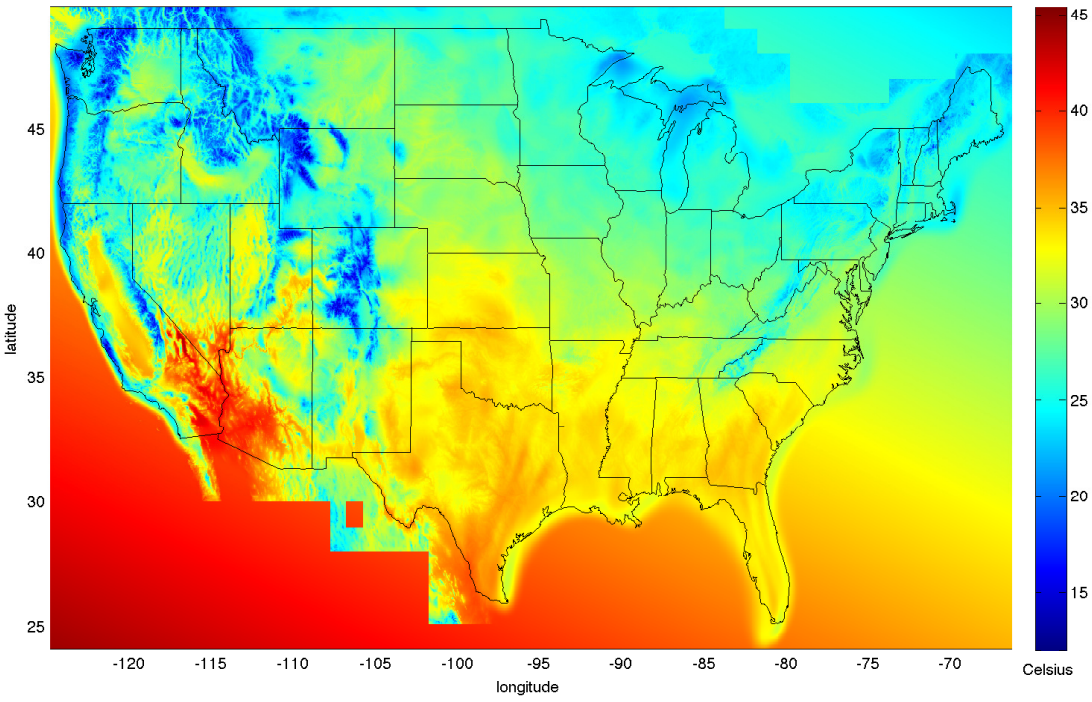

Figure 4: The Kriged surface of mean summer daily maximum temperatures (shown in Figure 2) evaluated on a 621×1405 grid using the nonstationary taper shown in Figure 3.

warping not too distorted but sufficiently dynamic to produce good tapering neighborhoods. We set the bandwidth parameter $b = 0.2$ so that 95% of the station taper neighborhoods contain at least 50 monitor locations.

Both the estimate $\hat{\boldsymbol{\beta}}$ and the kriging solution (8) only depend on σ^2 and τ^2 through the ratio σ^2/τ^2. Indeed one can re-write the estimate $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \{X^\dagger(\Omega + (\sigma^2/\tau^2)I_n)^{-1}X\}^{-1}X^\dagger(\Omega + (\sigma^2/\tau^2)I_n)^{-1}\boldsymbol{z}$$

where $\Omega = \left[C(|t_i - t_j|)T_{\hat{\varphi}}(t_i, t_j)\right]_{i,j=1}^n$ is the tapered correlation matrix, and I_n is the $n \times n$ identity matrix. The ratio σ^2/τ^2 is chosen to minimize the following cross-validation criterion $CV(\sigma^2/\tau^2)$:

$$CV(\sigma^2/\tau^2) := \frac{||\boldsymbol{z} - X\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{f}}||^2}{\{1 - \mathrm{tr}(A)/n\}^2}$$

where $A := \Omega[\Omega + (\sigma^2/\tau^2)I_n]^{-1}$ is called the projection matrix and $\hat{\boldsymbol{f}} := A(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})$ is the kriging estimate of $\boldsymbol{z} - X\hat{\boldsymbol{\beta}}$ at the sampling locations $t_1, \ldots, t_n$ (see Wahba (1990) for a discussion of cross-validation). Once again, tapering reduces the computational burden associated with the computation of $[\Omega + (\sigma^2/\tau^2)I_n]^{-1}$ and $\mathrm{tr}(A)$. For example, the trace of the matrix A can be approximated by

$$\mathrm{tr}(A) = \mathrm{E}(\mathbf{e}^\dagger A\mathbf{e}) \approx \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_k^\dagger A\mathbf{e}_k = \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_k^\dagger\mathbf{e}_k - \lambda(L^{-1}\mathbf{e}_k)^\dagger(L^{-1}\mathbf{e}_k), \tag{9}$$

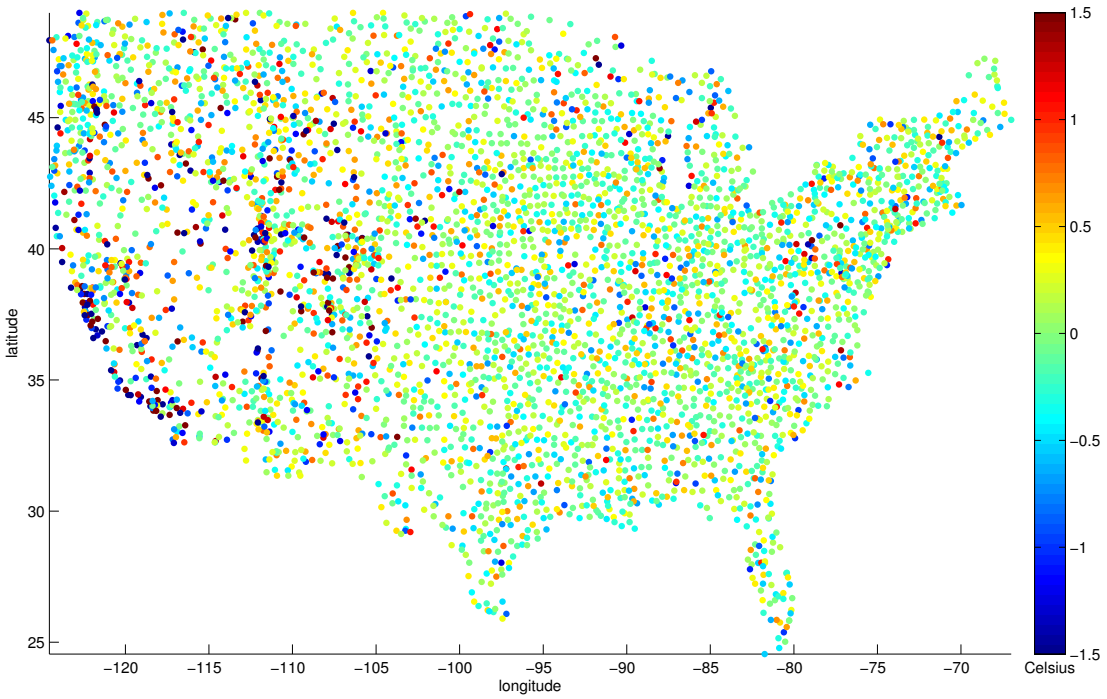

Figure 5: Residuals at each monitor location after kriging.

where $\mathbf{e}_1, \ldots, \mathbf{e}_K$ are *iid* standard Gaussian vectors of length n and $LL^\dagger$ is the Choleski decomposition of $[\Omega + (\sigma^2/\tau^2)I_n]^{-1}$. Increasing the number of samples $\mathbf{e}_1, \ldots, \mathbf{e}_K$ has the effect of improving the approximation (9) by the strong law of large numbers. The cross validation criterion yields an estimated ratio σ^2/τ^2 of 0.15 which corresponds to a signal to noise ratio of ~ 6.7.

Figure 4 displays the kriged map of temperatures on a grid of size 621×1405 using the nonstationary taper described above. The residuals, after kriging, are shown in Figure 5. The map tracks the observed temperature data while smoothing the noise and/or micro-scale variation. Indeed, the remaining noise at the sampling locations show significant reduction in amplitude and spatial de-correlation compared to the observations in Figure 2. In Figure 7 we show the separate contributions to the kriging surface from the estimated linear term (top) and the estimated smooth field f (bottom). These diagrams emphasize the necessity of the smooth field f in the model (6) for correcting the insufficiency of the linear model $\sum_{k=1}^{m} \beta_k \theta_k(t)$ near the ocean and at higher altitudes.

Producing the kriging surface using the nonstationary taper shown in Figure 4, takes 47 minutes on a desktop computer with 2×2.93 GHz 6-Core Intel Xeon processor and 16 GB of ram. To compare with the stationary taper we use two different bandwidths: one that bounds the number of neighbors, at each monitor, from above (to ensure computational efficiency); the other bounds the number of neighbors, at each monitor, from below (to ensure enough prediction points). The upper bound uses a stationary bandwidth of 0.116 (in the scaled longitude and latitude space) which is obtained by matching the 95^{th} percentile of the number of neighbors at each monitor with the same

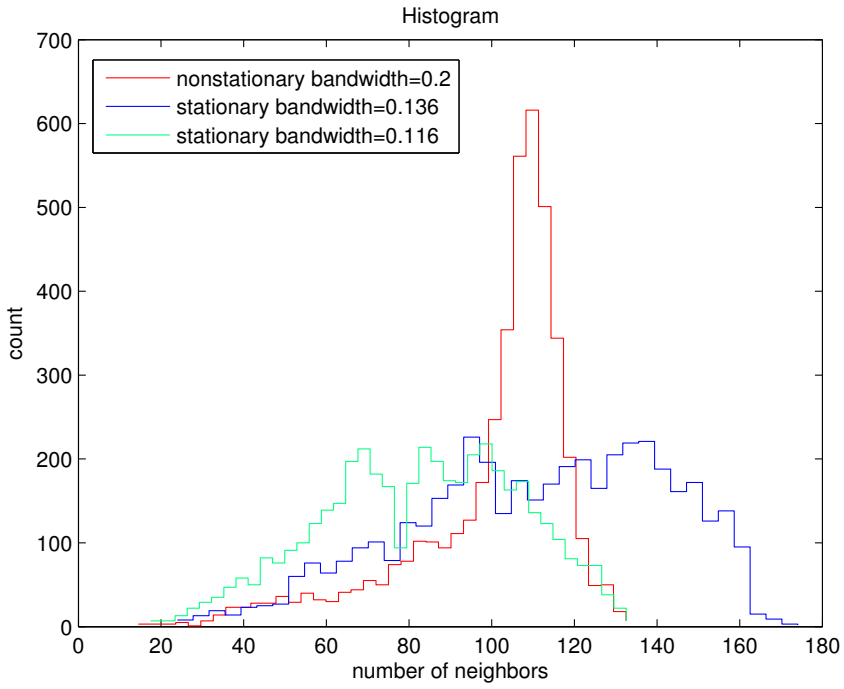

Figure 6: Histograms of the number of neighbors at each monitor using three different tapers: the nonstationary taper in red; a stationary taper with bandwidth 0.136 in blue; a stationary taper with bandwidth 0.116 in green. See Section 5 for a discussion.

quantity for our nonstationary taper. The second stationary bandwidth is 0.136 which matches the 5^{th} percentile to the same nonstationary taper quantity. The results are summarized in Figure 6 which show three histograms of the number of local taper neighbors at each monitor location, using the nonstationary taper and the two different stationary tapers. Notice that the distribution of the number of nonstationary taper neighbors is concentrated at the mean neighborhood size 98.5, whereas the stationary tapers are diffuse with mean neighborhood sizes of 81.5 (green) and 109.2 (blue). Indeed this is what the nonstationary taper is designed for: reducing the maximum number of neighbors for computational tractability while increasing the minimum number of neighbors for prediction accuracy.

6 Discussion

Tapering covariance matrices for estimation and prediction in spatial statistics is a useful and sometime necessary numerical technique which relieves some of the computational difficulties associated matrix inversion, Cholesky decompositions, et cetera. In this paper we explore using quasiconformal mappings or warpings for generating nonstationary tapers which are locally adapted to the spatial monitoring density and apply these nonstationary tapers to the problem of kriging. The nonstationary tapers we generate have spatially varying tapering neighborhoods which are designed to be inversely proportional to monitor density. The result is that the expected num-

ber of monitor locations stays relatively constant over different prediction neighborhoods. This simultaneously relieves the computational burden of over-dense prediction neighborhoods while also increasing the number of local predictors in sparsely observed regions. The quasiconformal maps used to generate these nonstationary tapers are estimated using a penalized log likelihood approach which provides a smooth objective function for which gradient based techniques are available for numerical optimization.

Using warpings or transformations to model nonstationary processes were first introduced to the spatial statistics literature in the paper by Sampson and Guttorp (1992). Their work, as well as that of subsequent authors (see for example, Perrin and Meiring (1999), Damian et al. (2001), Clerc and Mallat (2002), Schmidt and O'Hagan (2003), Iovleff and Perrin (2004), Anderes and Stein (2008), Anderes and Sourav (2009)) use warpings to model physical features in the data which, in principle, can be estimated from the values recorded at the monitor locations. . The work presented here is different in two respects. First, like stationary tapering, the warped stationary tapers are a purely numerical device. The nonstationarity is not intended to relate to any physical feature of the data. Rather, the nonstationary is designed to add tapering flexibility in the presence of highly irregular monitor locations. Secondly, this paper uses a new optimization procedure developed in Anderes and Coram (2011) for estimating a warping that produces taper neighborhoods which are inversely proportional to monitor density. The key feature of this optimization procedure is the use of a likelihood to generate these warpings. This circumvents problems associated with generating warpings by maximizing functionals based on monitor counts in taper neighborhoods. The difficulty with these functionals is that the objective function is typically not smooth. This makes optimization problematic. In contrast, the log likelihood approach developed here (and in Anderes and Coram (2011)) provides a differentiable objective function which accomplishes the same goal as those based on monitor counts.

In this paper we illustrate our nonstationary tapers through simulation studies and a numerical example. The simulation results demonstrate that nonstationary tapering can drastically improve prediction MSE in neighborhoods with relatively few monitor locations. There is some evidence for the computational savings provided by nonstationary versus stationary tapers, although the improvements were not as drastic as the MSE. Notice, however, that the simulations done in this paper were not designed to test the true computational gains when the number of data points is so large that stationary tapering becomes infeasible. Indeed, to be able to produce an adequate number of simulations for comparison it is necessary to consider a relatively small number of monitor locations which could be adequately solved by both stationary and nonstationary tapering. Regardless, the improvement in MSE points to clear advantages provided by nonstationary tapering. Moreover, the increased flexibility provided by nonstationary tapers contributes to the statistician's numerical tool box for estimating and predicting spatial random fields.

References

Ahlfors, L. (2006). *Lectures on Quasiconformal Mappings (with additional chapters by C. J. Earle, I. Kra, M. Shishikura, J.H. Hubbard).* University Lecture Series **38**, Amer. Math. Soc., Providence, RI..

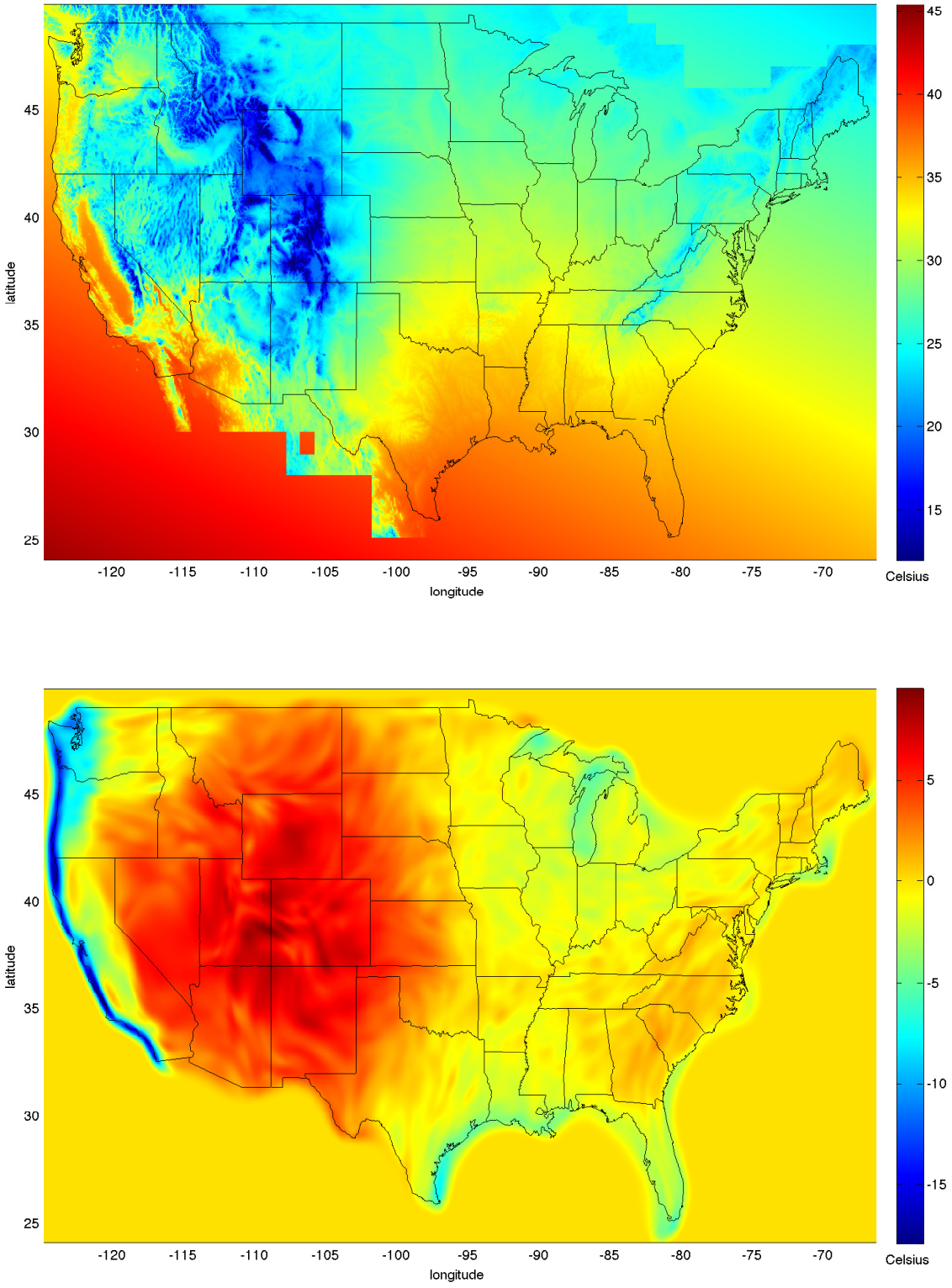

Figure 7: Top: Kriging estimate of the linear trend, $\sum_{k=1}^{m} \hat{\beta}_k \varphi_k(t)$, from model (6). Bottom: Estimated smooth field $\hat{f}(x)$ from (6).

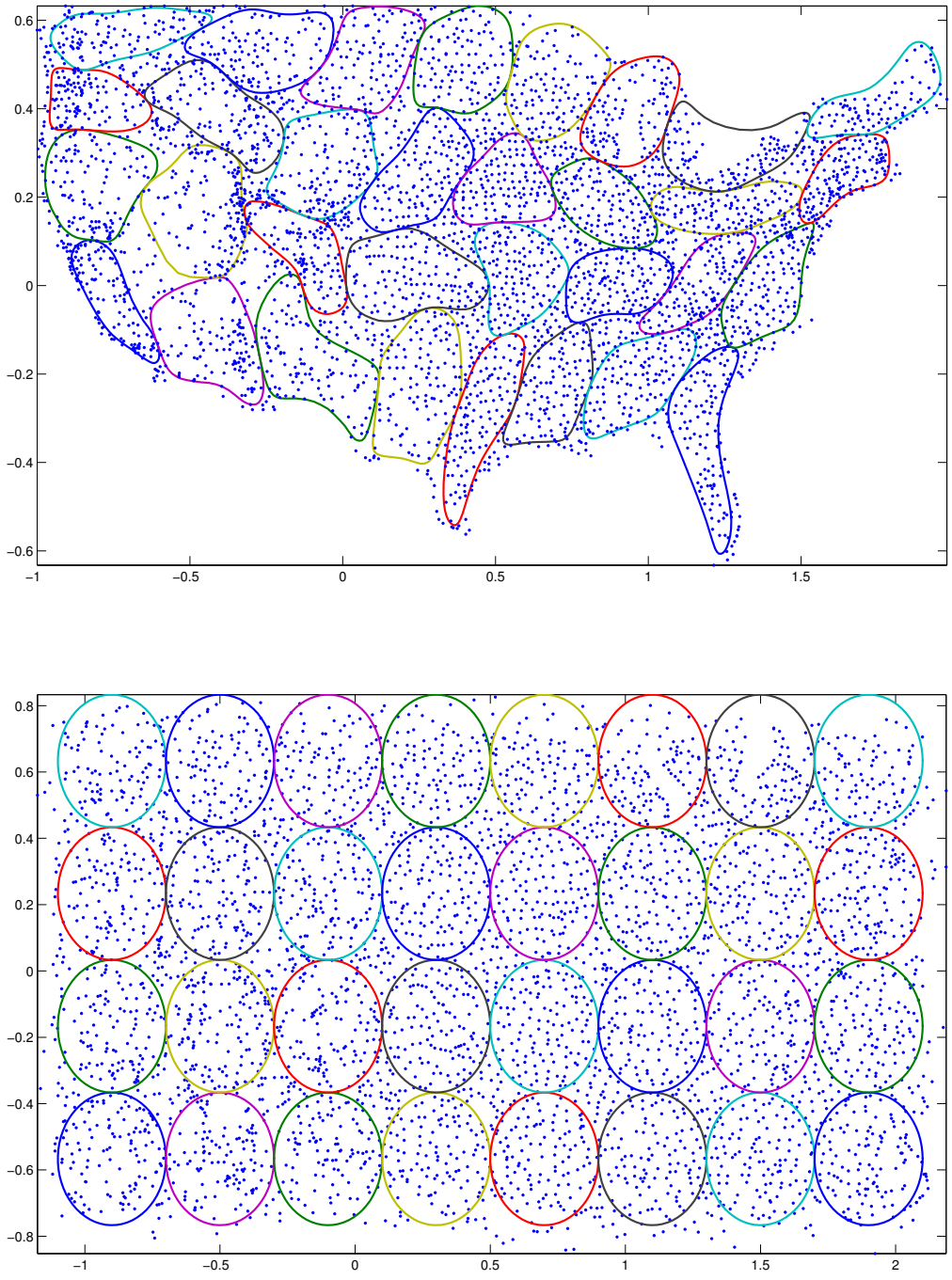

Figure 8: Top: The colored loops show the boundary of nonstationary taper neighborhoods $\{t\colon T_{\hat{\varphi}}(s,t) > 0\}$ at different observation locations across the USA. Bottom: This diagram shows the same neighborhoods but in the transformed space, i.e. $\hat{\varphi}\big(\{t\colon T_{\hat{\varphi}}(s,t) > 0\}\big)$.

Anderes, E. and Coram, M. (2011). Two dimensional density estimation using smooth invertible transformations. *Journal of Statistical Planning and Inference* **141**, 1183–1193.

Anderes, E. and Sourav, C. (2009). Consistent estimates of deformed isotropic Gaussian random fields on the plane. *Ann. Statist.* **37**, 2324–2350.

Anderes, E. and Stein, M. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *Ann. Statist.* **36**, 719–741.

Clerc, M and Mallat, S. (2002). Estimating deformations of stationary processes. *Ann. Statist.* **31**, 1772–1821.

N. Cressie (1990). The origins of Kriging. *Mathematical Geology* **22**, 239–252.

Damian, C., Sampson, P. and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics* **12**, 161–178.

Du, J., Zhang, H., and Mandrekar, V. (2009) Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.* **37**, 3330–3361.

Furrer, R., Genton, M., and Nychka, D. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics* **15**, 502–523.

Iovleff, S. and Perrin, O. (2004) Estimating a nonstationary spatial structure using simulated annealing *J. Comput. Graph. Statist.* **13**, 90–105.

Kaufman, C., Schervish, M., and Nychka, D. (2009). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Datasets. *J. Amer. Stat. Assoc.* **103**, 1545-1555.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246–1266.

McCann, R. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80**, 309–323.

Perrin, O. and Meiring, W. (1999) Identifiability for non-stationary spatial structure. *J. Appl. Probab.* **36**, 1244–1250.

Sampson, P. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108–119.

Schmidt, A. and O'Hagan, A.(2003) Bayesian inference for nonstationary spatial covariance structure via spatial deformations. *J. Roy. Statist. Soc. Ser. B* **65**, 745–758.

Shaby, B. and Ruppert, D. (2011) Tapered Covariance: Bayesian Estimation and Asymptotics. *J. Comp. Graph. Statist.*, in press.

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

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics* **4** 389–396.

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory* **93**, 258–272.

Zhang, H. and Du, J. (2008). Covariance tapering in spatial statistics. In *Positive definite functions: From Schoenberg to space-time challenges*, Mateu, J. and Porcu, E.(eds) University Jaume I: Castelló, de la Plana, Spain; 181-196