




## Nonstationary Positive Definite Tapering On The Plane

Ethan Anderes , Raphaël Huser , Douglas Nychka & Marc Coram


To cite this article: Ethan Anderes , Raphaël Huser , Douglas Nychka & Marc Coram (2013) Nonstationary Positive Definite Tapering On The Plane, Journal of Computational and Graphical Statistics, 22:4, 848-865, DOI: [10.1080/10618600.2012.729982](https://doi.org/10.1080/10618600.2012.729982)

To link to this article: <http://dx.doi.org/10.1080/10618600.2012.729982>

 View supplementary material 

 Accepted online: 19 Sep 2012.

 Submit your article to this journal 

 Article views: 180

 View related articles 

 Citing articles: 1 View citing articles 



# Nonstationary Positive Definite Tapering On The Plane

Ethan ANDERES, Raphaël HUSER, Douglas NYCHKA, and Marc CORAM

A common problem in spatial statistics is to predict a random field  $f$  at some spatial location  $t_0$  using observations  $f(t_1), \dots, f(t_n)$  at  $t_1, \dots, t_n \in \mathbb{R}^d$ . Recent work by Kaufman et al. and Furrer et al. studies the use of tapering for reducing the computational burden associated with likelihood-based estimation and prediction in large spatial datasets. Unfortunately, highly irregular observation locations can 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 article, we show how to generate *nonstationary* covariance tapers  $T(s, t)$  such that the number of observations in  $\{t : T(s, t) > 0\}$  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 quasi-conformal theory can be used. Supplementary materials for the article are available online.

**Key Words:** Covariance tapering; Kriging; Optimization; Random fields.

## 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  $f(t_1), \dots, f(t_n)$  at  $t_1, \dots, t_n \in \mathbb{R}^d$ . 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 estimated covariance. Kriging is the geostatistical term for optimal unbiased linear interpolation (coined by Matheron 1963). It has been the

---

Ethan Anderes is Assistant Professor, Statistics Department, University of California at Davis, Davis, CA 95616 (E-mail: [anderes@stat.ucdavis.edu](mailto:anderes@stat.ucdavis.edu)). Raphaël Huser is Ph.D. candidate, École Polytechnique Fédérale De Lausanne, Lausanne 1015, Switzerland (E-mail: [Raphael.Hauser@maths.ox.ac.uk](mailto:Raphael.Hauser@maths.ox.ac.uk)). Douglas Nychka is Director of CISL's Institute for Mathematics Applied to Geosciences (IMAGe), National Center for Atmospheric Research, Boulder, CO 80305 (E-mail: [nychka@ucar.edu](mailto:nychka@ucar.edu)). Marc Coram is Assistant Professor, Department of Health Research and Policy (Biostatistics), Stanford University, Stanford, CA 94305 (E-mail: [mcoram@stanford.edu](mailto:mcoram@stanford.edu)).

© 2013 *American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America*

*Journal of Computational and Graphical Statistics*, Volume 22, Number 4, Pages 848–865

DOI: [10.1080/10618600.2012.729982](https://doi.org/10.1080/10618600.2012.729982)

predominant prediction tool in spatial statistics since its introduction in the 1960s (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 article, 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, Schervish, and Nychka (2008) and Furrer, Genton, and Nychka (2006) studies the use of tapering for reducing the computational burden associated with likelihood-based estimation and prediction in large spatial datasets. Instead of using the covariance matrix  $\Sigma$  directly, the authors use a tapered matrix  $\Sigma_T$  that is obtained by component-wise multiplication of a correlation matrix constructed from a compactly supported stationary autocorrelation function. The ensuing sparsity in  $\Sigma_T$  often makes the Cholesky tractable. Kaufman, Schervish, and Nychka (2008) showed 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, Zhang, and Mandrekar (2009), and Shaby and Ruppert (2012). Furrer, Genton, and Nychka (2006) studied tapering from the kriging perspective and demonstrated 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 article, 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  $\{t : T(s, t) > 0\}$  are bounded and have a prespecified size. Of particular interest is the construction of nonstationary tapers  $T(s, t)$  that satisfy

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

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  $\{t : T(s, t) > 0\}$  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 article, we focus specifically on tapering in two dimensions. The reason is our method uses results from the theory of quasi-conformal 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 quasi-conformal 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 that is only available for stationary tapers. Second, we do not actually measure the density  $\rho$  in (1), just a finite number of

samples from  $\rho$ . This introduces an inherent ill-posedness for solving (1). In Section 2, we outline how to handle these difficulties using quasi-conformal theory and in Section 3, we discuss how tapering can be used for fast kriging prediction. We finish the article with Sections 4 and 5 that 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,  $\rho$ , as a transformed uniform density on a bounded region in  $\mathbb{R}^2$  and show how this transformation, denoted  $\varphi$ , can be used to solve (1). In Section 2.1, we construct the estimate of  $\varphi$  using a penalized maximum likelihood approach. Finally, Section 2.2 contains a discussion of the advantages provided by quasi-conformal theory.

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

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

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

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

where  $|\cdot|$  denotes the Euclidean norm on  $\mathbb{R}^2$ . Notice that the taper neighborhood for  $T_\varphi$  centered at  $s$  is the pullback of the circular neighborhood for  $K$  centered at  $\varphi(s)$ . In particular  $\{t : T_\varphi(s, t) > 0\} = \varphi^{-1}\{v : K(|\varphi(s) - v|) > 0\}$ . Therefore, by changing variables and letting  $r$  denote the tapering radius of  $K$ , we get

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

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

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

If the density,  $\rho$ , that governs the sampling distribution of the observations were known, estimating  $\varphi$  would be a matter of numerical approximation. What makes the problem more difficult is that we only have access to  $\rho$  through a finite number of independent samples

from  $\rho$ . The following section discusses the technique of penalized maximum likelihood estimation for estimating  $\varphi$ .

## 2.1 PENALIZED MAXIMUM LIKELIHOOD ESTIMATION OF $\varphi$

By representing the true, but unknown, monitoring distribution (governed by density  $\rho$ ) as a deformed uniform distribution on  $\Omega' \subset \mathbb{R}^2$ , one can essentially treat the map  $\varphi$  as an unknown statistical parameter. The log-likelihood function for this problem, considering the observation locations  $t_1, \dots, t_n$  as random draws from  $\rho$ , has the form  $\ell(\varphi) = \sum_{k=1}^n \log \det D\varphi(t_k) + c$ , where the constant  $c$  does not depend on  $\varphi$ . 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  $\varphi$  is defined by

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

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  $\varphi$  that have large amounts of distortion (see Section 2.2 for more details). The estimation of  $\rho$  by  $\det D\hat{\varphi}/\text{area } \Omega'$  has been recently studied in the case of nonparametric density estimation (Anderes and Coram 2011). In that article, the objective is to estimate  $\rho$  and the actual map  $\hat{\varphi}$  is unimportant. This article 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  $\varphi$ . In particular, let  $\varphi_{\epsilon}(t) = \varphi(t) + \epsilon u(\varphi(t)) + o(\epsilon)$  be a perturbation of  $\varphi$  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  $\dot{\ell}[\varphi](u) := \lim_{\epsilon \rightarrow 0} \frac{\ell(\varphi_{\epsilon}) - \ell(\varphi)}{\epsilon}$  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 QUASI-CONFORMAL MAPS

The theory of quasi-conformal maps provides a flexible theoretical framework for approximating  $\varphi$  (see Ahlfors 2006 for a classic reference on quasi-conformal maps). In two dimensions, an important object is the complex dilatation  $\mu : \Omega \rightarrow \mathbb{D}$  (here  $\Omega$  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,  $\mu$  characterizes quasi-conformal maps up to post-composition with a conformal map. This, along with the fact that the only requirement on  $\mu$  is measurability and  $|\mu|_{\infty} < 1$ , will allow us to construct rich families of maps for the transformation  $\varphi$ . Another advantage of the complex dilatation is that there is a constructible relationship between perturbations of  $\mu$  and the corresponding perturbation of  $\varphi$ . This allows us to design a gradient ascent algorithm for estimating  $\mu$ . In our algorithm, we restrict the class of maps  $\varphi$  to be quasi-conformal on all of  $\mathbb{R}^2$  with complex dilatations that

are twice continuously differentiable and have support in some known compact domain  $\Omega$  that 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 quasi-conformal map  $\varphi$  in terms of its coordinate functions  $(\varphi^1, \varphi^2)$ , each mapping  $\mathbb{R}^2 \rightarrow \mathbb{R}$ . We then define  $\partial_z \varphi$  and  $\partial_{\bar{z}} \varphi$  as functions mapping  $\mathbb{R}^2 \rightarrow \mathbb{C}$  such that

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

The complex dilatation for  $\mu$  is now defined as

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

The complex dilatation  $\mu$  can be interpreted as measuring the ellipticity and inclination of the local ellipse, which gets mapped to a local circle under the quasi-conformal map that it characterizes. In particular, the closer  $\mu_\varphi$  to 0, the more circular the local neighborhoods of  $T_\varphi$  are. This is important since by penalizing the departure of  $\mu_\varphi$  from 0 (using the smoothness penalty  $\mathcal{P}$ ), one can force  $T_\varphi$  to prefer circular local neighborhoods. For the remainder of the article, we drop the subscript and write  $\mu$  when the dependency on  $\varphi$  is clear in context. We also let  $\varphi^\mu$  denote the unique quasi-conformal map with complex dilatation  $\mu$  that maps  $\mathbb{R}^2$  onto  $\mathbb{R}^2$  and fixes 0 and 1. All other maps with complex dilatation  $\mu$  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  $\mu$  and  $\varphi^\mu$ . Let  $\mu$  be a complex dilatation that has support in  $\Omega$ ,  $|\mu|_\infty < 1$ , and  $\mu \in C^2$  (the differentiability assumption is not critical and is mainly to forgo some technicalities). Perturbing  $\mu$  in direction  $v \in L_\infty$  (also with compact support in  $\Omega$ ), so that  $\mu_\epsilon = \mu + \epsilon v$  for sufficiently small  $\epsilon > 0$ , leads to the ordinary differential equation characterization of  $\varphi^{\mu_\epsilon}$  (the map associated with  $\mu_\epsilon$ ) by

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

where  $o(\epsilon)/\epsilon \rightarrow 0$  uniformly on compact subsets as  $\epsilon \rightarrow 0$ . Note that we are using “ $\circ$ ” to denote composition of functions in Equation (3). The vector field  $u$  depends on both  $\mu$  and  $v$  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

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

with boundary condition  $u_0(z) = o(1)$  as  $|z| \rightarrow \infty$  (theorem 5, Ahlfors 2006, p. 61). 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 allows one to use linear basis functions to generate flexible classes of complex dilatations. Second, by rewriting the log-likelihood  $\ell$  and the smoothness penalty  $\mathcal{P}$  as a function of the complex dilatation  $\mu$ , instead of the associated map  $\varphi^\mu$ , the

directional derivative of  $\ell(\mu) - \lambda\mathcal{P}(\mu)$  in direction  $v \in L_\infty$  can be written as

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

where  $u(\zeta) = u_0(\zeta) - \zeta u_0(1) + (\zeta - 1)u_0(0)$ ,  $u_0$  is the solution to (4) and  $\operatorname{div}$  denotes the divergence operator. The advantage here is that  $\dot{\ell}[\mu](v)$  and  $\dot{\mathcal{P}}[\mu](v)$  are linear over the reals in argument  $v$ , so that one can easily design a gradient ascent algorithm for estimating the linear coefficients in the basis expansion of  $\mu$  (see Anderes and Coram 2011 for details). In addition, a basis expansion of  $\mu$  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  $\mathcal{P}(\mu)$  that essentially corresponds to  $\int |\nabla^2 \frac{\mu(z)}{1-|\mu(z)|}|^2 dx dy$ , where  $\nabla^2$  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  $\{z(t) : t \in \mathbb{R}^d\}$

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

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  $\{f(t) : t \in \mathbb{R}^d\}$  is spatially correlated and  $\epsilon(t)$  is Gaussian white-noise error with variance  $\sigma^2$ . The observation locations of  $z$  are denoted by  $t_1, \dots, t_n$  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  $\mathbf{z} := (z(t_1), \dots, z(t_n))'$  has the following multivariate Gaussian representation

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

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

$$\hat{\boldsymbol{\beta}} = \{X' \Sigma^{-1} X\}^{-1} X' \Sigma^{-1} \mathbf{z}. \quad (7)$$

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

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

where  $\Sigma_0$  is the vector  $[\text{cov}(f(t_0), f(t_j))]_{j=1}^n$ .

Notice that the estimate (8) requires computation of  $X' \Sigma^{-1} X$ ,  $\Sigma^{-1} z$ , and  $\Sigma^{-1}(z - X\hat{\beta})$ . Nominally, the inversion  $\Sigma^{-1}$  takes  $\mathcal{O}(n^3)$  operations, which is computationally prohibitive for even moderate  $n$ . However, if  $\Sigma$  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  $\Sigma = LL'$ . Now to compute  $\Sigma^{-1} z$ , for example, one simply performs two triangular linear solves: first solving the triangular system  $L x_1 = z$  to obtain  $x_1$ , then a second triangular system  $L' x_2 = x_1$  that gives  $x_2 = \Sigma^{-1} z$ . Similar triangular systems can be solved to obtain the other two quantities  $X' \Sigma^{-1} X$  and  $\Sigma^{-1}(z - X\hat{\beta})$ . Unfortunately,  $\Sigma$  is rarely sparse itself. This leads to the process of tapering  $\Sigma$ .

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

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

The advantage is that  $\Sigma_{T_{\hat{\phi}}}$  is sparse when the diameter of the compact support of  $K$  is small. Now replacing  $\Sigma^{-1}$  in (8) and (7) with  $\Sigma_{T_{\hat{\phi}}}^{-1}$  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. 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 with ordinary kriging; and finally, record prediction error and computational times. All of the following simulation times are recorded from the same computer with  $2 \times 2.93$  GHz 6-Core Intel Xeon processors and 16 GB of ram.



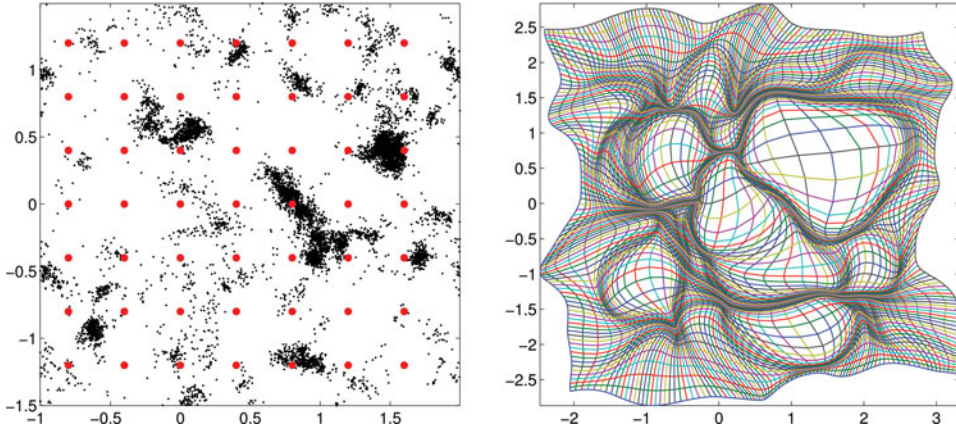


Figure 1. Left: randomly generated monitor locations (black) and the prediction grid used for evaluation (red). Right: plot of the estimated transformation  $\hat{\phi}$  used for nonstationary tapering on the monitor locations shown at left. The online version of this figure is in color.

The simulations are done in the region  $(-1.5, 1.5] \times (-1, 2]$  that contains 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. To generate random densities with sufficiently high resolution, we use a discrete fast Fourier transform 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. A realization of the random monitor locations is shown in the left plot of Figure 1 as black dots.

On each set of monitor and prediction locations, we simulated a mean zero Gaussian random field with Matérn autocovariance function given by

$$C(|h|) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left( \frac{2\sqrt{\nu}}{\rho} |h| \right)^\nu \mathcal{K}_\nu \left( \frac{2\sqrt{\nu}}{\rho} |h| \right), \quad (9)$$

where  $\Gamma(\cdot)$  is the Gamma function and  $\mathcal{K}_\nu(\cdot)$  is the modified Bessel function of second kind of order  $\nu$ . The parameter  $\rho$  controls the range of spatial correlation and  $\nu$  controls sample path smoothness (see Stein 1999). 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). 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. The random field values simulated on the prediction grid were set aside and used only for prediction evaluation.

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

$\nu = 3/2$	Stationary	Nonstationary
Min 20 points	0.055 (28.66 sec)	0.043 (26.95 sec)
Max 100 points	0.619 (8.69 sec)	0.236 (8.68 sec)

We used a fairly aggressive warping estimate  $\hat{\varphi}$  for the nonstationary tapers based on the Wendland<sub>2</sub> stationary taper (see Wendland 1995) defined as

$$\text{Wendland}_2(r) = (1 - r)_+^6(1 + 6r + 35r^2/3). \quad (10)$$

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 hr. 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 be computed multiple times. Therefore, for very large datasets, 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 mean square prediction error (MSE) over 50 independent simulations with computational time reported in parentheses (the results are similar for mean absolute error). 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 every prediction neighborhood 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 reduction in average MSE when using a nonstationary taper. 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 decrease in average MSE when using a nonstationary taper: the nonstationary taper fills sparse prediction neighborhoods that 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 overdense regions begins to dominate the computational time.

Table 2 shows the corresponding simulation results when smoothness parameter sets 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

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

$\nu = 1/2$	Stationary	Nonstationary
Min 20 points	0.137 (29.52 sec)	0.142 (28.80 sec)
Max 100 points	0.686 (8.58 sec)	0.335 (8.53 sec)

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  $t_1, \dots, t_n$  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 nonuniformity of the monitoring locations makes stationary tapers problematic, yielding some neighborhoods with too few predictive points and others with far too many for computational efficiency. This motivates using nonstationary tapers.

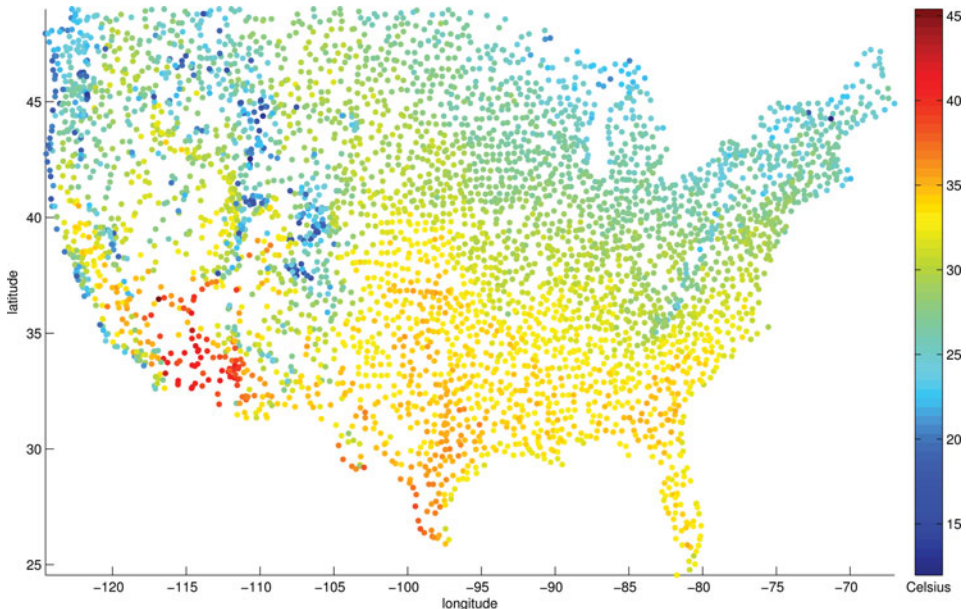


Figure 2. Mean summer daily maximum temperatures for June–August 1990.

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 four spatial covariates:  $\theta_1(t) = 1$ ;  $\theta_2(t) = \text{altitude of location } t$ ;  $\theta_3(t) = \text{longitude of location } t$ ; and  $\theta_4(t) = \text{latitude of location } t$ . We suppose  $f(t)$  is a smooth isotropic random field on  $\mathbb{R}^2$  with variance  $\tau^2$  and Matérn autocorrelation function  $C$ , defined in Equation (9), so that  $\text{cov}(f(s), f(t)) = \tau^2 C(|s - t|)$ . In this example, we set  $\rho = 0.3$ ,  $\nu = 1.5$  and estimate the ratio  $\sigma^2/\tau^2$  by cross-validation. The quantity  $\sigma^2/\tau^2$  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 a mean square differentiable random field. We remark that for modest changes in  $\nu$  and  $\rho$ , our kriging results are similar.

Due to the nonuniformity 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<sub>2</sub> autocorrelation function defined in (10),  $\hat{\varphi}$  is the estimated transformation discussed in Section 2.1, and  $b$  is the tapering bandwidth. The Wendland<sub>2</sub> autocorrelation has four continuous derivatives at the origin, compared with  $C(|h|)$  that has two. Indeed, a smoother taper than process covariance is a fundamental assumption for most of the asymptotic results found in the literature (see, e.g., Furrer, Genton, and Nychka 2006). The estimated map  $\hat{\varphi}$ , shown in Figure 3, transforms the approximate density of the monitor

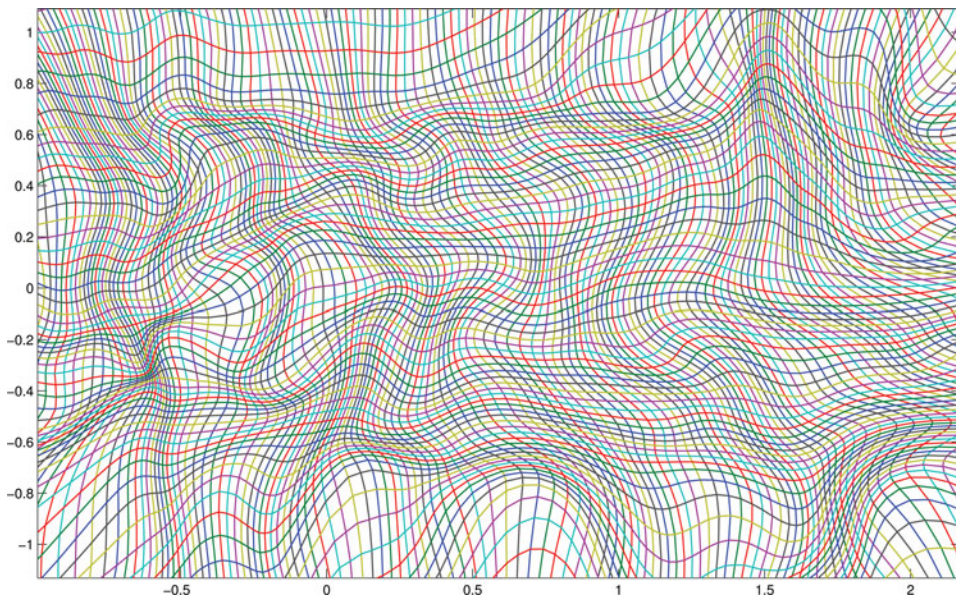


Figure 3. Plot of the estimated transformation  $\hat{\varphi}$  using the methodology described in Section 2.1. The map  $\hat{\varphi}$  transforms the monitor locations to an approximate uniform density on  $[-1, 2] \times [-2/3, 2/3]$ . The monitor locations before and after the transformation are shown in Figure 4. 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  $\lambda = 10$ ).



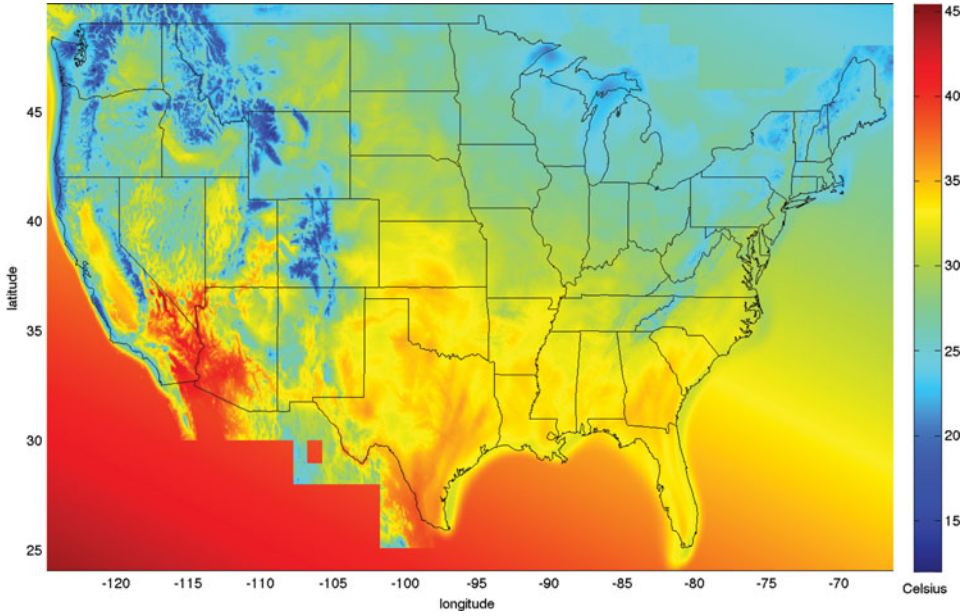


Figure 4. The kriged surface of mean summer daily maximum temperatures (shown in Figure 2) evaluated on a  $621 \times 1405$  grid using the nonstationary taper shown in Figure 3.

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 4. We used the same truncated Fourier basis and regularization penalty  $\mathcal{P}(\varphi)$  found in Anderes and Coram (2011, sec. 6). The tuning parameter  $\lambda$  was set to 10, which was picked by visual inspection of the resulting map  $\hat{\varphi}$ . A more systematic approach for choosing  $\lambda$  is possible, using cross-validation, for example. However, since we are viewing the nonstationary tapering methodology as essentially a numerical device, we used simple trial and error: looking for the largest  $\lambda$  that produced a warping 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  $\hat{\beta}$  and the kriging solution (8) only depend on  $\sigma^2$  and  $\tau^2$  through the ratio  $\sigma^2/\tau^2$ . Indeed one can rewrite the estimate  $\hat{\beta}$  as

$$\hat{\beta} = \{X'(\Xi + (\sigma^2/\tau^2)I_n)^{-1}X\}^{-1}X'(\Xi + (\sigma^2/\tau^2)I_n)^{-1}z,$$

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

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

where  $A := \Xi[\Xi + (\sigma^2/\tau^2)I_n]^{-1}$  is called as the projection matrix and  $\hat{f} := A(z - X\hat{\beta})$  is the kriging estimate of  $z - X\hat{\beta}$  at the sampling locations  $t_1, \dots, t_n$  (see Wahba 1990 for

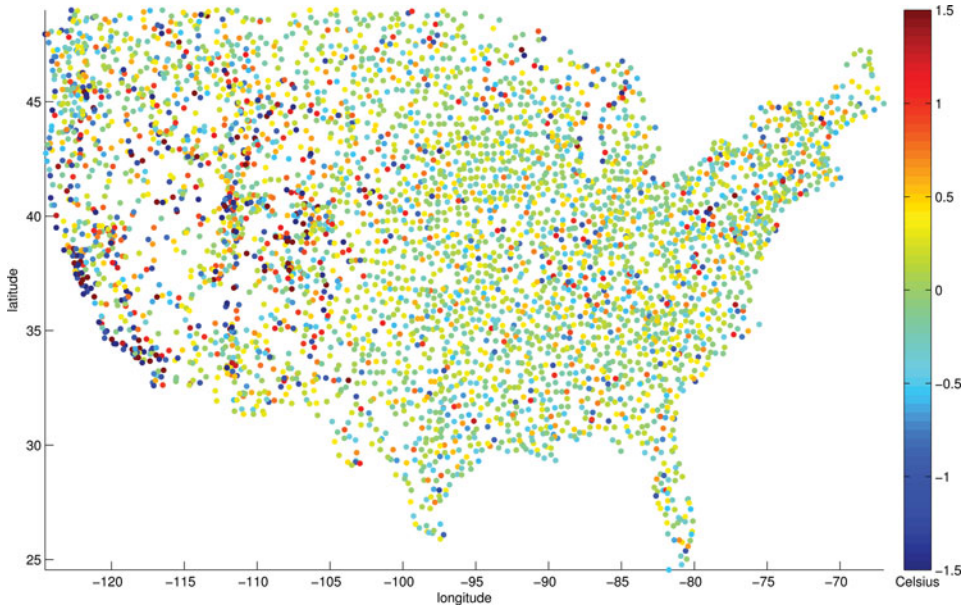


Figure 5. Residuals at each monitor location after kriging.

a discussion of cross-validation). Once again, tapering reduces the computational burden associated with the computation of  $[\Xi + (\sigma^2/\tau^2)I_n]^{-1}$  and  $\text{tr}(A)$ . For example, the trace of the matrix  $A$  can be approximated by

$$\text{tr}(A) = E(\mathbf{e}' A \mathbf{e}) \approx \frac{1}{K} \sum_{k=1}^K \mathbf{e}_k' A \mathbf{e}_k = \frac{1}{K} \sum_{k=1}^K \mathbf{e}_k' \mathbf{e}_k - \lambda (L^{-1} \mathbf{e}_k)' (L^{-1} \mathbf{e}_k), \quad (11)$$

where  $\mathbf{e}_1, \dots, \mathbf{e}_K$  are iid standard Gaussian vectors of length  $n$  and  $LL'$  is the Choleski decomposition of  $[\Xi + (\sigma^2/\tau^2)I_n]^{-1}$ . Increasing the number of samples  $\mathbf{e}_1, \dots, \mathbf{e}_K$  has the effect of improving the approximation (11) by the strong law of large numbers. The cross-validation criterion yields an estimated ratio  $\sigma^2/\tau^2$  of 0.15, which corresponds to a signal-to-noise ratio of  $\sim 6.7$ .

Figure 5 displays the kriged map of temperatures on a grid of size  $621 \times 1405$  using the nonstationary taper described above. The residuals, after kriging, are shown in Figure 6. The map tracks the observed temperature data while smoothing the noise and/or micro-scale variation. Indeed, the remaining noise at the sampling locations shows significant reduction in amplitude and spatial decorrelation compared with 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 high altitudes.

Producing the kriging surface using the nonstationary taper shown in Figure 5 takes 47 min on a desktop computer with  $2 \times 2.93$  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)

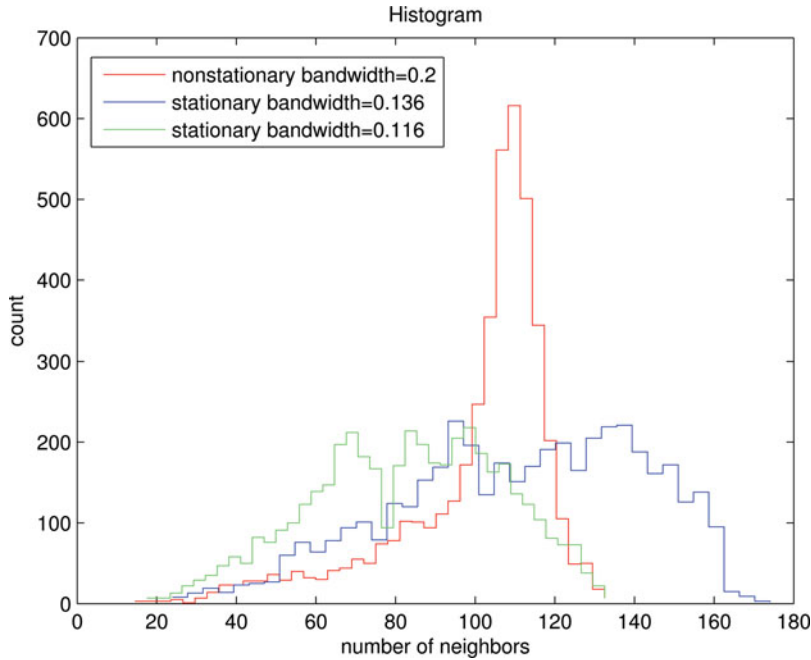


Figure 6. Histograms of the number of neighbors at each monitor using three different tapers: the nonstationary taper in red; a stationary taper with bandwidth 0.136 in blue; and a stationary taper with bandwidth 0.116 in green. See Section 5 for a discussion.

and the other that 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) that is obtained by matching the 95th percentile of the number of neighbors at each monitor with the same quantity for our nonstationary taper. The second stationary bandwidth is 0.136 that matches the 5th percentile to the same nonstationary taper quantity. The results are summarized in Figure 8 that shows 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 that relieves some of the computational difficulties associated matrix inversion, Cholesky decompositions, etc. In this article, we explore using quasi-conformal mappings or warpings for generating nonstationary tapers

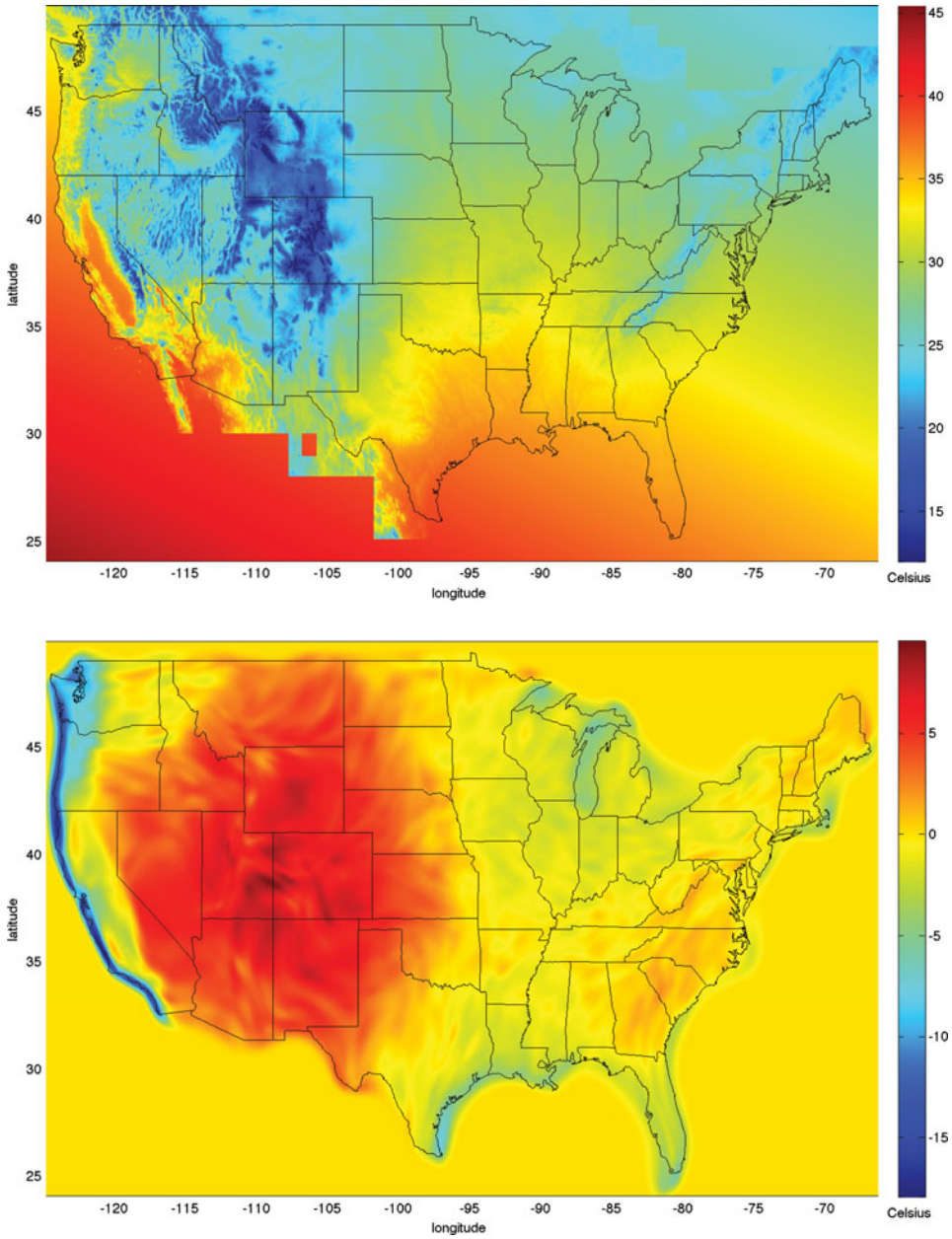


Figure 7. Top: kriging estimate of the linear trend,  $\sum_{k=1}^m \hat{\beta}_k \varphi_k(t)$ , from model (6). Bottom: estimated smooth field  $\hat{f}(x)$  from (6).

that 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 that are designed to be inversely proportional to monitor density. The result is that the expected number of monitor locations stays relatively constant



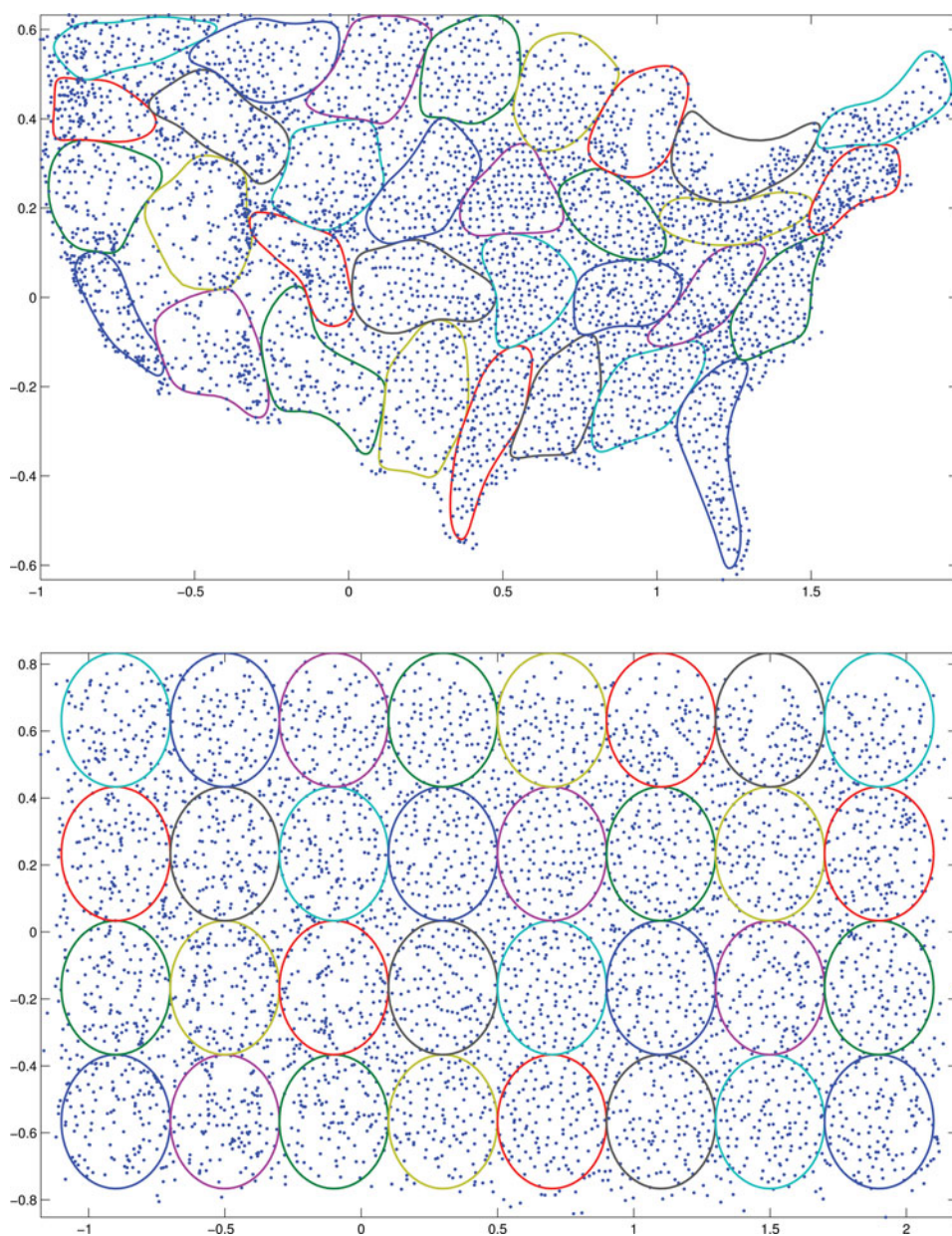


Figure 8. Top: the colored loops show the boundary of nonstationary taper neighborhoods  $\{t : T_{\hat{\phi}}(s, t) > 0\}$  at different observation locations across the USA. Bottom: this diagram shows the same neighborhoods but in the transformed space, that is,  $\hat{\phi}(\{t : T_{\hat{\phi}}(s, t) > 0\})$ .

over different prediction neighborhoods. This simultaneously relieves the computational burden of overdense prediction neighborhoods while also increasing the number of local predictors in sparsely observed regions. The quasi-conformal maps used to generate these nonstationary tapers are estimated using a penalized log-likelihood approach that provides a

smooth objective function for which gradient-based techniques are available for numerical optimization.

Using warpings or transformations to model nonstationary processes was first introduced to the spatial statistics literature in the article by Sampson and Guttorp (1992). Their work, as well as that of subsequent authors (see, e.g., Perrin and Meiring 1999; Damian, Sampson, and Guttorp 2001; Clerc and Mallat 2002; Schmidt and O'Hagan 2003; Iovleff and Perrin 2004; Anderes and Stein 2008; Anderes and Chatterjee 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. Second, this article uses a new optimization procedure developed in Anderes and Coram (2011) for estimating a warping that produces taper neighborhoods that 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 that accomplishes the same goal as those based on monitor counts.

In this article, we illustrate our nonstationary tapers through simulation studies and a numerical example. The simulation results demonstrate that nonstationary tapering can 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 article 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 that could be adequately solved by both stationary and nonstationary tapering. Regardless, the improvement in MSE points to 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.

## SUPPLEMENTARY MATERIALS

**Matlab code:** The supplemental files for this article include Matlab mfiles that can be used to produce the warping estimates described in Section 2 and the nonstationary kriging procedure described in Section 3. The supplemental files also contain all the datasets and mfiles used to produce the warping and kriging plots shown in Section 5 of this article. Please see the file README contained in the zip file for more details. (code.tar.gz, GNU zipped tar file)

## ACKNOWLEDGMENTS

Ethan Anderes is supported by NSF grant DMS-1007480. Raphaël Huser is supported by the Competence Center Environment and Sustainability (CCES) and the Swiss National Science Foundation (SNF).

[Received January 2011. Revised August 2012.]

## REFERENCES

- Ahlfors, L. (2006), *Lectures on Quasiconformal Mappings: University Lecture Series 38* (with additional chapters by eds. C. J. Earle, I. Kra, M. Shishikura, and J.H. Hubbard), Providence, RI: Amer. Math. Soc. [851,852]
- Anderes, E., and Chatterjee, S. (2009), “Consistent Estimates of Deformed Isotropic Gaussian Random Fields on the Plane,” *The Annals of Statistics*, 37, 2324–2350. [864]
- Anderes, E., and Coram, M. (2011), “Two Dimensional Density Estimation Using Smooth Invertible Transformations,” *Journal of Statistical Planning and Inference*, 141, 1183–1193. [851,852,853,856,864]
- Anderes, E., and Stein, M. (2008), “Estimating Deformations of Isotropic Gaussian Random Fields on the Plane,” *The Annals of Statistics*, 36, 719–741. [864]
- Clerc, M., and Mallat, S. (2002), “Estimating Deformations of Stationary Processes,” *The Annals of Statistics*, 31, 1772–1821. [864]
- Cressie, N. (1990), “The Origins of Kriging,” *Mathematical Geology*, 22, 239–252. [849]
- Damian, C., Sampson, P., and Guttorp, P. (2001), “Bayesian Estimation of Semi-Parametric Non-Stationary Spatial Covariance Structures,” *Environmetrics*, 12, 161–178. [864]
- Du, J., Zhang, H., and Mandrekar, V. (2009), “Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators,” *The Annals of Statistics*, 37, 3330–3361. [849]
- 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. [849,854,858]
- Iovleff, S., and Perrin, O. (2004), “Estimating a Nonstationary Spatial Structure Using Simulated Annealing,” *Journal of Computational and Graphical Statistics*, 13, 90–105. [864]
- Kaufman, C., Schervish, M., and Nychka, D. (2008), “Covariance Tapering for Likelihood-Based Estimation in Large Spatial Datasets,” *Journal of the American Statistical Association*, 103, 1545–1555. [849]
- Matheron, G. (1963), “Principles of Geostatistics,” *Economic Geology*, 58, 1246–1266. [848]
- McCann, R. (1995), “Existence and Uniqueness of Monotone Measure-Preserving Maps,” *Duke Mathematical Journal*, 80, 309–323. [850]
- Perrin, O., and Meiring, W. (1999), “Identifiability for Non-Stationary Spatial Structure,” *Journal of Applied Probability*, 36, 1244–1250. [864]
- Sampson, P., and Guttorp, P. (1992), “Nonparametric Estimation of Nonstationary Spatial Covariance Structure,” *Journal of the American Statistical Association*, 87, 108–119. [864]
- Schmidt, A., and O’Hagan, A. (2003), “Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations,” *Journal of the Royal Statistical Society, Series B*, 65, 745–758. [864]
- Shaby, B., and Ruppert, D. (2012), “Tapered Covariance: Bayesian Estimation and Asymptotics,” *Journal of Computational and Graphical Statistics*, 21, 433–452. [849]
- Stein, M. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer. [855]
- Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [849,858,859]
- Wendland, H. (1995), “Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree,” *Advances in Computational Mathematics*, 4, 389–396. [850,856]
- Zhang, H., and Du, J. (2008), “Covariance Tapering in Spatial Statistics,” in *Positive Definite Functions: From Schoenberg to Space-Time Challenges*, eds. J. Mateu and E. Porcu, Castelló, de la Plana, Spain: University Jaume I, pp. 181–196. [849]