Supplementary materials for this article are available at 10.1007/s13253-010-0050-6.

Spatial Regression Using Kernel Averaged Predictors

Matthew J. HEATON and Alan E. GELFAND

Traditional spatial linear regression models assume that the mean of the response is a linear combination of predictors measured at the same location as the response. In spatial applications, however, it seems plausible that neighboring predictors can also inform about the response. This article proposes using unobserved kernel averaged predictors in such regressions. The kernels are parametric introducing additional parameters that are estimated with the data. Properties and challenges of using kernel averaged predictors within a regression model are detailed in the simple case of a univariate response and a single predictor. Additionally, extensions to multiple predictors and generalized linear models are discussed. The methods are demonstrated using a data set of dew duration and shrub density. Supplemental materials are available online.

Key Words: Block averaging; Circular neighborhoods; Distributed covariates; Multivariate Gaussian process; Spatial linear model.

1. INTRODUCTION

Statistical methods and models for spatial data are plentiful, and panoramic overviews of spatial methodology can be found in the books by Cressie (1993), Chiles and Delfiner (1999), Wackernagel (2003), and Banerjee, Carlin, and Gelfand (2004). In spatial problems, a response variable is measured at various locations over a region $\mathcal{D}$. The response variable is correlated in space such that observations closer in space tend to be more correlated than observations farther away in space. For continuous responses, a normality assumption (perhaps on a transformation of the response variable) is typically used, and spatial correlation is induced via Gaussian processes. For binary or count data, spatial correlation is typically induced via random effects on the mean function as in Diggle, Tawn, and Moyeed (1998). This article primarily focuses on continuous response data, but extensions to binary or count data are discussed in Section 5.

Matthew J. Heaton is Ph.D. Student (E-mail: *matt@stat.duke.edu*) and Alan E. Gelfand (✉) is Professor (E-mail: *alan@stat.duke.edu*), Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251, USA

Journal of Agricultural, Biological, and Environmental Statistics, Volume 16, Number 2, Pages 233–252
DOI: 10.1007/s13253-010-0050-6

In many spatial applications, a vector of predictors (covariates) is also available, and the goal is to build a regression model for the response while accounting for spatial correlation in the response, the predictors, or both. For such spatial regression models, predictors are most commonly incorporated linearly via the mean function such that $\mathbb{E}(Y(\boldsymbol{s}) \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{X}(\boldsymbol{s})) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}(\boldsymbol{s})$, where $Y(\boldsymbol{s})$ is a univariate response at location $\boldsymbol{s} \in \mathcal{D}$, $\boldsymbol{X}(\boldsymbol{s}) = (X_1(\boldsymbol{s}), \ldots, X_p(\boldsymbol{s}))^T$ is a vector of p predictor variables measured at location $\boldsymbol{s}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of coefficients. An extension of the spatial regression model introduces local linearity through spatially varying coefficients as in Gelfand et al. (2003), who assume that $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\beta}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})) = \beta_0(\boldsymbol{s}) + \boldsymbol{\beta}^T(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})$ so that the coefficients are different for all $\boldsymbol{s}$.

In formulating a spatial regression model it is often necessary to assume that $\boldsymbol{X}(\boldsymbol{s})$ is random. This enables handling of the full dataset including observations where an X or Y is missing and hence handling *misalignment* in the dataset (see Banerjee, Carlin, and Gelfand 2004, Chapter 6). It also enables handling of the inverse problem, learning about $\boldsymbol{X}(\boldsymbol{s}_0)$ for an observed $Y(\mathbf{s}_0)$. This implies a joint model for $\boldsymbol{Z}(\boldsymbol{s}) = (Y(\boldsymbol{s}), \boldsymbol{X}^T(\boldsymbol{s}))^T$ as developed, for instance, in Royle and Berliner (1999), Berliner (2000), or Gelfand et al. (2004). Analogous to the case where the $\boldsymbol{X}(\boldsymbol{s})$ are viewed as fixed, the local spatial regression is usually defined in terms of the conditional distribution of $Y(\boldsymbol{s})$ given $\boldsymbol{X}(\boldsymbol{s})$ (see Banerjee, Carlin, and Gelfand 2004, Chapter 7) in the form of the conditional mean $E(Y(\boldsymbol{s})|\boldsymbol{X}(\boldsymbol{s})) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}(\boldsymbol{s})$, as above. However, in spatial applications, *neighboring* $\boldsymbol{X}(\boldsymbol{s}')$ can be expected to inform about $Y(\boldsymbol{s})$ particularly when the distance between $\boldsymbol{s}$ and $\boldsymbol{s}'$ is small. To cite a few examples, precipitation can affect the water table (the depth at which soil and pour spaces become completely saturated with water) at not only where the precipitation fell but also surrounding locations due to run off and changes in slope from uneven ground surfaces. Similarly, the concentration of ozone (O_3) is affected by pollutants, ultraviolet rays, and temperature within a neighborhood of location $\boldsymbol{s}$. Thus, a mean specification which only includes $\boldsymbol{X}(\boldsymbol{s})$ and not neighboring $\boldsymbol{X}(\boldsymbol{s}')$ may not adequately capture the process.

A mean specification which incorporates neighboring predictors might take the form $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \boldsymbol{X}) = \beta_0(\boldsymbol{s}) + \int_{\mathcal{D}} \boldsymbol{\beta}(\boldsymbol{s}, \boldsymbol{u}|\boldsymbol{\theta})^T \boldsymbol{X}(\boldsymbol{u})\, d\boldsymbol{u}$, where $\boldsymbol{\beta}(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\theta})$ is a coefficient model capturing the effect of $\boldsymbol{X}(\boldsymbol{s}')$ on $Y(\boldsymbol{s})$ with parameters $\boldsymbol{\theta}$. In the interest of clarification, with, say n observations $\boldsymbol{Z} = (\boldsymbol{Z}(\boldsymbol{s}_1), \ldots, \boldsymbol{Z}(\boldsymbol{s}_n))^T$, we might consider the conditional distribution, $[Y(\boldsymbol{s}_i) \mid \boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)]$, where $[\cdot]$ denotes a generic probability distribution. In the Gaussian setting, $[Y(\boldsymbol{s}_i) \mid \boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)]$ would contain n coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i, \boldsymbol{s}_1), \ldots, \boldsymbol{\beta}(\boldsymbol{s}_i, \boldsymbol{s}_n)$ which describe the loading of $\boldsymbol{X}(\boldsymbol{s}_j)$ on $Y(\boldsymbol{s}_i)$ for $j = 1, \ldots, n$. Such a choice is analogous to spatial prediction or kriging and is too limited for our goal because the explained relationship between the response and predictor is dependent on the number of sampling locations (n) and the arrangement of those locations in $\mathcal{D}$. It also fails to capture the idea of local $\boldsymbol{X}(\boldsymbol{s}')$ informing about $Y(\boldsymbol{s})$.

Below, we argue that kernels are an appropriate choice for $\boldsymbol{\beta}(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\theta})$ and present associated models. We offer full distributional detail and properties, show how to fit these models, and illustrate benefits using real and simulated data. We close this section with three remarks. First, we are not interested in what would be called co-kriging, i.e., prediction of $Y(\boldsymbol{s}_0)$ given the entire set of observed X's and Y's. When used for spatial prediction, the models considered here offer only minimal additional benefit from what would be

obtained using customary multivariate Gaussian processes. This is not surprising. We are introducing unobserved covariates into our regression specification, but we are still fitting a multivariate Gaussian process model to the data. We offer further discussion of this point in Section 2.2 below. Second, what we are doing differs from what is usually referred to as employing functional covariates (see Baíllo and Grané 2009, and the references therein). Functional covariates envision a function at location $\boldsymbol{s}$, say, $\boldsymbol{X}(\boldsymbol{s}, t)$ for $t \in (0, T]$, and seek to reduce this to a single covariate at $\boldsymbol{s}$ to explain $Y(\boldsymbol{s})$. Usually, this is achieved through some integration of the function wherein the integration is over t rather than over $\boldsymbol{s}$ as we seek. Third, we do not seek to build process models for $\boldsymbol{\beta}(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\theta})$. We provide process modeling for $(Y(\mathbf{s}), \boldsymbol{X}(\boldsymbol{s}))$ but specify $\boldsymbol{\beta}$ as a parametric function. Specifying the latter using processes will provide poorly-identified, over-fitted models.

Section 2 develops the methodology for the simplest case of a univariate response and a single predictor, and Section 3 presents the associated model fitting. Section 4 uses this methodology on simulated and field data collected over the Negev Desert in Israel. Section 5 discusses extensions of the methodology to multiple predictors and general (e.g., count or binary) response variables. Conclusions and future work are discussed in Section 6.

2. SPATIAL REGRESSION WITH A SINGLE AVERAGED COVARIATE

2.1. THE MODEL

Let $Y(\boldsymbol{s})$ and $X(\boldsymbol{s})$ denote a univariate response variable and a single covariate at location $\boldsymbol{s} \in \mathcal{D} \subseteq \mathbb{R}^d$ for $d \in \mathbb{Z}$, respectively. Furthermore, assume that $X(\boldsymbol{s})$ follows a Gaussian process (GP) of the form

$$X(\boldsymbol{s}) = \mu_X(\boldsymbol{s}) + \sigma_X w_X(\boldsymbol{s}), \tag{2.1}$$

where $\mu_X(\boldsymbol{s})$ is the mean surface at location $\boldsymbol{s}$, and $w_X(\boldsymbol{s})$ is a mean 0, variance 1 GP with correlation function $\operatorname{corr}(w_X(\boldsymbol{s}), w_X(\boldsymbol{s}')) = \rho_X(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\phi}_X)$, where $\boldsymbol{\phi}_X$ denotes the (possibly vector-valued) parameter associated with ρ_X. For example, if ρ_X is the Matérn (Matérn 1986) correlation function, then $\boldsymbol{\phi}_X = (\psi_X, \nu_X)^T$, where ψ_X is the decay parameter, and ν_X is the smoothness parameter. Notice that (2.1) defines a purely spatial covariate process. In some applications, however, covariates are measured with error. In these cases, $X(\boldsymbol{s})$ can be thought of as the "true" underlying covariate process, and the observed covariate is $H(\boldsymbol{s}) = X(\boldsymbol{s}) + \epsilon_X(\boldsymbol{s})$, where $\epsilon_X(\boldsymbol{s})$ is an iid white noise process. Distributional results for $H(\boldsymbol{s})$ will differ from the those of $X(\boldsymbol{s})$ by an additive nugget variance term only, so, for simplicity, we assume that $X(\boldsymbol{s})$ is observed directly for the remainder of this article. Moreover, we still want to use $X(\boldsymbol{s})$ to explain $Y(\boldsymbol{s})$, i.e., we want the regression of $Y(\boldsymbol{s})$ to be on the true $X(\boldsymbol{s})$.

Following the discussion in the Introduction, we define the unobserved local covariate at $\boldsymbol{s}$ incorporating neighbor information as $\widetilde{X}(\boldsymbol{s})$ using a kernel function, i.e.,

$$\widetilde{X}(\boldsymbol{s}) \equiv \frac{1}{K(\boldsymbol{s} \mid \boldsymbol{\xi})} \int_{\mathcal{D}} K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi}) X(\boldsymbol{u})\, d\boldsymbol{u}, \tag{2.2}$$

where $K(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi})$ is a kernel defining a weight on the distance between $\boldsymbol{s}$ and $\boldsymbol{s}'$ with parameters $\boldsymbol{\xi}$ and $0 < K(\boldsymbol{s} \mid \boldsymbol{\xi}) = \int_{\mathcal{D}} K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi})\, d\boldsymbol{u} < \infty$. The parameter $\boldsymbol{\xi}$, most commonly, consists of *scale* parameters such as the entries of a covariance matrix but can also include *location* parameters as in Higdon (1998) and Xu, Wikle, and Fox (2005). Choices for K are discussed below. By (2.2), $\widetilde{X}(\boldsymbol{s})$ is the kernel averaged value of $X(\boldsymbol{s})$ over the domain $\mathcal{D}$. Again, we note that the integration is over $\mathcal{D}$, i.e., $X(\boldsymbol{s})$ is not viewed as a functional covariate. The latter would take the form, say $X(\boldsymbol{s}, t)$, whence the integration would be over t.

Because a valid GP was defined for $X(\boldsymbol{s})$, $\widetilde{X}(\boldsymbol{s})$ also is a valid GP with mean $\widetilde{\mu}_X(\boldsymbol{s}) = K(\boldsymbol{s} \mid \boldsymbol{\xi})^{-1} \int_{\mathcal{D}} K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi}) \mu_X(\boldsymbol{u})\, d\boldsymbol{u}$ and

$$\begin{aligned} \mathbb{C}\mathrm{ov}(\widetilde{X}(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}')) &= \frac{\sigma_X^2}{K(\boldsymbol{s} \mid \boldsymbol{\xi}) K(\boldsymbol{s}' \mid \boldsymbol{\xi})} \int_{\mathcal{D}} \int_{\mathcal{D}} K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi}) K(\boldsymbol{s}', \boldsymbol{v} \mid \boldsymbol{\xi}) \rho_X(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{v}\, d\boldsymbol{u} \\ &\equiv \sigma_X^2 \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}'). \end{aligned} \tag{2.3}$$

Not only are $X(\boldsymbol{s})$ and $\widetilde{X}(\boldsymbol{s})$ marginally Gaussian processes, but a valid bivariate GP is induced for the pair $(X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))^T$. Specifically, $(X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))^T$ follows a bivariate GP with mean $(\mu_X(\boldsymbol{s}), \widetilde{\mu}_X(\boldsymbol{s}))^T$ and $\mathbb{C}\mathrm{ov}((X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))^T, (X(\boldsymbol{s}'), \widetilde{X}(\boldsymbol{s}'))^T)$ given by

$$\begin{aligned} \boldsymbol{\Sigma}_{X\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') &= \sigma_X^2 \begin{pmatrix} \rho_X(\boldsymbol{s}, \boldsymbol{s}') & \frac{1}{K(\boldsymbol{s}' \mid \boldsymbol{\xi})} \int_{\mathcal{D}} K(\boldsymbol{s}', \boldsymbol{u} \mid \boldsymbol{\xi}) \rho_X(\boldsymbol{s}, \boldsymbol{u})\, d\boldsymbol{u} \\ \frac{1}{K(\boldsymbol{s} \mid \boldsymbol{\xi})} \int_{\mathcal{D}} K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi}) \rho_X(\boldsymbol{u}, \boldsymbol{s}')\, d\boldsymbol{u} & \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \end{pmatrix} \\ &\equiv \sigma_X^2 \begin{pmatrix} \rho_X(\boldsymbol{s}, \boldsymbol{s}') & \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \\ \rho_{\widetilde{X},X}(\boldsymbol{s}, \boldsymbol{s}') & \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \end{pmatrix} \end{aligned} \tag{2.4}$$

for any other location $\boldsymbol{s}' \in \mathcal{D}$.

To account for effects of $\{X(\boldsymbol{s}') : \boldsymbol{s}' \in \mathcal{D}\}$ on $Y(\boldsymbol{s})$, consider the linear model defined by

$$Y(\boldsymbol{s}) \mid X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}) = \beta_0 + \beta_1 \widetilde{X}(\boldsymbol{s}) + \sigma_Y w_Y(\boldsymbol{s}) + \epsilon_Y(\boldsymbol{s}), \tag{2.5}$$

where $w_Y(\boldsymbol{s})$ is defined analogously to $w_X(\boldsymbol{s})$ in (2.1) but with correlation function ρ_Y, and $\epsilon_Y(\boldsymbol{s})$ is a Gaussian white noise process with variance τ_Y^2. Intuitively, the kernel $K(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi})$ is used here to describe how the effect of the covariate $X(\boldsymbol{s})$ propagates to the response. In contrast, the kernels used in the integro-difference equation models of Wikle (2002) and Xu, Wikle, and Fox (2005) describe how the spatio-temporal process, say $Y(\boldsymbol{s}, t)$, propagates in time.

Notice that the bivariate GP for $(X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))^T$ above, together with (2.5), provides a joint specification of a trivariate GP for $\boldsymbol{Z}(\boldsymbol{s}) = (X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}), Y(\boldsymbol{s}))^T$. Such sequential specification is discussed in Royle and Berliner (1999), Berliner (2000), and Gelfand et al. (2004). Specifically, $\boldsymbol{Z}(\boldsymbol{s})$ follows a valid trivariate GP with mean $\boldsymbol{\mu}(\boldsymbol{s}) = (\mu_X(\boldsymbol{s}), \widetilde{\mu}_X(\boldsymbol{s}), \beta_0 + \beta_1 \widetilde{\mu}_X(\boldsymbol{s}))^T$ and cross-covariance

$$\sigma_X^2 \begin{pmatrix} \rho_X(\boldsymbol{s}, \boldsymbol{s}') & \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') & \beta_1 \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \\ \rho_{\widetilde{X},X}(\boldsymbol{s}, \boldsymbol{s}') & \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') & \beta_1 \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \\ \beta_1 \rho_{\widetilde{X},X}(\boldsymbol{s}, \boldsymbol{s}') & \beta_1 \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') & \frac{\sigma_Y^2}{\sigma_X^2} \rho_Y(\boldsymbol{s}, \boldsymbol{s}') + \beta_1^2 \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}') \end{pmatrix}. \tag{2.6}$$

Table 1. Examples of kernel functions $K(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi})$ and their parameters. $\mathbb{I}_{\{\mathcal{A}\}}$ is the indicator of a set $\mathcal{A}$.

Kernel	$K(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi})$	Parameters ($\boldsymbol{\xi}$)
Uniform	$\mathbb{I}_{\{\|\boldsymbol{s}-\boldsymbol{s}'\| \le \xi\}}$	ξ
Epanechnikov	$(\xi^2 - \|\boldsymbol{s}-\boldsymbol{s}'\|^2)\mathbb{I}_{\{\|\boldsymbol{s}-\boldsymbol{s}'\| \le \xi\}}$	ξ
Component Wise Gaussian	$\prod_{i=1}^{d} \xi_i^{-1} \exp\{-(s_i - s_i')^2/(2\xi_i^2)\}$	$\xi_i,\ i = 1, \ldots, d$
Oriented Gaussian	$\lvert\boldsymbol{\Xi}\rvert^{-1/2} \exp\{-\frac{(\boldsymbol{s}'-\boldsymbol{s})^T \boldsymbol{\Xi}^{-1}(\boldsymbol{s}'-\boldsymbol{s})}{2}\}$	$\boldsymbol{\Xi} = \{\xi_{ij}\}$

The joint distribution of $\boldsymbol{Z}(\boldsymbol{s})$ is useful for evaluating the properties of (2.5) such as induced correlations between $Y(\boldsymbol{s})$ and $(X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))$ (see Section 2.2). Note also that the induced bivariate process model for $(X(\boldsymbol{s}), Y(\boldsymbol{s}))$ is not a usual bivariate process specification; the kernel appears in the covariance structure.

Analogous to the discussion in Section 1, while we have a joint process model for $\boldsymbol{Z}(\boldsymbol{s})$, as above, we focus on the *local* spatial regression in terms of the conditional distribution of $Y(\boldsymbol{s})$ given $\widetilde{X}(\boldsymbol{s})$ in the form of the conditional mean $E(Y(\boldsymbol{s})|\widetilde{X}(\boldsymbol{s})) = \beta_0 + \beta_1 \widetilde{X}(\boldsymbol{s})$. Thus, given $\widetilde{X}(\boldsymbol{s})$, concepts such as R^2, mean square error, variable selection, shrinkage, etc. are applicable. Of course, all are random and would be averaged over the distribution of $\widetilde{X}(\boldsymbol{s})$ in order to interpret them. Evidently, the potentially complex relationship between $\{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ and $Y(\boldsymbol{s})$ is captured through a single parameter (β_1). However, from (2) we see that, effectively, we are introducing a coefficient weighting of the entire surface $X(\boldsymbol{s})$ to explain $Y(\boldsymbol{s})$. That is, the coefficient of $X(\boldsymbol{s}')$ is $\beta_1 K(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi})/K(\boldsymbol{s} \mid \boldsymbol{\xi})$. The normalization by $K(\boldsymbol{s} \mid \boldsymbol{\xi})$ identifies β_1. Extending β_1 to $\beta_1(\boldsymbol{s})$ using spatially varying coefficients modeling as described in Gelfand et al. (2003) could be envisioned to create an arbitrarily rich specification. However, such models are expected to be poorly identified since they introduce the product form $\beta(\boldsymbol{s})\widetilde{X}(\boldsymbol{s})$ where each variable comes from a process model but neither is observed. Our proposed form through β_1 and $\boldsymbol{\xi}$ is parsimonious and well identified.

The choice of K with the data informing about $\boldsymbol{\xi}$ enables a fairly rich regression specification while attractively reducing to a simple linear regression model in $\widetilde{X}(\boldsymbol{s})$. Examples of kernels are given in Table 1, and kernel selection is explored in more detail in Section 4.2. We note that computation of $K(\boldsymbol{s} \mid \boldsymbol{\xi})$ for general regions $\mathcal{D}$ varies with $\boldsymbol{s}$ and can be computationally expensive when done repeatedly over MCMC iterations. For much of the following, the kernel is taken to be $K(\boldsymbol{s}, \boldsymbol{s}' \mid \xi) = \mathbb{I}_{\{\|\boldsymbol{s}-\boldsymbol{s}'\| \le \xi g(\boldsymbol{\phi}_X)\}}$, where $\mathbb{I}_{\{\mathcal{A}\}}$ is the indicator for a set $\mathcal{A}$, $\|\cdot\|$ denotes Euclidean distance, $\xi \in (0, 1)$, and $g(\boldsymbol{\phi}_X)$ is the effective spatial range associated with ρ_X; for example, the exponential correlation function has $g(\boldsymbol{\phi}_X) \approx 3/\phi_X$, where ϕ_X is the spatial decay parameter. Intuitively, the kernel $K(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$ is a disk centered at location $\boldsymbol{s}$ with radius $r = \xi g(\boldsymbol{\phi}_X)$. The scale parameter r can be loosely interpreted as a hard threshold "decay" parameter in that the effect of $X(\boldsymbol{s}')$ on $Y(\boldsymbol{s})$ is negligible if $\|\boldsymbol{s} - \boldsymbol{s}'\| \ge r$. Given $K(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$ as above, $K(\boldsymbol{s} \mid \xi) = |\{\boldsymbol{s}' : \|\boldsymbol{s} - \boldsymbol{s}'\| \le r\} \cap \mathcal{D}|$, where $|A|$ denotes the area of A. For model fitting, if we make $\mathcal{D}$ large enough so that each $\boldsymbol{s}_i$ in the sample and for all r of interest the disks are contained in $\mathcal{D}$, then $K(\boldsymbol{s}_i \mid \xi) = \pi r^2$. This facilitates model fitting, but if we seek to

interpolate $\mathbb{E}(Y(\boldsymbol{s}_0) \mid \widetilde{X}(\boldsymbol{s}_0))$ for $\boldsymbol{s}_0$ near the boundary of $\mathcal{D}$, then we will have to use the general form $K(\boldsymbol{s} \mid \xi)$. Moreover, if r is large enough so that the disks contain $\mathcal{D}$, then $\widetilde{X}(\boldsymbol{s}) = |\mathcal{D}|^{-1} \int_{\mathcal{D}} X(\boldsymbol{u})\, d\boldsymbol{u}$. In other words, for large r, $\widetilde{X}(\boldsymbol{s}) \approx \widetilde{X}(\boldsymbol{s}')$ for all $\boldsymbol{s}, \boldsymbol{s}'$ yielding a highly collinear regression.

With the choice of $K(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$ above, r and $\boldsymbol{\phi}_X$ are strongly associated parameters; this is evident from the forms in (2.6). For illustration, let $\rho_X(\cdot) = \exp\{-\phi_X \|\boldsymbol{s} - \boldsymbol{s}'\|\}$ be the exponential correlation function. Figure 1 displays contour plots of $\operatorname{corr}(Y(\boldsymbol{s}), X(\boldsymbol{s}))$ and $\operatorname{corr}(Y(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))$ for various values of ϕ_X and r with $\beta_1 = \sigma_X^2 = \sigma_Y^2 = 1$. In general, holding the other parameters fixed, $\operatorname{corr}(Y(\boldsymbol{s}), X(\boldsymbol{s}))$ and $\operatorname{corr}(Y(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))$ decrease non-linearly as r and ϕ_X increase, but the rate of decay is greater for $\operatorname{corr}(Y(\boldsymbol{s}), X(\boldsymbol{s}))$. Indeed, Figure 1 suggests that plausible values for scale parameters of kernels depend on ϕ_X as well as $\mathcal{D}$. For the K above, parameterizing the scale parameter as $r = \xi g(\boldsymbol{\phi}_X)$ where $\xi \in (0, 1)$ removes this dependency such that the prior distribution for ξ and ϕ_X can be taken as $[\xi][\phi_X]$ and restricts K to be within the effective spatial range of ρ_X. In addition, we choose $[\xi]$ and $[\phi_X]$ to be discrete to help with identifiability (see Zhang 2004) and to make computation easier (see Section 3). Similar parameterizations of scale parameters can be envisioned for other kernels (see Section 4.2).

2.2. USING $X(\boldsymbol{s})$ INSTEAD OF $\widetilde{X}(\boldsymbol{s})$

We now consider what (2.5) implies about using $X(\boldsymbol{s})$ instead of $\widetilde{X}(\boldsymbol{s})$ in the conditional mean for $Y(\boldsymbol{s}) \mid X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s})$ assuming that $\mathbb{E}(Y(\boldsymbol{s}) \mid X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s})) = \beta_0 + \beta_1 \widetilde{X}(\boldsymbol{s})$. Similarly, we could investigate the consequences of assuming model (2.5) when, in fact, $\mathbb{E}(Y(\boldsymbol{s}) \mid X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s})) = \beta_0 + \beta_1 X(\boldsymbol{s})$. We focus on the former model misspecification here, but results for the latter are given in the supplementary materials.

Using the fact that $\boldsymbol{Z}(\boldsymbol{s}) = (X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}), Y(\boldsymbol{s}))^T$ follows a trivariate GP with known mean and covariance given by (2.6), multivariate normal theory gives that $Y(\boldsymbol{s}) \mid X(\boldsymbol{s})$ is also normally distributed with mean and variance,

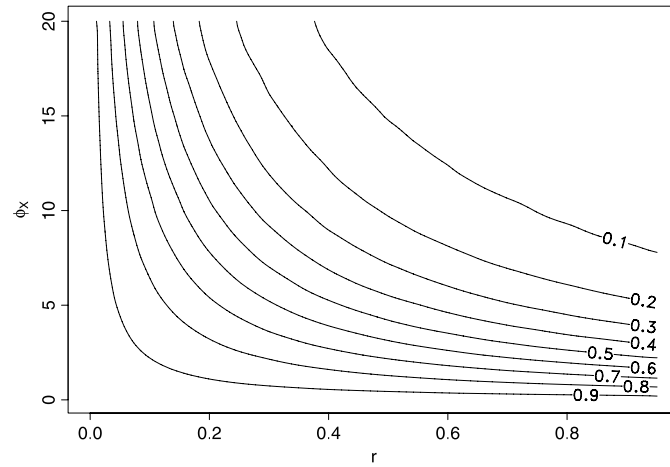
$$\mathbb{E}(Y(\boldsymbol{s}) \mid X(\boldsymbol{s})) = \beta_0 + \beta_1(\widetilde{\mu}_X(\boldsymbol{s}) - \mu_X(\boldsymbol{s})\rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s})) + \beta_1 \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}) \times X(\boldsymbol{s}), \tag{2.7}$$

$$\mathbb{V}\mathrm{ar}(Y(\boldsymbol{s}) \mid X(\boldsymbol{s})) = \tau_Y^2 + \sigma_Y^2 + \beta_1^2 \sigma_X^2 \big(\rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}) - \rho_{X,\widetilde{X}}^2(\boldsymbol{s}, \boldsymbol{s})\big), \tag{2.8}$$

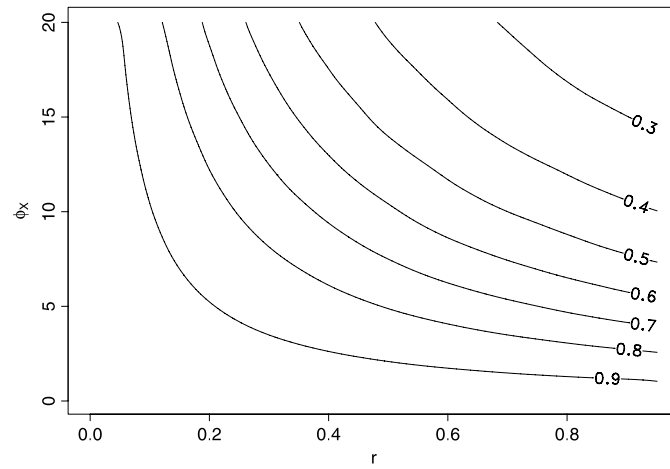
and

$$\mathbb{C}\mathrm{ov}(Y(\boldsymbol{s}), Y(\boldsymbol{s}') \mid X(\boldsymbol{s}), X(\boldsymbol{s}')) = \sigma_Y^2 \rho_Y(\boldsymbol{s}, \boldsymbol{s}') + \beta_1^2 \sigma_X^2 \rho_{\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}'). \tag{2.9}$$

Notice that when the model given by (2.5) holds, then the change in $Y(\boldsymbol{s})$ as a result of a unit change in $X(\boldsymbol{s})$ is $\beta_1 \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s})$ as opposed to β_1 when using $\widetilde{X}(\boldsymbol{s})$. Hence, under the trivariate Gaussian process model, i.e., when the assumptions of (2.5) hold, using $X(\boldsymbol{s})$ implies that the effect of the covariate on $Y(\boldsymbol{s})$ is shrunk towards zero; $|\beta_1 \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s})| \leq |\beta_1|$ because $0 \leq \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s}) \leq 1$. This result is not surprising in that if other $X(\boldsymbol{s}')$ in $\mathcal{D}$ besides $X(\boldsymbol{s})$ affect $Y(\boldsymbol{s})$, then the effect due to $X(\boldsymbol{s})$ is expected to diminish. The amount of shrinkage is determined by the kernel parameters, $\boldsymbol{\xi}$, as well as $\boldsymbol{\phi}_X$. Figure 2 displays $1 - \rho_{X,\widetilde{X}}(\boldsymbol{s}, \boldsymbol{s})$ where $\mathcal{D} = \mathbb{R}^2$, $K(\boldsymbol{s}, \boldsymbol{s}' \mid \xi, \phi_X) = \mathbb{I}_{\{\|\boldsymbol{s} - \boldsymbol{s}'\| \leq r\}}$ with $r = \xi g(\phi_X)$ and $\xi \in (0, 1)$, ρ_X is the exponential correlation function, and $\boldsymbol{\phi}_X = 10$ is the spatial decay parameter.



(a)



(b)

Figure 1. Contour plots of (a) $\text{corr}(Y(\boldsymbol{s}), X(\boldsymbol{s}))$ and (b) $\text{corr}(Y(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}))$ as functions of r and ϕ_X. Both correlations decrease as r and ϕ_X increase, but the rate of decay is much faster for $\text{corr}(Y(\boldsymbol{s}), X(\boldsymbol{s}))$.

Figure 2 shows that the amount of shrinkage could be severe depending on the value of ξ. To appreciate the implications of this, suppose, for example, that $Y(\boldsymbol{s})$ is ozone and $X(\boldsymbol{s})$ is temperature (since high temperatures are known to encourage the formation of ozone). Then using $X(\boldsymbol{s})$ in the model could, potentially, underestimate the change in $Y(\boldsymbol{s})$ as a result from a unit change in temperature, leading to underestimation of the production of ozone.

A second consequence of using $X(\boldsymbol{s})$ instead of $\widetilde{X}(\boldsymbol{s})$ when (2.5) holds is that the percent of variation in $Y(\boldsymbol{s})$ explained by $X(\boldsymbol{s})$ is less than the percent of variation in $Y(\boldsymbol{s})$ explained by $\widetilde{X}(\boldsymbol{s})$ for many common covariance functions as detailed by the following result.

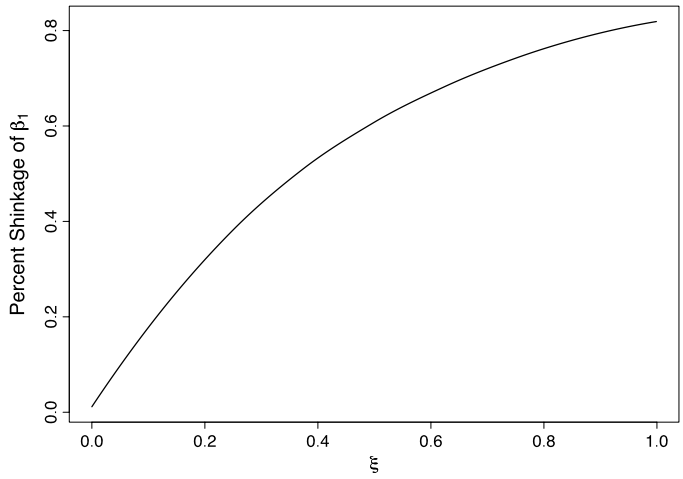


Figure 2. Percent shrinkage of β_1 against ξ assuming $\phi_X = 10$. The rapid decay of $\rho_{X,\widetilde{X}}(\boldsymbol{s},\boldsymbol{s})$ as ξ increases leads to a smaller estimated effect of $X(\boldsymbol{s})$ on the response.

Result 1. *Let $\rho^2_{Y|X}$ and $\rho^2_{Y|\widetilde{X}}$ be the population coefficients of determination for the linear models defined by (2.7) and (2.5), respectively. If ρ_X is an isotropic, log-concave correlation function, then $\rho^2_{Y|X} \le \rho^2_{Y|\widetilde{X}}$.*

Proof: For convenience, write $\mathbb{V}\mathrm{ar}(Y(\boldsymbol{s})) = \kappa^2$. Using the variance–covariance matrix of $(X(\boldsymbol{s}), \widetilde{X}(\boldsymbol{s}), Y(\boldsymbol{s}))^T$ given in (2.6), we then have

$$\begin{aligned}
\rho^2_{Y|X} &= \frac{\sigma_X^2\beta_1^2}{\kappa^2}\rho^2_{X,\widetilde{X}}(\boldsymbol{s},\boldsymbol{s}) \\
&= \frac{\sigma_X^2\beta_1^2}{\kappa^2 K(\boldsymbol{s}\mid\boldsymbol{\xi})^2}\int_{\mathcal{D}}\int_{\mathcal{D}} K(\boldsymbol{s},\boldsymbol{u}\mid\boldsymbol{\xi})\rho_X(\|\boldsymbol{s}-\boldsymbol{u}\|)K(\boldsymbol{s},\boldsymbol{v}\mid\boldsymbol{\xi})\rho_X(\|\boldsymbol{v}-\boldsymbol{s}\|)\,d\boldsymbol{u}\,d\boldsymbol{v} \\
&\le \frac{\sigma_X^2\beta_1^2}{\kappa^2 K(\boldsymbol{s}\mid\boldsymbol{\xi})^2}\int_{\mathcal{D}}\int_{\mathcal{D}} K(\boldsymbol{s},\boldsymbol{u}\mid\boldsymbol{\xi})K(\boldsymbol{s},\boldsymbol{v}\mid\boldsymbol{\xi})\rho_X(\|\boldsymbol{v}-\boldsymbol{u}\|)\,d\boldsymbol{u}\,d\boldsymbol{v} \\
&= \frac{\sigma_X^2\beta_1^2}{\kappa^2}\rho_{\widetilde{X}}(\boldsymbol{s},\boldsymbol{s}) \\
&= \rho^2_{Y|\widetilde{X}},
\end{aligned}$$

where the inequality comes from the log-concavity assumption. □

The class of log-concave covariance functions includes the powered exponential, $\rho(\boldsymbol{s},\boldsymbol{s}') = \exp\{-\phi\|\boldsymbol{s}-\boldsymbol{s}'\|^\alpha\}$ $0 \le \alpha \le 2$, by direct calculation. Also included are closed-form Matérn models, i.e., those with smoothness parameter ν of the form $\nu = k + 1/2$ for $k \in \{0, 1, 2, \dots\}$ again by direct calculation with an indication that this is the case for arbitrary ν (see Majumdar and Gelfand 2007). Also, it is easy to argue that convolution of covariance functions produces valid covariance functions. In fact, convolution of log-concave functions produces log-concave functions (for further examples, see, e.g., Gardner 2002).

Thus, for such covariance functions, using $X(\boldsymbol{s})$ instead of $\widetilde{X}(\boldsymbol{s})$ results in less variation in $Y(\boldsymbol{s})$ explained. Additionally, notice that marginalizing over $\widetilde{X}(\boldsymbol{s})$ adds extra spatial variation to $Y(\boldsymbol{s})$. To see this, notice that the covariance given by (2.9) has the added variance term $\beta_1^2\sigma_X^2\rho_{\widetilde{X}}(\boldsymbol{s})$ (see the remark below expression (2.6) in this regard).

We now turn to Bayesian kriging at an unobserved location $\boldsymbol{s}_0$ under the model given by (2.5). Bayesian kriging obtains the predictive distribution $[Y(\boldsymbol{s}_0) \mid \boldsymbol{Y}, \boldsymbol{X}]$, where $\boldsymbol{Y} = (Y(\boldsymbol{s}_{y,1}), \ldots, Y(\boldsymbol{s}_{y,n}))^T$ is the vector of responses observed at spatial locations $\boldsymbol{s}_{y,1}, \ldots, \boldsymbol{s}_{y,n}$, and $\boldsymbol{X} = (X(\boldsymbol{s}_{x,1}), \ldots, X(\boldsymbol{s}_{x,k}))^T$ is the vector of covariates observed at the locations $\boldsymbol{s}_{x,1}, \ldots, \boldsymbol{s}_{x,k}$, where $\{\boldsymbol{s}_{y,i}\}$ and $\{\boldsymbol{s}_{x,j}\}$ are, potentially, misaligned (see Section 3). Mathematically,

$$\begin{aligned}[Y(\boldsymbol{s}_0) \mid \boldsymbol{Y}, \boldsymbol{X}] &= \int_{\boldsymbol{\Theta}} [Y(\boldsymbol{s}_0) \mid \boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{X}][\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}]\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \int_{\mathbb{R}} [Y(\boldsymbol{s}_0) \mid \widetilde{X}(\boldsymbol{s}_0), \boldsymbol{\theta}, \boldsymbol{Y}][\widetilde{X}(\boldsymbol{s}_0) \mid \boldsymbol{\theta}, \boldsymbol{X}] \cdots \\ &\quad \times [\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}]\, d\widetilde{X}(\boldsymbol{s}_0)\, d\boldsymbol{\theta}, \end{aligned} \tag{2.10}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are the model parameters. Based on (2.10), kriging under model (2.5) must occur in two stages. First, using $\boldsymbol{\theta} \sim [\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}]$, a kriged value for $\widetilde{X}(\boldsymbol{s}_0)$ is drawn from the conditional distribution $[\widetilde{X}(\boldsymbol{s}_0) \mid \boldsymbol{\theta}, \boldsymbol{X}]$. Due to the normality assumption on $X(\boldsymbol{s})$, $[\widetilde{X}(\boldsymbol{s}_0) \mid \boldsymbol{\theta}, \boldsymbol{X}]$ is also normal with mean and variance obtained using multivariate normal theory. Second, the kriged value of $\widetilde{X}(\boldsymbol{s}_0)$ is used to draw $Y(\boldsymbol{s}_0) \sim [Y(\boldsymbol{s}_0) \mid \widetilde{X}(\boldsymbol{s}_0), \boldsymbol{\theta}, \boldsymbol{Y}]$, where $[Y(\boldsymbol{s}_0) \mid \widetilde{X}(\boldsymbol{s}_0), \boldsymbol{\theta}, \boldsymbol{Y}]$ is also Gaussian. Repeating the above steps M times will produce M draws $\{Y^{(1)}(\boldsymbol{s}_0), \ldots, Y^{(M)}(\boldsymbol{s}_0)\}$ from $[Y(\boldsymbol{s}_0) \mid \boldsymbol{Y}, \boldsymbol{X}]$.

As noted in the Introduction, we emphasize that the models proposed in this article are motivated by the objective of describing a local regression for $\mathbb{E}(Y(\boldsymbol{s}))$ using neighboring covariate values. However, kriging under the model with kernel averaged predictors is not expected to differ much from that under the customary bivariate process spatial regression model. To see why this is the case, notice from (2.6) that the induced joint model for $(Y(\boldsymbol{s}), X(\boldsymbol{s}))$ is a GP with a mean that is linear in $X(\boldsymbol{s})$ and a covariance function which differs from that ordinarily used for spatial regression problems. Thus, co-kriging under the two models will only differ in the specification of the cross-covariance function. In practice, predictions under the two models will be similar as is demonstrated in Section 4.1 by Table 3.

On a related note, it is evident that model comparison between the local regression given $X(\boldsymbol{s})$ and given $\widetilde{X}(\boldsymbol{s})$ cannot be made in predictive space; the models will be essentially indistinguishable. Rather, comparison resides in performance with regard to parametric inference using criteria such as mean square error and likelihood-based goodness-of-fit criterion such as AIC or DIC (see Section 4.1 for examples). In this regard, suppose that the kernel allows arbitrarily rapid decay, i.e., given $\boldsymbol{s}$ and $\boldsymbol{s}'$ such that $\|\boldsymbol{s} - \boldsymbol{s}'\| \leq \epsilon$, there exists a δ such that $K(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi}) < \delta$. Then, we can view using $X(\boldsymbol{s})$ vs. $\widetilde{X}(\boldsymbol{s})$ in the conditional mean for $Y(\boldsymbol{s})$ as reduced and full models, respectively, where the difference in dimension between the models is the dimension of $\boldsymbol{\xi}$. But then, using within sample goodness of fit, the full model will always fit at least as well as the reduced model.

3. MODEL FITTING

Let $\boldsymbol{Y} = (Y(\boldsymbol{s}_{Y,1}), \ldots, Y(\boldsymbol{s}_{Y,n}))^T$ and $\boldsymbol{X} = (X(\boldsymbol{s}_{X,1}), \ldots, X(\boldsymbol{s}_{X,k}))^T$ be the vectors of observed responses and covariates, respectively. The locations of the observed responses $\{\boldsymbol{s}_{y,i}\}_{i=1}^n$ need not be aligned with the locations of the observed covariates $\{\boldsymbol{s}_{x,j}\}_{j=1}^k$ (e.g. ozone monitoring stations are much more sparse than, say, temperature monitoring stations). Thus, in the general case, the data are the two vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ with some locations in common as opposed to n $(Y(\boldsymbol{s}_i), X(\boldsymbol{s}_i))$ pairs.

Assuming that $\mu_X(\boldsymbol{s}) = \mu_X$ for simplicity, a general Gibbs sampler for fitting (2.5) would proceed as follows:

1. sample $\widetilde{\boldsymbol{X}} = (\widetilde{X}(\boldsymbol{s}_1), \ldots, \widetilde{X}(\boldsymbol{s}_n))^T \sim [\widetilde{\boldsymbol{X}} \mid -, \boldsymbol{X}, \boldsymbol{Y}]$,
2. sample $\mu_X \sim [\mu_X \mid -, \widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{Y}]$,
3. sample $(\beta_0, \beta_1) \sim [\beta_0, \beta_1 \mid -, \widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{Y}]$,
4. sample $(\boldsymbol{\xi}, \phi_X, \sigma_X^2) \sim [\boldsymbol{\xi}, \phi_X, \sigma_X^2 \mid -, \widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{Y}]$,
5. sample $(\phi_Y, \sigma_Y^2, \tau_Y^2) \sim [\phi_Y, \sigma_Y^2, \tau_Y^2 \mid -, \widetilde{\boldsymbol{X}}, \boldsymbol{X}, \boldsymbol{Y}]$,

where "$-$" denotes all other model parameters, and $\widetilde{\boldsymbol{X}} = (\widetilde{X}(\boldsymbol{s}_{y,1}), \ldots, \widetilde{X}(\boldsymbol{s}_{y,n}))^T$ is the vector of $\{\widetilde{X}(\boldsymbol{s}_{y,i})\}_{i=1}^n$ aligned with the n locations where $Y(\boldsymbol{s})$ was observed. Steps 1, 2, and 3 above can be done directly as the complete conditional distributions for $\widetilde{\boldsymbol{X}}$, μ_X, and $(\beta_0, \beta_1)^T$ are all Gaussian, but steps 4 and 5 possibly require Metropolis–Hastings (MH) updates. For misaligned data, step 1 still has closed form because $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})^T$ follows a normal distribution with mean $\mathbf{1}_{k+n}\mu_X$ and covariance matrix

$$\mathbb{V}\mathrm{ar}\begin{pmatrix}\boldsymbol{X}\\ \widetilde{\boldsymbol{X}}\end{pmatrix} = \boldsymbol{\Sigma}_{X\widetilde{X}} = \sigma_X^2\begin{pmatrix}\boldsymbol{R}_0 & \boldsymbol{R}_1\\ \boldsymbol{R}_1^T & \boldsymbol{R}_2\end{pmatrix}, \tag{3.1}$$

where $\mathbf{1}_n$ is a length n vector of ones, $\boldsymbol{R}_0 = \{\rho_X(\boldsymbol{s}_{x,i}, \boldsymbol{s}_{x,j})\}_{i,j=1}^k$, $\boldsymbol{R}_1 = \{\rho_{X,\widetilde{X}}(\boldsymbol{s}_{x,i}, \boldsymbol{s}_{y,j})\}_{i=1,j=1}^{k,n}$, and $\boldsymbol{R}_2 = \{\rho_{\widetilde{X}}(\boldsymbol{s}_{y,i}, \boldsymbol{s}_{y,j})\}_{i,j=1}^n$. Thus, because $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})^T$ is normally distributed, $[\widetilde{\boldsymbol{X}} \mid -, \boldsymbol{X}]$ and $[\widetilde{\boldsymbol{X}} \mid -, \boldsymbol{X}, \boldsymbol{Y}]$ are also Gaussian distributions.

Computationally, step 1 above can be expensive for large n and/or k as it involves the inversion of an $n \times n$ and a $k \times k$ matrix which take order $\mathcal{O}(n^3)$ and $\mathcal{O}(k^3)$ operations. For discussion on handling such "large n" problems see Banerjee et al. (2008).

Oftentimes, the values of $\widetilde{X}(\boldsymbol{s})$ are not of direct interest. In these cases, $\widetilde{X}(\boldsymbol{s})$ can be integrated out, and the user can fit the marginal model of $(X(\boldsymbol{s}), Y(\boldsymbol{s}))^T$. By working with the marginal model, the dimensionality of the posterior distribution is decreased by n. However, this dimension reduction comes at a price in that the complete conditional distribution for (β_0, β_1) is no longer available in closed form. Assuming that $\sigma_Y^2 = 0$ would also simplify computations because τ_Y^2 would then have a closed-form complete conditional distribution. If σ_Y^2 is assumed to be zero, the implication is that the marginal spatial correlation in $Y(\boldsymbol{s})$ is a result of spatial correlation in $X(\boldsymbol{s})$.

In general, the stochastic integrations in (2.6) cannot be done analytically, and, yet, evaluation of these integrals is required to either evaluate the likelihood or obtain a

draw from a complete conditional distribution. As suggested in Banerjee, Carlin, and Gelfand (2004), Monte Carlo (MC) integration becomes a good candidate for approximate computation of these integrals due to the relation $\rho_{X,\widetilde{X}}(\boldsymbol{s},\boldsymbol{s}') = \mathbb{E}_{\boldsymbol{u}}(\rho_X(\boldsymbol{s},\boldsymbol{u}))$, where $\boldsymbol{u} \sim K(\boldsymbol{s}',\boldsymbol{u} \mid \boldsymbol{\xi})/K(\boldsymbol{s}' \mid \boldsymbol{\xi})$. For kernels where $\boldsymbol{u}$ can be sampled from easily, such as those in Table 1, Monte Carlo integration is computationally feasible but challenging for a few reasons. First, even for the easiest case of an aligned data set of n $(Y(\boldsymbol{s}_i), X(\boldsymbol{s}_i))$ pairs, a continuous prior for (ξ, ϕ_X) would require at least $n(n+1)$ such MC integrations at each iteration of the MCMC algorithm! This problem is compounded for cases where $\|\boldsymbol{s}-\boldsymbol{s}'\|$ is small and thousands, if not millions, of draws of $\boldsymbol{u} \sim K(\boldsymbol{s}',\boldsymbol{u} \mid \boldsymbol{\xi})/K(\boldsymbol{s}' \mid \boldsymbol{\xi})$ are required such that $\widehat{\mathbb{V}\mathrm{ar}}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X})$ is positive definite. The remedy we adopt is to use a discrete prior for $(\boldsymbol{\xi}, \boldsymbol{\phi}_X)$ such that the associated variance–covariance matrices can be computed prior to the MCMC algorithm, saved, and the appropriate one retrieved at each iteration. Beyond computational convenience, specifying a discrete prior for $(\boldsymbol{\xi}, \boldsymbol{\phi}_X)$ is also theoretically warranted. Results in Zhang (2004) clarify that, with a weak prior on σ_X^2, an informative prior is needed to identify $\boldsymbol{\phi}_X$.

Second, for complex regions $\mathcal{D}$, drawing $\boldsymbol{u} \sim K(\boldsymbol{s}', u \mid \boldsymbol{\xi})/K(\boldsymbol{s}' \mid \boldsymbol{\xi})$ may be slow when the latter is a nonstandard distribution or a standard distribution truncated to $\mathcal{D}$. Additionally, $K(\boldsymbol{s} \mid \boldsymbol{\xi})$ varies spatially, so a different distribution is sampled for each $\boldsymbol{s}_i$ for each iteration. Here, we take $\mathcal{D} = \mathbb{R}^2$ so that using disks, $K(\boldsymbol{s} \mid \xi) = \pi r^2$ with $r = \xi g(\phi_X)$ for all $\boldsymbol{s} \in \mathcal{D}$ and $\boldsymbol{u}$ can be drawn from a uniform distribution on the disk with center $\boldsymbol{s}$ and radius r.

For well-behaved model fitting, K cannot have arbitrarily high concentration around $\boldsymbol{s}$. Equivalently, choosing between models which use $X(\boldsymbol{s})$ or $\widetilde{X}(\boldsymbol{s})$ in $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{X}(\boldsymbol{s}), X(\boldsymbol{s}))$ cannot be done with priors for $\boldsymbol{\xi}$ that allow K to have arbitrarily high concentration. To illustrate, let K be as defined in Section 2.1 with $\xi \in (0,1)$. As ξ approaches zero, $\boldsymbol{\Sigma}_{X\widetilde{X}}$ (as in (3.1)) will be approximately singular leading to an unstable likelihood for $[\boldsymbol{X}, \widetilde{\boldsymbol{X}} \mid \cdot]$. For the K we use, simulation experiments have shown that a lower bound of $\xi = 0.05$ works well in practice but more experience is needed to understand the sensitivity of the posterior to the prior on $\boldsymbol{\xi}$ for other choices of kernels.

4. APPLICATIONS OF THE SINGLE KERNEL AVERAGED PREDICTOR MODEL

This section applies the modeling of Section 2 to simulated and field data. The field data set consists of measurements of dew duration and shrub density over the Negev Desert in Israel and was previously analyzed by Banerjee and Gelfand (2002).

4.1. MODEL PERFORMANCE ON SIMULATED DATA

Data were simulated under model (2.5) with $K(\boldsymbol{s},\boldsymbol{s}' \mid \xi) = \mathbb{I}_{\{\|\boldsymbol{s}-\boldsymbol{s}'\| \le r\}}$ and $r = \xi g(\phi_X)$ as defined above, in a full factorial design for $n \in \{50, 100\}$, $\phi_X = 10$, $\xi \in \{0, 0.1, 0.3, 0.5\}$, $\sigma_X^2 = 1$, and $\boldsymbol{\beta} = (0,1)^T$. Values for τ_Y^2 were chosen such that the population coefficient of determination (ρ^2) was 0.3, 0.6, and 0.9, respectively, for the true model. We assumed that the locations of observed responses and predictors were aligned and $\{\boldsymbol{s}_i, i = 1, \ldots, n\}$

were confined to the unit square in $\mathbb{R}^2$ but $\mathcal{D}$ was taken to be all of $\mathbb{R}^2$ so as to avoid the difficulty in dealing with locations near the boundary. The case of $\xi = 0$ implies $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{X}}, \boldsymbol{X}) = \beta_0 + \beta_1 X(\boldsymbol{s})$. For convenience, σ_Y^2 was assumed to be zero. For each combination of (n, ξ, τ_Y^2), 30 data sets were simulated with additional n values of $Y(\boldsymbol{s})$ left as a hold-out sample to determine predictive performance of the fitted models. Define $\mathrm{B}(\widehat{\beta}_1)$, $\mathrm{MSE}(\widehat{\beta}_1)$, $\mathrm{CIC}(\beta_1)$, and $\mathrm{CIW}(\beta_1)$ as the observed bias of the posterior mean $\widehat{\beta}_1$, mean square error of $\widehat{\beta}_1$, empirical 95% credible interval coverage for β_1, and empirical 95% credible interval width for β_1, respectively. Additionally, define $\mathrm{MSE}(\widehat{y}) = \sum_{j=1}^{30} \sum_{i=1}^{n} (y^{(j)}(\boldsymbol{s}_i) - \widehat{y}^{(j)}(\boldsymbol{s}_i))^2/(n \times 30)$ to be the average mean square error for the n training sample points, where $y^{(j)}(\boldsymbol{s}_i)$ is the ith observation of the jth data set, and $\widehat{y}^{(j)}(\boldsymbol{s}_i)$ is the posterior mean of $\mathbb{E}(y^{(j)}(\boldsymbol{s}_i) \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{X}}, \boldsymbol{X})$. Finally, to assess predictive performance, define PMSE, PIC, and PIW as the average predictive mean square error, predictive interval coverage, and predictive interval width across all of the n hold-out values.

Two models were fit for comparison. The first model was the kernel averaged predictor (KAP) model given by (2.5), where $\sigma_Y^2 = 0$, and K was defined in Section 2.1 The second model was a point predictor (PtP) model with $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{X}(\boldsymbol{s}), X(\boldsymbol{s})) = \beta_0 + \beta_1 X(\boldsymbol{s})$, where $X(\boldsymbol{s})$ was defined in (2.1). The exponential correlation function with $\rho_X(\boldsymbol{s}, \boldsymbol{s}' \mid \phi_X) = \exp(-\phi_X \|\boldsymbol{s} - \boldsymbol{s}'\|)$ was used for both models. Discrete prior distributions for ϕ_X and ξ were used with mass at $(30, 60, 300)$ and $(0.1, 0.3, 0.5)$, respectively. Vague, but proper, conjugate prior distributions were assumed for the remaining parameters. Chains were run for an initial burn in period of 10,000 draws, and the following 10,000 were retained as draws from the posterior distribution.

To avoid redundancy, only the simulation results for the case of $\rho^2 = 0.9$ are displayed. Table 2 displays the results of the model fit along with values of DIC (Spiegelhalter et al. 2002) and the associated effective number of parameters (p_d). First, when the "true" model is $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) = \beta_0 + \beta_1 X(\boldsymbol{s})$, the KAP model has poor coverage ($\mathrm{CIC}(\beta_1)$) of 33% and 3% for $n = 50$ and $n = 100$, respectively. However, the KAP model still has favorable values for $\mathrm{MSE}(\widehat{y})$. When $\xi > 0$, the KAP model performs well, while the performance of the PtP model declines rapidly as ξ increases. For example, notice that for the moderate value of $\xi = 0.1$, the coverage probability for the PtP model falls to 17% for $n = 50$ and 0% for $n = 100$. In terms of model choice, DIC correctly chooses the PtP model when $\xi = 0$ and the KAP model when $\xi > 0$. The simulation results for $\rho^2 \in \{0.3, 0.6\}$ were similar except that the discrepancy in DIC between the KAP and PtP models was smaller.

Table 3 displays the predictive results for the two models on the n hold-out values. The predictive performance of the models is similar in terms of PMSE and PIC. In terms of PIW, however, the KAP model seems to give smaller intervals. The comparable predictive performance between the two models was expected and clarified in Section 2.2.

4.2. KERNEL SELECTION SIMULATION STUDY

The simulation study here focuses on the problem of kernel selection. 50 data sets of $\boldsymbol{Y} = (Y(\boldsymbol{s}_{Y,1}), \ldots, Y(\boldsymbol{s}_{Y,n}))^T$ and $\boldsymbol{X} = (X(\boldsymbol{s}_{X,1}), \ldots, X(\boldsymbol{s}_{X,k}))^T$ with $n = 50$ and $k = 150$ (to introduce misalignment) were simulated from the KAP model (2.5) using $K(\boldsymbol{s}, \boldsymbol{s}' \mid$

Table 2. Model fit diagnostics of the kernel averaged predictor (KAP) and the point predictor (PtP) models on simulated data with $\rho^2 = 0.9$. The KAP model estimates β_1 well when $\xi > 0$.

n		50				100			
ξ		0	0.1	0.3	0.5	0	0.1	0.3	0.5
$B(\widehat{\beta}_1)$	KAP	0.17	−0.02	−0.18	−0.23	0.15	−0.03	−0.16	−0.18
	PtP	0.01	−0.19	−0.43	−0.60	−0.01	−0.19	−0.46	−0.62
$MSE(\widehat{\beta}_1)$	KAP	0.03	0.01	0.04	0.06	0.03	0.00	0.04	0.05
	PtP	0.00	0.04	0.19	0.36	0.00	0.04	0.21	0.38
$CIC(\beta_1)$	KAP	0.33	0.97	0.93	0.90	0.03	0.90	0.93	0.97
	PtP	0.97	0.17	0.00	0.00	0.93	0.00	0.00	0.00
$CIW(\beta_1)$	KAP	0.33	0.40	0.66	0.73	0.21	0.22	0.60	0.69
	PtP	0.22	0.27	0.28	0.26	0.15	0.17	0.19	0.18
$MSE(\widehat{y})$	KAP	0.03	0.08	0.09	0.06	0.02	0.06	0.08	0.05
	PtP	0.11	0.19	0.21	0.20	0.11	0.19	0.22	0.21
p_d	KAP	22.65	18.95	17.25	19.34	52.87	43.96	39.84	46.32
	PtP	2.41	2.70	2.81	2.86	2.68	2.84	2.91	2.94
DIC	KAP	49.75	62.57	62.98	50.69	80.11	112.09	117.31	95.06
	PtP	37.07	64.57	69.95	65.59	68.51	121.66	138.35	130.28

Table 3. Predictive diagnostics of the kernel averaged predictor (KAP) and the point predictor (PtP) models on simulated data with $\rho^2 = 0.9$. The predictive properties of the KAP model are similar to those of the PtP model, but the KAP model gives slightly smaller predictive intervals.

n		50				100			
ξ		0	0.1	0.3	0.5	0	0.1	0.3	0.5
PMSE	KAP	1.12	0.86	0.56	0.38	1.08	0.91	0.54	0.40
	PtP	1.12	0.85	0.55	0.36	1.09	0.90	0.52	0.36
PIC	KAP	0.93	0.94	0.96	0.95	0.92	0.95	0.95	0.95
	PtP	0.95	0.95	0.97	0.96	0.96	0.96	0.96	0.95
PIW	KAP	3.89	3.61	2.91	2.39	3.69	3.51	2.74	2.26
	PtP	4.35	3.80	3.00	2.45	4.10	3.69	2.84	2.37

$\xi) = \mathbb{I}_{\{\|s-s'\| \le r\}}$, where $r = \xi g(\phi_X)$, ρ_X is the exponential correlation function, $g(\phi_X) = 3/\phi_X$, $\phi_X = 10$, $\xi = 0.5$, $\sigma_X^2 = 1$, and $\boldsymbol{\beta} = (0, 1)^T$. The value for τ_Y^2 was chosen such that $\rho^2_{Y|\widetilde{X}} = 0.5$. The locations for the $n = 50$ data points $\{\boldsymbol{s}_{Y,1}, \ldots, \boldsymbol{s}_{Y,50}\}$ were drawn uniformly on the unit square in $\mathbb{R}^2$. The $\{\boldsymbol{s}_{X,j}\}$ were drawn uniformly on the larger square $(-0.5, 1.5) \times (-0.5, 1.5)$. Misalignment between the response and predictors is common in environmental studies but handled easily using the methods described in Section 3.

Three KAP models with kernels,

$$K_1(\boldsymbol{s}, \boldsymbol{s}' \mid \xi) = \mathbb{I}_{\{\|\boldsymbol{s}-\boldsymbol{s}'\| \le \xi g(\phi_X)\}},$$

$$K_2(\boldsymbol{s}, \boldsymbol{s}' \mid \xi) = \frac{\phi_X}{\xi} \exp\left\{ -\frac{(\boldsymbol{s}' - \boldsymbol{s})^T (\boldsymbol{s}' - \boldsymbol{s})}{2\xi^2/\phi_X^2} \right\},$$

Table 4. Model fit diagnostics for the kernel selection simulation study. The kernels K_1 and K_2 produce similar results because they are of similar center and scale, but the misspecified kernel K_3 produces unsatisfactory results in terms of bias and credible interval coverage. This difference in performance is captured by both $\mathbb{E}(\rho^2_{Y|\widetilde{X}} \mid \boldsymbol{Y})$ and DIC.

Kernel	$\mathrm{B}(\widehat{\beta}_1)$	$\mathrm{CIC}(\beta_1)$	$\mathrm{CIW}(\beta_1)$	$\mathbb{E}(\rho^2_{Y\|\widetilde{X}} \mid \boldsymbol{Y})$	DIC
$K_1(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$	−0.12	0.96	0.84	0.46	91.95
$K_2(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$	−0.21	0.92	0.90	0.46	93.82
$K_3(\boldsymbol{s}, \boldsymbol{s}' \mid \xi)$	−0.63	0.50	1.06	0.29	103.72

and

$$K_3(\boldsymbol{s}, \boldsymbol{s}' \mid \xi) = \frac{\phi_X}{\xi} \exp\left\{ -\frac{(\boldsymbol{s}' - (\boldsymbol{s} + \boldsymbol{\theta}))^T (\boldsymbol{s}' - (\boldsymbol{s} + \boldsymbol{\theta}))}{2\xi^2/\phi_X^2} \right\},$$

were fit to each simulated data set, where $\xi \in (0, 1)$ for all kernels, and $\boldsymbol{\theta} = (-0.2, 0.2)^T$. Notice that kernel K_1 is the true kernel and kernels K_2 and K_3 are simply Gaussian kernels with means $\boldsymbol{s}$ and $\boldsymbol{s} + \boldsymbol{\theta}$, respectively, and common standard deviation ξ/ϕ_X. The standard deviation of ξ/ϕ_X was chosen to confine the bulk of the mass of K_2 and K_3 to be within the spatial range $g(\phi_X)$ so as to prevent strong collinearity in the $\widetilde{X}(\boldsymbol{s})$ (see Section 2.1). Notice that under this parameterization, the spatial range $g(\phi_X)$ is assumed to be, at most, 3 standard deviations away from the mean of the kernel in any direction. Each model assumed discrete priors for $[\phi_X]$ and $[\xi]$ with equal mass at $(30, 10, 5)$ and $(0.1, 0.5, 0.8)$, respectively. Here, $\boldsymbol{\theta}$ is assumed to be known because we wish to investigate the consequences of using an inappropriate kernel. Table 4 displays $\mathrm{B}(\widehat{\beta}_1)$, $\mathrm{CIC}(\beta_1)$, $\mathrm{CIW}(\beta_1)$, $\mathbb{E}(\rho^2_{Y|\widetilde{X}} \mid \boldsymbol{Y})$, and DIC for each KAP model.

$\mathbb{E}(\rho^2_{Y|\widetilde{X}} \mid \boldsymbol{Y})$ and DIC can be used to select an appropriate kernel. That is, different kernels yield different predictors in the linear regression given by (2.5). Thus, comparing different kernels is essentially equivalent to comparing predictors. As such, DIC is a familiar criterion. As noted in Section 2, $\rho^2_{Y|\widetilde{X}} \mid \boldsymbol{Y}$ is also an appropriate (but random) measure for the performance of a kernel averaged predictor model. Hence, its posterior mean provides a sensible kernel model comparison criterion.

The results in Table 4 give a few encouraging insights into kernel selection. First, similar kernels yield similar performance. Notice that KAP models using K_1 and K_2 exhibited acceptable performance despite the subtleties in the differences of their shapes while K_3, which was shifted by a factor of $\boldsymbol{\theta}$, exhibited inferior bias, credible interval coverage, and credible interval width. Interestingly, when using K_2, the value of ξ with the highest posterior probability was most frequently 0.8 across the data sets. Thus, despite the subtle difference in shape, the scale of K_2 was properly adjusted to capture the scale of the true kernel.

4.3. NEGEV DESERT FIELD DATA

The Negev Desert data set contains measurements of dew duration time in hundredths of an hour starting from 8 a.m. and shrub density (density of shrubs within a 5 meter ×

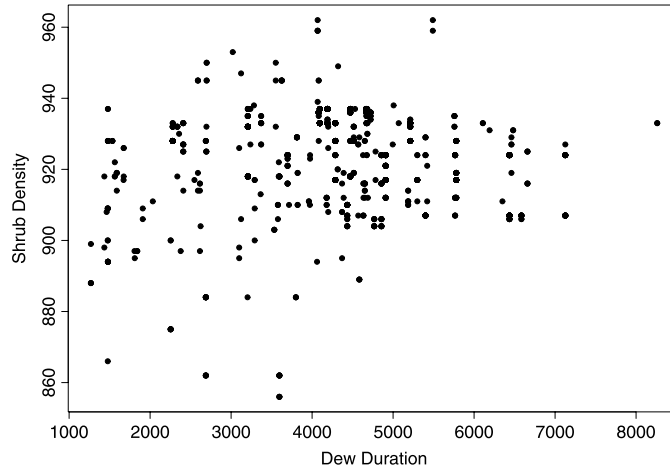


Figure 3. Scatter plot of dew duration by shrub density.

5 meter block) at $n = 1129$ aligned locations in the Negev Desert in Israel. As the Negev Desert is very arid, condensation can contribute considerably to the annual water levels. The analysis here is to determine the effect of shrub density on dew duration. Figure 3 displays the scatter plot of dew duration by shrub density, and Figure 4 displays contour plots of dew duration and shrub density over the region with sampling locations overlaid. All 1129 locations were used for the following analysis.

Let $Y(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_i)$ represent the measurement of dew duration and shrub density for UTM coordinate $\boldsymbol{s}_i$, $i = 1, \ldots, n$. Again, the KAP and PtP models described in the previous simulation study were fit to the shrub data. Because the primary interest of this study is the effect of shrub density on dew duration (β_1), both models assumed that the response process had no residual spatial correlation. A discrete uniform prior distribution for $g(\phi_X) = 3/\phi_X$ was used in both models, where mass was placed at 0.1, 0.3, and 0.5 times $\max_{(i,j)} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$. A discrete uniform prior distribution for ξ was used where masses were placed at the points $(0.1, 0.2, 0.3, 0.4)$. Both models used vague, but proper, normal prior distributions for $[\beta_0, \beta_1]$ and vague, but proper, inverse-gamma distributions for $[\sigma_X^2, \tau_Y^2]$.

Markov chain Monte Carlo algorithms similar to that presented in Section 3 were run for an initial burn-in period of 10,000 iterations, and 20,000 draws were kept after this burn-in period. The KAP model took an average of 5.29 seconds per iteration compared to 1.66 seconds per iteration for the PtP model. Analysis of the chains using trace plots and the convergence diagnostics of Raftery and Lewis (1992) and Geweke (1992) showed that the chains had reasonable convergence properties and mixed well.

Comparing the KAP and PtP models, the KAP model had a DIC value of DIC $= 23{,}846$ versus DIC $= 25{,}111$ for the PtP model. The effective numbers of parameters were $p_d = 977.81$ and $p_d = 2.74$ for the KAP and PtP models, respectively.

Table 5 displays posterior quantiles for parameters that are common to both models, Table 6 lists the posterior distribution of $g(\phi_X)$ for both models, and the posterior distribution for ξ was $[\xi \mid \boldsymbol{X}, \boldsymbol{Y}] = (0.0, 0.97, 0.03, 0.0)$ for $\xi \in (0.1, 0.2, 0.3, 0.4)$, respectively. According to Table 5, all the quantiles of $[\beta_1 \mid \boldsymbol{X}, \boldsymbol{Y}]$ under the KAP model are less than

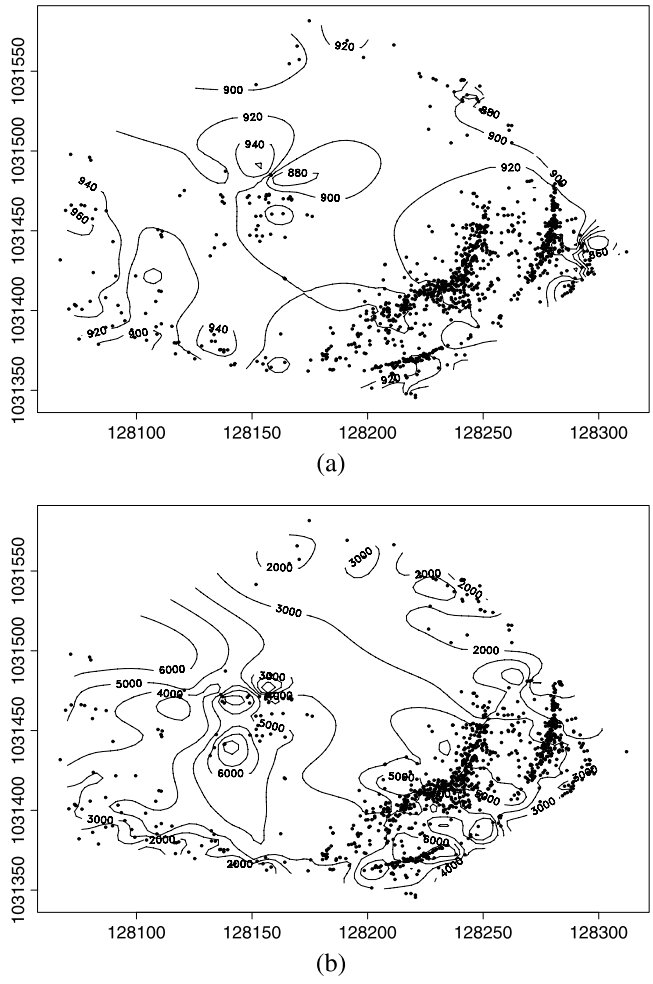


Figure 4. Contour plots of (a) dew duration and (b) shrub density over the Negev desert. The points correspond to the $n = 1129$ locations in universal transverse Mercator coordinates.

the associated quantiles under the PtP model. Thus, as expected, the effect of shrub density on dew duration is shrunk toward zero for the unblocked predictor model. Additionally, because the estimate of β_1 was shrunk toward zero, the estimate of β_0 differs between the two models as well. Notice also that $\rho^2_{\mathrm{KAP}} > \rho^2_{\mathrm{PtP}}$, indicating that $\widetilde{X}(\boldsymbol{s})$ is a better predictor of dew duration than $X(\boldsymbol{s})$.

For many ecological processes, response at a site would be expected to be influenced by covariate information from neighbors. The value of our approach is that we can incorporate this notion and quantify its benefit without actually observing the neighboring covariate levels. This is the primary finding of our reanalysis of the Negev desert data; that $\widetilde{X}(\boldsymbol{s})$ was found to be a better predictor of $Y(\boldsymbol{s})$ than $X(\boldsymbol{s})$. This finding suggests new ecological interpretation with regard to the relationship between shrub density and dew duration within the Negev desert. Additionally, we find a stronger negative association between shrub density and dew duration than previous analyses had discovered Though

Table 5. Posterior quantiles of parameters for the kernel averaged predictor (KAP) and point predictor (PtP) models used in the Negev Desert data application. As expected, $|\beta_1|$ and ρ^2 are greater for the KAP model.

Parameter	Model	Posterior Quantile				
		0.025	0.25	0.50	0.75	0.975
β_0	KAP	9.26e2	9.28e2	9.29e2	9.30e2	9.32e2
	PtP	9.25e2	9.26e2	9.27e2	9.28e2	9.29e2
β_1	KAP	−0.016	−0.012	−0.010	−0.008	−0.004
	PtP	−0.011	−0.008	−0.007	−0.005	−0.003
ρ^2	KAP	0.002	0.008	0.013	0.018	0.031
	PtP	0.002	0.006	0.009	0.014	0.024
μ_X	KAP	5.19e2	5.27e2	5.32e2	5.36e2	5.44e2
	PtP	5.19e2	5.27e2	5.32e2	5.36e2	5.44e2
σ_X^2	KAP	5.21e4	5.49e4	5.64e4	5.81e4	6.13e4
	PtP	5.28e4	5.56e4	5.73e4	5.89e4	6.23e4
τ_Y^2	KAP	2.59e2	2.73e2	2.81e2	2.89e2	3.05e2
	PtP	2.60e2	2.74e2	2.82e2	2.90e2	3.06e2

Table 6. Posterior distribution of the effective spatial range $g(\phi_X)$ for the shrub data using both a kernel averaged predictor (KAP) and point predictor (PtP) model.

Model	$g(\phi_X)$		
	24.82	74.44	124.07
KAP	0.93	0.04	0.03
PtP	0.33	0.34	0.33

this somewhat counterintuitive, the negative association between dew duration and shrub density is evident upon examination of the scatterplot in Figure 3.

5. EXTENSIONS

5.1. MULTIPLE KERNEL AVERAGED PREDICTORS

Extending the methodology of Section 2 to multiple predictors involves inducing a multivariate spatial process for $\boldsymbol{X}(\boldsymbol{s}) = (X_1(\boldsymbol{s}), \ldots, X_p(\boldsymbol{s}))^T$. Once a multivariate process has been specified for $\boldsymbol{X}(\boldsymbol{s})$, the extension follows by defining

$$\widetilde{X}_j(\boldsymbol{s}) = \frac{1}{K_j(\boldsymbol{s})} \int_{\mathcal{D}} K_j(\boldsymbol{s}, \boldsymbol{u} \mid \boldsymbol{\xi}_j) X_j(\boldsymbol{u})\, d\boldsymbol{u},$$

where $K_j(\boldsymbol{s}, \boldsymbol{s}' \mid \boldsymbol{\xi}_j)$ is the jth kernel with parameters $\boldsymbol{\xi}_j$ and extending (2.5) to be

$$\begin{aligned} Y(\boldsymbol{s}) \mid \boldsymbol{X}(\boldsymbol{s}), \widetilde{\boldsymbol{X}}(\boldsymbol{s}) &= \beta_0 + \sum_{j=1}^{p} \beta_j \widetilde{X}_j(\boldsymbol{s}) + \sigma_Y w_Y(\boldsymbol{s}) + \epsilon_Y(\boldsymbol{s}) \\ &= \beta_0 + \boldsymbol{\beta}^T \widetilde{\boldsymbol{X}}(\boldsymbol{s}) + \sigma_Y w_Y(\boldsymbol{s}) + \epsilon_Y(\boldsymbol{s}), \end{aligned} \tag{5.1}$$

where $w_Y(\boldsymbol{s})$ and $\epsilon_Y(\boldsymbol{s})$ are defined the same as in (2.5). It seems sensible that the choice of kernel would depend upon the predictor. For example, with K defined in Section 2.1, r should be allowed to vary with the covariate. Details and distributional results for the multiple kernel averaged predictor model are given in the supplementary materials where the multivariate spatial process for $\boldsymbol{X}(\boldsymbol{s})$ is induced by the methods of coregionalization as described in Gelfand et al. (2004).

The multiple blocked predictor model in (5.1) poses significant computational challenges beyond those of the single kernel averaged predictor model in (2.5). Most noticeably, an MCMC algorithm to fit (5.1) requires that the length np vector $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1^T, \ldots, \widetilde{\boldsymbol{X}}_p^T)^T$ be drawn at each iteration. Obtaining a draw of $\widetilde{\boldsymbol{X}}$ involves inversion of a $kp \times kp$ and an $np \times np$ matrix which can be challenging even for moderately sized n, k, and p. Marginalization over $\widetilde{\boldsymbol{X}}(\boldsymbol{s})$ will remove the necessity to sample $\widetilde{\boldsymbol{X}}$ at each iteration but conjugacy of full conditional distributions is lost. However, this marginalization allows a marked reduction in dimension of np. A problem which cannot be eluded via marginalization is the need to calculate integrals which are typically not available in closed form. Thus, to emphasize, a discrete prior for the parameters of the spatial correlation function and of the kernels $K_j(\cdot)$ becomes necessary for computational tractability.

We note that not all covariates are appropriate to kernel average. For instance, with categorical covariates, Gaussian process modeling will not be sensible, and kernel averaging will not provide an interpretable covariate level. Also, if trend surfaces are introduced to explain $Y(\boldsymbol{s})$, such surfaces would not be viewed as random; moreover, it may not be sensible to kernel average such surfaces.

5.2. Generalized Linear Models With Kernel Averaged Predictors

Obviously, applications exist where the response variable is not continuous so that the use of a Gaussian distribution as a likelihood is inappropriate. The methods proposed herein naturally extend to noncontinuous $Y(\boldsymbol{s})$ through the use of link functions. Specifically, if $Y(\boldsymbol{s}) \sim [Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{X}}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})]$, where $[\cdot]$ is some distribution function with parameters $\boldsymbol{\theta}$, then a generalized linear model with kernel averaged predictors would be $\mathbb{E}(Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{X}}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})) = h^{-1}(\beta_0 + \boldsymbol{\beta}^T \widetilde{\boldsymbol{X}}(\boldsymbol{s}))$, where h is an appropriate link function. For example, if $Y(\boldsymbol{s})$ is a count, then $[Y(\boldsymbol{s}) \mid \boldsymbol{\theta}, \widetilde{\boldsymbol{X}}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s})]$ is typically assumed to be the Poisson distribution, and $h^{-1}(\beta_0 + \boldsymbol{\beta}^T \widetilde{\boldsymbol{X}}(\boldsymbol{s})) = \exp(\beta_0 + \boldsymbol{\beta}^T \widetilde{\boldsymbol{X}}(\boldsymbol{s}))$.

6. DISCUSSION AND FUTURE DIRECTIONS

We have presented an approach to account for neighboring predictor values in a spatial regression setting. The need for such an approach is apparent in that, to explain the response at a particular location, levels of certain predictors (e.g., precipitation, temperature) at neighboring locations can affect the response. The kernel averaged predictor (KAP) model elucidated in this article used kernel averaging to define a new predictor $\widetilde{X}(\boldsymbol{s})$ for use in a spatial regression model. The parameters of the kernel control the extent to which neighboring predictors influence the response. An important result under the KAP model is that a regression model using $X(\boldsymbol{s})$ instead of $\widetilde{X}(\boldsymbol{s})$ will diminish the effect of the predictor

on the response. This result was found to be the case for a real dataset of dew duration and shrub density over the Negev desert.

According to application, different kernels, perhaps oriented or location-shifted could be employed. Such kernels could be constructed to align with physical aspects such as meteorological conditions, elevation, or differential equations. For more complex kernels, more experience is needed to learn how well-associated parameters can be estimated.

Often, the MCMC algorithms used to estimate model parameters had to be run for a substantial amount of time in order to reach the limiting distribution. Analysis of the Markov chains run for the example above showed that an initial burn-in of 10,000 was necessary. For large n problems where each iteration requires the inversion of $n \times n$ matrices, the time needed to perform these iterations can compound quickly. Alternative computational methods such as reparameterization or, perhaps, Laplace approximations to the posterior as in Rue, Martino, and Chopin (2009) can be investigated.

Our development here has been exclusively for the static spatial setting. However, extension to incorporate dynamics—both explanatory and response variables recorded over time—opens a rich variety of process modeling opportunities. These possibilities are currently under investigation and will be reported on in future work.

SUPPLEMENTARY MATERIALS

Results for model misspecification when $\widetilde{X}(\boldsymbol{s})$ is used when $X(\boldsymbol{s})$ is the appropriate predictor. (PDF)

Distributional results for the case of multiple predictors. (PDF)

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant number 5R01-ES-014843-01A2. Any opinions, findings, and conclusions or recommendations expressed in this publications are those of the authors and do not necessarily reflect the views of the National Institute of Health. This research was also supported in part by NSF grant DMS-0914906 to the National Institute of Statistical Sciences. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[Published Online January 2011.]

REFERENCES

Baíllo, A., and Grané, A. (2009), "Local Linear Regression for Functional Predictor and Scalar Response," *Journal of Multivariate Analysis*, 100, 102–111.

Banerjee, S., and Gelfand, A. E. (2002), "Prediction, Interpolation, and Regression for Spatially Misaligned Data Points," *Sankhya A*, 64, 227–245.

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

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Datasets," *Journal of the Royal Statistical Society, Series B*, 70, 825–848.

Berliner, L. M. (2000), "Hierarchical Bayesian Modeling in the Environmental Sciences," *Journal of the German Statistical Society*, 84.

Chiles, J. P., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley Interscience.

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

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics," *Applied Statistics*, 47, 299–350.

Gardner, R. J. (2002), "The Brunn–Minkowski Inequality," *Bulletin of the American Mathematical Society*, 39, 355–405.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary Multivariate Process Modeling Through Spatially Varying Coregionalization" (with discussion), *Test*, 12, 1–50.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 169–194.

Higdon, D. (1998), "A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

Majumdar, A., and Gelfand, A. E. (2007), "Multivariate Spatial Modeling for Geostatistical Data Using Convolved Covariance Functions," *Journal of Mathematical Geology*, 39, 225–245.

Matérn, B. (1986), *Spatial Variation* (2nd ed.), Berlin: Springer.

Raftery, A. E., and Lewis, S. (1992), "How Many Iterations in the Gibbs Sampler," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 763–773.

Royle, J. A., and Berliner, L. M. (1999), "A Hierarchical Approach to Multivariate Spatial Modeling and Prediction", *Journal of Agricultural, Biological, and Environmental Statistics*, 4.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations" (with discussion), *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Wackernagel, H. (2003), *Multivariate Geostatistics*, Berlin: Springer.

Wikle, C. K. (2002), "A Kernel Based Spectral Method for Non-Gaussian Spatio-Temporal Processes," *Statistical Modelling*, 2, 299–314.

Xu, K., Wikle, C. K., and Fox, N. L. (2005), "A Kernel-Based Spatio-Temporal Dynamical Model for Nowcasting Weather Radar Reflectivities," *Journal of the American Statistical Association*, 100, 1133–1144.

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261.