Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 203-216

Soil Property Estimation and Design for Agroecosystem Management using Hierarchical Geospatial Functional Data Models

Christopher K. Wikle[1], Scott H. Holan[1], Kenneth A. Sudduth[2] and D. Brenton Myers[3]
[1]*Department of Statistics, University of Missouri, Columbia, MO 65211*
[2]*USDA-ARS- Cropping System and Water Quality Unit, Columbia, MO 65211*
[3]*Department of Plant Sciences, 214 Waters Hall, Columbia, MO 65211*

Received 30 September 2013; Revised 05 February 2014; Accepted 16 April 2014

SUMMARY

Sustainable agriculture requires a site-specific approach to address crop management problems and environmental degradation processes that are spatially and temporally variable. These issues lead to production losses (water stress, low fertility, pest problems), soil degradation (erosion, soil organic carbon losses, compaction), and water quality degradation (sediment, nutrients, agrochemicals) - often at the sub-field scale. Management solutions must be implemented at the resolution of the problems; however, changes require information on the magnitude and extent of the issue. Unfortunately, landscape processes and properties can change at a finer spatial resolution than can be practically analyzed with lab methods due to time and cost of sampling and analysis. Thus, it is increasingly important to augment lab methods with field-sensor methods that can accurately characterize within-field variability at a more reasonable cost and with reliability and timeliness. These instruments can produce large data profiles and require calibration and prediction methods that can accommodate "big data." We consider a functional spatial approach to perform calibration, spatial prediction, and design in this big data context. Specifically, using hierarchical Bayesian methodology we develop a signal/feature extraction approach for visible and near-infrared (VNIR) spectroscopic data that facilitates prediction of cation exchange capacity (CEC) over space. This methodology is also used to develop optimal spatial sampling locations to minimize the mean squared prediction error corresponding to a predicted spatial surface of this CEC response variable.

Keywords: Adaptive design, Bayesian, DRS, Functional data, Optimal spatial design, Principal components, Stochastic search variable selection, VNIR.

1. INTRODUCTION

Knowledge of the characteristics of soils is important in many contexts, including development of economically and environmentally sustainable agricultural production systems and for understanding the potential of soils to remediate environmental problems through processes such as carbon sequestration. Historically, this knowledge has been gained through laborious field soil surveys along with careful chemical, physical, and biological laboratory analysis of collected samples. Better and more timely soil information is needed to be able to respond to complex agricultural and environmental issues, both today and into the future. In recent years, the use of soil sensing technology has increased greatly in research and to a lesser extent in agricultural production. Soil sensors have improved the efficiency of obtaining soil information, which is particularly important where spatially dense data are needed – characterizing soil properties for precision agriculture, digital soil mapping, and spatio-temporal modeling (Viscarra

Corresponding author: Christopher K. Wikle
E-mail address: wiklec@missouri.edu

Rossel *et al*. 2011). Among soil sensing technologies, optical diffuse reflectance spectroscopy (DRS) in the visible and near-infrared (VNIR) wavelength ranges (~400-2500 nm) stands out for its ability to be simultaneously calibrated to a range of soil properties (Sudduth *et al*. 1997, Viscarra Rossel *et al*. 2006, Stenberg *et al*. 2010). To be consistent with the DRS literature, we use the term calibration to describe the process of estimating spectroscopic models.

1.1 Calibration Issues with VNIR Sensing

The key to obtaining useful soil information with VNIR spectroscopy is a valid and robust calibration methodology. In contrast to longer-wavelength (*e.g*., mid-infrared) spectra where specific peaks may be assigned to a particular constituent, VNIR soil spectra are largely nonspecific due to overlapping effects of the different soil constituents. The lack of specificity is compounded by scatter effects due to soil structure and mineralogy. Thus, multivariate analysis techniques are required to relate reflectance data at multiple wavelengths to the soil constituents of interest (Stenberg *et al*. 2010). Since its first application to VNIR soil data by Sudduth and Hummel (1991), partial least squares (PLS) regression has become the de facto standard, and has been used in the majority of recent studies. Other methods used to a lesser extent include stepwise multiple linear regression (SMLR; Dalal and Henry 1986, Lee *et al*. 2009), neural networks (Daniel *et al*. 2003, Fidencio *et al*. 2002), multivariate adaptive regression splines (MARS; Shepherd and Walsh 2002), and boosted regression trees (Brown *et al*. 2006). In studies where multiple methods have been compared, no single approach has emerged as best in all cases. Thus, further research on VNIR calibration methodology development is warranted, particularly such methodology that can accommodate "big data" covariates.

Although most VNIR soil sensing to date has been carried out in the lab, field prototype systems were developed by Shonk *et al.* (1991) and Sudduth and Hummel (1993). Since then, other mobile systems have been developed by Shibusawa *et al*. (2001), Mouazen *et al*. (2005), Stenberg *et al*. (2007) and Christy (2008), who described a commercially available mobile VNIR system. These field-ready sensors all collected data at a single depth at the soil surface while moving across a field. As mobile systems have made it easier to collect data with high spatial resolution, issues of calibration (model estimation) under spatial dependence have become paramount. Although most studies have developed calibrations using samples obtained from a broad area, ranging from continental (Shepherd and Walsh 2002), to regional (Chang *et al*. 2001, Sudduth and Hummel 1996, Lee *et al*. 2009), to statewide (Sudduth and Hummel 1991), some have reported field-specific calibrations (Viscarra Rossel *et al*. 2006, Ge *et al*. 2007, Christy 2008, Lee *et al*. 2010, Sudduth *et al*. 2010). If only a few within-field sampling sites are employed in a field, it may be valid to assume little or no spatial dependence. However, a denser sampling scheme (*e.g*., grid sampling) may result in significant spatial autocorrelation. Not including these effects in the model violates the usual assumption of statistical independence of the calibration model residuals, and can render the model suboptimal (Ge *et al*. 2007). Current methodologies for incorporating spatial dependence in VNIR DRS analysis are generally multi-step and ignore various sources of uncertainty; thus, opportunities exist for developing an integrated spatial analysis framework through hierarchical statistical modeling (Cressie and Wikle 2011).

Interaction of the calibration dataset and methodology can have a significant effect on the accuracy of soil property estimates. Specifically, choice of a calibration dataset (*i.e*., training dataset) requires consideration of the trade-off between the accuracy required and the resources available to develop the calibration. If the highest accuracy is needed and resources are not a limitation, then individual field calibration with a significant number of lab-measured calibration samples will likely provide the best results. In this case, the main consideration is where to obtain calibration samples for highest accuracy, and various methods have been proposed (Lesch 2005, Minasny and McBratney 2006, Christy 2008). On the other hand, resource or practicality issues may impose limitations on within-field calibration sampling, requiring approaches ranging from a global or "factory" calibration to various methods that combine pre-existing calibration data with limited within-field sampling (Brown 2007, Sankey *et al*. 2008, Lee *et al*. 2010). This suggests the need to determine an optimum framework for combining local and pre-existing calibration data along with how to optimally select local calibration sites.

1.2 Functional Data Analysis

One promising framework for achieving robust multisite soil characterization using high-dimensional "big data" spectral covariate data from VNIR DRS and auxiliary sensors is hierarchical geospatial functional data analysis (FDA; Ramsay and Silverman 2005). The basic premise of FDA is that one considers infinite dimensional observations in the form of curves (*e.g*., diffuse reflectance spectra) with some intrinsic ordering of observations (*e.g*., time ordering for curves and spatial ordering for images). Including functional predictors in a setting where the responses are spatially-dependent can complicate FDA. Indeed, spatial functional analysis is a developing area of research in spatial statistics and FDA (see the reviews in Delicado *et al*. 2010, Ruiz-Medina 2012a, and Kokoszka 2012). Most geostatistical FDA developments have been in the context of cokriging approaches (*e.g*., Goulard and Voltz 1993, Monestiez and Nerini 2008, Giraldo *et al*. 2010, 2012) as well as the general theory of spatial autoregressive and moving average Hilbertian processes (Ruiz-Medina 2011, Ruiz-Medina and Montes 2011, Ruiz-Medina 2012b, Ruiz-Medina and Espejo 2013). In the spirit of traditional FDA, Gromenko *et al*. (2012) and Gromenko and Kokoszka (2013) consider a functional principal component approach for estimating mean functions. With the exception of Baladanayuthapani *et al*. (2008), almost all of the work in spatial FDA has been from a classical perspective. Recently, Yang *et al*. (2014) consider a fully hierarchical Bayesian approach for the analysis of spatially-dependent functional responses with spatially-dependent multi-dimensional functional predictors that relies on stochastic search variable selection (SSVS) methods to accommodate modeling in the presence of very high-dimensional covariates.

Our prediction methodology below could be considered a special case of Yang *et al*. (2014), with the significant addition of an optimal spatial design approach. Specifically, in this work we develop hierarchical spatial statistical models for environmental outcomes conditional on functional predictors; *i.e*., spectral measurements. Conditional on a set of spectral and non-spectral (*e.g*., topography, soil apparent electrical conductivity (ECa)) predictors, our methodology provides estimates and measures of uncertainty of various soil properties and allow for digital soil mapping. Further, by using the Bayesian SSVS implementation, this methodology also extracts signals/features from the high-dimensional spectral measurements that are important predictors of specific environmental outcomes. Importantly, our approach can be applied to the problem of choosing optimal sampling locations for taking measurements in future studies, thus improving the efficiency of future data collection (*e.g*., Wikle and Royle 1999, 2005; Hooten *et al*. 2009, Holan and Wikle 2012, Hooten *et al*. 2012).

Section 2 describes the hierarchical model and methodology for signal extraction and spatial prediction given high-dimensional spectroscopic signals. In addition, this section outlines the optimal spatial design methodology derived from this model. Section 3 then contains a real-world data example related to the prediction of spatially referenced cation exchange capacity given spectral measurements of soil conductivity and elevation, as well as the selection of optimal sampling locations. We then conclude with a brief discussion in Section 4.

2. METHODOLOGY

Over the last three decades or more, researchers have estimated soil properties using visible and near-infrared (VNIR) diffuse reflectance spectroscopy (DRS), with varying degrees of success. More importantly, many of the previous modeling attempts using VNIR and DRS have neglected to provide measures of uncertainty associated with particular estimates. Critically, in order to reduce sampling costs while simultaneously improving the quality of the estimates an adaptive/dynamic design strategy is needed to choose optimal sampling locations.

The details of the methodology are described in the following subsections. We use standard hierarchical statistical model notation; for two random variables X and Y, we denote the distribution of X and conditional distribution of X given Y as $[X]$ and $[X|Y]$, respectively. Then, heuristically, for hierarchical geospatial functional data models, like the ones considered here, the idea is to approach the problem by breaking it into several stages (Berliner 1996, Wikle *et al*. 1998, Wikle 2003, Cressie and Wikle 2011):

- *Stage* 1. Data Model: [*data* | *process, parameters*]
- *Stage* 2. Process Model: [*process* | *parameters*]
- *Stage* 3. Parameter Model: [*parameters*].

The first stage is concerned with the observational process or "data model," which specifies the distribution of the data given the true values (process of interest) and parameters that describe the data model. The second stage then describes the process of true values, conditional on other parameters. Finally, the last stage accounts for the uncertainty in the parameters. Ultimately, our main interest is in the distribution of the process of true values and the parameters updated by the data. We obtain this so-called "posterior" distribution via Bayes' rule:

$$[\textit{process, parameters} \mid \textit{data}]$$
$$\propto [\textit{data} \mid \textit{process}, \textit{parameters}]$$
$$[\textit{process} \mid \textit{parameters}] \times [\textit{parameters}].$$

The Bayesian hierarchical approach serves as the basis for a flexible framework consisting of fundamental methodology that can be adapted to handle many core substantive applied spatial problems of interest, such as those encountered here (*e.g*., Holan *et al*. 2008, 2009, 2010; Wikle and Berliner 2005; Wikle and Hooten 2010, Wikle 2010a). Importantly, the Bayesian hierarchical framework allows one to include multiple data sources conditional upon the true process of interest, thus, typically simplifying the data-model dependence structure. Perhaps more importantly, one then uses the spatial dependence in the process stage to borrow strength across spatial areas, thereby reducing variability in the true process estimates and in predictions. In many instances, exogenous information and additional dependence structure can also be added into the models for parameters associated with these higher stages (*i.e*., data and process), providing a more flexible, yet probabilistically coherent, way to combine information.

2.1 Hierarchical Geospatial Spectroscopic Functional Models

Consider the case of one field for which we have observations. Specifically, assume there are m observations of a spatial process (soil property), denoted by $\{Y_i : i = 1, ..., m\}$ and define a latent spatial vector, $\breve{\mathbf{u}} = (\breve{u}_1, ..., \breve{u}_m)'$, where $\breve{u}_i$ is from a Gaussian spatial process and the index i is associated with a spatial location within the given field (assumed to be in some subset of two-dimensional Euclidean space). In general, the locations corresponding to the latent vector $\breve{\mathbf{u}}$ may not coincide with the observation locations (*i.e*., the observation and prediction locations need not be the same). Nevertheless, we will assume that the observations and latent spatial process have the same point-level support. Further, let $s_i(\omega)$ denote a location-specific mean centered spectroscopic curve for location i and wavelength ω. Then

$$Y_i = \int s_i(\omega) b(\omega) d\omega + \mathbf{x}_i'\boldsymbol{\delta} + \breve{u}_i + \breve{\varepsilon}_i, \qquad (1)$$

where the regression parameter $b(\cdot)$ is assumed smooth and square integrable and $\mathbf{x}_i$ is a p-dimensional vector of non-spectral covariates observed at location i having p × 1 vector of parameters, δ.

Equation (1) can be viewed as an infinite dimensional regression, at location i, with a separate predictor for every wavelength, ω, in the continuous band. Now, if $\{\phi_k(\omega)\}$ represents a complete orthonormal basis then s_i and $b(\cdot)$ have the unique representations $s_i(\omega) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(\omega)$ and $b(\omega) = \sum_{k=1}^{\infty} \beta_k\phi_k(\omega)$. Substituting the above expansions into (1) yields

$$\begin{aligned} y_i &= \int \sum_{k=1}^{\infty} \xi_{ik}\phi_k(\omega) \sum_{k=1}^{\infty} \beta_k\phi_k(\omega) d\omega + \mathbf{x}_i'\boldsymbol{\delta} + \breve{u}_i + \breve{\varepsilon}_i \\ &= \sum_{k=1}^{\infty} \xi_{ik}\beta_k + \mathbf{x}_i'\boldsymbol{\delta} + \breve{u}_i + \breve{\varepsilon}_i \\ &= \sum_{k=1}^{K} \xi_{ik}\beta_k + \mathbf{x}_i'\boldsymbol{\delta} + u_i + \varepsilon_i \\ &= \xi_i'\boldsymbol{\beta} + \mathbf{x}_i'\boldsymbol{\delta} + u_i + \varepsilon_i, \end{aligned} \qquad (2)$$

where $s_i(\omega)$ and $\mathbf{x}'_i$ are assumed known for $i = 1, ..., m$. Note that the truncation error $\sum_{k=K+1}^{\infty} \xi_{ik}\beta_k$ can be accounted for through flexible specification of the error distributions; *i.e*., by appropriately modifying the error distributions for $\breve{u}_i$ and $\breve{\varepsilon}_i$ (notationally we denote this modification by removing the "$\breve{}$" symbol). See Wikle (2010b) for a detailed discussion. We remark that, in practice, similar to Holan *et al*. (2010, 2012), Wikle and Holan (2011) and Yang *et al*. (2014), $\sum_{i=1}^{K} \beta_k\xi_{ik}$ will undergo further dimension reduction through stochastic search variable selection (SSVS) as described below (George and McCulloch 1993, 1997).

In matrix form, (2) can be equivalently expressed as

$$\mathbf{Y} = \Xi\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\delta} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \qquad (3)$$

where $\mathbf{Y}$ has dimension $m \times 1$ and $\Xi = (\xi_1, ..., \xi_m)'$ and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)'$ have dimension $m \times K$ and $m \times p$, respectively. In this context, the matrix Ξ is known since $\boldsymbol{s}_i = (s_i(\omega_1), ..., s_i(\omega_q))' = \Phi\xi_i$, with $\Phi = (\phi(\omega_1), ..., \phi(\omega_q))'$, where the signal $\boldsymbol{s}_i$ is measured at q discrete wavelengths and Φ and ξ_i have dimensions $q \times K$ and $K \times 1$, respectively. Thus, it follows that $\xi_i = \Phi' \boldsymbol{s}_i$ and, therefore, $\Xi = \boldsymbol{S}\Phi$ where $\boldsymbol{S} = (\boldsymbol{s}'_1, ..., \mathbf{s}'_m)'$ has dimension $m \times q$. In terms of a general hierarchical framework, our geospatial functional model can be specified as

$$\text{Data Model: } [\mathbf{Y}|\beta, \delta, \mathbf{u}, \sigma^2_\epsilon] = N(\Xi\beta + \mathbf{X}\delta + \mathbf{u}, \sigma^2_\epsilon \mathbf{I}) \quad (4)$$

$$\text{Process Model: } [\mathbf{u}|\theta_u] = N(\mathbf{0}, \Sigma(\theta_u)) \quad (5)$$

$$\text{Parameter model: } [\beta, \delta, \sigma^2_\epsilon, \theta_u], \quad (6)$$

where σ^2_ϵ corresponds to an independent, small-scale (nugget) spatial effect and/or measurement error process ϵ, and θ_u is a vector of parameters associated with the latent spatial process $\mathbf{u}$.

It is important to note that no forms for the orthonormal expansions were specified. In fact, there are many choices for these expansions among the class of orthogonal (or biorthogonal) bases (*e.g.*, Fourier, splines, wavelets, empirical orthogonal functions (EOFs; *i.e.*, functional principal components), among others). The choice is somewhat subjective, but there can be advantages and disadvantages depending on the specific application. Additionally, the residual error term, ϵ, is typically assumed to have covariance matrix $\sigma^2_\epsilon \mathbf{I}$. However, a more general covariance matrix Σ_ϵ could be specified to help account for differences between the complete basis expansion and its low-dimensional representation. See Wikle (2010b) for a comprehensive discussion.

The model given by (4)-(6) can be equivalently expressed by combining the data and process stages into one stage (*i.e.*, by integrating out the random latent process, $\mathbf{u}$, leading to more complicated marginal dependence). Specifically,

$$\text{Data Model: } [\mathbf{Y}|\beta, \delta, \sigma^2_\epsilon, \theta_u] = N(\Xi\beta + \mathbf{X}\delta, \Sigma(\theta_u) + \sigma^2_\epsilon \mathbf{I}) \quad (7)$$

$$\text{Parameter model: } [\beta, \delta, \sigma^2_\epsilon, \theta_u]. \quad (8)$$

Also, note that it is usually the case that the parameter distributions in (8) are considered to be independent, $[\beta, \delta, \sigma^2_\epsilon, \theta_u] = [\beta][\delta][\sigma^2_\epsilon][\theta_u]$. Clearly, other choices are available and one should verify that posterior inference is not overly sensitive to the choice of parameter distributions. Such sensitivity analyses are important when conducting Bayesian analysis and are undertaken as part of our model development. The connection between the model given by the fully hierarchical form (4)-(6) and the marginal form (7)-(8) is exactly the same as in the traditional linear mixed model setting. When inference is concerned with β and δ, it may be more convenient to proceed with the so-called marginal form (*i.e.*, "spatial functional regression analysis").

In most traditional geostatistical settings, interest resides in predictions of the spatial process at unobserved locations. To facilitate "big data" spatial prediction, we consider an additional basis function expansion. For this purpose, let $\xi_{ik} = \psi'_i v_k$, where ψ_i and v_k both have dimension $r \times 1$. This further basis function decomposition allows us to estimate ξ_{ik} at any spatial location i given that we know v_k. Additionally, we write the $m \times K$ matrix $\Xi = \Psi\mathbf{V}$, where $\Psi_{m \times r} = (\psi_1, ... \psi_m)'$ and $\mathbf{V}_{r \times K} = [v_1, v_2, ..., v_K]$. Note that if the spatial basis functions Ψ are known, then v_k can be obtained by $\mathbf{V} = \Psi' \Xi = \Psi' \mathbf{S} \Phi$. Alternatively, letting $\breve{\gamma}_k = (\gamma_{1k}, ..., \gamma_{mk})'$ we could impose spatial structure by specifying $\breve{\gamma}_k \sim N(\mathbf{0}, \Sigma(\theta_{\gamma k}))$ or $\breve{\gamma}_k = \Psi v_k$ with $v_k \sim N(\mathbf{0}, \Sigma_{v,k})$. We take the former approach here.

Thus, if we are interested in the process Y_i at n locations, say $\tilde{\mathbf{Y}}$, by taking the spatial basis function approach, we can use the model

$$\tilde{\mathbf{Y}} = \tilde{\Psi}\Psi'\mathbf{S}\Phi\beta + \tilde{\mathbf{X}}\delta + \tilde{\mathbf{u}} + \tilde{\epsilon},$$

where $\tilde{\Psi}_{n \times r} = (\psi_1 ... , \psi_n)'$, $\tilde{\mathbf{X}}_{n \times p} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$, $\tilde{\mathbf{u}}_{n \times 1} = (u_1, ..., u_n)'$, $\tilde{\epsilon} = (\epsilon_1, ..., \epsilon_n)'$ and $\mathbf{x}_i$ is assumed known for $i = 1, ..., n$. Migration to the case where $\mathbf{x}_i$ is not known for all $i = 1, ..., n$ is relatively straightforward given a mechanism for dealing with unobserved spectroscopic (functional) predictors (*e.g.*, Yang *et al*. 2014). Furthermore, let $\tilde{\Psi}\Psi'\mathbf{S}\Phi = \mathbf{M}_{n \times K}$, then

$$\tilde{\mathbf{Y}} = \mathbf{M}\beta + \tilde{\mathbf{X}}\delta + \tilde{\mathbf{u}} + \tilde{\epsilon}. \quad (9)$$

Finally, in order to generate predictions of the spatial process at unobserved locations we must specify the prior distributions of $\beta, \delta, \tilde{\mathbf{u}}$ and $\tilde{\epsilon}$ in (9). For this purpose, we assume $\beta \sim$ SSVS prior (as described below), $\delta \sim N(\mathbf{0}, \Sigma(\theta_\delta))$ (or an SSVS prior), $\tilde{\mathbf{u}} \sim N(\mathbf{0}, \Sigma(\theta_u))$ and $\tilde{\epsilon} \sim N(\mathbf{0}, \sigma^2_\epsilon \mathbf{I})$.

It is important to note that the above models can be readily extended to include a third spatial dimension – depth. In this case, instead of observing a DRS curve at each spatial location, a DRS image is observed (see Fig. 1 for an example image). The DRS image

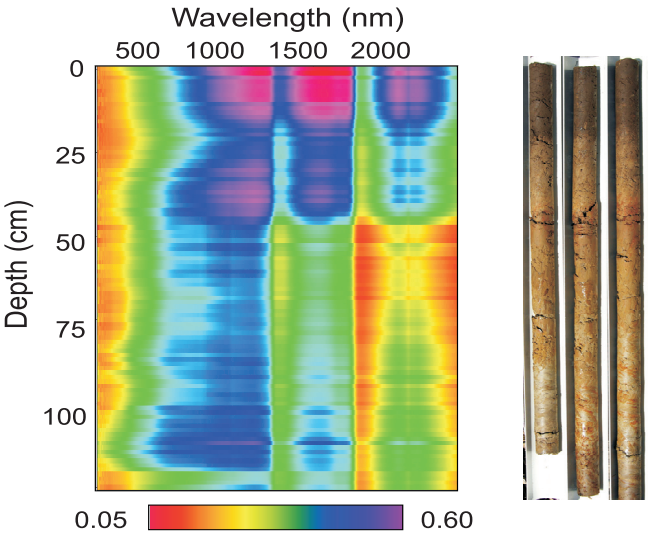

Fig. 1. This image consists of approximately 120 individual spectra sampled every 1 cm on a 120-cm core. Color scale represents variation in decimal reflectance.

predictors can then be modeled similar to the DRS curves, once the image has been vectorized (*e.g.*, see Holan *et al*. 2010, 2012 and Yang *et al*. 2014). Additionally, the above models can be readily extended to include multiple sites and/or multivariate responses. In each of these cases the hierarchical statistical approach proposed here facilitates the additional model complexity; see Cressie and Wikle (2011) for a comprehensive discussion.

2.1.1 Stochastic Search Variable Selection (SSVS)

In general, Bayesian SSVS algorithms provide an effective means of model selection when interest lies in considering a large number of potential submodels (*e.g.*, see George 2000, OHara and Sillanpää 2009, for a detailed overview). More specifically, as popularized by George and McCulloch (1993, 1997), SSVS provides a quick and efficient way of performing variable selection within the hierarchical Bayesian framework. For example, assume we have a basic multiple regression problem with $Z_i \mid \beta, \sigma^2 \sim N(\mathbf{x}_i'\beta, \sigma^2)$, where β is an n_β-dimensional vector and n_β is quite large. One standard implementation then assumes the following prior specification for the elements of β:

$$\beta(j) \mid \gamma_j \sim \gamma_j N(0, c_j\tau_j^2) + (1-\gamma_j)N(0,\tau_j^2), \quad j = 1, \ldots, n_\beta, \qquad (10)$$

where $\{\gamma_j\}$ are specified at the next level of the hierarchy to have independent Bernoulli distributions with probability parameters $\{\pi_j\}$. In this case, π_j can be viewed as the prior probability that $\beta(j)$ should be included in the model. Furthermore, note that $\gamma_j = 1$ indicates that the j-th variable is included in the model. Typically, c_j, τ_j, and π_j are fixed hyperparameters. George and McCulloch (1993, 1997) provide various alternatives for the specification of these hyperparameters. They suggest that one would like τ_j to be small so that, when $\gamma_j = 0$, it is reasonable to specify an effective prior for $\beta(j)$ that is near zero. In addition, one typically wants c_j to be large (greater than 1) so that, if $\gamma_j = 1$, then our prior would favor a non-zero $\beta(j)$.

To facilitate further dimension reduction of (2) we allow the possibility of both the spectral and non-spectral regression parameters to have SSVS priors. In doing so, heuristically, the hierarchical geospatial spectroscopic model can be viewed as spatial regression model in which the regression produces an extremely low-dimensional "filtered" representation of the infinite dimensional spectral and finite dimensional non-spectral predictors. Regardless of the specific orthogonal basis functions chosen, this approach has the distinct advantage that, similar to partial least squares, it will facilitate dimension reduction by choosing the basis components (most salient spectroscopic components) and non-spectral predictors that best predict the response; see Holan *et al*. (2010, 2012), Wikle and Holan (2011), and Yang *et al.* (2014) for a detailed overview on using SSVS to choose effective basis expansions and variable selection.

2.2 Adaptive Design for Spatial Spectroscopic Sampling

The types of environmental processes of general interest here include variability over space and possibly

time, and given the opportunity to sample over different spatial locations through time, or to sample in space adaptively, it is possible to improve the design efficiency. There has been substantive work in the statistics literature devoted to the issue of obtaining spatial designs for environmental processes (Le and Zidek 2006, Wikle and Royle 1999, 2005, Hooten *et al*. 2009, Holan and Wikle 2012, Hooten *et al*. 2012). For a comprehensive overview of optimal spatial designs see Müller (2007), Mateu and Müller (2012) and the references therein. As demonstrated in Wikle and Royle (1999), one can gain significant improvement in efficiencies by allowing the design to change at each sampling time. This is typically referred to as adaptive (or dynamic) design, which is not to be confused with adaptive sampling (*e.g*., Thompson and Seber 1996), although the notions are similar. The primary difference is that the former is model based, whereas the latter occurs in a traditional sample survey framework, but with sample unit selection modified as observations are made.

In general, the design problem is to decide where spectroscopic samples should be taken at time $t + 1$ based on observations through time t. In particular, let Y denote a spatial or spatio-temporal process such as cation exchange capacity (CEC) and suppose that the process, spectroscopic measurements and exogenous variables are sampled at a set of n locations (the "design") denoted as $\mathbf{D} = (d_1, ..., d_n)$ where $d_i \in \mathcal{R} \subset \mathcal{F}$ and $\mathcal{F}$ is the agricultural field of interest. The design objective is to locate $m << n$ sampling locations in an optimal fashion, meaning the design minimizes some variance criterion. Common criteria include the average prediction variance, maximum prediction variance, minimum root mean-square prediction error (RMSPE), or the variance of regression parameter estimates.

The adaptive design problem we consider is challenging. The principal difficulty is that model-based approaches for spatio-temporal spectroscopic (functional) data are relatively complicated and, for design applications, must be specified in such a way as to allow for prediction of the spectroscopic covariates at unobserved locations. Any model lacking this facet will be inadequate for constructing designs.

For the application presented here, we consider the purely spatial case corresponding to (3) and note that the spatio-temporal case follows easily given a specified model; see Wikle and Royle (1999), Wikle and Royle (2005), Cressie and Wikle (2011), Holan and Wikle (2012), and the references therein for comprehensive details. The goal in the purely spatial case is to choose an optimal design of size $m < n$ sampling locations based on the current observations and to use them at the next sampling occasion (assuming close temporal proximity in sampling occasions). For this purpose we define the $n \times m$ incidence matrix, $\mathbf{H}$, that maps the observations to the candidate subset sampling locations. In order to choose an optimal design we then consider the following model

$$\mathbf{Y} = \mathbf{H}\mu + \in, \in \sim N(\mathbf{0}, \sigma^2_{\in}\mathbf{I}),$$
$$\mu = \Xi\beta + \mathbf{X}\delta + \mathbf{u}, \mathbf{u} \sim N(\mathbf{0}, \Sigma(\theta_u)).$$

Assuming that the parameters $\beta, \delta, \sigma^2_{\in}$ and θ_u are known, the posterior distribution of $\mu|\mathbf{Y}$ is given by

$$\mu|\mathbf{Y} \sim N((\Xi\beta + \mathbf{X}\delta) + \mathbf{R}[\mathbf{Y} - \mathbf{H}(\Xi\beta + \mathbf{X}\delta)], (\mathbf{I}–\mathbf{RH})\ \Sigma(\theta_u)),$$

where $\mathbf{R} = \Sigma(\theta_u)\mathbf{H}'\ (\sigma^2_{\in}\mathbf{I} + \mathbf{H}\Sigma(\theta_u)\mathrm{H}')^{-1}$. In practice, one plugs in estimates of the required parameters to evaluate this distribution, where those estimates are obtained from the posterior distribution for the model implementation given in Section 2.1.

The optimal design could be chosen as the one that reduces the prediction variance associated with the locations (corresponding to the rows of "**H**") that minimize the trace of the prediction variance $(\mathbf{I} - \mathbf{RH})\Sigma(\theta_u)$. Or, in the case where one is seeking to compare to a validation sample (as in our example), one could select the locations that lead to the best predictions at the validation locations and use a RMSPE criterion. In either case, to find these locations, we need an efficient way to go through "all" possible combinations of locations. In practice, this is problematic due to the enormous number of possibilities to consider. To carry out this minimization we propose a simple exchange algorithm. The basic idea is that the criterion is evaluated successively for different designs, and the design is updated by exchanging bad sampling locations for better sampling locations. Such algorithms are widely used in practice and many variations on the basic theme exist (Cook and Nachtsheim 1980, Atkinson and Federov 1988, Nychka *et al.* 1997). Although these algorithms are somewhat greedy and tend to find local optima, for relatively small problems, like the ones considered here, experience indicates that they do find the global optimum, or a design very close to the optimum. For larger problems, the solutions tend to be arbitrarily close to the global optimum depending on how long the algorithm is allowed to run.

3. EXAMPLE: MODELING SOIL CATION EXCHANGE CAPACITY

To illustrate the effectiveness of our methodology, in terms of prediction and optimal spatial design, we present an example using the methodology described above. Data were collected on a 35 hectare (ha) research field located near Centralia, Missouri on a nominal 30-m grid. See Sudduth *et al*. (2010) for comprehensive discussion surrounding the study site, soil sampling and spectral data acquisition.

3.1 Illustration of SSVS and Spatial Prediction

We considered the response variable cation exchange capacity (CEC), with covariates consisting of reflectance spectra shown in Fig. 2, soil ECa measured by a Geonics EM38 sensor operated in vertical dipole mode (Sudduth *et al*. 2005), and elevation (height above

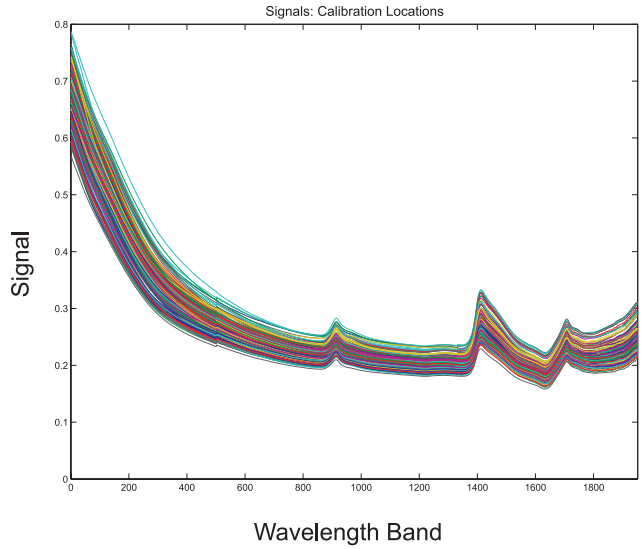

Fig. 2. Spectroscopic signals for the 285 training data locations.

sea level (*m*)). We used 285 locations for calibration (model estimation) and 69 locations for validation (prediction). The signals were projected onto a functional principal component basis set. Specifically, we considered the first 20 principal components (accounting for over 99.9% of the variance in the signals) as potential covariates.

In terms of modeling choices, the nugget variance, σ^2_ϵ, was fixed at 1 mEq/100 g; this value comes from independent repeated measurements that were specifically taken to determine the error associated with soil sampling and laboratory analysis. Further, the spatial process **u** was assumed to have an exponential covariance structure with σ^2_u specified as an inverse gamma distribution with a prior mean of 0.9 and variance 10, and θ_u was specified to have a discrete uniform prior with bounds from 0.01 to 2.0 (in this study, distances range from 0 to 876 meters). Note, the results presented here are not sensitive to these choices of prior distributions.

The spectroscopic covariate parameters were assigned a SSVS prior as defined in (10) with $\tau = .01$, $c = 10$ and $\pi = .25$. These were selected based on a sensitivity analysis that considered all possible combinations of $\tau = \{0.01, 0.05, 0.1, 0.15, 0.2, 0.5\}$, $c = \{0.1, 0.5, 1, 5, 10, 20\}$, and $\pi = \{0.1, 0.25, 0.5\}$. Note that the choice of $\pi = .25$ allowed more shrinkage to 0 and thus, more parsimony. Spatial predictions of the ξs are based on bi-harmonic splines in order to obtain the curves at the validation locations. Finally, the predictions at the validation locations are performed at each iteration of the Markov chain Monte Carlo (MCMC) procedure (after appropriate burn-in), thus giving a "model averaged" prediction; see Holan *et al*. (2010, 2012) for complete details. All full-conditional distributions were of conjugate form and sampled via Gibbs steps. The MCMC algorithm was run for 10,000 iterations after a 1,000 iteration burn-in. There was no evidence of lack-of-convergence in the sample chains.

For the types of models considered in this illustration it is informative to examine a plot of the posterior probability of inclusion for the principal component coefficients based on the SSVS prior distribution (see Fig. 3). In this case, the β coefficients

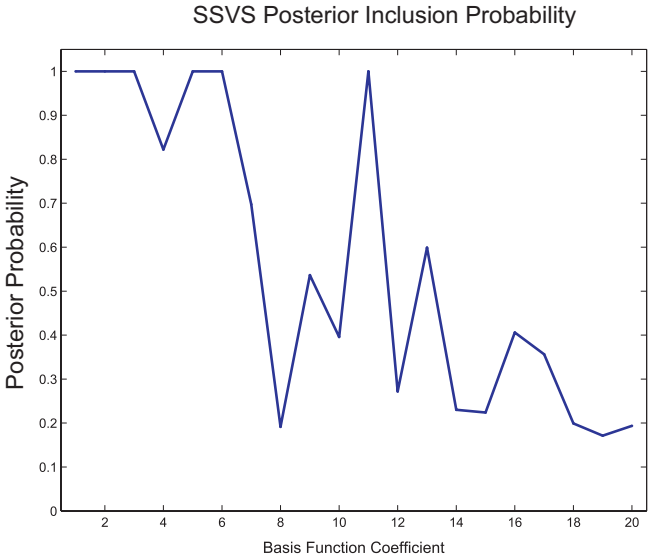

Fig. 3. Posterior probability of inclusion for the functional principal component coefficients based on SSVS.

associated with functional principal component 1, 2, 3, 5, 6, and 11 were always included in the model. The coefficients associated with principal component 9 and 13 were included over 50% of the time, with the remaining coefficients included less than 40% of the time. This figure provides a visualization tool for evaluating which basis coefficients are important for predicting CEC.

As depicted in Fig. 4 the posterior distribution of $\beta(1)$ is unimodal, whereas the posterior distribution for $\beta(13)$ is bi-modal, which is a result of the SSVS prior distribution. This is important since it demonstrates that the Bayesian hierarchical modeling approach gives us the flexibility to find bimodality in the parameters. In terms of the other variables, Fig. 4 also displays the posterior histogram of σ_u^2 and θ_u. Note, as seen from the posterior distribution of θ_u, the posterior mean estimate of θ_u (0.028) is relatively small, indicating weak spatial dependence after accounting for the covariates, yet there is a reasonable spread in this distribution. Note that we have not attempted to account for potential spatial confounding in this analysis since the goal is prediction (*e.g.*, Reich *et al*. 2006).

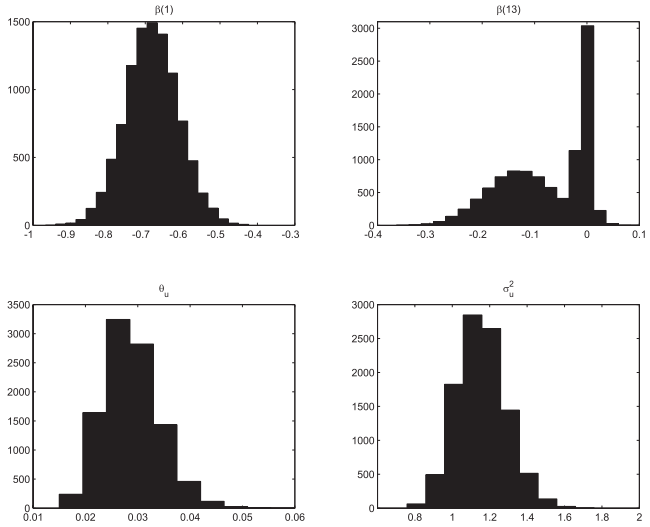

Fig. 4. Histogram of the posterior distribution for relevant model parameters. Top left: $\beta(1)$, Top right: $\beta(13)$, Bottom left: θ_u, right: σ_u^2.

Another informative plot is that of the posterior prediction distributions for all of the 69 validation locations (Fig. 5). This figure illustrates the posterior prediction distributions for the 69 locations as blue box-plots with the "true" values superimposed in red. Importantly, this demonstrates the ability of the hierarchical approach to capture (quantify) the uncertainty in our estimates. This uncertainty quantification is critical when using these measurements for making economic and/or managerial decisions and will be of immediate use to the end-user. Note that most of the predictive distributions cover the truth, with a few clear outliers.

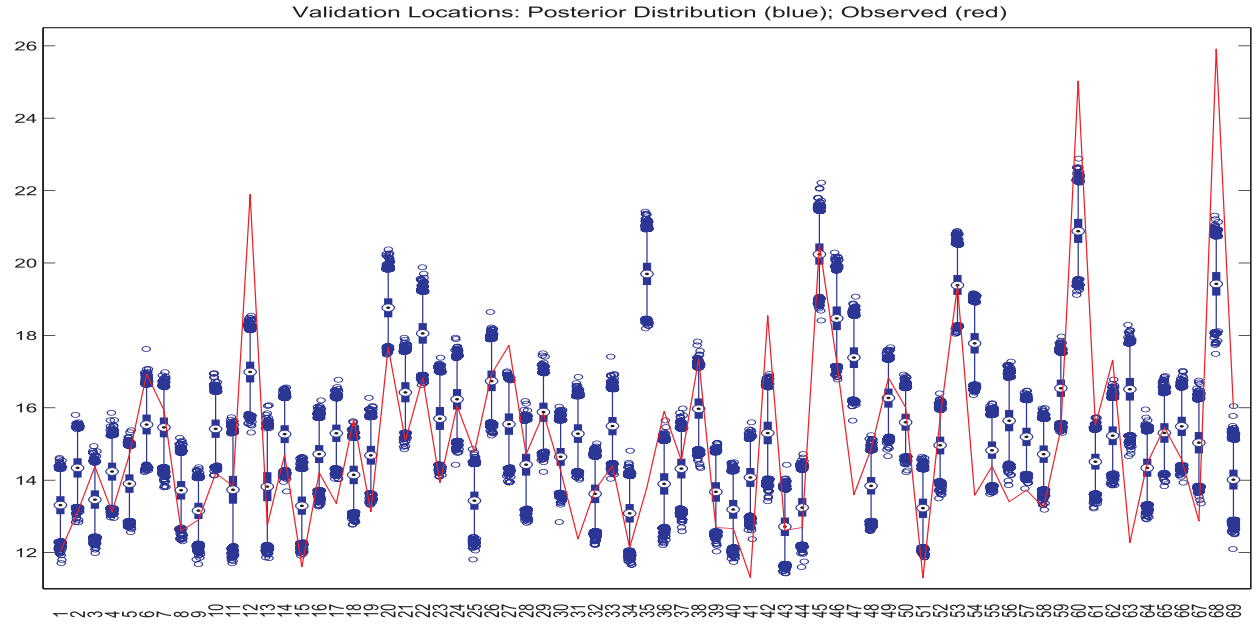

Fig. 5. Posterior prediction distribution for each of the 69 validation locations. Blue box-plots denotes the posterior prediction distributions whereas the "true" values are denoted in red.

3.2 Illustration of Optimal Spatial Design

As an illustration of the spatial design methodology, we again consider the Centralia dataset used for the prediction analysis in Section 3.1. Specifically, we demonstrate that, for given number of locations, using our approach leads to designs with significantly reduced root mean-square prediction error versus what would be obtained utilizing the best design from 10,000 random arrangements. For this illustration, we utilized the estimated posterior mean for spatial parameters, $\hat{\sigma}_u^2 = 1.16$ and $\hat{\theta}_u = 0.028$, from our previous Bayesian analysis. Similarly, the measurement error variance was again fixed at 1 mEq/100 g. For this example, we used the functional principal components that were always selected in the SSVS procedure (*i.e.*, 1, 2, 3, 5, 6, 11). This accounted for over 99.6% of the variation in the signals. For a given design, we estimated the signal and non-signal parameters using generalized least squares. Subsequently, we interpolated the spectral coefficients at the validation sites (as previously discussed) based on the design locations.

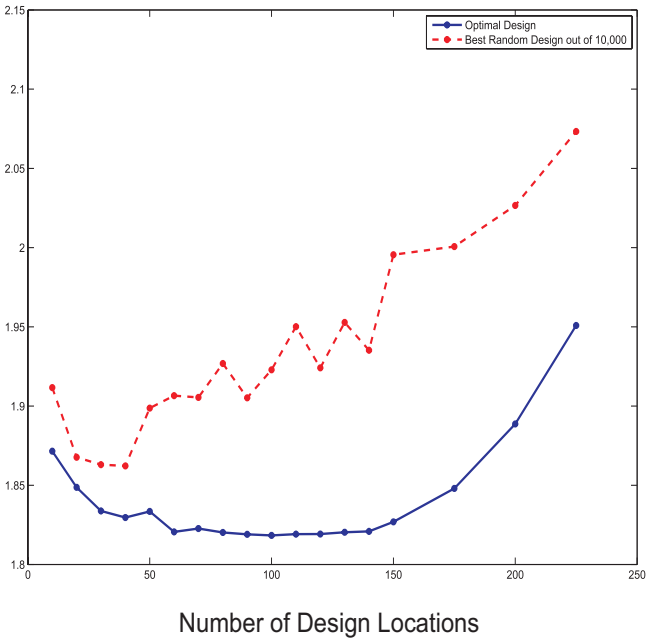

Fig. 6. RMSPE for the optimal design (blue solid line) and "best" design out of 10,000 random designs (red dashed line) as a function of the number of design locations.

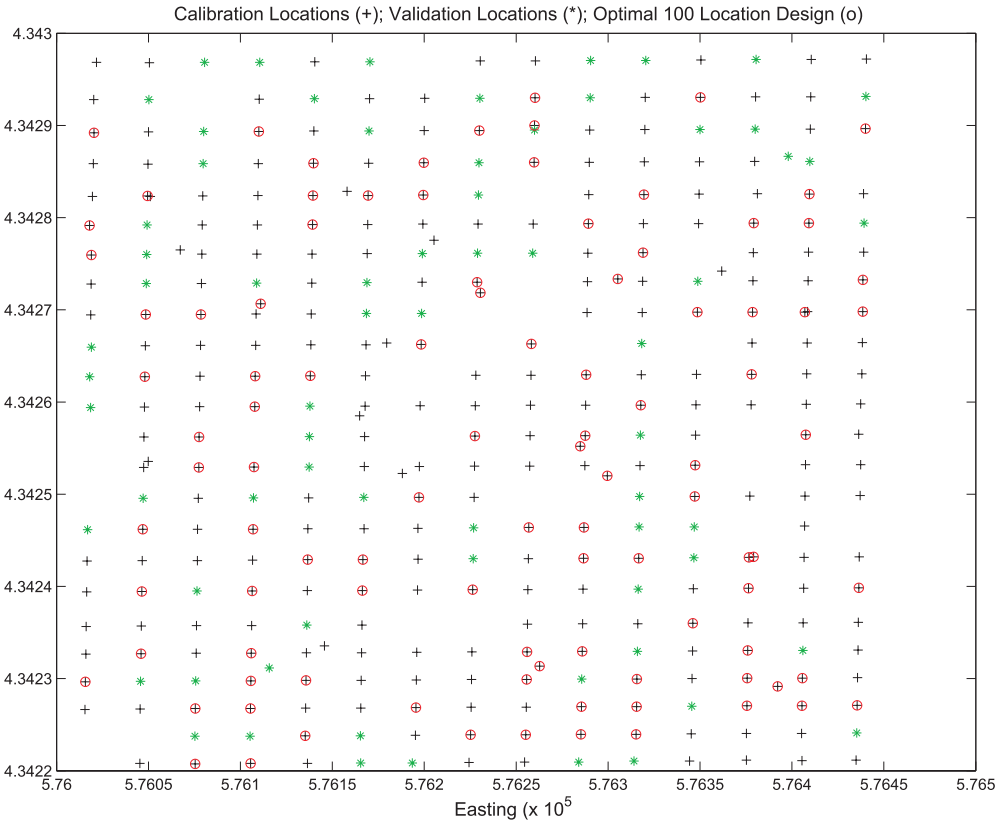

Fig. 7. Calibration locations (blue +), validation locations (green *) and the optimal 100 design locations (red o).

Note, for this illustration, the nonspectral covariates were assumed known at the validation sites but not the spectral curves. Next, we estimated the mean, $\Xi\beta + \mathbf{X}\delta$. Using Kriging formulae (Cressie and Wikle 2011) we obtained estimates of the mean and variance-covariance matrix of $\mu|\mathbf{Y}$, as described above. Based on the RMSPE design criterion we then found an optimal spatial design as function of the number of sites. Fig. 6 shows the RMSPE for the optimal design and the best design out of 10,000 random designs for 10 to 225 design locations. For comparison, note that the RMSPE for the validation locations based on using all of the sample locations from the posterior prediction was 2.02. It is interesting to note that using all the sampling locations results in an overfit, as the optimal spatial designs for fewer sample locations all produced smaller RMSPE values. Fig. 6 shows that the minimum RMSPE (1.819) occurred for 100 sampling locations, if they were chosen based on the optimal design algorithm, but there was not much difference in RMSPE in any of the optimal designs for sizes from 60-140 locations. In addition, the RMSPE was significantly smaller for all network sizes than the corresponding "best" design based on 10,000 random designs of the same size. Finally, Fig. 7 shows the optimal sampling network of size 100 (red circles) superimposed on all of the possible network locations.

It is important to emphasize that our approach is extremely flexible on several grounds. First, in the case were we have no prior spectroscopic data to include in the model, our formulation could be modified to choose optimal starting points for spectroscopic sampling based on other environmental covariates, including digital elevation maps (DEMs) or remote sensing images. Additionally, our approach is easily adapted to the case of spatiotemporal design in which subsequent designs are determined from the dynamics governing the evolution of the environmental outcome of interest conditional on the previously imposed designs. Finally, when including spectroscopic predictors, our designs are conditional on the estimates β. These parameters are given an SSVS prior distribution and thus the resulting design includes information on which principal component features are important for spatial prediction.

4. DISCUSSION

Sustainable agriculture requires a site-specific approach to address crop management problems and environmental degradation processes that are spatially and temporally variable. These issues lead to production losses (water stress, low fertility, pest problems), soil degradation (erosion, soil organic carbon losses, compaction), and water quality degradation (sediment, nutrients, agrochemicals) – often at the sub-field scale (Kitchen *et al*. 2005, Lerch *et al*. 2005). Unfortunately, landscape processes and properties can change at a finer spatial resolution than can be practically analyzed with lab methods due to time and cost of sampling and analysis (Sudduth *et al*. 1997). Thus, it is increasingly important to augment lab methods with field-sensor methods such as VNIR that can accurately characterize within-field variability at a more reasonable cost and with reliability and timeliness.

The body of work on VNIR soil analysis, as reviewed by Malley *et al*. (2004), Viscarra Rossel *et al*. (2006) and Stenberg *et al*. (2010) includes a multiplicity of separate solutions for the suite of problems with calibration of VNIR sensors. Some of these problems are data preprocessing and smoothing (Igne *et al.* 2010), data reduction and wavelength selection (Sudduth and Hummel 1991, Lee *et al*. 2009), local and global calibration (within and between regions) (Brown *et al*. 2005, Lee *et al*. 2010), spatio-temporal autocorrelation in calibration data (Ge *et al*. 2007, Sudduth *et al*. 2010), and the final step of interpolating the predicted data.

The hierarchical statistical methodology presented here can provide solutions for all of these problems in a single framework (Wikle *et al*. 1998). To quantify uncertainty, it is of principal importance to properly handle and propagate error (uncertainty) between all of these previously independent steps. Also, accounting for these uncertainties and the spatial dependency allow for the possibility of developing efficient monitoring designs for critical parameters such as soil organic carbon, soil nutrients, and soil moisture. In particular, the adaptive/dynamic design capabilities of the method will also enable more accurate calibration and interpretation of spectral data at a reduced cost. This is achieved through stochastic selection of functional principal components for spatial prediction and the corresponding spatial locations for sampling (to optimize sampling costs and prediction accuracy). Further, these hierarchical statistical models readily extend to multiple predicted variables and can incorporate additional inexpensive and rapidly collected geospatial covariate information to simultaneously improve VNIR calibration and spatial predictions.

We expect that the methods and approaches presented here will be of interest to a wide range of users. They could be directly used by sensor developers and researchers who may imbed them in commercial or experimental data analysis systems. This could in turn enable use by educators, consultants, and producers as part of an overall sensing and analysis system.

ACKNOWLEDGEMENTS

Funding for the methodological research was partially provided through National Science Foundation grant DMS-1049093.

REFERENCES

Atkinson, A.C. and Federov, V.V. (1988). Optimum design of experiments. In: *Encyclopedia of Statistics*. (Supplemental Volume), (eds) Kotz, S., Johnson, NI, Wiley, New York, 107-114.

Baladandayuthapani, V., Mallick, B., Young Hong, M., Lupton, J., Turner, N., and Carroll, R. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics,* **64**, 64-73.

Berliner, L. (1996). Hierarchical Bayesian time-series models. In: *Fundamental Theories of Physics*, **79,** 15–22.

Brown, D.J. (2007). Using a global vis-NIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma*, **140,** 444-453.

Brown, D.J., Bricklemyer, R.S. and Miller, P.R. (2005). Validation requirements for diffuse reflectance soil characterization Models with a case study of VNIR soil C prediction in Montana. *Geoderma,* **129,** 251-267.

Chang, C.W., Laird, D.A., Mausbach, M.J. and Hurburgh Jr., C.R. (2001). Near-infrared reflectance spectroscopy - principal component regression analysis of soil properties. *Soil Sci. Soc. Amer. J.*, **65,** 480-490.

Christy, C.D. (2008). Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Comp. Elect. Agric.,* **61**, 10-19.

Cook. R.D. and Nachtsheim, C.J. (1980). A comparison of algorithms for constructing the exact D-optimal designs. *Technometrics*, **22,** 315-324.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data.* John Wiley and Sons.

Dalal, R.C. and Henry, R.J. (1986). Simultaneous determination of moisture, organic carbon, and total nitrogen by near-infrared reflectance spectrophotometry. *Soil Sci. Soc. Amer. J.*, **50,** 120-123.

Daniel, K.W., Tripathi, N.K. and Honda, K. (2003). Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of lop buri (Thailand). *Austr. J. Soil Res.,* **41**, 47-59.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, **21,** 224-239.

Fidencio, P.H., Poppi, R.J. and De Andrade, J.C. (2002). Determination of organic matter in soils using radial basis function networks and near-infrared spectroscopy. *Analytica Chimica Acta*, **453,** 125-134.

Ge, Y., Thomasson, J.A., Morgan, C.L. and Searcy, S.W. (2007). VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. *Trans. SABE*, **50,** 1081-1092.

George, E. (2000). The variable selection problem. *J. Amer. Statist. Assoc.*, **95**, 452.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88,** 881-889.

George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7,** 339-374.

Giraldo, R., Delicado, P. and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *J. Agric. Biol. Environ. Statist.*, **15,** 66-82.

Giraldo, R., Delicado, P. and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, **66,** 40-421.

Goulard, M. and Voltz, M. (1993). Geostatistical interpolation of curves: A case study in soil science. In: *Geostatistics Tróia92*, pages 805-816. Springer.

Gromenko, O. and Kokoszka, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Comput. Statist. Data Anal.*, **59**, 82-94.

Gromenko, O., Kokoszka, P., Zhu, L. and Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Ann. Appl. Statist.*, **6,** 669-696.

Holan, S.,Wang, S., Arab, A., Sadler, E. and Stone, K. (2008). Semiparametric geographically weighted response curves with application to site-specific agriculture. *J. Agril. Biol. Environ. Statist.,* **13**, 424-439.

Holan, S., Davis, G., Wildhaber, M., DeLonay, A. and Papoulias, D. (2009). Hierarchical Bayesian Markov

switching models with application to predicting spawning success of shovelnose sturgeon. *J. Roy. Statist. Soc., Series C (Applied Statistics)*, **58,** 47-64.

Holan, S., Wikle, C., Sullivan-Beckers, L. and Cocroft, R. (2010). Modeling complex phenotypes: Generalized linear models using spectrogram predictors of animal communication signals. *Biometrics,* **66**, 914-924.

Holan, S. and Wikle, C.K. (2012) Semiparametric dynamic design of monitoring networks for non-Gaussian spatio-temporal data. In: *Spatio-temporal Design: Advances in Efficient Data Acquisition,* (Eds.) Jorge Mateu and Werner Muller, 269-284, Wiley.

Holan, S., Yang, W., Matteson, D. and Wikle, C. (2012). An approach for identifying and predicting economic recessions in real-time using time-frequency functional models. In: *Applied Stochastic Models in Business and Industry*, **28**, 485-499.

Hooten, M., Ross, B. and Wikle, C. (2012). Optimal spatio-temporal monitoring designs for characterizing population trends. *J. Veg. Sci.*, **20,** 639-649.

Hooten, M.,Wikle, C., Sheriff, S. and Rushin, J. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. In: *Design and Analysis of Long-Term Ecological Monitoring Studies,* (Eds.) R.A. Gitzen, J.J. Millspaugh, A.B. Cooper, and D.S. Licht, 443-459, Cambridge University Press.

Igne, B., Reeves III, J. B., McCarty, G., Hively, W. D., Lund, E., and Hurburgh Jr., C. R. (2010). Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. *J. Near Infrared Spectroscopy*, **18,** 167-176.

Kitchen, N.R., Sudduth, K.A., Myers, D.B., Massey, R.E., Sadler, E.J., Lerch, R.N., Hummel, J.W. and Palm, H.L. (2005). Development of a conservation-oriented precision agriculture system: Crop production assessment and plan implementation. *J. Soil Water Conserv.,* **60,** 421-430.

Kokoszka, P. (2012). Dependent functional data. *ISRN Probab. Statist.,* 2012.

Le, N. and Zidek, J. (2006). *Statistical Analysis of Environmental Space-Time Processes.* Springer Verlag.

Lee, K.S., Lee, D.H., Sudduth, K.A., Chung, S.O., Kitchen, N. R. and Drummond, S.T. (2009). Wavelength identification and diffuse reflectance estimation for surface and profile soil properties. *Trans. ASABE*, **52**, 683-695.

Lee, K.S., Sudduth, K.A., Drummond, S.T., Lee, D.H., Kitchen, N.R. and Chung, S.O. (2010). Calibration methods for soil property estimation using reflectance spectroscopy. *Trans. ASABE*, **53**, 675-684.

Lerch, R.N., Kitchen, N.R., Kremer, R.J., Donald, W.W., Alberts, E.E., Sadler, E.J., Sudduth, K.A., Myers, D.B. and Ghidey, F. (2005). Development of a conservation-oriented precision agriculture system: Water and soil quality assessment. *J. Soil Water Conserv.,* **60**, 411-421.

Lesch, S.M. (2005). Sensor-directed spatial response surface sampling designs for characterizing spatial variation in soil properties. *Comput. Elect. Agric.*, **46,** 153–180.

Malley, D.F., Martin, P.D. and Ben-Dor, E. (2004). *Near Infrared Spectroscopy in Agriculture*. Chap 26: Application in Analysis of Soils. ASA, CSSA, and SSSA.

Mateu, J. and Müller, W.G. (2012). Collecting spatio-temporal data. In: *Spatio-Temporal Design*: *Advances in Efficient Data Acquisition*, 1-36.

Minasny, B., and McBratney, A.B. (2006). A conditioned latin hypercube method for sampling in the presence of ancillary information. *Comput. Geosci.*, **32,** 1378-1388.

Monestiez, P. and Nerini, D. (2008). A cokriging method for spatial functional data with applications in oceanology. In: *Functional and Operatorial Statistics*, 237-242. Springer.

Mouazen, A. M., de Baerdemaeker, J. and Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res.,* **80,** 171-183.

Müller, W.G. (2007). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Springer.

Nychka, D., Yang, Q. and Royle, J.A. (1997). Constructing spatial designs using regression subset selection. In: *Statistics for the Environment 3: Pollution Assessment and Controls*, eds. V. Barnett and K.F. Turkman, Wiley, New York, 131-154.

OHara, R. and Sillanpää, M. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, **4,** 85-118.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. Springer-Verlag.

Reich, B.J., Hodges, J.S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62,** 1197-1206.

Ruiz-Medina, M.D. (2011). Spatial autoregressive and moving average Hilbertian processes. *J. Multivar. Anal.*, **102**, 292-305.

Ruiz-Medina, M.D. (2012a). New challenges in spatial and spatiotemporal functional statistics for high-dimensional data. In: *Spatial Statistics*, **1,** 82-91.

Ruiz-Medina, M.D. (2012b). Spatial functional prediction from spatial autoregressive Hilbertian processes. *Environmetrics*, **23,** 119-128.

Ruiz-Medina, M.D. and Montes, R. (2011). Incorporating spatial interaction between large dimensional temperature series in atmosphere-ocean modelling of global climate change. *Procedia Environ. Sci.*, **7**, 2-7.

Ruiz-Medina, M.D. and Espejo, R. (2013). Integration of spatial functional interaction in the extrapolation of ocean surface temperature anomalies due to global warming. *Intt. J. Appl. Earth Observ. Geoinfor.*, **22,** 27-39.

Sankey, J.B., Brown, D.J., Bernard, M.L. and Lawrence, R. L. (2008). Comparing local vs global visible and near-Infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay. In: *Organic C and Inorganic C. Geoderma*, **148,** 149-158.

Shepherd, K.D. and Walsh, M.G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Sci. Soc. Amer. J.*, **66,** 988-998.

Shibusawa, S., Made Anom, S.W., Sato, H.P. and Sasao, A. (2001). Soil mapping using the real-time soil spectrometer. *Proceedings of the 3rd European Conference on Precision Agriculture,* **2**.

Shonk, J.L., Gaultney, L.D., Schulze, D.G. and Van Scoyoc, G.E. (1991). Spectroscopic sensing of soil organic matter content. *Trans. ASAE,* **34**, 1978-1984.

Stenberg, B., Rogstrand, G., Bolenius, E. and Arvidsson, J. (2007). On-line soil NIR spectroscopy: Identification and treatment of spectra influenced by variable probe distance and residue contamination. *Proceedings of the 6th European Conference on Precision Agriculture*.

Stenberg, B, Viscarra Rossel, R.A., Mouazen, A.M. and Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. In: *Advances in Agronomy*, **107**, 163-215.

Sudduth, K.A. and Hummel, J.W. (1991). Evaluation of reflectance methods for soil organic matter sensing. *Trans. ASAE*, **34,** 1900-1909.

Sudduth, K.A. and Hummel, J.W. (1993). Soil organic matter, CEC, and moisture sensing with a portable NIR spectrophotometer. *Trans. ASAE*, **36**, 1571-1582.

Sudduth, K.A. and Hummel, J.W. (1996). Geographic operating range evaluation of a NIR soil sensor. *Trans. ASAE*, **39**, 1599-1604.

Sudduth, K.A., Hummel, J.W. and Birrell, S.J. (1997). *The State of Site-Specific Management for Agriculture,* Chap. 10: Sensors for site-specific management. ASA, CSSA, and SSSA.

Sudduth, K.A., Kitchen, N.R., Wiebold, W.J., Batchelor, W. D., Bollero, G. A., Bullock, D.G., Clay, D.E., Palm, H. L., Pierce, F.J., Schuler, R.T. and Thelen, K.D. (2005). Relating apparent electrical conductivity to soil properties across the North-Central USA. *Comput. Elect. Agric.*, **46,** 263-283.

Sudduth, K.A., Kitchen, N.R., Sadler, E.J., Drummond, S.T. and Myers, D.B. (2010). *Proximal soil sensing*, Chap. 13: VNIR spectroscopy estimates of within-field variability in soil properties. Springer.

Thompson, S.K. and Seber, A.F. (1996). *Adaptive Sampling*. JohnWiley & Sons Inc., New York.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. and Skjemstad, J.O. (2006). Visible, near-infrared, mid-infrared, or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, **131,** 59-75.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J. and Lobsey, C. (2011). Proximal soil sensing: An effective approach for soil measurements in space and time. *Adv. Agronomy*, **113**, 237-282.

Wikle, C.K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, **84**, 1382-1394.

Wikle, C.K. (2010a). Hierarchical modeling with spatial data. In: *Handbook of Spatial Statistics*. (eds.) A.Gelfand, P. Diggle, M. Fuentes, P. Guttorp, Chapman and Hall. 89-106.

Wikle, C.K. (2010b). Low rank representations as models for spatial processes. In: *Handbook of Spatial Statistics*. (eds.) A.Gelfand, P. Diggle, M. Fuentes, P. Guttorp, Chapman and Hall. 107-118.

Wikle, C.K. and Berliner, L. (2005). Combining information across spatial scales. *Technometrics*, **47,** 80-91.

Wikle, C.K., Berliner, L. and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environ. Ecol. Statist.*, **5,** 117-154.

Wikle, C.K. and Holan, S. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *J. Time Series Anal.*, **32,** 339-350.

Wikle, C.K. and Hooten, M. (2010). A general science-based framework for spatiotemporal dynamical models. *Test*, **19**, 417-451.

Wikle, C.K. and Royle, J. (1999). Space: time dynamic design of environmental monitoring networks. *J. Agril. Biol. Environ. Statist.*, 489-507.

Wikle, C.K. and Royle, J. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics*, **16,** 507-522.

Yang, W.-H., Wikle, C.K., Holan, S.H., Myers, D.B., and Sudduth, K.A. (2014). Bayesian analysis of spatially-dependent functional responses with spatially-dependent multi-dimensional functional predictors. *Statitsica Sinica*, to appear.