A Multivariate Spatial Model for Soil Water Profiles

Stephan R. Sain,[1] Shrikant Jagtap,[2] Linda Mearns,[3] and Doug Nychka[4]

July 20, 2004

SUMMARY: Pedotransfer functions are classes of models used to estimate soil water holding characteristics based on commonly measured soil composition data as well as other soil characteristics. These models are important on their own but are particularly useful in modelling agricultural crop yields across a region where only soil composition is known. In this paper, an additive, multivariate spatial process model is introduced that offers the flexibility to capture the complex structure typical of the relationship between soil composition and water holding characteristics. A new form of pedotransfer function is developed that models the entire soil water profile. Further, the uncertainty in the soil water characteristics is quantified in a manner to simulate ensembles of soil water profiles. Using this capability, a small study is conducted with the CERES maize crop model to examine the sources of variation in the yields of maize. Here it is shown that the interannual variability of weather is a more significant source of variation in crop yield than the uncertainty in the pedotransfer function for specific soil textures.

KEY WORDS: Additive and mixed models, smoothing, pedotransfer functions, crop models.

[1]Geophysical Statistics Project, National Center of Atmospheric Research and Department of Mathematics, University of Colorado at Denver, P.O. Box 173364, Denver, CO 80217-3364, ssain@math.cudenver.edu. Corresponding author.

[2]Department of Agricultural and Biological Engineering, University of Florida

[3]Environmental and Societal Impacts Group, National Center for Atmospheric Research

[4]Geophysical Statistics Project, National Center for Atmospheric Research

1 Introduction

Soil scientists have long been interested in *pedotransfer functions* that are used to estimate soil water holding characteristics from commonly measured soil composition data. These water holding characteristics are necessary, along with other inputs such as weather, for use with crop models such as the Crop Environment Resources Synthesis (CERES) models that predict yields of maize, wheat, etc. Although crop yields are of interest in their own right, our interest is in using these models to assess the impacts of the interannual variability in climate as well as climate change. Crop yields are a useful integration of the weather during a growing season and provide a meaningful summary measure of the meteorology at a specific location. Considering the difference in the average yields predicted by the crop models under weather simulated from present and a possible future climate is a metric for climate change with agricultural and economic import. However, the uncertainty and variation in the predicted yields are also of vital concern to scientists, researchers, and policy makers. In particular, it is important to quantify how variation in the soil characteristics influence the crop models. Understanding the uncertainty in the soil characteristics as well as other inputs to the crop models will further our understanding of the uncertainty in the predicted crop yields and the impact of climate change.

1.1 Soil Characteristics

The water holding characteristics of a soil are generally characterized by measurements of the drained upper limit (DUL) and wilting point or lower limit (LL). It is these water holding characteristics that are inputs for the CERES crop model. The DUL is defined as the amount of water that a particular soil can hold after drainage is virtually complete. The LL is defined as the smallest amount of water that plants can extract from a particular soil. These values are often measured at different depths at a single location, yielding an entire profile of water holding characteristics.

The direct measurement of DUL and LL is difficult and is only available for a limited number of soils. In order to apply the crop model to a wide variety of soils, it is necessary

to infer DUL and LL from more commonly measured soil characteristics based on soil composition and it is these measurements of the percentages of clay, sand, and silt that are key components of most pedotransfer functions. Other variables are also at times incorporated into pedotransfer functions. These include, for example, bulk density (the weight of dry soil per unit volume of soil) and the amount of organic carbon in the soil.

1.2 Modeling Approaches for Pedotransfer Functions

A number of different approaches to the development of pedotransfer functions have appeared in the literature. Multiple regression, nonlinear regression (e.g. neural networks), and nonparametric methods such as nearest-neighbor regression have been used along with methods based on established physical relationships and differential equations. Several reviews have appeared in the literature discussing the various forms of pedotransfer functions including Rawls et al. (1991), Timlin et al. (1996), Minasny et al. (1999), Pachepsky et al. (1999), and Gijsman et al. (2002). However, there are still issues in applying modern statistical models to this problem and, just as important, in characterizing the uncertainty of the estimated pedotransfer function.

1.3 Impacts for Soil and Statistical Science

We propose a flexible procedure based on an additive, multivariate spatial process model that simultaneously models the entire soil water profile of LL and DUL yielding a new and innovative form of pedotransfer function. While accounting for the spatial dependence in the LL and DUL as a function of depth, this model also allows for smooth contributions of important covariates related to the soil composition as well as the inclusion of additional covariates.

These types of models are complex and direct estimation of parameter values through, for example, maximum likelihood or restricted maximum likelihood (REML) can be problematic. Beyond the specific contributions of such models in the development of new forms of pedotransfer functions, we also introduce an iterative approach for fitting such mod-

els. Inspired by traditional approaches to geostatistics and algorithms like backfitting and expectation-maximization (EM), it is demonstrated in this situation that our approach produces parameter values consistent with those obtained via REML but with substantial computational savings.

Although these models can be complicated, the additive structure of the model allows the uncertainty in the predicted LL and DUL to be easily characterized. Further, the model gives a framework for generating random realizations of soil water profiles for particular collections of soil composition characteristics. These realizations can be interpreted as random samples from the posterior distribution of the soil water profiles given the database of observed soil water profiles. Of course, the variation in these random samples reflects the uncertainty in the soil water profiles and the subsequent use as inputs to a crop model will produce variation in crop yields.

1.4 Outline

The following section gives details about the data used in this work and Section 3 outlines the construction of the model as well as estimation. Section 4 presents the results of the model fitting and Section 5 presents a discussion of the prediction error in the model and the generation of soil water profiles. Finally, Section 6 discusses an application utilizing the CERES maize crop model to examine the behavior of the predicted crop yields for different soil types.

2 Soil Data

The soil database used in this work consists of 272 individual measurements on 63 soil samples selected by Gijsman et al. (2002) from a national field study of Ratliff et al. (1983) and Ritchie et al. (1987). See also Jagtap et al. (2003). To our knowledge, this is one of the most extensive soil databases for use in the development and validation of pedotransfer functions.

Individual measurements from the original study that were beyond the reach of the

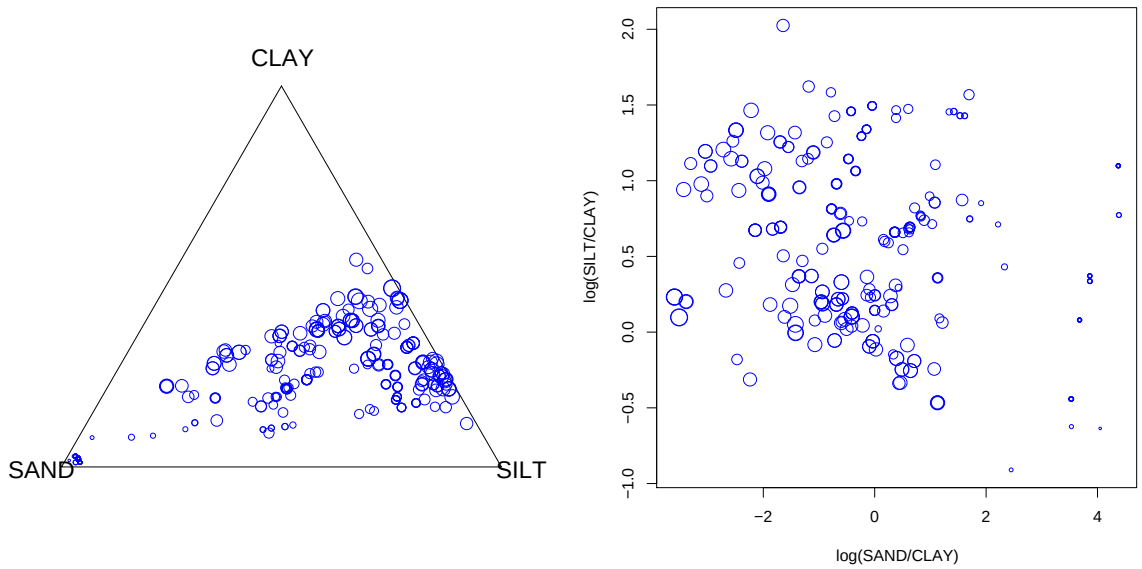

Figure 1: Scatterplot of soil compositional data in the standard soil-texture triangle (left frame) as well as a scatterplot of the transformed compositional data (right frame). The size of the plot character is related to the measured value of the LL.

roots were excluded as well as samples in the top layers of soil that dry out much more than the rest of the soil. The data set includes information on depth, soil composition and texture (percentages of clay, sand, and silt), bulk density, and organic matter, as well as field measured values of LL and DUL.

2.1 Soil Composition

The soils in the study represent a broad range of soil textures of interest in agriculture. A scatterplot of the soil composition data is given in Figure 1 transformed to the standard soil-texture triangle. The plot character size in the scatterplot is related to the measured value of the LL. A point in the center of the triangle represents equal amounts of clay, sand, and silt. A point nearer one of the three labelled corners in the triangle represents a soil that is dominated by that component.

The percentages of clay, sand, and silt for each soil are constrained to sum to 100 (or, equivalently, the proportions sum to one). Thus, knowledge of any two of the composition

percentages completely defines the third. So there are not three distinct variables in the soil composition, rather the measurements inhabit a lower-dimensional structure. Aitchison (1987) suggests transforming such data using the additive log-ratio transform. Consider constructing two new variables defined as

$$X_1 = \log\left(\frac{\text{Silt}}{\text{Clay}}\right) \qquad X_2 = \log\left(\frac{\text{Sand}}{\text{Clay}}\right).$$

A scatterplot of X_1 and X_2 is also shown in Figure 1. The choice of which variable to use in the denominator is somewhat arbitrary, and, for these data, the clay component is used since that choice yields the smallest correlation among the transformed variables X_1 and X_2.

2.2 Soil Water Profiles

The data exhibit a nested structure in that each of the 63 different soil samples have measurements of soil composition, LL, DUL, etc. taken at different depths. For a particular soil, this collection of LL and DUL measurements at different depths make up the soil water profile. Examples of the LL and DUL profiles for four particular soil texture types are displayed in Figure 2. The soil texture classification is based on the relative percentages of clay, sand, and silt. The data set includes soils with the following soil textures: silty loam (SIL), silty clay loam (SICL), loam (L), clay loam (CL), sand (S), silty clay (SIC), sandy loam (SL), sandy clay loam (SCL), clay (C) and loamy sand (LS).

As is shown by Figure 2, the LL and DUL measurements are constrained in the sense that $0 < LL < DUL < 1$. Since DUL is typically much less than 1, this suggests the transformation

$$Y_1 = \log(LL) \qquad Y_2 = \log(\Delta)$$

where $\Delta = DUL - LL$. Given the measurements, the statistical problem now is to accurately reproduce the profile structure in the Y_1 and Y_2 as a function of depth and across the range of possible soil compositions and texture classes.

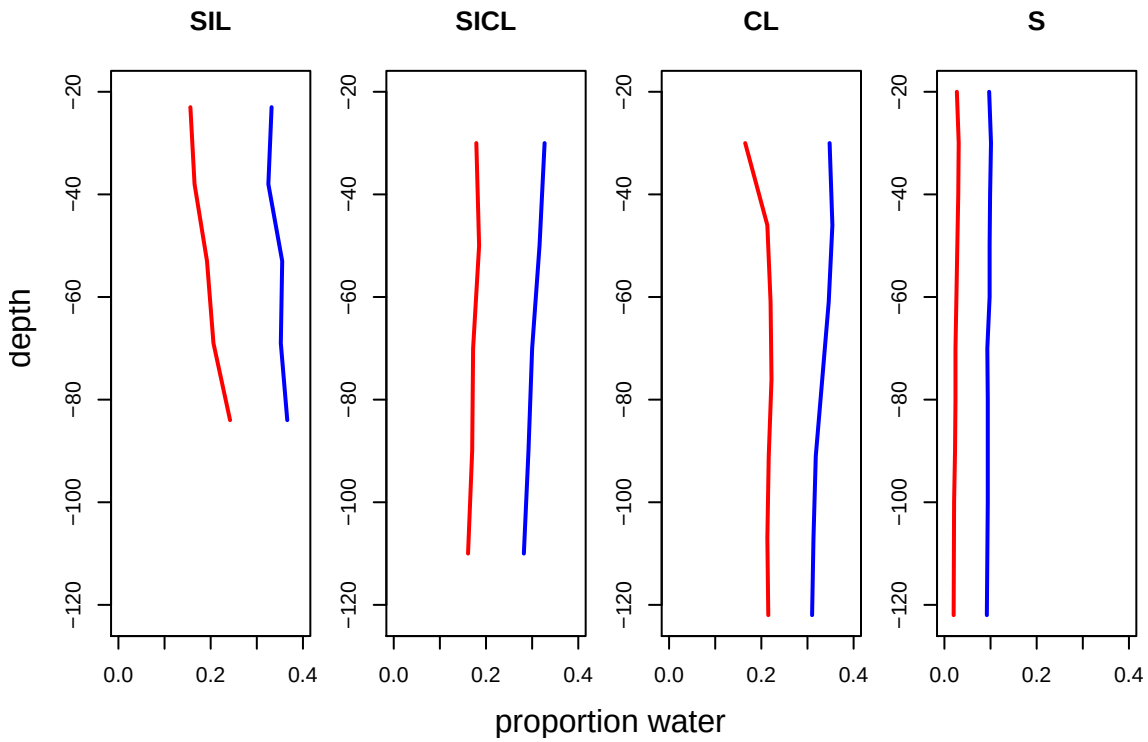

Figure 2: Examples of soil water holding profiles for four soil textures: silty loam (SIL), silty clay loam (SICL), clay loam (CL), and sand (S). Dashed (red) lines indicate the lower limit (LL) while dotted (blue) lines represent the drained upper limit (DUL).

3 A Flexible Multivariate Model

It is assumed that the data has the form $(\mathbf{Y}_i, \mathbf{X}_i, \{\mathbf{Z}_{ki}\}_{k=1}^{p}, S_i, D_i)$ for $i = 1, \ldots, n$. The S_i indicate the soil sample from which the measurements are taken, while the D_i record the particular depth. The p-vector $\mathbf{Y}_i$ is composed of the $\log LL$ and $\log \Delta$ measurements for soil S_i and at depth D_i. The q-vector $\mathbf{X}_i$ is composed of the bivariate (transformed) soil composition data, also for soil S_i and at depth D_i. While it is expected that that the soil composition data, $\mathbf{X}_i$, are used as covariates for all of the measurements in $\mathbf{Y}_i$, the $\{\mathbf{Z}_{ki}\}$ are q_k-vectors of additional covariates specific to the kth measurement in $\mathbf{Y}_i$.

3.1 Multivariate Regression Models

Construct the $n \times p$ matrix $\mathbf{Y} = [\mathbf{Y}_1 \dots \mathbf{Y}_n]'$ and the $n \times q$ matrix $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_n]'$, and, for the moment, ignore the nested structure of the data, depth, and the additional covariates $\{\mathbf{Z}_{ki}\}$. The traditional approach to multivariate response regression suggests a model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\beta}$ is a $q \times p$ matrix of regression coefficients and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1 \dots \boldsymbol{\epsilon}_n]'$ is a $n \times p$ matrix of errors. The standard assumptions on the error terms posit that, for each $i = 1, \dots, n$, $E[\boldsymbol{\epsilon}_i] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\epsilon}_i] = \Sigma$. Also, the rows of $\boldsymbol{\epsilon}$ are assumed to be independent. Then the least-squares estimates of $\boldsymbol{\beta}$ are equivalent to performing p individual regressions on the columns of $\mathbf{Y}$.

A second formulation of the multivariate response regression model in (1) can be constructed by stacking the columns of $\mathbf{Y}$, yielding

$$\text{vec}\mathbf{Y} = (\mathbf{I}_p \otimes \mathbf{X})\text{vec}\boldsymbol{\beta} + \text{vec}\boldsymbol{\epsilon}, \tag{2}$$

where vec$\mathbf{Y}$ is now a $np \times 1$ vector given by

$$\text{vec}\mathbf{Y} = [Y_{11} \cdots Y_{1n} Y_{21} \cdots Y_{2n}]',$$

and Y_{1i} and Y_{2i} represent the measurements of $\log LL$ and $\log \Delta$, respectively. Also, $\otimes$ is the Kronecker product, vec$\boldsymbol{\beta}$ is a $qp \times 1$ vector of regression coefficients, and vec$\boldsymbol{\epsilon}$ is a $np \times 1$ vector of errors. This formulation changes the behavior of the error structure somewhat. For example, although $E[\text{vec}\boldsymbol{\epsilon}] = \mathbf{0}$, $\text{Var}[\text{vec}\boldsymbol{\epsilon}] = \Sigma \otimes \mathbf{I}_n$. In the following, we will drop the vec notation and simply consider, for example, $\mathbf{Y}$ as a $np \times 1$ vector while retaining the knowledge that the first n elements of $\mathbf{Y}$ refer to the first response variable, that the second set of n elements refer to the second response variable, etc.

3.2 An Extended Multivariate Model

Preliminary analysis as well as a close examination of the data in Figure 2 suggest that a model such as those in (1) and (2) is inappropriate. These models fail because the response

is not linear and the covariance structure does not follow the simple Kronecker form. We seek to generalize these simple models in several ways. First, through a type of additive model, flexibility will be incorporated into the regression function, in particular, to model the relationship between the transformed soil composition data and the transformed soil water variables. Secondly, additional covariates, including those specific to just LL and Δ will be included. Finally, there is a certain smoothness in the profiles in Figure 2 suggesting that the measurements are correlated with depth. Hence, the model should be able to account for dependence not only within measurements for each observation (e.g. between LL and Δ at the same depth for a particular soil) but also possible spatial dependence between the observations (e.g. across LL and Δ measurements at different depths for a particular soil).

Consider expanding the multivariate response regression model as

$$\mathbf{Y}_i = P(\mathbf{X}_i, \{\mathbf{Z}_{ki}\}_{k=1}^{p}) + h(\mathbf{X}_i) + \boldsymbol{\epsilon}_{S_i}(D_i), \tag{3}$$

where P is a fixed polynomial function, h is a random function of the transformed soil composition, and $\boldsymbol{\epsilon}_{S_i}$ is a random error process accounting for both the variation between measurements for a particular soil at some fixed depth and the spatial dependence across measurements for a particular soil at different depths. The inclusion of h in the model is equivalent to a form of nonparametric regression.

Assuming that the errors have a normal distribution, the model in (3) implies that $\mathbf{Y}$ is multivariate normal with

$$E[\mathbf{Y}] = \mathbf{T}\boldsymbol{\beta} \qquad \text{Var}[\mathbf{Y}] = \Sigma_h + \Sigma_{\boldsymbol{\epsilon}}. \tag{4}$$

The regression matrix $\mathbf{T}$ is a block diagonal matrix with p blocks. The kth diagonal block has the form $[\mathbf{1} \quad \mathbf{X} \quad \mathbf{Z}_k]$. Here, $\mathbf{1}$ is a n-vector of 1's (intercept), $\mathbf{X}$ is the $n \times q$ matrix of variables common to all response variables and $\mathbf{Z}_k$ is the $n \times q_k$ matrix of explanatory variables specific to the kth response variable. This structure implies that $\mathbf{T}$ and $\boldsymbol{\beta}$ have $p + nq + \sum_k q_k$ columns and elements, respectively.

3.3 Covariance Structures

The covariance matrix Σ_h in (4) is associated with random process h in (3) while $\Sigma_{\boldsymbol{\epsilon}}$ is the covariance associated with the error process, $\boldsymbol{\epsilon}_S$. These covariance matrices have the form

$$\Sigma_h = \text{diag}(\rho_1, \ldots, \rho_p) \otimes \mathbf{K} \qquad \Sigma_{\boldsymbol{\epsilon}} = \mathbf{W} \otimes \mathbf{B}, \tag{5}$$

where diag indicates a diagonal matrix. We assume that h is a stationary process and so the matrix $\mathbf{K}$ contains elements of the form $K_{ij} = \text{Cov}[h(\mathbf{X}_i), h(\mathbf{X}_j)] = C(||\mathbf{X}_i - \mathbf{X}_j||)$ for $i, j = 1, \ldots, n$. A flexible class of covariance functions is the Matern family (Stein, 1999) given by

$$C(d) = \sigma^2 \frac{2(\theta d/2)^\nu K_\nu(\theta d)}{\Gamma(\nu)} \tag{6}$$

where $d = ||\mathbf{X}_i - \mathbf{X}_j||$, σ^2 is a scale parameter, θ represents the range, ν controls the smoothness, and K_ν is a modified Bessel function of order ν. We use this family for modelling the regression function relating soil parameters to composition. In this application, the covariance function for $\mathbf{K}$ is fixed with $\sigma^2 = 1$ (the parameters ρ_i adjust for different variances), $\nu = 1$, and θ is taken to be on the order of the range of the data.

The matrices $\mathbf{W}$ and $\mathbf{B}$ represent the within observation covariance (fixed depth) and the across observation covariance (across depths), respectively. The $p \times p$ matrix $\mathbf{W}$ has elements $w_{ij} = \text{Cov}[\epsilon_{ki}, \epsilon_{kj}]$ for all $k = 1, \ldots, n$ and $i, j = 1, \ldots, p$. The matrix $\mathbf{B}$ is a $n \times n$ matrix whose structure is more complex as it reflects the nested structure in the data, i.e. the effect that measurements within a particular soil water profile are related across depth. The elements of $\mathbf{B}$ are given by

$$b_{ij} = \begin{cases} \text{Cov}[\epsilon_{ik}, \epsilon_{jk}] = C(||D_i - D_j||) & S_i = S_j \\ 0 & \text{otherwise,} \end{cases}$$

for all $k = 1, \ldots, p$ and $i, j = 1, \ldots, n$ where again we assume that the process modelling dependence on depth is stationary. The covariance function used here is the exponential covariance given by

$$C(d) = \exp(-(d/\theta)^2), \tag{7}$$

where $d = ||D_i - D_j||$. Note that that the exponential covariance function is in the Matern family where $\nu = 0.5$. Typically, the observations will be grouped by soil sample, indicated by the S_i. Hence, $\mathbf{B}$ will have a block diagonal structure and all of the off-diagonal elements will be zero reflecting the independence of measurements from different soil samples.

Our choices for fixed and free parameters in the covariance functions is an attempt to balance model flexibility with the amount of information available in the data. The covariance function for h has only the scale parameters free to be estimated from the data, while the covariance function for $\boldsymbol{\epsilon}_S$ has both the scale parameters as well as the range parameters free. This difference in the parameterization of the two covariance matrices is due to important distinctions between the roles of the two random terms in the model and the structure of the data used for estimation. As mentioned earlier, h is included in the model as a form of nonparametric regression. Just as the bandwidth is more important than the functional form of the kernel for optimal kernel smoothing (Scott, 1992, and Wand and Jones, 1994), the scale parameter (or the ratio of the scale parameter to the range in the case of an exponential covariance function) plays a similar role for optimal prediction with kriging (Stein, 1999). With a single realization of the soil compositional data, it is difficult to estimate both scale and range parameters for h. On the other hand, the nested structure of the soil data is essentially comparable to multiple realizations which should greatly improve our ability to estimate both sets of parameters for the covariance of $\boldsymbol{\epsilon}_S$. In this case, different soil profiles are combined to yield information about the correlation of water holding capacity with depth.

3.4 Smoothing

Smoothing the data to estimate the surface represented by h is accomplished by finding the appropriate balance between the respective components of the overall covariance matrix. However, it is also desirable to allow different amounts of smoothing across the different variables. Formally, the full covariance matrix can be written as

$$\Sigma_h + \Sigma_{\boldsymbol{\epsilon}} \quad = \quad \text{diag}(\rho_1, \ldots, \rho_p) \otimes \mathbf{K} + \mathbf{W} \otimes \mathbf{B}$$

$$
\begin{aligned}
&= w_{11}\left[\text{diag}(\eta_1, \ldots, \eta_p) \otimes \mathbf{K} + \mathbf{V} \otimes \mathbf{B}\right] \\
&= w_{11}\boldsymbol{\Omega}, \qquad (8)
\end{aligned}
$$

where $\eta_i = \rho_i / w_{11}$ and the elements of $\mathbf{V}$ are of the form $v_{ij} = w_{ij}/w_{11}$. Assuming that $\mathbf{V}$, $\mathbf{B}$, and $\mathbf{K}$ are known, then smoothing can be accomplished by choosing the appropriate values of $\eta_1, \ldots, \eta_p$. Larger values of η_i leads to rougher estimates.

3.5 Estimation I: REML

Given covariances for h and $\boldsymbol{\epsilon}_S$ in (3), a standard estimator is the best linear unbiased method (commonly known as universal kriging) which has the form

$$
\hat{\mathbf{Y}} = \mathbf{T}\hat{\boldsymbol{\beta}} + (\text{diag}(\eta_1, \ldots, \eta_p) \otimes \mathbf{K})\hat{\boldsymbol{\delta}}. \qquad (9)
$$

More traditional geostatistical methods for parameter estimation and prediction often use the data twice. First, spatial covariance parameters are estimated by, for example, fitting covariance models to sample covariance functions. Then, the data are used for prediction at locations where data are not observed. See, for example, Cressie (1991). However, it is difficult to fit covariance models in this fashion with fixed-effects terms present.

An alternative is to estimate model parameters using REML (Kitanidis, 1997; Stein, 1999; see also Wahba, 1990a, and Nychka, 2000). To demonstrate, take the QR decomposition of $\mathbf{T}$, i.e. write $\mathbf{T}$ as

$$
\mathbf{T} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2] \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix},
$$

where the matrix $\mathbf{R}$ is upper-triangular and $\mathbf{Q} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2]$ is an orthogonal matrix with $\mathbf{Q}_1$ having columns that span the column space of $\mathbf{T}$ and $\mathbf{Q}_2$ having columns that span the space orthogonal to $\mathbf{T}$. Then $\mathbf{Q}_2'\mathbf{Y}$ has zero mean and covariance matrix given by $\mathbf{Q}_2'(\Sigma_h + \Sigma_{\boldsymbol{\epsilon}})\mathbf{Q}_2$. Hence, the likelihood corresponding to the transformed $\mathbf{Q}_2'\mathbf{Y}$ only involves covariance parameters, which in this case includes $\{\rho_i\}$, $\mathbf{W}$, and θ from (7). The likelihood is complex and closed form solutions for the parameter estimates are intractable.

However, the likelihood can be maximized via numerical methods. Then, given the covariance parameters, estimates of $\boldsymbol{\beta}$ and predictions for $\boldsymbol{\delta}$ are obtained directly as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{T})^{-1}\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y} \\ \hat{\boldsymbol{\delta}} &= \hat{\boldsymbol{\Omega}}^{-1}(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\beta}}),\end{aligned}$$

where $\hat{\boldsymbol{\Omega}}$ is an estimate of Ω in (8). Note that the form $\hat{\boldsymbol{\beta}}$ is equivalent to the generalized least-squares estimator and that $\hat{\boldsymbol{\delta}}$ is equivalent to the best linear unbiased predictor (BLUP).

3.6 Estimation II: An Iterative Approach

In practice, the size and complexity of data sets may lead to computational difficulties when directly maximizing the likelihood. To counter such problems, we propose an iterative approach to parameter estimation that is inspired by traditional approaches to geostatistical data analysis as well as such procedures as backfitting (Breiman and Friedman, 1985; Buja et al., 1989) and the expectation-maximization algorithm (Dempster et al., 1977). The procedure cycles between a REML step to update the smoothing parameters, an estimation/prediction step to update $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$, and a final step examining residuals to update estimates of $\mathbf{W}$ and $\mathbf{B}$. We conjecture that, at convergence, the algorithm produces approximate likelihood estimators that are comparable to those obtained from the REML.

The algorithm is as follows:

0. Initialize: Compute $\mathbf{K}$ and set $\mathbf{W} = \mathbf{I}_p$ and $\mathbf{B} = \mathbf{I}_n$.

1. Using a modified form of the REML discussed Section 3.5, estimate $\eta_1, \ldots, \eta_p$. Because the matrices $\mathbf{K}$, $\mathbf{W}$, and $\mathbf{B}$ are fixed at this stage, profiling the log-likelihood with respect to w_{11} from (8) leaves the log-likelihood as a function of only the $\eta_1, \ldots, \eta_p$. Estimates of the η_i are then found through a simple grid search.

2. Estimate $\boldsymbol{\beta}$ and predict $\boldsymbol{\delta}$ via

$$\hat{\boldsymbol{\beta}} = (\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{T})^{-1}\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\delta}} \;=\; \hat{\boldsymbol{\Omega}}^{-1}(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\beta}}).$$

3. Compute residuals as $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ where $\hat{\mathbf{Y}}$ is given in (9).

   a. Update $\mathbf{W}$ ($\mathbf{B}$ fixed).

   b. Update $\mathbf{B}$ ($\mathbf{W}$ fixed).

4. Repeat steps 1-3 until parameter estimates converge.

Some remarks on the third step in the procedure are in order. Residuals are computed by subtracting estimates of the fixed effects and predictions of the random effects. To justify the use of such residuals to estimate the covariance matrix of the error process ($\mathbf{W}$ and $\mathbf{B}$), let $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon}$ denote a general mixed effects model. Then, assuming $\mathbf{h}$ and $\boldsymbol{\epsilon}$ are independent Gaussian random variables, the conditional distribution of $\mathbf{Y} - \boldsymbol{\mu} - \mathbf{h}$ given $\mathbf{h}$ is a zero mean Gaussian with the covariance matrix of $\boldsymbol{\epsilon}$. See, for example, McCulloch and Searle (2000).

Considering the updating of $\mathbf{W}$ and $\mathbf{B}$, note that the log-likelihood associated with the residuals $\mathbf{R}$ is proportional to

$$-\frac{n}{2}|\mathbf{W}| - \frac{p}{2}|\mathbf{B}| - \mathbf{R}'(\mathbf{W}^{-1} \otimes \mathbf{B}^{-1})\mathbf{R} \tag{10}$$

where the determinant and inverse of $\Sigma_{\boldsymbol{\epsilon}}$ follow directly from the properties of Kronecker products and $\mathbf{R} = [\mathbf{r}_1' \cdots \mathbf{r}_p']'$ with $\mathbf{r}_i$ denoting the n-vector of residuals for the ith variable. Note that the quadratic form in (10) can be rewritten as

$$\mathrm{tr}(\mathbf{W}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} b^{ij} \mathbf{u}_j \mathbf{u}_i')$$

where tr indicates the trace of a matrix, b^{ij} is the ijth element of $\mathbf{B}^{-1}$ and $\mathbf{u}_i$ indicates the p-vector of residuals for the ith observation constructed by unstacking the residual vector, $\mathbf{R}$, into a $n \times p$ matrix. Thus, an update of $\mathbf{W}$ can be obtained via

$$\begin{aligned} \widehat{\mathbf{W}} \;&=\; \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} b^{ij} \mathbf{u}_j \mathbf{u}_i' \\ &=\; \frac{1}{n} \mathbf{U}' \mathbf{B}^{-1} \mathbf{U} \end{aligned}$$

	η_1	η_2	W_{11}	W_{22}	W_{12}	θ
REML	5.84	1.66	0.0765	0.0483	-0.0222	134.6
Iterative	5.74	2.21	0.0697	0.0445	-0.0217	144.2

Table 1: A comparison of parameter estimates from REML (Section 3.5) and the iterative approach (Section 3.6).

where $\mathbf{U}$ indicates the $n \times p$ matrix of (unstacked) residuals. See, for example, Theorem 4.2.1 from Mardia et al. (1979).

Unfortunately, there is no easy update for $\mathbf{B}$. However, $\mathbf{B}$ is simply a function of the range θ from (7), and, given $\mathbf{W}$, (10) can be computed easily and maximized via a simple grid search.

4 Results

The methods presented in Section 3 were used to fit the model in (3) and (9). The matrix $\mathbf{X}$ contains the transformed composition data, which are common to both $\log LL$ and $\log \Delta$. A measure of organic matter in the soil was included as an additional covariate for $\log LL$ ($Z_{i1},\ i = 1, \ldots, n$). Also, the profiles in Figure 2 suggest that Δ is smaller for deeper soils. Hence, both linear and quadratic terms were included as additional covariates for $\log \Delta$ by setting $\mathbf{Z}_{i2} = [(D_i - 70) \quad (D_i - 70)^2]'$ for $i = 1, \ldots, n$ where the midpoint of the values of depth in the data is 70. Approximately ten iterations of the algorithm in Section 3.6 were required for fitting the model and convergence was monitored by examining parameter estimates and predicted values.

Table 4 shows a comparison of the parameter estimates of the covariance matrices obtained from directly maximizing the likelihood using REML and the iterative algorithm from Section 3.6. The parameter estimates are quite close and yield virtually identical predictions of $\log LL$ and $\log \Delta$.

The partial regression functions for the transformed composition data and $\log LL$ and $\log \Delta$ are given Figures 3 and 4. The partial regression functions after back transforming the composition data and displaying the fits on the standard soil-texture triangle are shown

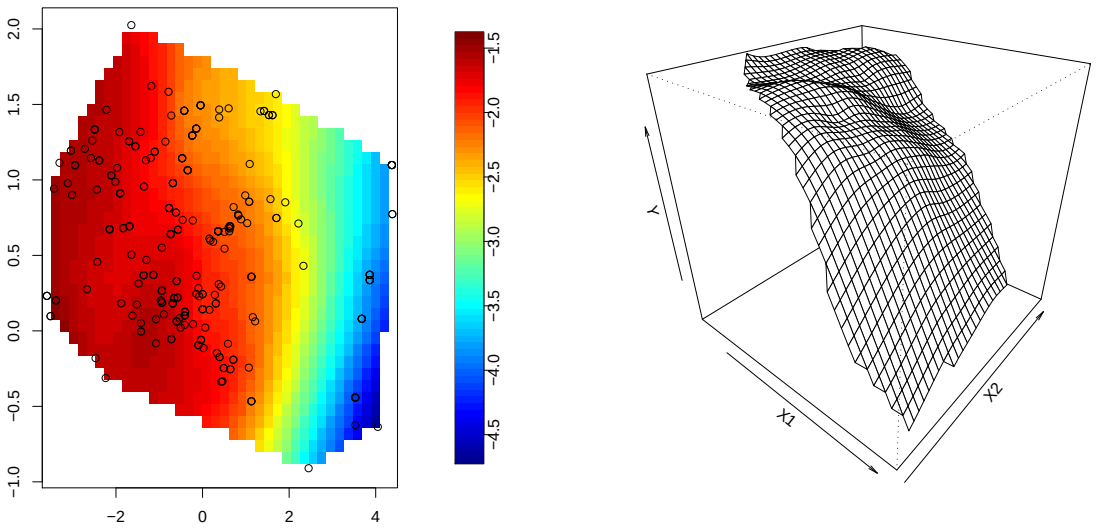

Figure 3: The partial regression function for $\log LL$ as a function of the transformed composition data.

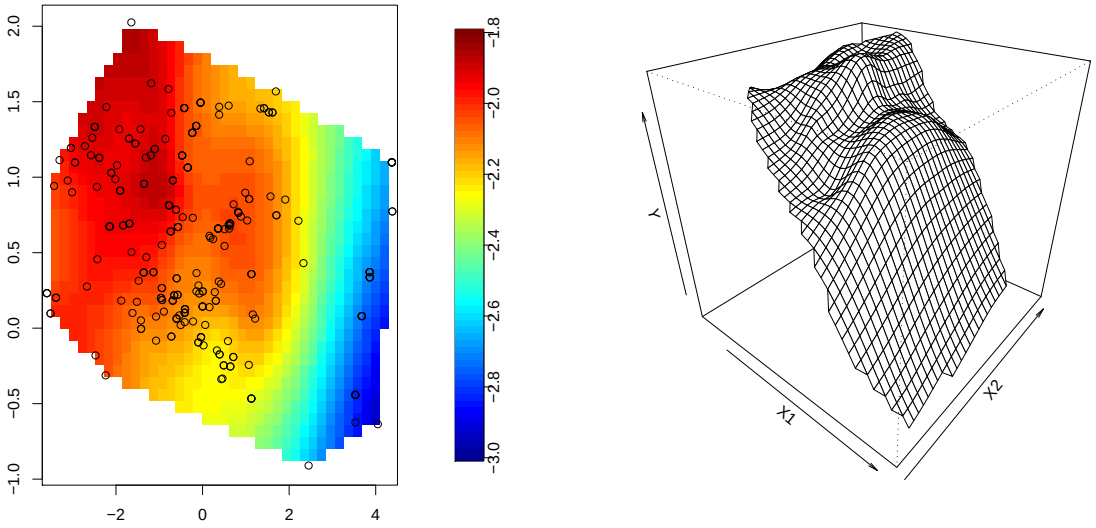

Figure 4: The partial regression function for $\log \Delta$ as a function of the transformed composition data.

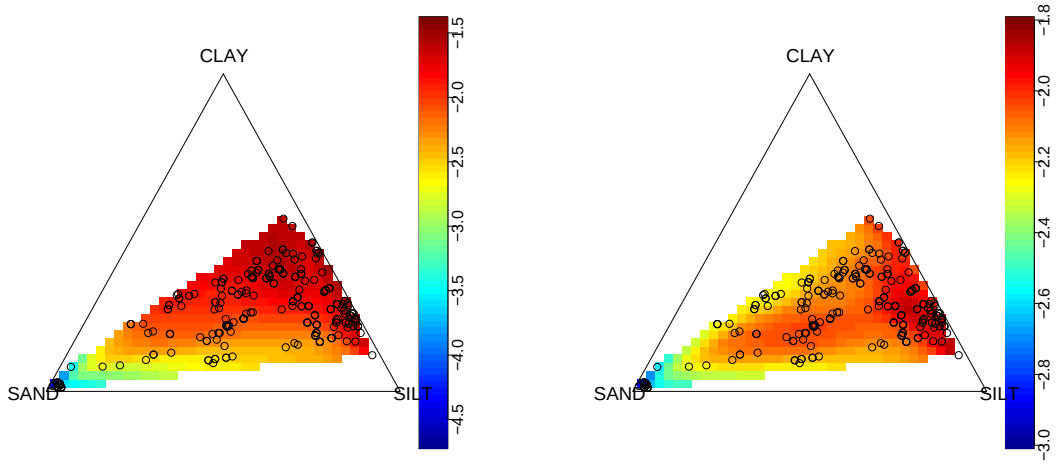

Figure 5: Partial regression functions for $\log LL$ (left frame) and $\log \Delta$ (right frame) as a function of the transformed composition data after transforming back to the soil composition triangle.

in Figure 5. In general, the fits show considerable structure with smaller $\log LL$ and $\log \Delta$ for larger values of $X_1 = \log(\text{Silt}/\text{Clay})$, increasing $\log LL$ and $\log \Delta$ as X_1 decreases, and a levelling out of $\log LL$ and $\log \Delta$ for larger $X_2 = \log(\text{Sand}/\text{Clay})$ values.

This phenomenon translates into sandy soils having the lowest $\log LL$ and soils with a stronger clay component having the highest $\log LL$. Sandy soils also had lower $\log \Delta$ values, suggesting that the LL and DUL are closer together. Interesting, the final regression fits for $\log \Delta$ seem to peak and plateau near more loamy soils. These results indicate that the random component h plays a significant role in determining the composition beyond just a simple polynomial relationship.

The contribution of organic matter to the $\log LL$ is shown in the left frame of Figure 6 along with two standard error bands (pointwise). There does not seem to be strong evidence of a significant contribution of organic carbon to $\log LL$. The contribution of depth to the $\log \Delta$ is shown in the right frame of Figure 6, also with two standard error bands (pointwise). Depth, on the other hand, does seem to have a significant impact on

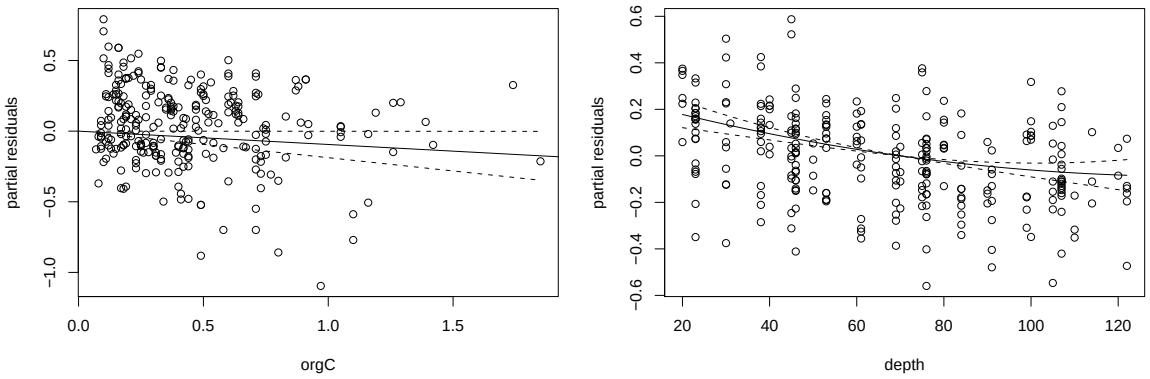

Figure 6: The partial regression function for $\log LL$ as a function of the organic matter in the soil is shown in the left frame while the right frame shows the partial regression function for $\log \Delta$ as a function of depth. Dotted lines represent two standard error bands (pointwise).

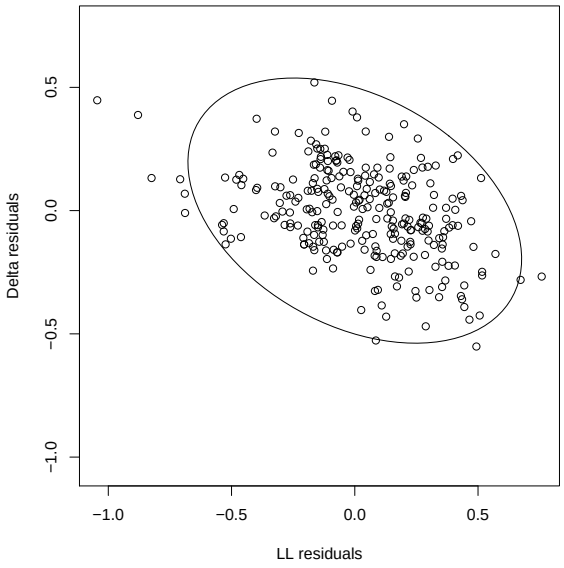

Figure 7: Final estimate of covariance matrix $\mathbf{W}$.

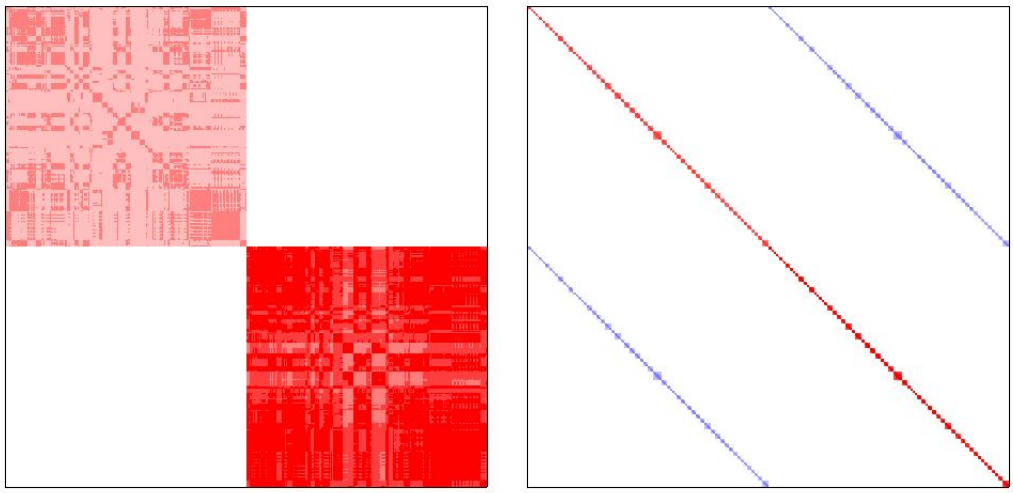

Figure 8: Final fitted covariance matrices Σ_h (left frame) and $\Sigma_{\boldsymbol{\epsilon}}$ (right frame)

.

$\log \Delta$ with LL and DUL getting closer together at greater depths.

The contour in Figure 7 represents the covariance matrix $\mathbf{W}$ and suggests a slight negative correlation between $\log LL$ and $\log \Delta$ at a fixed depth. See also Table 4.

The final fitted covariance matrices are shown in Figure 8. For Σ_h, the matrix is block diagonal with the two blocks representing the spatial covariance between the transformed composition parameters. The blocks are identical except for the differences in scale resulting from the different amounts of smoothing for the two variables. For $\Sigma_{\boldsymbol{\epsilon}}$, the block structure of the matrix due to the nested structure of the data is clear. The diagonal blocks represent the spatial correlation with depth for each soil and the off-diagonal blocks represent the correlation between $\log LL$ and $\log \Delta$ for a fixed depth. Again, the differences in scaling due to the different amounts of smoothing are clear.

5 Prediction Error and Conditional Simulation

Ultimately, the interest is in quantifying the uncertainty in the soil water holding characteristics and generating simulated water profiles for soils with specific compositions and depth profiles. Noting that the estimator in (9) is linear in $\mathbf{Y}$, predicted values as well as

the prediction error are easily found. Let $(\mathbf{X}_{0i}, \{\mathbf{Z}_{0ki}\}_{k=1}^{p}, S_{0i}, D_{0i})$ for $i = 1, \ldots, n_0$ denote the data structure for a new soil where predictions of the water holding characteristics are desired. Further, let $\mathbf{X}_0$ denote the $n_0 \times q$ matrix of explanatory variables common to all the response variables and $\{\mathbf{Z}_{0k}\}$ the $n_0 \times q_k$ matrix of explanatory variables specific to the kth response variable. Predicted values are obtained via

$$\begin{aligned} \hat{\mathbf{Y}}_0 &= \mathbf{T}_0\hat{\boldsymbol{\beta}} + \mathbf{K}_0\hat{\boldsymbol{\delta}} \\ &= \mathbf{A}_0\mathbf{Y} \end{aligned} \tag{11}$$

where the matrix $\mathbf{T}_0$ is given by

$$\mathbf{T}_0 = \begin{bmatrix} \mathbf{1} & \mathbf{X}_0 & \mathbf{Z}_{01} & \mathbf{0} & & \\ \mathbf{0} & & & \mathbf{1} & \mathbf{X}_0 & \mathbf{Z}_{02} \end{bmatrix},$$

$\mathbf{K}_0$ is a $n_0 \times n$ matrix with elements $K_{ij} = C(||\mathbf{X}_{0i} - \mathbf{X}_j||)$ and C is the covariance function defined in (6). Letting

$$\mathbf{M} = (\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{T})^{-1}\mathbf{T}'\hat{\boldsymbol{\Omega}}^{-1},$$

then

$$\mathbf{A}_0 = \mathbf{T}_0\mathbf{M} + \mathbf{K}_0\hat{\boldsymbol{\Omega}}^{-1}(\mathbf{I} - \mathbf{M}).$$

The prediction error is obtained via

$$\begin{aligned} \mathrm{Var}(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0) &= \mathrm{Var}(\mathbf{Y}_0 - \mathbf{A}_0\mathbf{Y}) \\ &= \mathrm{Var}(\mathbf{Y}_0) + \mathbf{A}_0\mathrm{Var}(\mathbf{Y})\mathbf{A}_0' - 2\mathbf{A}_0\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}_0). \end{aligned} \tag{12}$$

The variance of $\mathbf{Y}$ is easily estimated by plugging in the parameter estimates for Σ_h and $\Sigma_{\boldsymbol{\epsilon}}$. Similarly, an estimate for the variance of $\mathbf{Y}_0$ can also be easily constructed. The covariance between $\mathbf{Y}$ and $\mathbf{Y}_0$ is implied by the random function h in (3) and is obtained via the covariance function based on the transformed compositional data.

Figure 9 shows ten simulated LL and DUL profiles for two soils of interest, silty loam (SIL) and sand (S) based on the estimates obtained from the multivariate spatial model.

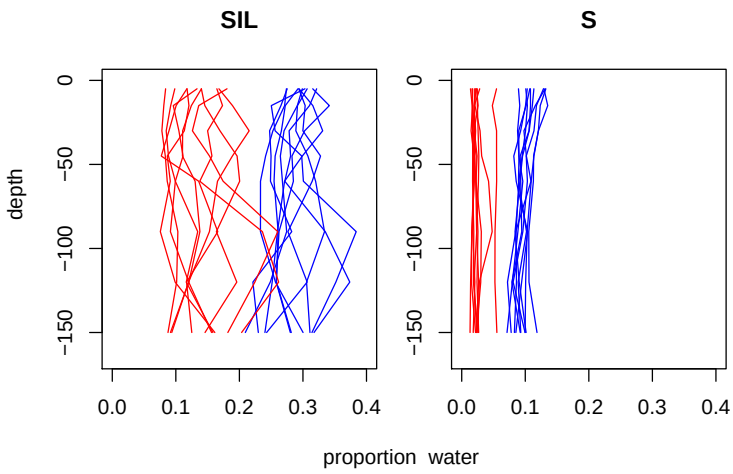

Figure 9: Simulated LL and DUL profiles for SIL and S soils.

Realizations for the $\log LL$ and $\log \Delta$ were generated from a multivariate normal distribution with mean and covariance based on the estimates of the parameters in (11) and (12) obtained from the fits in the previous section. Predicted values for the mean are based on an average soil composition profile, computed from the database, for SIL and S soils where soil composition was assumed to be constant across all depths, which in this case are $D = 5, 15, 30, 45, 60, 90, 120, 150$. Again, these simulated profiles represent a draw from the distribution of soil water profiles based on information on the mean and covariance structure and the uncertainty gleaned from the data.

6 Application

To demonstrate the value of simulated soil water profiles, a simple study using the CERES maize crop model was conducted. One hundred realizations of soil water profiles for SIL and S soils were generated and each was used as input to the CERES maize crop model utilizing the same twenty years of weather taken from a station in North Carolina, U.S.A. from 1960 to 1979. The results of the simulation experiment are summarized in Figure 10. In both the top and bottom frames, bands are displayed indicating the average yield for each year using the SIL (light) and S (dark) soils. The width of the bands at each year represents

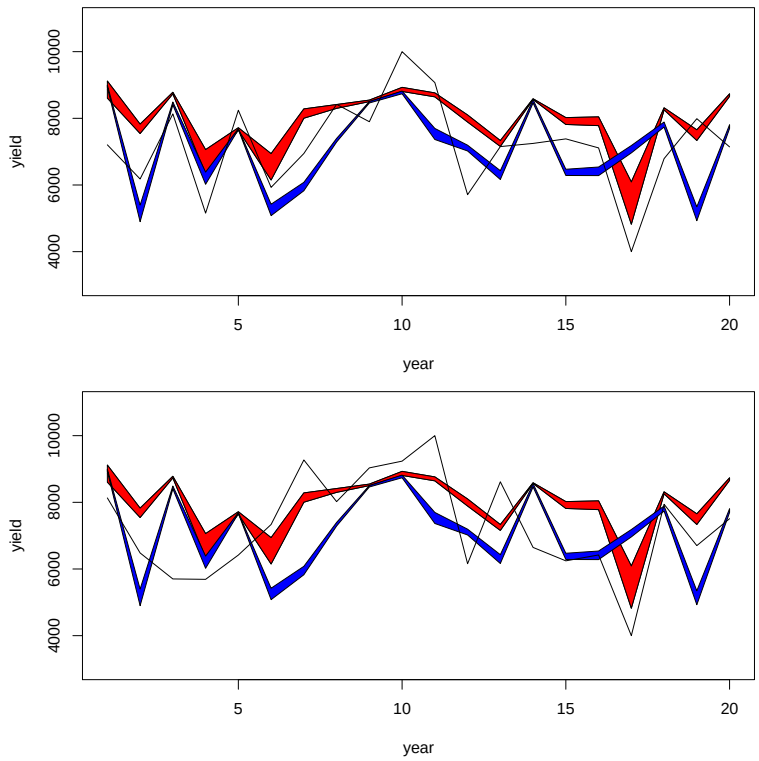

Figure 10: Output of 100 runs of the CERES maize crop model based on 100 simulated soil water profiles for SIL and S soils. SIL soils are shown in red and S soils are shown in blue. The width of the line for each year represents two standard errors in the average yield. Solid line in the top line represents total yearly precipitation while in the bottom plot the solid line represents average temperature.

two standard errors in average yield for that year. Also displayed are a representation of the total precipitation (top frame) and average temperature (bottom frame) for each year.

In general, results for both soils seem to track the weather characteristics. Further, yields are typically higher for SIL soils than that for S soils. There appears to be large differences in yield between broad classes of soil textures. However, the year-to-year variation in yields attributable to the weather dominates the within-year variation inherited from the soils.

One other feature of note appears in the results. In year 17, a dramatic difference between the yields for the two soils appears, with S soils having substantially larger average yields. On closer inspection, 57 of the 100 model runs for SIL failed to produce crops. This

year is characterized by low temperatures and low precipitation. Because the SIL requires a greater amount of water in the soil (refer to Figures 2 and 9), the lack of precipitation leads to decreases in yields and more frequent failed crops.

7 Concluding Remarks

An additive, multivariate regression model is introduced that includes smooth contributions of the explanatory variables common to all of the response variables as well as traditional linear contributions of additional covariates specific to each response variable. It should be noted that variations on this basic structure are easily accommodated. The model is also able to account for complex error structures, including spatial and other forms of dependence.

The model presented here shares a connection with thin-plate smoothing splines (de Boor, 1978; Wahba, 1990a; and Green and Silverman, 1994) and spatial models including universal kriging (Cressie, 1991). This model also offers an alternative to the semiparametric smoothing approach of Ruppert et al. (2003). On the surface, these regression techniques are quite different and are motivated from different perspectives. However, there has been much effort on establishing the clear connection between the spline smoothing and spatial models (Wahba, 1990b; Cressie, 1990; Kent and Mardia, 1994; and Nychka, 2000). This connection lies in the additive structure of these models that includes fixed, polynomial components as well as random components whose correlation structure effectively controls the amount of smoothing.

The model parameters can be fit using established techniques such as REML. However, for complex data structures and regression functions this may be impractical and an iterative procedure is introduced to reduce the computational burden. In this setting, parameter estimates very similar to those obtained from REML are obtained using the iterative approach. While we believe this algorithm is effective at producing reasonable parameter estimates, more study is certainly warranted.

The model is used to develop a new type of pedotransfer function for estimating soil

water holding characteristics based on soil composition data as well as other covariates. This model is unique in that the entire soil water profile as a function of depth is estimated. Of perhaps more importance is that the model offers the ability to capture the uncertainty in the soil characteristics as well as the ability to simulate complete soil water profiles.

Using this model, a small simulation study was conducted in which soil water profiles were simulated for two soil texture classes and yields computed using the CERES maize crop model. This initial study suggests that there are differences in yields between soil texture classes (based on composition), but that variation in yields due to the variation in weather dominate that due to variation in soils and uncertainty in the pedotransfer function.

Finally, this work represents a preliminary result that is a part of a larger collaboration between statisticians and scientists who are studying global climate change by using statistical models to assess the sources of uncertainly in large, complicated and typically deterministic models used to describe natural phenomenon.

Acknowledgements

This research is supported in part by the National Science Foundation under grant DMS 9815344.

References

Breiman, L. and Friedman, J.H. (1985), "Estimating optimal transformations for multiple regression and correlations (with discussion)," *Journal of the American Statistical Association*, **80**, 580-619.

Buja, A., Hastie, T.J., and Tibshirani, R.J. (1989), "Linear smoothers and additive models (with discussion)," *Annals of Statistics*, **17**, 453-555.

Cressie, N.A.C. (1990), Reply to Wahba's letter, *American Statistician*, **44**, 256-258.

Cressie, N.A.C. (1991), *Statistics for Spatial Data*, New York: John Wiley.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society, Ser. B*, **39**, 1-37.

Gijsman, A.J., Jagtap, S.S., and Jones, J.W. (2002), "Wading through a swamp of complete confusion: how to choose a method for estimating soil water retention parameters for crop models," *European Journal of Agronomy*, **18**, 75-105.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, London: Chapman and Hall.

Kent, J.T. and Mardia, K.V. (1994), "The link between Kriging and thin plate spline splines," In Kelly, F.P. (ed.), *Probability, Statistics, and Optimization*, New York: John Wiley.

Kitandidis, P.K. (1997), *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press.

Jagtap, S.S., Lall, U., Jones, J.W., Gijsman, A.J., and Ritchie, J.T. (2003), "A dynamic nearest neighbor method for estimating soil water parameters," In press.

McCulloch, C.E. and Searle, S.R. (2000), *Generalized, Linear, and Mixed Models*, New York: John Wiley.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press.

Minansy, B., McBratney, A.B., and Bristow, K.L. (1999), "Comparison of different approaches to the development of peodtransfer functions for water retention curves," *Geoderma 93*, 225-253.

Nychka, D.W. (2000), "Spatial-process estimates as smoothers," In Schimek, M.G. (ed.), *Smoothing and Regression: Approaches, Computation, and Application*, New York: John Wiley.

Pachepsky, Y.A., Rawls, W.J., and Timlin, D.J. (1999), "The current status of pedotransfer functions: their accuracy, reliability and utility in field- and region-scale modeling," In Corwin, D.L., Loague, K.M., and Ellsworth, T.R. (ed.), *Assessment of Non-Point Source Pollution in the Vadose Zone*, Geophysical Monograph 108, American Geophysical Union, Washington, D.C.

Ratliff, L.F., Ritchie, J.T., and Cassel, D.K. (1983), "Field-measured limits of soil water availability as related to laboratory-measured properties," *Soil Science Society of America Journal*, **47**, 770-775.

Rawls, W.J., Gish, T.J., and Brakensiek, D.L. (1991), "Estimating soil water retention from soil physical properties and characteristics," *Advances in Agronomy*, **16**, 213-234.

Ritchie, J.T., Ratliff, L.F., and Cassel, D.K. (1987), "Soil laboratory data, field descriptions and filed measured soil water limits for some soils of the United States," ARS Technical Bulletin, ARS Agriculture Research Service and Soil Conservation Service / Stat Agriculture Experimental Stations, Temple, Texas.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

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

Timlin, D.J., Pachepsky, Y.A., Acock. B., and Whisler, F. (1996), "Indirect estimation of soil hydraulic properties to predict soybean yield using GLYCIM," *Agricultural Systems*, **52**, 331-353.

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

Wahba, G. (1990b), Letter to the editor, *American Statistician*, **44**, 256.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.