A data generator for evaluating spatial issues in precision agriculture*

Dragoljub Pokrajac[1], Tim Fiez[2], Zoran Obradovic[1]

[1]Temple University, Center for Information Science and Technology, Philadelphia PA 19122
{pokie, zoran}@ist.temple.edu
[2] CAMO Inc., Corvallis OR 97339
tfiez@camo.com

* Partial support by the INEEL University Research Consortium project No.C94-175936 to T. Fiez and Z. Obradovic is gratefully acknowledged. The authors express their gratitude to the anonymous reviewers whose comments significantly improved the simulation method and overall quality of the paper.

All correspondences to: Z. Obradovic, zoran@ist.temple.edu, phone: +(215) 204 – 6265, fax: +(215) 204 - 5082

A data generator for evaluating spatial issues in precision agriculture

Abstract

With the rapid rise in site-specific data collection, many research efforts have been directed towards finding optimal sampling and analysis procedures. However, the absence of widely available high quality precision agriculture data sets makes it difficult to compare results from separate experiments and to assess the optimality and applicability of procedures. To provide a tool for spatial data experimentation, we have developed a spatial data generator that allows users to produce data layers with given spatial properties and a response variable (e.g. crop yield) dependent upon user specified functions. Differences in response functions within fields can be simulated by assigning different models to regions in coordinate- (x and y) or feature space (multidimensional space of attributes that may have an influence on response). Noise, either unexplained variance or sensor error, can be added to all spatial layers. Sampling and interpolation error is modeled by sampling a continuous data layer and interpolating values at unsampled locations. The program has been successfully tested for up to 15000 grid points, 10 features and 5 models. As an illustration of the potential uses of generated data, the effect of sampling density and kriging interpolation on neural network prediction of crop yield was assessed. Yield prediction accuracy was highly related (correlation coefficient 0.98) to the accuracy of the interpolated layers indicating that unless data are sampled at very high densities relative to their geostatistical properties, one should not attempt to build highly accurate regression functions using interpolated data. By allowing users to generate large amounts of data with controlled complexity and features, the spatial data generator should facilitate the development of improved sampling and analysis procedures for spatial data.

A data generator for evaluating spatial issues in precision agriculture

I. Introduction

The advent of global positioning systems and new sensors enables agricultural producers to collect spatial data more easily than ever before. However, this ability raises many questions regarding how to best collect spatial data and how to properly interpret and analyze the information. The answers to these questions often cannot be proven by theory, but must be inferred from multiple experiments. However, diversity in data sets, data collection procedures, and analysis approaches makes it difficult to compare results from separate experiments and will likely slow the discovery of optimal procedures.

Every experimental data set is unique and reflects the underlying spatial properties of the experimental site. McBratney and Pringle (1997) presented semivariograms for pH, clay content, carbon, potassium (K), and phosphorus (P) obtained by various researchers, and for every variable, there were differences in the nugget, sill variance and range. Furthermore, they showed that the sampling intensity and the size of the sampled area influenced semivariogram estimation, so even if the true underlying spatial properties are equal among experimental sites, differences in sampling schemes and methods can result in different estimates of spatial properties. Data collection procedures also vary greatly. Some attributes are measured directly by established analytical procedures while some attributes are measured indirectly from measurements of related properties such as reflected electromagnetic radiation (Ben-Dor and Banin, 1995; Moran et al., 1997) or soil electrical conductivity (Doolittle et al., 1994). Sampling density can range from one sample per few acres (grid soil fertility sampling) to one sample for every 9 m^2 (100 ft^2) or less (on-the-go sensors, high resolution remote sensing). Often, indirect measurements allow greater sampling density at the cost of some error in attribute values. Depending on the sensing system, a single sample can represent a point estimate of the true underlying value (single soil sample, some instantaneous sensor readings) or the aggregate value of the attribute over some area (remote sensing, yield monitors).

Post-collection data processing further complicates the comparison of data from different sources and the results of different analytical routines. Interpolation procedures are frequently used to adjust the resolution of various features to one common grid resolution. However, the interpolation of sparsely

sampled data can introduce large estimation errors (Vucetic et al., 2000). Post-collection data processing procedures can also identify and remove outliers from large data sets and compensate for yield variability that is contributed by the sensors and measurement systems and therefore is unexplainable by observed features (Blackmore and Marshall, 1996; Saunders et al., 1996; Stafford et al., 1996).

The joint effect of all of these factors is that precision agriculture data sets will contain varying levels of error or noise, which are often not quantified. Furthermore, the number of features and their importance in determining the measured response, most often yield, can vary. Thus, it is not possible to fully assess the applicability of a single procedure if it is tested on one or a few data sets or to equate the performance of two procedures if each is tested on different data sets.

As data become more available, researchers can test knowledge discovery procedures on multiple data sets allowing the formation of more general statements. As an example, Kravchenko and Bullock (1999) used 30 fields in Illinois, Indiana, and Iowa in their comparative study of interpolation methods, while Gotway et al. (1996) only used two fields in their comparison of kriging and inverse distance interpolation. While increasing the number of data sets used in sampling and data analysis studies makes the results of these studies less data set specific, it is still not possible to truly compare the results of separate studies unless the same data or data from the same population are used. One solution to this is to evaluate procedures using standard widely available data sets. For example, the knowledge discovery in databases (KDD) and machine learning research communities have established data repositories (Bay, 1999; Blake and Mertz, 1998) to enable researchers to use reference data for testing purposes. Unfortunately, these types of data repositories do not exist for precision agriculture.

Data generators, programmable and controllable mathematical procedures for simulating data, provide another solution to the problem of data availability and sharing. Data generators provide the ability to create unlimited amounts of data, and sophisticated generators allow the user to control the complexity of the data, add known amounts of noise, and generate response variables dependent upon generated driving variables. Thus, one can generate feature layers that match the spatial statistics of attributes determined through real world sampling. Investigators can use these simulated attribute layers to test sampling schemes, post-data collection processing procedures, and other spatial processes. For example, Russo and Jury (1988) and Webster and Oliver (1992) used unconditioned stochastic simulation to

compare various soil-sampling networks for estimating spatial variability and Bierkens and Burrough (1993) proposed a conditioned indicator simulation technique to map categorical soil data. Englund (1993) and Rossi et al. (1993) demonstrated the use of conditioned simulation methods as a tool for risk management in environmental applications, which Desbarats (1996) successfully applied in geology. Pachepsky and Acock (1998) applied stochastic simulation to produce available soil water capacity maps and assess variability and uncertainty of crop yield estimates.

To enable testing of regression procedures, a data generator must provide a way to formulate a response variable based on a set of driving attributes. A data generator could use a mechanistic crop model to estimate yield. Corá et al. (1999), Fraisse et al. (1999), and Paz et al. (1999) tested the ability of the CERES-Maize model to predict observed within-field yield variability. In many cases, the predictions were close to observed results indicating that mechanistic crop models can produce simulated crop yields that reflect real world variability. However, crop models can require many input variables complicating the data generating process. In addition, the response of the model to a given input is hard-coded in the model. To arrive at a given output response, say a curvilinear nitrogen (N) response, one might have to derive the appropriate input values for multiple interacting features. While many crop simulation models try to model the true biological complexity, a data generator could simply use deterministic mathematical functions to compute yield for a given input. Combination of individual function outputs could simulate the effect of multiple driving attributes.

By combining sequential simulation techniques (Goovaerts, 1997; Deutsch and Journel, 1998) and deterministic response functions, we have developed a spatial data generator that can generate spatial layers satisfying given statistical properties and a response variable dependent upon user-controlled equations. Our intention is to provide a user-friendly yet powerful way to simulate different aspects of data generation, acquisition, and processing. The generator enables the testing of the effects of sampling density, sensor error, presence or absence of particular attributes, and field heterogeneity on sampling procedures, data acquisition systems, and KDD algorithms in a controllable way (Pokrajac et al., 2000). Furthermore, by using this generator, it is possible to examine the behavior of different machine learning algorithms and to determine their robustness and potential effectiveness with complex spatial data. Also, due to the modular structure of the generator, it is possible to gradually increase the

complexity of generated data to simulate prospective emerging aspects of precision agriculture technology and practice. The following sections discuss the method and approach for data generation, the software implementation, and an analysis of interpolation effects on non-linear neural network prediction using generated data.

II Method and approach

The data generation process consists of two steps: feature generation and yield simulation.

II.1 Feature generation

Features are generated through a multistep process of 1) grid determination, 2) generation of independent features, 3) feature correlation and 4) feature verification.

II.1.1 Spatial grid determination

The first step in the data generation process is to specify the shape of the simulated field and the grid spacing on which data will be generated. Non-rectangular fields can be obtained by masking a rectangular field with other shapes. Symbolically, denote with S the set of spatial points $\boldsymbol{s}_i=(x_i, y_i)$ $i=1,\ldots,n$ on which yield and features are generated.

II.1.2. Generation of independent features

Each independent feature $f_j(\boldsymbol{s}_i)$ is a random function of spatial points $\boldsymbol{s}_i$. A feature is characterized by its spatial correlation structure and probability density function. To generate a feature $f_j(\boldsymbol{s}_i)$, we first simulate an intermediate second-order stationary (Chilès and Delfiner, 1999) standardized feature $f^*(\boldsymbol{s})$ with a zero-mean, unit-variance Gaussian probability distribution (Devore, 1995) and a spatial correlation specified by a theoretic semivariogram $\gamma(\boldsymbol{h})$ (Chilès and Delfiner, 1999), a conditionally negative definite function (Cressie, 1993) measuring the average dissimilarity between data samples separated in space by a vector $\boldsymbol{h}=(h_x,h_y)$ (Goovaerts, 1997).

We generate a standardized feature f^* ($\boldsymbol{s}$) using a modification of the sequential Gaussian simulation algorithm (Deutsch and Journel, 1998), which samples from an unconditional distribution at the first simulated point. This procedure is as follows:

1. Define a random sequence $\boldsymbol{s}^{(0)}$, $\boldsymbol{s}^{(1)}$,…, $\boldsymbol{s}^{(n-1)}$ of visiting spatial grid points $\boldsymbol{s}_i$, i =1,…,n.
2. At spatial location $\boldsymbol{s}^{(0)}$, generate a feature value $f^*(\boldsymbol{s}^{(0)})$ as a sample of a standardized Gaussian distribution.
3. For i = 1,…,n-1 (while a feature value is not determined on all grid points):
    a. Take the next grid point $\boldsymbol{s}^{(i)}$ from the sequence of visiting defined in step 1;
    b. Using a set S' of up to N_s (a pre-specified constant) previously simulated values of the standardized feature f^*, determine the mean μ_K and variance σ^2_K of a Gaussian distribution to be utilized for sampling the feature value $f^*(\boldsymbol{s}^{(i)})$ at the location $\boldsymbol{s}^{(i)}$;
    c. Generate $f^*(\boldsymbol{s}^{(i)})$ as a sample from the Gaussian distribution with parameters μ_K and σ^2_K.

Since the simulated spatial processes f^* should obey a standard Gaussian distribution with a zero mean and unit variance, by exploiting a relationship between semivariograms and spatial covariance (Chilès and Delfiner, 1999) values μ_K (kriging estimation) and σ^2_K (kriging variance) can be determined using simple kriging (Deutsch and Journel, 1998) as:

$$\mu_K = \sum_{\substack{for\ all \\ s^{(\alpha)}\ in\ S}} \lambda_\alpha f^*\left(\boldsymbol{s}^{(\alpha)}\right)$$
$$\sigma^2_K = 1 - \sum_{\substack{for\ all \\ s^{(\alpha)}\ in\ S'}} \lambda_\alpha \left(1 - \gamma\left(\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(\alpha)}\right)\right) \quad (1)$$

where kriging coefficients λ are obtained from the following system of equations:

$$\sum_{\substack{for\ all \\ s^{(\beta)}\ in\ S'}} \lambda_\beta \left(1 - \gamma\left(\boldsymbol{s}^{(\beta)} - \boldsymbol{s}^{(\alpha)}\right)\right) = 1 - \gamma\left(\boldsymbol{s}^{(i)} - \boldsymbol{s}^{(\alpha)}\right), \text{ for all } \alpha \text{ such that } \boldsymbol{s}^{(\alpha)} \text{ is in set S'.} \quad (2)$$

Since a semivariogram is conditionally negative definite function, a kriging system, eq. (2), has a unique solution assuming that for the particular semivariogram γ ($\boldsymbol{h}$), a corresponding spatial covariance function is defined (Chilès and Delfiner, 1999).

Semivariograms $\gamma(\boldsymbol{h})$ are specified by the type (e.g. exponential, spherical), nugget c_0, sill c_e, and range a, and can be isotropic or anisotropic. Isotropic semivariograms depend only on the modulus $h \equiv |\boldsymbol{h}| = \sqrt{h_x^2 + h_y^2}$ of the separation vector $\boldsymbol{h}$, while the spatial variability at anisotropic semivariograms depends on both the modulus and direction of $\boldsymbol{h}$. The character (geometric or zonal), intensity (ratio) and direction of anisotropy are specified by a *transformation matrix* (Chilès and Delfiner, 1999).

When the simulated feature $f_j(\boldsymbol{s}_i)$ has a Gaussian probability distribution function specified by mean μ_j and standard deviation σ^2_j, its values can be computed from generated values $f^*(\boldsymbol{s}_i)$ of a standardized feature using the following transformation:

$$f_j(\boldsymbol{s}_i) = f^*(\boldsymbol{s}_i) \cdot \sigma_j + \mu_j, i = 1,\ldots,n \quad (3)$$

To generate a feature that has an arbitrary distribution q(y), feature values can be obtained by using a backward normal score transformation (Goovaerts, 1997). Given a realization of the random variable x that follows a standardized Gaussian distribution p(x), an equivalent value of y is identified such that values of the cumulative distribution functions $cdf_p(x)$ and $cdf_q(y)$ are equal (i.e. y corresponds to the same quantile as x). Observe that $cdf_p(x)$, the cumulative distribution function for p(x), can be obtained analytically since p(x) is Gaussian, while $cdf_q(y)$ can be estimated from (previously decorrelated) real world features using the technique described in Deutsch and Journel (1998).

II.1.3 Feature correlation

All features are simulated independently, but the user may desire that they be correlated according to some pre-defined specifications. To obtain features with a specified covariance matrix $\mathbf{C_V}$ (which can be estimated from real agricultural data as a sample covariance matrix of features, see Appendix or (Flury, 1997)), simulated independent features that should have a unit variance are multiplied by the Cholesky transformation (Flury, 1997) of $\mathbf{C_V}$. In this case, semivariogram parameters of independent features should be estimated on decorrelated real-world features transformed to have a standard Gaussian distribution (zero mean, unit standard deviation). Features can be decorrelated using the following eigenvalue projection method (Duda et al., 2000) (a generalization of the principal

component transformation that Desbarats (1996) used for a two-variable case). If **A** is an n×F matrix of F simulated independent features and $\mathbf{C_A}$ is the sample covariance matrix (Flury, 1997), pre-multiplication of **A** with the eigenmatrix of $\mathbf{C_A}$ (Golub and Loan, 1989) decorrelates the features. Having been decorrelated, features can be sequentially (one by one) transformed into features with a standard Gaussian histogram using a normal score transformation inverse to the backward transformation described above (Goovaerts, 1997).

II.1.4 Feature verification

The final step in the feature generation process is to verify that generated features match user-specified properties. Univariate statistics such as means, medians, and variance of particular features can be determined using standard statistical techniques (Devore, 1995). Features distribution and inter-attribute correlation can be evaluated through estimated histograms, quantile-quantile plots and bivariate scatterplots (Deutsch and Journel, 1998). Spatial characteristics of simulated data can be verified by calculating estimated semivariograms and fitting model semivariograms.

To estimate isotropic semivariograms $\hat{\gamma}(h)$, the relevant range of distances h is divided into L equally spaced lags [h_l -D_h, h_l +D_h], l=1,…,L, where $h_l - h_{l-1}$ is the lag spacing and D_h is the lag tolerance. Then, for each l=1,…,L, all n_l pairs of points whose distance belongs to the l-th bin are used to calculate an average squared difference of function values applying the method of moments (Cressie, 1993) or robust estimation (Hawkins and Cressie,1984). An analog procedure for estimation of anisotropic variograms is described in (Deutsch and Journel, 1998).

To fit an isotropic semivariogram model $\tilde{\gamma}(h_l, c_e, c_0, a_s)$, the parameters c_e, c_0, a_s that "best" fit the estimated semivariogram must be determined. This can be done by visual assessment or by minimizing the sum:

$$\sum_{l=1}^{L} [\hat{\gamma}(h_l) - \tilde{\gamma}(h_l, c_e, c_0, a_s)]^2 \qquad (4)$$

Since the accuracy of the estimated semivariogram depends on the number of points n_l in l-th bin, an alternative is to minimize the sum (Cressie, 1985):

$$\sum_{l=1}^{\mathrm{L}} n_l \left[\frac{\hat{\gamma}(h_l)}{\tilde{\gamma}(h_l, \mathrm{c_e}, \mathrm{c_0}, \mathrm{a_s})} - 1 \right]^2 \quad (5)$$

The procedure for fitting anisotropic semivariograms includes determination of the anisotropy transformation matrix and other applicable semivariogram parameters by minimization of sums analog to (4) and (5) computed for all relevant discrete values of distances and directions of anisotropy (Deutsch and Journel, 1998).

II. 2 Simulation of crop yield

Yield is generated from the feature data using linear, loglinear, linear plateau, or exponential plateau models $g(\boldsymbol{s}_i)$. For the linear model, yield is generated as the weighted sum of the feature values $f_j(\boldsymbol{s}_i)$:

$$g(\boldsymbol{s}_i) = \mathrm{u}_0 + \sum_{j=1}^{\mathrm{F}} \mathrm{u}_j f_j(\boldsymbol{s}_i) \quad (6)$$

For the loglinear model, the logarithm of yield is computed as the weighted sum of the logarithms of the feature values:

$$\log(g(\boldsymbol{s}_i)) = \mathrm{u}_0 + \sum_{j=1}^{\mathrm{F}} \mathrm{u}_j \log(f_j(\boldsymbol{s}_i)) \quad (7)$$

To apply this model, all features must be positive, which can be achieved by normalization.

For the plateau models, yield is generated proportional to the product of plateau functions $\pi_j(\boldsymbol{s}_i)$ for each feature used to generate yield:

$$g(\boldsymbol{s}_i) = \mathrm{G} \prod_{j=1}^{\mathrm{F}} \pi_j\left(f_j(\boldsymbol{s}_i)\right) \quad (8)$$

where G is the coefficient of proportionality.

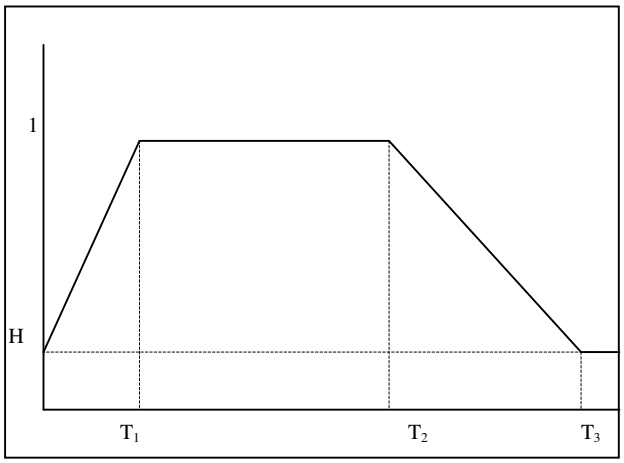
Plateau functions can be linear:

$$\pi_j(x)=\begin{cases} H_j, & \text{if } x \le 0 \\ H_j + x(1-H_j)/T_{1,j}, & \text{if } 0 < x \le T_{1,j} \\ 1, & \text{if } T_{1,j} < x \le T_{2,j} \\ 1-(x-T_{2,j})\cdot(1-H_j)/(T_{3,j}-T_{2,j}), & \text{if } T_{2,j} < x \le T_{3,j} \\ H_j, & \text{if } x > T_{3,j} \end{cases} \quad (9)$$

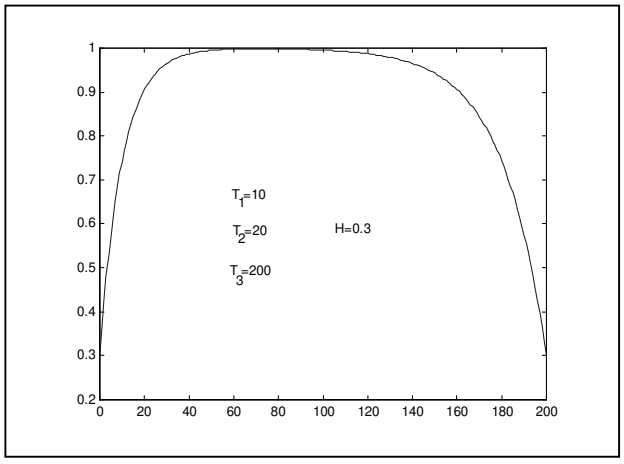
or exponential:

$$\pi_j(x)=\begin{cases} H_j+[1-\exp(-(T_{3,j}-x)/T_{2,j})]\cdot[1-\exp(-x/T_{1,j})]\cdot(1-H_j), & \text{if } 0<x\le T_{3,j} \\ H_j, & \text{otherwise} \end{cases} \quad (10)$$

In these formulas, $T_{1,j}$, $T_{2,j}$ and $T_{3,j}$ determine the slope and range of the plateau functions while $H_j > 0$ determines the floor of the particular feature's influence on yield (Fig. 1). The effect of H_j is quantified by calculating the influence strength mod_j defined as $mod_j = (1-H_j) / (1+H_j)$. Mod_j has a maximum value of 1 when feature *j* has its maximum influence on yield. Alternatively, when $mod_j = 0$, yield does not depend on feature *j* since then $\pi_j(x)=1$ for all values *x* of the feature.



a) Linear plateau model.



b) Exponential plateau model.

Figure 1. Plateau models for yield generation.

Model coefficients can be determined from actual data using regression techniques or taken from sources of expert knowledge such as fertilizer recommendation guides. Not all generated features must be used to generate yield. To introduce feature irrelevancy, one can set corresponding model coefficients to values [u_j in (6), (7) to 0, and H_j in (9), (10) to 1], which cancel the influence of the feature on calculated yield.

II.2.1 Modeling of spatial and feature-space heterogeneity

To simulate situations where the yield response to features varies over space, or varies due to differences in feature values, yield can be modeled such that the yield at grid point $\boldsymbol{s}_i$ is the weighted sum of yields generated by M models $g_m(\boldsymbol{s}_i)$:

$$y(\boldsymbol{s}_i) = \sum_{m=1}^{M} w_{m,i}\, g_m(\boldsymbol{s}_i), \qquad where \ \sum_{m=1}^{M} w_{m,i} = 1 \tag{11}$$

This gives the spatial data generator the ability to simulate situations such as a difference in N response due to soil drainage or a difference in P response due to soil pH. Different models can be assigned to different spatial areas or to different regions of feature values (e.g. poor and well-drained soil, or acid, neutral, and basic pH). The use of spatial position for model assignment can simulate conditions where variables that contribute to yield determination are missing from the data set.

Coefficients $w_{m,i}$ for each grid point $\boldsymbol{s}_i$ define the proportion of the total yield contributed by model $g_m(\boldsymbol{s}_i)$. We use the term "hard" generation if only one of M models contributes to yield at each point $\boldsymbol{s}_i$, and the term "smooth" generation when multiple models contribute to yield. Smooth generation allows gradual transitions from one model to another.

The determination of $w_{m,i}$ to vary models with spatial location is as follows. For each point $\boldsymbol{s}_i = (x_i, y_i)$ in S, generate a realization of a Gaussian variable v with zero mean and unit variance. For each (x_i, y_i), compute v' as the average of all values v for the points in the rectangle bounded by points $x_i \pm \Delta x$, $y_i \pm \Delta y$ where Δx, Δy are user-defined parameters. Generate v^* by discretizing the value of v' over S into M equal bins. Compute coefficients for "hard" yield generation such that:

$$w_{\text{hard},m,i} = \begin{cases} 1, if\ v^*(\boldsymbol{s}_i) = m \\ 0, if\ v^*(\boldsymbol{s}_i) \neq m \end{cases} \quad (12)$$

Compute coefficients for "smooth" yield generation by averaging values of $w_{\text{hard},m,i}$ for all points in rectangle $x_i \pm \Delta x_{\text{soft}}$, $y_i \pm \Delta y_{\text{soft}}$ (Δx_{soft}, Δy_{soft} are adjustable parameters).

In addition to this stochastic approach, when the partition of a field into subfields assigned to particular models is known in advance (as an input to the simulation procedure), each sample point $\boldsymbol{s}_i$ can be assigned a value v^* equal to the label of a subfield to which the point belongs. Afterward, coefficients for "hard" and "smooth" yield generation can be determined using eq. (12) and the previously explained averaging technique.

To vary models due to differences in feature values, assign a model label c_i to each point $\boldsymbol{s}_i$ and generate coefficients $w_{\text{hard},m,i}$ such that $w_{\text{hard},m,i} = 1$ if $c_i = m$ and 0 otherwise. The smooth coefficient for point $\boldsymbol{s}_i$ is generated by averaging $w_{\text{hard},m,i}$ for all points in a specified neighborhood around point $\boldsymbol{s}_i$ in feature space. Model labels c_i take integer values between 1 and M and are determined by a pre-specified partitioning of feature space into regions that correspond to particular models. The existence of multivariate indexes (data-structures to facilitate fast access, range search and retrieval of spatial and multivariate data, described in Bertino et al. (1997)) in feature space can dramatically improve the performance of the algorithm.

II. 2.2 Simulation of unexplained variance, sensor error, and sampling and interpolation

Unexplained variance in yield is modeled by adding Gaussian noise with zero mean and variance such that the ratio of explainable yield variability (determined by feature values and response functions) to total yield variability is equal to a user specified value. Sensor error with a symmetric probability distribution around zero and a magnitude proportional to the actual value of the measured variable (features or yield) is modeled by multiplying values with Gaussian noise. This multiplicative noise has unit mean and standard deviation σ_n dependent on the specified sensor precision and the confidence in calibration. For instance, a sensor which has an error not larger than 5% of a true measured value in 95% of time is modeled with a multiplicative noise having a standard deviation that satisfies $2\sigma_n=0.05$ (since in

the Gaussian framework, a 95% two-sided confidence interval corresponds approximately to two standard deviations around the mean, (Devore, 1995)).

Finally, the error through the interpolation of sampled values is introduced by emulating the actual sampling process. Data are sampled from the generated data grids and used to interpolate values at unsampled locations. Uniform and random sampling (Webster and Oliver, 1990) and inverse distance and kriging interpolations (Chilès and Delfiner, 1999) are supported.

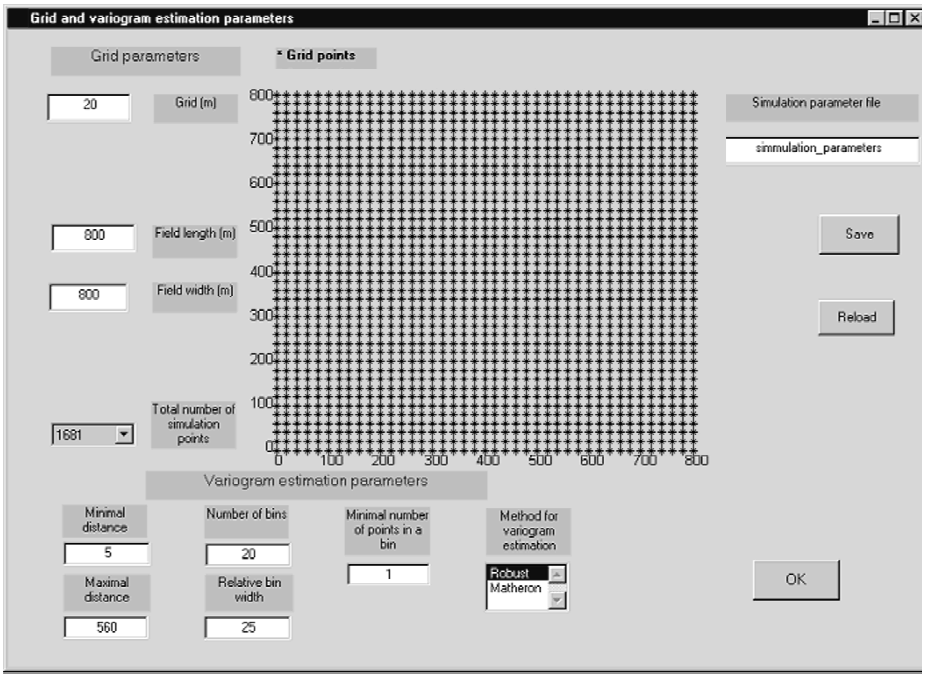
III Implementation

The spatial data generator can be implemented in any programming environment that facilitates embedding of the geostatistical and matrix functions and provides a user-friendly graphical interface. The current version of our software has been implemented in MATLAB® (The MathWorks, Inc., version 5.3) (Hanselman and Littlefield, 1997;Middleton, 2000), but the same functionality can be achieved using GSLIB software (Deutsch and Journel, 1998) supplemented by auxiliary routines written in a high-level language (such as Microsoft Visual Basic™ described in (Halvorson, 1998) or Microsoft Visual C++™ described in (Horton, 1998)). The software provides three modes: i) generation of features, ii) generation of yield, and iii) generation of unexplained variance and sensor error and the simulation of sampling and interpolation.

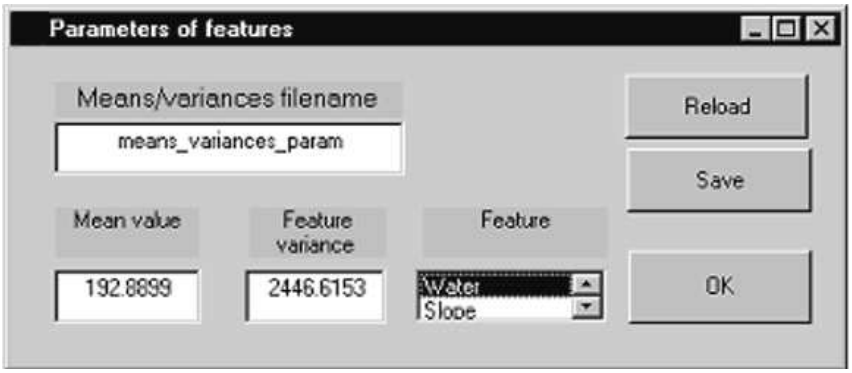
In the feature generation mode, the user is presented with a number of parameter input screens. Grid and semivariogram estimation parameters include field shape, grid step size, the number of simulated points, minimal and maximal lag and bin width for semivariogram estimation (Fig. 2a). Feature parameters include the number of features to generate, feature means and variances (Fig. 2b), the parameters of their model semivariograms (sill, range, nugget, semivariogram type: exponential or spherical, and the presence and character of anisotropy, Fig. 2c), feature distribution specifications (histograms) and a feature correlation matrix. Parameters are stored in binary files and the user can load parameters from existing files in addition to keyboard entry.

Once the parameters have been entered, the program proceeds through an interactive feature generation process. The user is presented with the estimated and model semivariogram and normalized histogram for a feature (Fig. 3). The user can then accept the feature and continue with the next one, or

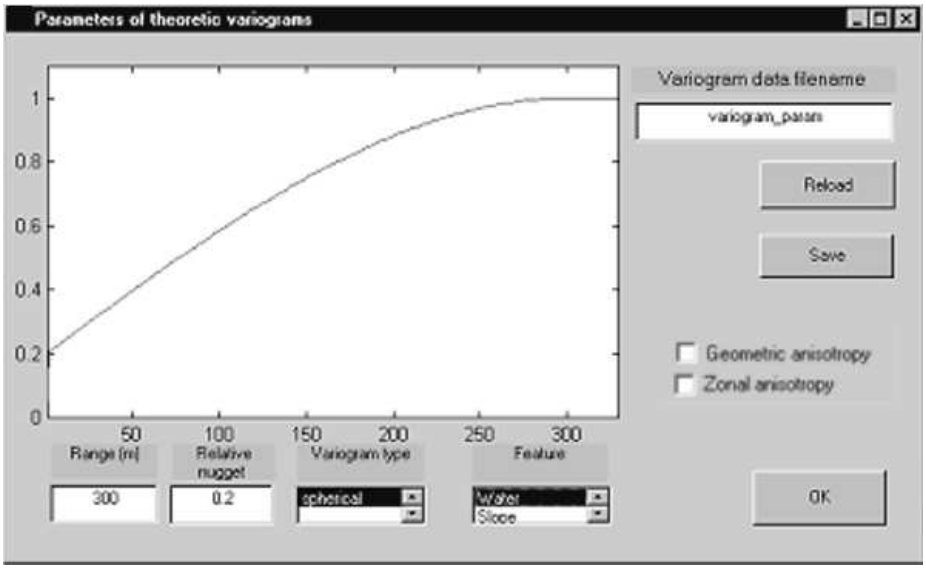
re-generate the current feature if unsatisfied with the generated statistics. After the features are generated, the user can view probability (quantile-quantile) plots to assess feature distributions and feature maps (value versus spatial coordinates) to assess spatial patterns.



a) Variogram estimation and grid parameters.



b) Feature parameters.



c) Model variogram parameters.

Figure 2. Feature generation input screens.

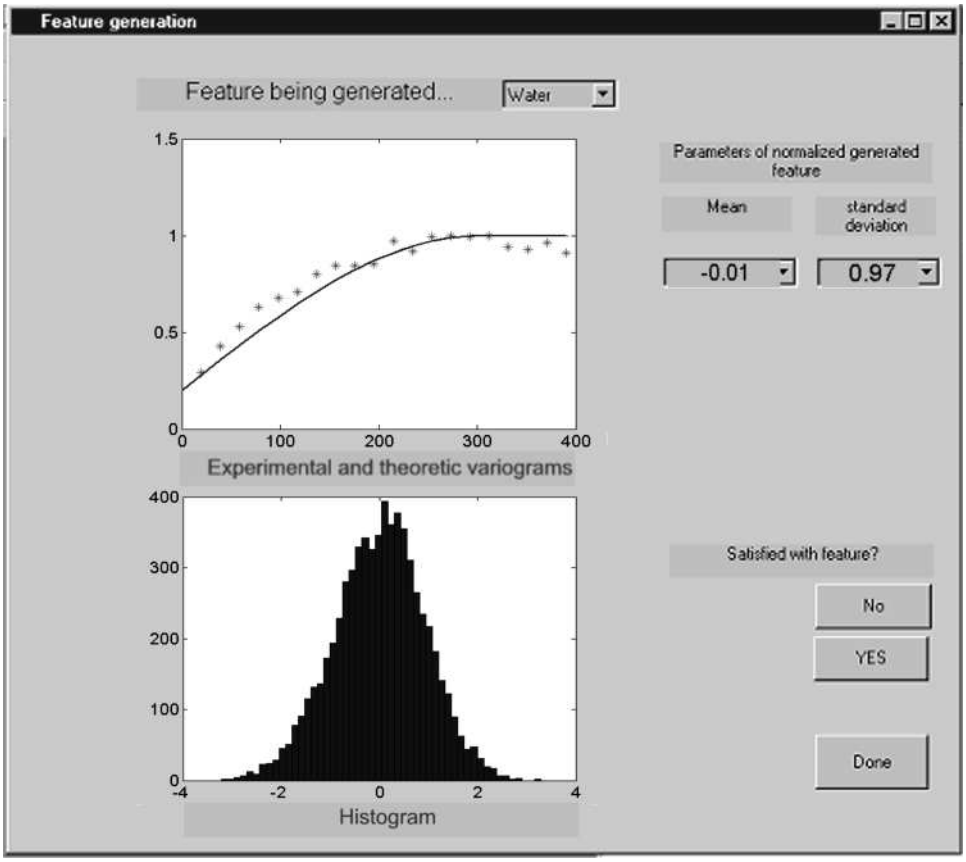


Figure 3. Screen for the interactive feature generation process showing the current parameters of the generated feature.

In the yield generation mode, the user selects the features upon which yield will depend, and the number and the type of models to apply (linear, loglinear or plateau). If the user selects more than one yield-generating model, he/she must also choose how to assign the different models; models can be assigned to regions in coordinate (*x* and *y*) or in feature space. Next, in an interactive process, the user sets parameters for each model and controls their influence on generated yield. An example of this process for a linear plateau model is shown in Fig. 4. For each model $m = 1,\ldots,M$ and for each feature $j=1,\ldots,F$, (e.g. N, P,…), the user enters parameters $T_{1,m,j}$, $T_{2\,m,j}$, $T_{3,m,j}$ (denoted by T1, T2 and T3 on the screen capture) and $mod_{m,j} = (1-H_{m,j})/(1+H_{m,j})$ (denoted with mod on screen) describing the shape of $g_{m,j}$ which is plotted along with the normalized histogram of the feature. By varying G, the user can also

specify a "ceiling," the maximum allowed yield. Also, for each model $m = 1,\ldots,M$, the user can visualize where the model is applied in x and y coordinate space (Fig. 5).

By inspecting mean, median, standard deviation, minimum and maximum values, the number of outliers (points where yield is not defined or is less than zero), and the normalized histogram for yield (Fig. 4), the user can tune model parameters to obtain generated yields matching desired requirements. After parameters are established for all features and models, the user can check statistics (mean, variance, and others) and view the histogram and probability plots (versus a normalized Gaussian distribution) of the overall generated yield layer.

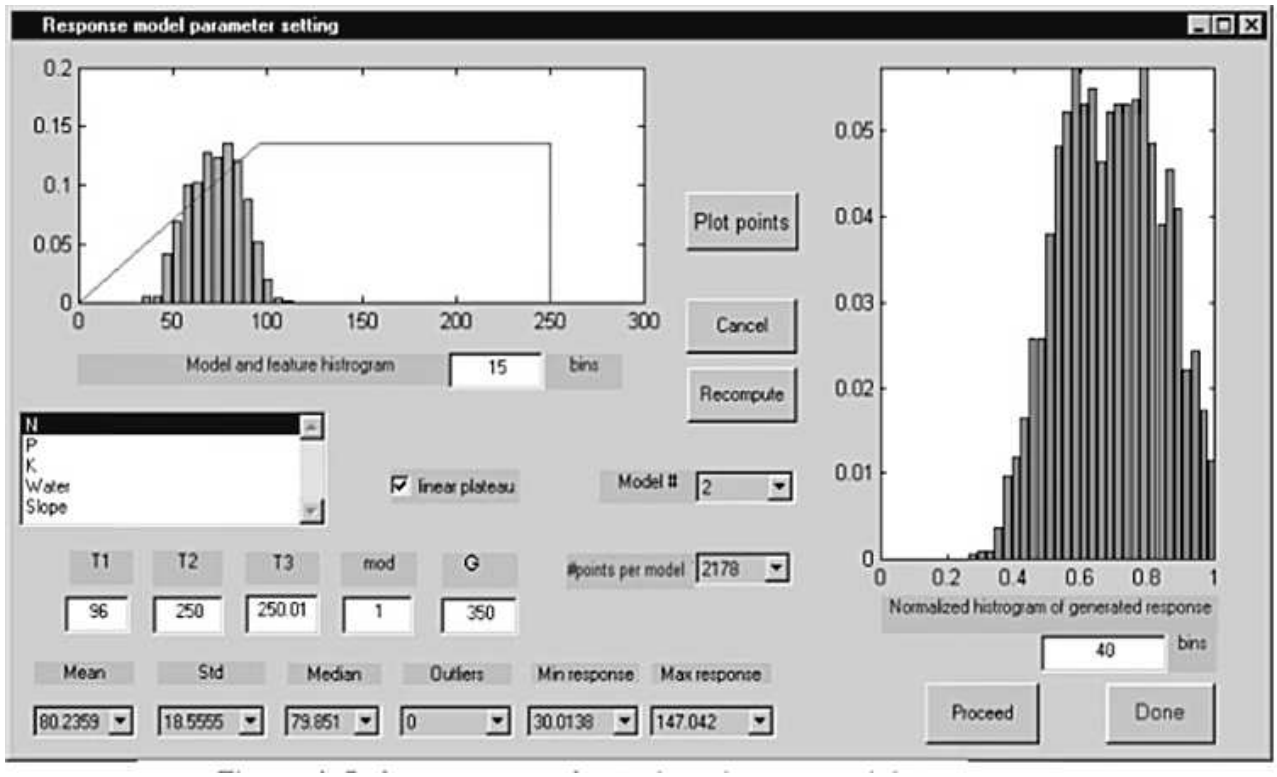


Figure 4. Software screen for tuning plateau model parameters.

In the postprocessing mode, unexplained variance and sensor error are added to yield and feature values and the effect of sampling is simulated. Figure 6 shows a yield layer before and after the addition of unexplained variance. To simulate sampling, data points are selected from a complete data grid at a user specified spacing and used to interpolate a new data grid (using ordinary kriging). The influence of different sampling distances is measured by ε, the interpolation mean squared error (squared

difference of a true feature value and a kriging estimate, averaged through the whole field) normalized with the variance of the original feature. Figure 7 compares an original feature sampled with a 10 $\times$10m^2 resolution to interpolated features with 10 $\times$10m^2 resolution generated from sampling the original feature using 40 $\times$40, 80 $\times$80 and 200 $\times$200m^2 sampling grids.

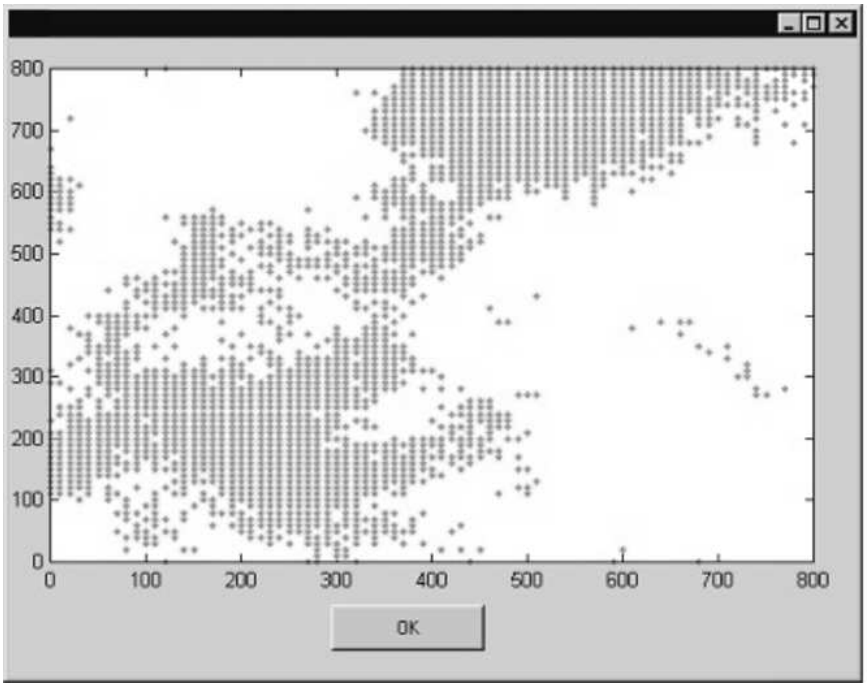
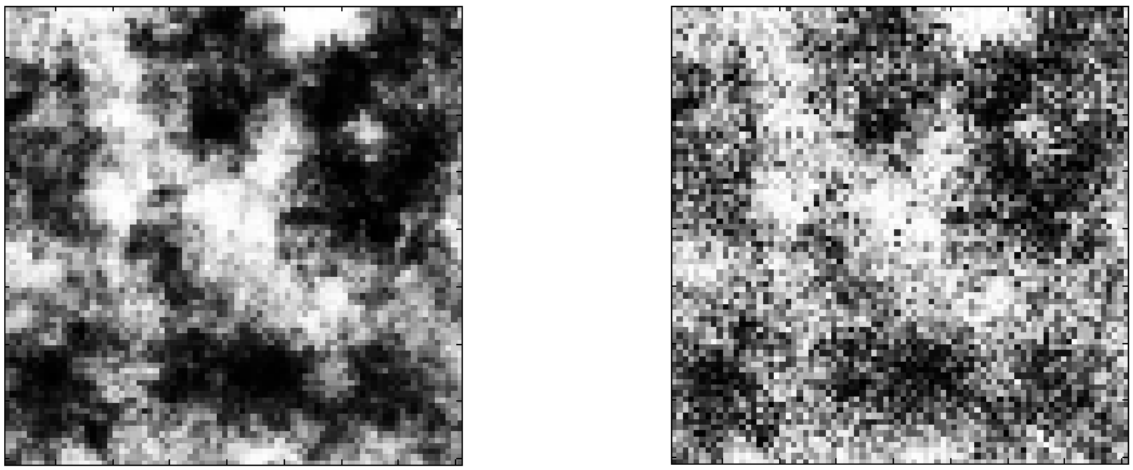


Figure 5. Plot of the assignment (denoted by symbols) of one model in a multi-model simulation in x- and y-coordinate space.

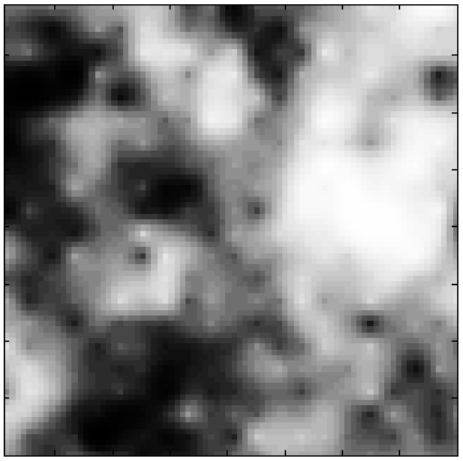


a) Before addition of unexplained variance.

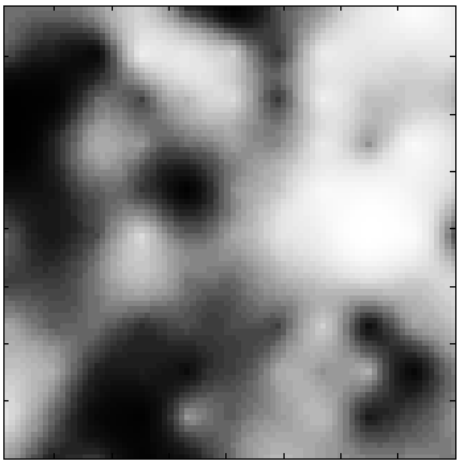
b) With 50% unexplained variance.

Figure 6. Generated yield.

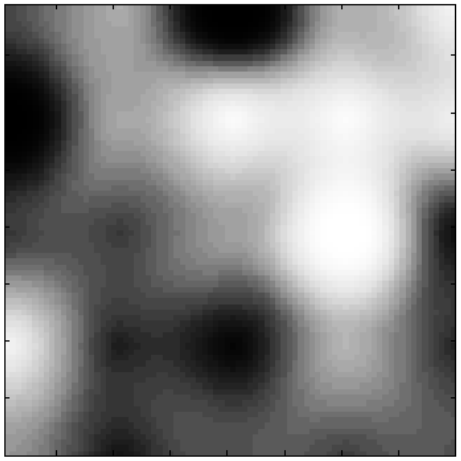
a) Original feature on a 10×10 m^2 grid.



b) Interpolated feature using samples from a 40×40 m^2 grid, $\varepsilon = 0.12$.



c) Interpolated feature using samples from a 80×80 m^2 grid, $\varepsilon = 0.25$.



d) Interpolated feature using samples from a 200×200 m^2 grid, $\varepsilon = 0.80$.

Figure 7. Influence of sampling density on interpolation error.

When the data generation process is complete, the simulated features and yield, as well as the simulation parameters are saved in a file specified by the user. Due to the interactive nature of the software, typical data generation sessions last 20 minutes to 1 hour. Intermediate results are saved during the data generation process in case program execution is interrupted before the process is complete.

The computational complexity of the data generation process establishes the practical limits for grid point, feature, and model numbers. The program has been successfully tested for up to n=15,000 grid points, F=10 features and M=5 models and has a linear complexity in F,M, and n.

IV Experiments with simulated data

There are many important spatial data issues in precision agriculture that can be tested with generated data. Consider building relationships between crop yield maps and gridded soil samples. Typically, grain yield is measured almost continuously (readings every 1 to 2 seconds) throughout a field resulting in very dense maps (Stafford et al., 1996), whereas soil parameters are often measured at very low sampling densities (Wollenhaupt et al., 1997). To create GIS overlays, the sparsely sampled soil data are often interpolated to a denser grid to match the resolution of the crop yield data. If one proceeds to model the relationship between yield and the interpolated soil variables, the soil data do not represent true sampled values but estimated values.

The questions whether and to what extent interpolation errors influence the modeling process are easily tested with generated data. Since stochastic simulations must be repeated to provide conclusive results (Journel, 1996; Deutch and Journel, 1998), we generated ten fields, each with five features representing soil N, P, and K and land slope and profile curvature. The semivariogram ranges of the features roughly corresponded to ranges obtained from a real-world data set (Hess and Hoskinson, 1996). The ranges for the soil fertility features (N, P and K) were all set to 200m. All features were approximately normally distributed. The semivariograms were estimated using the robust method with L=20 and $D_h / (h_l - h_{l-1})$=25% and approximated by minimizing the sum represented by equation (5), using 400 randomly chosen points for semivariogram estimation. All data were generated on a 10m grid and the size of each field was 800m $\times$800m. Yield was generated using a linear plateau model, whose parameters were set

using fertilizer recommendation guides (Brown, 1982) and regression results from a real-world data set (Hess and Hoskinson, 1996).

The generated data allowed us to compare the accuracy of the yield prediction obtained from using all the original data at the 10 $\times$10m^2 grid to that obtained from using sampled data interpolated back to the same grid. The soil fertility features were sampled at densities of 1 per 50 $\times$50, 100 $\times$100, 150 $\times$150 and 200 $\times$200m^2, which corresponded to sampling intervals of ¼, ½ , ¾ and 1 times the range of the features, and then interpolated to a 10 $\times$10m^2 grid using ordinary kriging (Chilès and Delfiner, 1999). For each sampling interval, we report average interpolation accuracy on the ten simulated fields. Interpolation accuracy was assessed using coefficient of determination R^2 (Devore, 1995), which is related to the normalized mean squared error ε defined in Section III, as R^2=1-ε.

We applied a feed-forward neural network model to predict yield as the response to all five simulated features. Feed-forward neural neural networks are highly non-linear, adaptive systems of sigmoidal functions (Haykin, 1999). By learning from examples, and using gradient-based coefficients estimation, these models can in theory provide an arbitrary close approximation of any non-linear step-wise continuous response function on a compact domain (Haykin, 1999). Previous agricultural modeling applications of neural networks include McBratney and Pringle's (1997) determination of relationships among features and the estimation of a Bermudagrass hay fertilizer response function by Joerding et al. (1994).

Yield prediction accuracy was measured using the coefficient of determination R^2 to quantify the part of a yield variance explained by a prediction model. Due to the stochastic character of the neural network training, we repeated experiments ten times on each of the ten simulated fields and as a final result reported averaged prediction accuracy for these 100 experiments. In each experiment, a simulated field was partitioned into two spatially disjoint subfields. A prediction model was trained on the first subfield and tested on the second subfield and the process was repeated using the second subfield as a training and the first as a test set. To compensate for the influence of the partitioning, prediction accuracy obtained through these two evaluations was averaged and used as the prediction accuracy of the experiment.

When trained on non-interpolated data, the neural network was in average able to explain 81% of the yield variability in the testing fields. However, interpolation errors, even those that occurred when the

soil features were sampled at a spacing equal to ¼ of their geostatistical range, decreased yield prediction accuracy (Fig. 8). In fact, yield prediction accuracy and interpolation accuracy were highly linearly related (correlation coefficient $r = 0.98$).

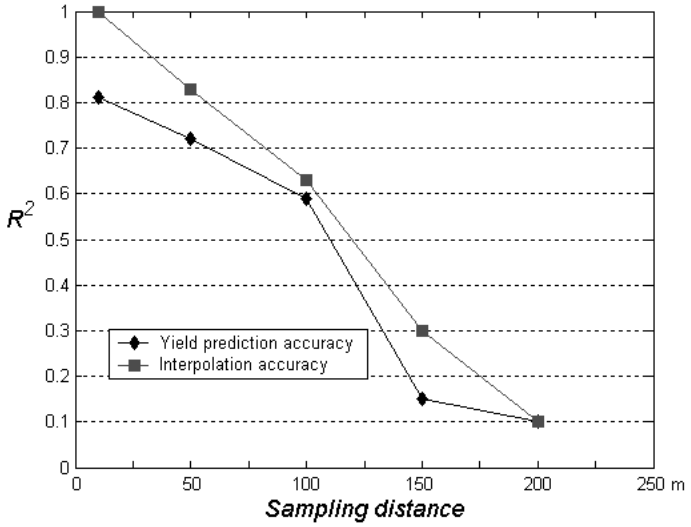


Figure 8. Average interpolation and yield prediction accuracy vs. sampling distance of simulated data.

This dependence could lead to new methods to predict the accuracy of yield prediction based on the estimation of interpolation accuracy (the latter could be performed using standard techniques for kriging error estimation). Conversely, it could be possible to determine the necessary sampling density to obtain a desired level of yield prediction accuracy. While a further range of situations need be tested, these results indicate that unless data are sampled at very high densities relative to their geostatistical properties, one should not attempt to build highly accurate regression functions using interpolated data.

V Discussion

In this paper, we employ unconditioned sequential simulation to produce spatial features values and generate the response (e.g. crop yield) as a deterministic function of attributes combined with random noise to model unexplained variance and sensor error. The current implementation of the simulator produces spatial data on uniform sampling grids (each location equidistant from all adjacent sampling

points) but the proposed sequential simulation technique can also generate samples from non-regular grids (Goovaerts, 1997).

In the proposed simulation technique, we are applying a variant of the sequential simulation algorithm where simulated feature values are not conditioned on real samples. Due to the theoretical reasons cited in Deutsch and Journel (1998), we have applied simple kriging with zero mean to compute the parameters of the standardized Gaussian distribution from which feature values are sampled. Since the application of simple kriging requires that the simulated spatial variable is second-order stationary, a spatial covariance function must be associated with the employed theoretical semivariograms $\gamma(\boldsymbol{h})$ (Chilès and Delfiner, 1999). This condition is automatically satisfied for isotropic exponential and spherical variograms with possible nugget effects (Chilès and Delfiner, 1999) as well as for anisotropic exponential and spherical variograms when anisotropy is determined by a *non-singular* transformation matrix (D.Pokrajac, unpublished work). The proposed method for feature generation can be easily modified to take into account existing feature values, using a conditioned sequential simulation technique described in Deutsch and Journel (1998). Although this conditioned framework cannot exactly match the actual response of a real-life field due to the inevitable imperfections of adopted deterministic response functions, the conditioned simulation technique can be helpful in the evaluation of field-specific knowledge discovery issues, in addition to the "global" assessment that can be achieved by the implemented unconditioned method. Also, by using a conditioning technique similar to that described in Chilès and Delfiner (1999), potential problems due to the use of smoothed (kriged) values as the basis of the Gaussian simulation may be compensated.

To simulate realistic attributes, the user must provide the data generator with proper simulation parameters. If these parameters are to be derived from real world features that are most likely correlated, users should first estimate the feature covariance matrix and de-correlate attributes. Then, histogram estimation and subsequent normal score transformation should be performed to produce independent Gaussian features for semivariogram estimation and fitting. This complex transformation and parameter estimation process is sensitive to the number of the data samples and their spatial resolution. The simulated data may have a limited resemblance to real-world data if simulation parameters were estimated from comparatively small real-life data sets.

For real-world features that have non-Gaussian probability distributions such as many types of soils data (Cambardella et al., 1994; Chien et al., 1997), the proposed simulator can generate features with probability distributions matching specified histograms or cumulative density functions. However, the applied sequential simulation technique does not allow for a significant spatial correlation of extremely large or small feature values (Goovaerts, 1997). For applications where a proper spatial correlation of such extreme feature values is important, sequential indicator simulation could be implemented (Goovaerts, 1997; Deutsch and Journel, 1998).

The data simulator can generate mutually correlated features using data transformations similar to ones proposed by Luster (1985) and Desbarats (1996). Such approaches are useful in the evaluation of knowledge discovery methods that exploit inter-attribute correlation. However, due to the linear character of the proposed transformation there is no guarantee that the non-linear properties of the multivariate attribute distributions of the simulated data (such as regions of homogeneity in feature space) actually reproduce those of a real-world data pattern. Further, there is no control on the reproduction of the spatial cross-correlations at non-zero lags (Goovaerts, 1997). When the aim is to meticulously mimic the multivariate distribution of real-world spatially-correlated data, more flexible but also more time-consuming simulation methods for joint simulation of multiple variables, such as simulated annealing (Goovaerts, 1997), ought to be considered instead of the proposed technique.

The simulator's feature generation method assumes that independent features are stationary in space and thus do not exhibit spatial trend in the sample means (Chilès and Delfiner, 1999). However, the studies on real-world data (Chung et al., 2001) suggest that the assumption of no trend for soil features (such as N,P,K) may be unrealistic. In addition, trend can be useful for modeling topographic features such as elevation (Wollenhaupt et al., 1997). Our simulation method can be easily modified to deal with trended features by introducing the following modifications of the sequential simulation algorithm. Generate the feature value at the first grid location $\boldsymbol{s}^{(0)}$ of the visiting sequence from a Gaussian process with mean determined by the coordinates of $\boldsymbol{s}^{(0)}$. Determine the mean of the Gaussian distribution at subsequent grid points using universal kriging (Chilès and Delfiner, 1999) with a proper trend function (usually a low degree polynomial function of the sample location coordinates), while the distribution variance is to be determined using simple kriging (Goovaerts, 1997).

The simulator provides deterministic linear, loglinear and plateau models to simulate the response dependence on driving attributes. The choice of a particular model may depend on the aimed characteristics of the generated data and the specific purpose of the simulation, such as exploration of a specific knowledge discovery algorithm or sampling strategy. Response model coefficients can be determined from real-world data using parametric (Devore, 1995) or non-parametric (Haykin, 1999) regression techniques. In addition, model parameters can be extracted from fertilizer recommendation guides. We should point out that the model specification process can also substantially benefit from the experience of a skilled agricultural practitioner.

The proposed simulator has a capability of modeling heterogeneous fields, where the yield response to features varies over space, or varies due to differences in feature values. To accomplish this, the user should provide the simulator with rules for model assignment. When the response function varies over space, in addition to a stochastic method, the model assignment can be performed according to pre-specified spatial subfield boundaries. When the response function depends on feature values, the user should provide applicable feature space regions for each model. Such regions can be obtained using algorithms for learning through competition (Avnimelech and Intrator, 1999) that gradually introduce new regression models to explain samples on which already trained regressors have a small prediction accuracy. In addition, when exploration of real-world data indicates the existence of homogeneous regions in feature space (e.g. manifested by distinct clouds in scatter plots), boundaries of such regions obtained using unsupervised machine learning algorithms (Lloyd, 1982;Sander et al., 1998; Lazarevic et al., 2000) can be used for model assignment.

The simulator uses a multiplicative noise model for sensor errors. This type of noise model is commonly used for sensor errors in remote sensing applications (Jain, 1989). However, due to the modular structure of our simulator, other sensor error models can be easily implemented. For instance, additive sensor error may be modeled similarly as unexplained variance. The current simulator could be significantly improved by implementing appropriate models for systematic errors in yield mapping that have known sources such as variable crop width entering the header of a combine, time lag of grain flow through threshing mechanisms and inaccuracy of global positioning systems (Blackmore and Marshall, 1996).

The sample interpolation error is modeled by emulating the actual sampling process and using ordinary kriging and inverse distance interpolation to produce interpolated values. However, we have not compared the implemented interpolation techniques on simulated potentially non-Gaussian features; see Warrick et al. (1988) and Gotway et al. (1996) for related sampling and interpolation studies based on real-world soil data.

Finally, we should point out that due to the stochastic character of applied geostatistical simulation technique, the user should not be satisfied with a single realization of a simulated field (Journel, 1996; Deutch and Journel, 1998). Since the realizations of simulated fields are random, a correct evaluation of knowledge discovery methods should involve fields from repeated simulations to establish the confidence in obtained results.

VI Conclusions

As evidenced in the provided experiment, a spatial data generator can facilitate many experiments regarding sampling and analysis of spatial data. Unlike "real" data, the data generator provides the ability to control data properties such as the spatial characteristics and the amount of noise. Furthermore, prediction accuracy can be easily assessed because the true answer (the value of a feature or the response variable) is known at every location. Finally, the modular nature of the simulation program allows easy implementation of new functionality needed to generate data scenarios that are beyond the original design scope.

References

1. R. Avnimelech and N. Intrator, "Boosting Regression Estimators," Neural Computation, **11**, 491-513 (1999).
2. S.D. Bay, The UCI KDD Archive [http://kdd.ics.uci.edu] (Department of Information and Computer Science, University of California, Irvine, CA, 1999).
3. E. Ben-Dor and A. Banin, "Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties", Soil Science, **159** (4), 364-372 (1995).
4. E. Bertino, B. Chin Ooi, R. Sacks-Davis and K.L.Tan, Indexing Techniques for Advanced Database Systems, (Kluwer Academic Publishers, Dordrecht, the Netherlands 1997).
5. M.F.O. Bierkens, P.A. Burrough, "The indicator approach to categorical soil data. Parts I and II," J. Soil Sci., **44** (2), 361-381 (1993).
6. S. Blackmore and C. Marshall, "Yield mapping, Errors and Algorithms", In Precission Agriculture, in Proc. of the 3rd Int. Conf. on Precision Agriculture, edited by P.C. Robert, R.H. Rust, and W.E. Larson (ASA, CSSA, SSSA, Madison, WI, 1996), pp. 403-415.
7. C.L. Blake and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. (Department of Information and Computer Science, University of California, Irvine, CA, 1998).

8. B. Brown, Idaho fertilizer guide-irrigated wheat, Current Information Series (CIS) No 373 (College of agriculture, Cooperative extension service, Agriculture experimental station, University of Idaho, Moscow, ID, 1982).
9. C.A. Cambardella, T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, R.F. Turco, and A.E. Konopka, "Field-scale variability of soil properties in central Iowa soils," Soil. Sci. Soc. Am. J., **58**, 1501-1511 (1994).
10. Y-J. Chien, D-Y. Lee, H-Y. Guo, and K-H. Houng, "Geostatistical analysis of soil properties of mid-west Taiwan soils," Soil Science, **162** (4), 291-298 (1997).
11. J-P. Chilès, P.Delfiner, Geostatistics-Modeling Spatial Uncertainty (John Wiley & Sons, New York 1999).
12. S. Chung, J. Sung, K.A. Sudduth, S.T. Drummond, and B. Hyun, "Spatial variability of yield, chlorophyll content, and soil properties in a Korean rice paddy field," In Proc. of the 5th Int. Conf. on Precision Agriculture, edited by P.C. Robert et al. (ASA, CSSA, SSSA, Madison, WI, 2001), in press.
13. J.E. Corá, F.J. Pierce, B. Basso, and J.T. Ritchie, "Simulation of within field variability of corn yield with Ceres-Maize model," In Proc. of the 4th Int. Conf. on Precision Agriculture, edited by P.C. Robert et al. (ASA, CSSA, and SSSA, Madison WI, 1999), pp.1309-1319.
14. N. Cressie, "Fitting variogram models by weighted least squares," Mathematical Geology, **17** (5), 563-586 (1985).
15. N. Cressie, Statistics for Spatial Data (John Wiley & Sons, New York 1993).
16. J.A. Desbarats, "Modeling spatial variability using geostatistical simulation," in Geostatistics for Environmental and Geotechnical Applications, ASTM STP 1283, edited by R.M.Srivastava et al. (American Society for testing and materials, West Conshohocken, PA, 1996), pp. 32-48.
17. C.V. Deutsch and A.G. Journel, GSLIB: Geostatistical Software Library and Users Guide, 2nd edn. (Oxford University Press, New York, 1998).
18. J.L.Devore, Probability and Statistics for Engineering and the Sciences, 4th edn. (International Thomson Publishing Company, Belmont, CA, 1995).
19. J.A Doolittle, K.A. Sudduth, N.R. Kitchen, and S.J. Indorante, "Estimating depths to claypans using electromagnetic induction methods," J. Soil Water Cons. **49**(6),572-575 (1994).
20. R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification and Scene Analysis: Pattern Classification, 2nd edn. (John Wiley & Sons, New York, 2000).
21. E. Englund, "Spatial simulation: environmental applications," In Environmental Modeling with GIS, edited by M.F. Goodchild, B.Q. Parks, end L.T. Steayaert, (Oxford Press, New York, 1993), pp. 432-437.
22. B. Flury, A First Course in Multivariate Statistics (Springer, New York, 1997).
23. C.W. Fraisse, K.A. Sudduth, and N.R. Kitchen. 1999. Evaluation of crop models to simulate site-specific crop development and yield," In Proc. of the 4th Int. Conf. on Precision Agriculture, edited by P.C. Robert et al. (ASA, CSSA, and SSSA, Madison WI, 1999), pp. 1297-1308.
24. G.H. Golub and C.F. van Loan, Matrix Computations, 2nd edn. (The John Hopkins University Press,Baltimore, MD, 1989).
25. P. Goovaerts, Geostatistics for Natural Resources Evaluation (Oxford University Press, New York, 1997).
26. C.A. Gotway, R.B. Ferguson, G.W. Hergert, and T.A. Peterson, "Comparison of kriging and inverse-distance methods for mapping soil parameters," Soil Sci. Am. J. **60** (4),1237-1247 (1996).
27. M.Halvorson, Microsoft Visual Basic 6.0 Professional Step-By-Step (Microsoft Press, Redmond, WA,1998).
28. D. C. Hanselman and B. C. Littlefield, Mastering MATLAB 5: A Comprehensive Tutorial and Reference (Matlab Curriculum Series) (Prentice Hall Englewood Cliffs, NJ, 1997).
29. D.M. Hawkins and N. Cressie," Robust kriging-a proposal," Journal of the International Association for Mathematical Geology, **16**(1), 3-18 (1984).
30. S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd edn. (Prentice Hall, Englewood Cliffs, NJ, 1999).
31. J.R. Hess and R.L. Hoskinson, "Methods for characterization and analysis of spatial and temporal variability for researching and managing integrated farming systems," In Proc. of the 3rd Int. Conf. on Precision Agriculture, edited by P.C. Robert, R.H. Rust, and W.E. Larson (ASA, CSSA, SSSA, Madison, WI, 1996), pp. 641 – 650.
32. I. Horton, Beginning Visual C++ 6 (Wrox Press, Birmingham, UK, 1998).

33. A.K. Jain, Fundamentals of Digital Image Processing (Prentice Hall, Englewood Cliffs, NJ, 1989).
34. W.H. Joerding, Y. Li, and D.L. Young, “Feedforward neural network estimation of a crop yield response function,” J. Agricultural and Applied Economics, **26**(1), 252-263 (1994).
35. A.G. Journel, “Modeling uncertainty and spatial dependence: Stochastic imaging,” Int. J. Geographical Information System, **10**(5), 517-522 (1996).
36. A. Kravchenko and D.G. Bullock, “A comparative study of interpolation methods for mapping soil properties,” Agron. J. **91**(3), 393-400 (1999).
37. A. Lazarevic, D. Pokrajac, and Z. Obradovic, "Distributed Clustering and Local Regression for Knowledge Discovery in Multiple Spatial Databases", Proc. European Symp. Artificial Neural Networks, (2000).
38. S.P. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Information Theory, **28** (2,Part 1), 129-137 (1982).
39. G.R. Luster, Raw materials for Portland cement: applications of conditional simulation of coregionalization, PhD thesis (Stanford University, Stanford, CA, 1985).
40. A.B. McBratney and M.J. Pringle (1997). Spatial variability in soil - implications for precision agriculture. In Precision Agriculture, edited by J.V. Stafford (Bios, Oxford, England, 1997), pp 3-32.
41. G.V. Middleton, Data Analysis in the Earth Sciences using MATLAB (Prentice Hall, Upper Saddle Hill, NJ, 2000).
42. M.S. Moran, Y. Inoue, and E.M. Barnes, “Opportunities and limitations for image-based remote sensing in precision crop management,” Remote Sensing Environ. **61**(3),319-346 (1997).
43. Y. Pachepsky and B. Acock, “Stochastic imaging of soil parameters to assess variability and uncertainty of crop yield estimates, “ Geoderma **8**, 213-229 (1998).
44. J.O. Paz, W D. Batchelor, T.S. Colvin, S.D. Logsdon, T.C. Kaspar, D.L. Karlen, B.A. Babcock, and G.R. Pautsch, “Model-based technique to determine variable rate nitrogen for corn,” In Proc. of the 4th Int. Conf. on Precision Agriculture, edited by P.C. Robert et al. (ASA, CSSA, and SSSA, Madison WI, 1999), pp. 1279-1289.
45. D.Pokrajac, Z. Obradovic, and T. Fiez, “Understanding the influence of noise, sampling density and data distribution on spatial prediction quality through the use of simulated data,” In Proc. 14th European Simulation Multiconference (ESM), (2000).
46. R. Rossi, P. Borth, and J. Toleerfson, “Stochastic simulation for characterizing ecological spatial patterns and appraising risk,” Ecological Applications, **3** (4), 719-735 (1993).
47. D. Russo and W. Jury, “Effect of the sampling network on estimates of the covariance function of stationary fields, “ Soil Sci. Soc. Amer. J., **52** (5), 1228-1234 (1988).
48. J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications,” Data Mining and Knowledge Discovery, **2** (2), 169-194 (1998).
49. S.P.Saunders, G. Larscheid, B.S.Blackmore, and J.V.Stafford, “A method for direct comparison of differential global positioning system suitable for precision farming,” In Proc. of the 3rd Int. Conf. on Precision Agriculture, edited by P.C. Robert, R.H. Rust, and W.E. Larson (ASA, CSSA, SSSA, Madison, WI, 1996), pp.663-680.
50. J.V. Stafford, B. Ambler, R.M. Lark, and J. Catt, “Mapping and interpreting the yield variation in cereal crops,” Computers and Electronics in Agriculture, **14** (2/3), 101-119 (1996).
51. S. Vucetic, T. Fiez, and Z. Obradovic, “Analyzing the influence of data aggregation and sampling density on spatial estimation,” Water Resources Research, **36** (12), 3721-3731 (2000).
52. A.W. Warrick, R. Zhang, M.K. El-Harris, and D.E. Myers, “Direct comparisons between kriging and other interpolators”, In Validation of Flow and Transport Models for the Unsaturated Zone, edited by P.J. Wierenga and D. Bachelet (New Mexico State Univ., Las Cruses, NM, 1988), pp. 505-515.
53. R. Webster and M.A.Oliver, Statistical Methods in Soil and Land Resource Survey (Oxford University Press, Oxford [England], New York, 1990).
54. R. Webster and M.A.Oliver, ”Sample adequacy to estimate variograms of soil properties, “ J. Soil Sci., **43** (1), 177-192 (1992).
55. N.C. Wollenhaupt, D.J. Mulla, and C.A. Gotway, “Soil sampling and interpolation techniques for mapping spatial variability of soil properties,“ In The State of Site Specific Management for Agriculture, edited by F.J. Pierce and E.J. Sadler (ASA/CSSA/SSSA, Madison, WI, 1997), pp. 19-54.

Appendix: Basic matrix algebra (Golub and van Loan 1989; Flury, 1997)

A. 1 Elementary definitions

A matrix is said to be square if it has the same number of rows and columns.

Matrix **B** with n×n elements is upper triangular if all its elements below the main diagonal $b_{11},\ldots,b_{nn}$ are equal to zero.

A diagonal matrix may have non-zero elements only on its diagonal. A diagonal matrix is the identity matrix if all diagonal elements are equal to 1.

Matrix **C** is the transposed matrix of matrix **B**, if rows of **C** correspond to columns of **B** and vice versa and we write

$$\mathbf{C} = trans(\mathbf{B}) \tag{A.1}$$

A matrix equal to its own transposed matrix is symmetric.

A. 2 Eigenvalues and eigenvectors

Eigenvalues *e* and eigenvectors ***v*** of a square matrix **B** are solutions of the following matrix equation:

$$\mathbf{B}\times\boldsymbol{v} = e\times\boldsymbol{v} \tag{A.2}$$

Matrix **B** with n×n elements has n not necessary different eigenvalues and eigenvectors.

Matrix **V,** containing eigenvectors $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ in its columns such that $\mathbf{V}=[\boldsymbol{v}_1\ldots\boldsymbol{v}_n]$, is called an eigenmatrix.

A matrix is positive definite if all its eigenvalues are strictly positive.

A matrix is non-singular if and only if it does not have eigenvalues equal to zero. For a non-singular matrix **B**, its inverse matrix inv (**B**) is defined such that inv (**B**)×**B** is the identity matrix.

A. 3 Covariance matrix

Let **A** be a n×F matrix that contains n samples of F features $x_1,\ldots,x_F$. The sample covariance matrix $\mathbf{C_A}$ is defined as:

$$\mathbf{C}_\mathbf{A} = \frac{1}{n-1} trans\left(\mathbf{A}-\overline{\mathbf{X}}\right)\times\left(\mathbf{A}-\overline{\mathbf{X}}\right) \tag{A.3}$$

where $\overline{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_F \\ \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_F \\ \ldots & \ldots & \ldots & \ldots \\ \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_F \end{bmatrix}$ is a n×F matrix of repeated estimated feature means $\bar{x}_1, \ldots, \bar{x}_F$.

Assuming that the variance of each feature is positive, the sample covariance is always a positive definite matrix. The sample covariance matrix $\mathbf{C_A}$ is diagonal when features in **A** are not correlated.

The multiplication of features matrix **A** with eigenmatrix **V** of its covariance matrix $\mathbf{C}_A$,

$$\mathbf{A}^* = \mathbf{A} \times \mathbf{V} \tag{A.4}$$

decorrelates features since it can be shown that the sample covariance matrix $\mathbf{C_{A^*}}$ of features $\mathbf{A}^*$ obtained by this "eigenvalue transformation" is diagonal.

A. 4 Cholesky transformation

The Cholesky transformation chol (**B**) of a square symmetric positive definite matrix **B,** is an upper triangular matrix **T** such that

$$\mathbf{B} = trans(\mathbf{T}) \times \mathbf{T} \tag{A.5}$$

If **A** is a matrix that contains n samples of F independent features with unit variance (and hence with an identity covariance matrix), it can be shown that the matrix of transformed features **A'** obtained by the multiplication of **A** with matrix **T**=chol (**B**),

$$\mathbf{A}' = \mathbf{A} \times \mathbf{T} \tag{A.6}$$

has a covariance matrix specified by **B**.

List of footnotes:

* Partial support by the INEEL University Research Consortium project No.C94-175936 to T. Fiez and Z. Obradovic is gratefully acknowledged. The authors express their gratitude to the anonymous reviewers whose comments significantly improved the simulation method and overall quality of the paper.

All correspondences to: Z. Obradovic, zoran@ist.temple.edu, phone: +(215) 204 – 6265, fax: +(215) 204 - 5082

List of figure captions:

*Figure 1***.** Plateau models for yield generation.
a) Linear plateau model.
b) Exponential plateau model.

Figure 2. Feature generation input screens.
a) Variogram estimation and grid parameters.
b) Feature parameters.
c) Model variogram parameters.

Figure 3. Screen for the interactive feature generation process showing the current parameters of the generated feature.

Figure 4. Software screen for tuning plateau model parameters.

Figure 5. Plot of the assignment (denoted by symbols) of one model in a multi-model simulation in x- and y-coordinate space.

Figure 6. Generated yield.
a) Before addition of unexplained variance.
b) With 50% unexplained variance.

Figure 7. Influence of sampling density on interpolation error.
a) Original feature on a $10 \times 10 m^2$ grid.
b) Interpolated feature using samples from a $40 \times 40 m^2$ grid, $\varepsilon = 0.12$.
c) Interpolated feature using samples from a $80 \times 80 m^2$ grid, $\varepsilon = 0.25$.
d) Interpolated feature using samples from a $200 \times 200 m^2$ grid, $\varepsilon = 0.80$.

Figure 8. Average interpolation and yield prediction accuracy vs. sampling distance of simulated data.

List of keywords

Precision agriculture, spatial data simulation, soil sampling, yield prediction