A TOOL FOR CONTROLLED KNOWLEDGE DISCOVERY IN SPATIAL DOMAINS*

Dragoljub Pokrajac
School of Electrical Engineering and Computer Science, Washington State University, Pullman WA 99164 USA
E-mail: dpokraja@eecs.wsu.edu

Zoran Obradovic
School of Electrical Engineering and Computer Science, Washington State University, Pullman WA 99164 USA
E-mail: zoran@eecs.wsu.edu

Tim Fiez
Department of Crop and Soil Sciences, Washington State University, Pullman WA 99164 USA
E-mail: tfiez@wsu.edu

KEYWORDS
Decision-support systems, AI-supported simulation, Stochastic, Statistical analysis, Agriculture

ABSTRACT

A data simulator that can facilitate the development of improved sampling and analysis procedures for spatial analysis is proposed. The simulator, implemented in MATLAB, provides a graphical user interface and allows users to generate data layers satisfying given spatial properties and a response variable dependent upon user specified functions. It has a modular structure and is capable of modeling response function heterogeneity (both in spatial coordinates and in driving attribute space) as well as unexplained variance, sensor error, spatial data sampling and interpolation. As an illustration of the potential uses of the simulator in precision agriculture, the effect of sampling density and interpolation on neural network prediction of crop yield was assessed.

INTRODUCTION

The advent of global positioning systems and new sensors has enabled the collection of large volumes of spatial data. However, this ability has given rise to many questions regarding both how to best collect 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, the variability in data sets, data collection procedures, and analysis approaches make it difficult to compare results from separate experiments and will likely slow the discovery of optimal procedures.

As spatial data have become more available, it has become easier to test knowledge discovery procedures on multiple data sets allowing the formation of more general statements. However, 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. Unfortunately, unlike in machine learning and knowledge discovery in some other domains (Bay 1999), (Blake and Mertz 1998), standardized spatial data repositories do not exist.

Another solution to the problem of data availability is to use data simulators. Not only do they provide the ability to create unlimited amounts of data, but sophisticated simulators can allow the user to control the complexity of the data, add known amounts of noise, and to generate response variables dependent upon simulated driving attributes. To test regression procedures, a simulator must provide a way to formulate a response variable based on a set of driving attributes. While mechanistic models (Corá et al. 1999) can simulate response variables that reflect real world variability, they 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. On the other hand, a data simulator could simply use deterministic mathematical functions to compute a response variable for a given input by combining individual effects of multiple driving attributes.

Using an extension of the geostatistical simulation techniques of (Desbrats 1996) and deterministic response functions, we have developed a spatial data simulator that can generate spatial layers satisfying given statistical properties and a response variable dependent upon user--controlled equations.

*Partial support by the INEEL University Research Consortium project No.C94-175936 to T. Fiez and Z. Obradovic is gratefully acknowledged.

Our intention is to provide a user-friendly yet powerful way to simulate different aspects of data generation, acquisition and processing. The simulator can be used to explore in a controllable way the effects of sampling density, sensor error, presence or absence of particular driving attributes, and layer heterogeneity on sampling procedures, data acquisition systems, and KDD algorithms (Pokrajac at al. 2000). Furthermore, by using this simulator, 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 simulator, it is possible to gradually increase the complexity of generated data to simulate prospective emerging aspects of technology and practice. The following sections discuss the method of data generation, the software implementation, and finally give an example of using simulated data for investigating interpolation effects on neural network yield prediction in precision agriculture.

METHOD

The data simulation process consists of two steps: driving attribute generation and response variable simulation.

Generation of Driving Attributes

Driving attributes are generated through a multistep process of 1) grid determination, 2) generation of independent attributes, 3) attribute correlation and cluster generation and 4) attribute verification.

Spatial Grid Determination

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

Generation of Independent Attributes

Each driving attribute $f_j(\mathbf{s})$ is considered as a random function of vectors $\mathbf{s}_i$. Driving attributes can be generated to have Gaussian or arbitrary distributions (Devore 1995). The attribute simulation procedure is similar to (Desbrats 1996). Attribute values are established such that the spatial statistic properties of the simulated attribute are described by the chosen theoretic isotropic semivariogram $\gamma(h)$, specified by type (exponential or spherical) and parameters (nugget, c_0, sill, c_e, and range a_s) (Cressie 1993).

The first step in generating $f(\mathbf{s})$ is to generate an attribute $f^*(\mathbf{s})$ satisfying the normalized semivariogram $\gamma^*(h)$:

$$\gamma^*(h) = \begin{cases} \gamma(h), & h = 0 \\ (\gamma(h) - c_o)/c_e, & h > 0 \end{cases} \quad (1)$$

Generation of $f^*(\mathbf{s})$ takes two steps: first, a set of seed points $S' \subset S$ is selected and for each seed point, an attribute value is simulated to satisfy the theoretical semivariogram $\gamma^*(h)$. Second, attribute values for the remaining points in S are obtained from the values of S' by spatial interpolation. Assuming the second-order stationarity is satisfied (Cressie 1993), attribute values at the seed points are a sample of a n' dimensional random vector with zero mean and a covariance matrix

$$R = (r_{ij})_{i,j=1,\ldots,n'} = \begin{cases} 1, & \text{if } i = j \\ 1 - \gamma^*(h_{i,j}), & \text{otherwise} \end{cases} \quad (2)$$

where $h_{i,j}$ is a distance between two distinct seed points.

Once attribute values are established for the seed points, attribute values for the rest of the data grid are determined through spatial interpolation using ordinary kriging (Cressie 1993). To incorporate the uncertainty of the kriged attribute values into the data layers, the interpolated value at each point in the data grid is substituted with a sample from a Gaussian variable having mean and variance equal to that predicted by kriging (Desbrats 1996). If the simulated driving attribute is not Gaussian, but instead has an arbitrary distribution, attribute values can be obtained using the distribution transformation approach (Desbrats 1996).

Finally, the data are transformed from $f^*(\mathbf{s})$ to the driving attribute $f(\mathbf{s})$ with required mean μ, sill c_e and nugget c_0:

$$f(\mathbf{s}) = f(\mathbf{s})^* \cdot \sqrt{c_e - c_0} + N(\mu, c_e) \tag{3}$$

where $N(\mu,c_e)$ is a Gaussian variable having mean μ and variance c_e.

Attribute Correlation and Cluster Generation

All driving attributes are simulated independently, but due to spatial correlation, they may be correlated to each other. In order to obtain attributes with a given correlation matrix C_V (which can be estimated as the sample covariance matrix of real-world data), attributes are de-correlated using eigenvalue projection (Dudal at al. 1999) and Cholesky transformation of C_V (Flury 1997).

The last phase of driving attribute simulation is the formation of clusters in attribute space. Using a subset Φ of the simulated driving attributes, a set of "cluster seeds", consisting of M points in Φ is chosen. Then, each point in the simulated data set is moved towards the nearest cluster seed. The intensity of the shift is proportional to the distance to the seed point and can be adjusted to control cluster aggregation. To avoid unnaturally clear separation between clusters, "perturbation" noise with variance proportional to that of the attributes in Φ is added to the clustered values. Finally, the driving attributes in Φ are renormalized to have specified means and variances. Since the clusters are used to assign multiple models during the simulation process of a response variable (see next section), a cluster label c_i is maintained for each point $\mathbf{s}_i$ in the simulated data grid.

Attribute Verification.

The final step in the attribute generation process is to verify that simulated driving attributes match user-specified spatial properties by calculating estimated semivariograms and fitting model semivariograms.

To compute estimated semivariograms, the relevant distance range is divided into equally spaced lags and for each lag, all pairs of points whose distances are within the lag are used to calculate an average squared difference of function values using the method of moments or robust estimation (Cressie 1993).

To fit a theoretic semivariogram model with the estimated semivariogram, the parameters c_e,c_0,a_s must be determined. This can be done by visual assessment, or by simple or weighted least squares (Cressie 1985).

Simulation of a Response Variable

Response variable $g(.)$ is generated from the driving attribute data using linear and loglinear models that are common in statistical literature, as well as plateau models, common in agriculture. In linear models, $g(.)$ is the weighted sum of the driving attribute values $f_j(\mathbf{s}_i)$. When a loglinear model is applied, the logarithm of the response variable is computed as the weighted sum of the logarithms of the attribute values.

In plateau models, response variable $g(.)$ is generated proportional to the product of the plateau functions $h_j(.)$ for each driving attribute used to generate the response variable:

$$g(s_i) = G\prod_{j=1}^{F} h_j\left(f_j(\mathbf{s}_i)\right) \tag{4}$$

where G is the coefficient of proportionality.
Plateau functions can be linear:

$$h_j(x) = \begin{cases} H_j, & \text{if } x \le 0 \\ H_j + x(1-H_j)/T_{1,j}, & \text{if } 0 < x \le T_{1,j} \\ 1, & \text{if } T_{1,j} < x \le T_{2,j} \\ 1-(x-T_{2,j})\cdot(1-H_j)/(T_{3,j}-T_{2,j}), & \text{if } T_{2,j} < x \le T_{3,j} \\ H_j, & \text{if } x > T_{3,j} \end{cases} \tag{5}$$

or exponential:

$$h_j(x) = \begin{cases} H_j + [1-\exp(-(T_{3,j}-x)/T_{2,j})]\cdot[1-\exp(-x/T_{1,j})]\cdot(1-H_j), & \text{if } 0 < x \le T_{3,j} \\ H_j, & \text{otherwise} \end{cases} \tag{6}$$

In these formulas, $T_{1,j}$, $T_{2,j}$ and $T_{3,j}$ determine the slope and range of the plateau models while H_j determines the floor of the particular attribute's influence on response variable. 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 driving attribute j has its maximum influence on the response variable. Oppositely, if $mod_j = 0$, the response variable does not depend on the driving attribute j, by which we can introduce attribute irrelevancy.

Model coefficients can be determined from actual data or taken from sources of expert knowledge such as fertilizer recommendation guides for applications in agriculture.

Modeling of Heterogeneity

To simulate situations where the influence of driving attributes to the response variable varies over space, or varies due to differences in driving attribute values, the response variable can be modeled such that the value of the response variable at grid point $\mathbf{s}_i$ is the weighted sum of values generated by M models $g_m(.)$:

$$y(s_i) = \sum_{m=1}^{M} w_{m,i}\, g_m(\mathbf{f}(\mathbf{s}_i)), \qquad \text{where} \sum_{m=1}^{M} w_{m,i} = 1 \tag{7}$$

Different models can be assigned to different spatial areas or to different driving attribute clusters. The use of spatial position for model assignment can simulate conditions where variables that contribute to the response variable determination are missing from the data set.

We use the term "hard" generation if only one of M models contributes to the response variable at each point $\mathbf{s}_i$, and the term "smooth" generation when multiple models contribute to the value of the response variable. Smooth generation allows gradual transitions from one model to another.

The determination of $w_{m,j}$ to vary models with spatial location is as follows. For each point $\mathbf{s}_i=(x_i,y_i)$ in S, generate 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 $x_i \pm \Delta x$, $y_i \pm \Delta x$ where Δx is a user-defined parameter. Generate v^* by discretizing the value of v' over S into M equal bins. Compute coefficients for "hard" response variable generation such that:

$$w_{hard,m,i} = \begin{cases} 1, & \text{if } v^*(\mathbf{s}_i) = m \\ 0, & \text{if } v^*(\mathbf{s}_i) \ne m \end{cases} \tag{8}$$

To compute coefficients for "smooth" response variable generation, one should average values of $w_{hard,m,i}$ for all points in rectangle $x_i \pm \Delta x_{soft}$, $y_i \pm \Delta y_{soft}$ (Δx_{soft}, Δy_{soft} are adjustable parameters).

To vary models due to differences in driving attribute values, assign class label c_i to each point $\mathbf{s}_i$ and generate coefficients $w_{hard,m,i}$ such that $w_{hard,m,i} = 1$ if $c_i = m$ and 0 otherwise. The smooth coefficient for point $\mathbf{s}_i$ is generated by averaging $w_{hard,m,i}$ for all points in a specified neighborhood around point $\mathbf{s}_i$ in attribute subspace Φ. Observe that the existence of spatial indexes in subspace Φ can dramatically improve the performance of this algorithm.

Modeling Unexplained Variance, Sensor Error, Sampling and Interpolation

Unexplained variance in the response variable is modeled by adding Gaussian noise with zero mean and variance such that the ratio of explainable response variability (determined by driving attribute values and response functions) to the total response variability is equal to a user specified value. The effect of measurement error on driving attributes and the response variable is modeled as multiplicative Gaussian noise.

Finally, the introduction of error through the interpolation of sampled values is simulated by emulating the actual sampling process. Data are sampled from the generated data grids and used to interpolate values at unsampled locations using inverse-distance or kriging interpolation (Cressie 1993).

IMPLEMENTATION

The spatial data simulator has been implemented in MATLAB (by The MathWorks, Inc.). The software can work in three modes: i) simulation of driving attributes, ii) simulation of response variable, and iii) simulation of unexplained variance and sensor error and the simulation of sampling and interpolation.

In the driving attribute simulation mode, the user is presented with a number of parameter input screens. Grid and semivariogram estimation parameters include layer shape, grid step size, number and positions of seed points, number of points, minimal and maximal lag and bin width for semivariogram estimation (Fig. 1). Driving attribute parameters include the number of driving attributes to be simulated, driving attribute means and variances, and the parameters of their model semivariograms. Also, the user can load parameters from existing files in addition to keyboard entry.

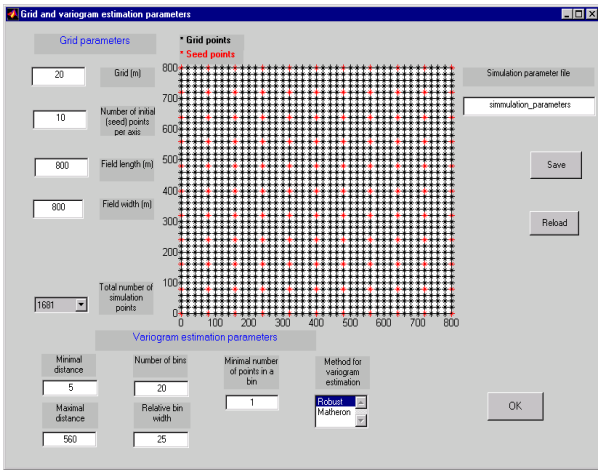


Figure 1: Input screen for grid and variogram estimation Parameters

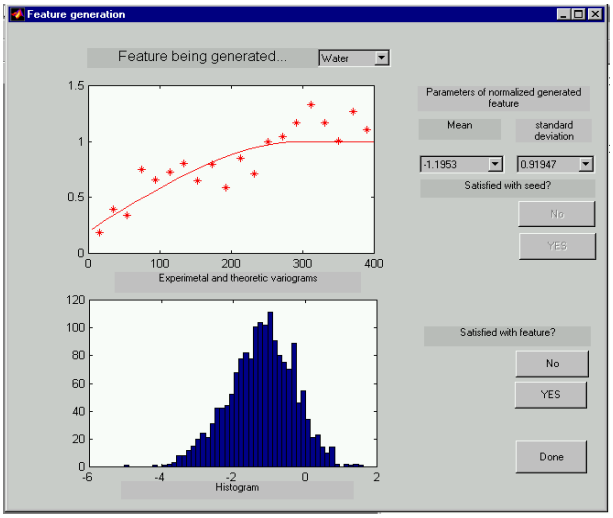
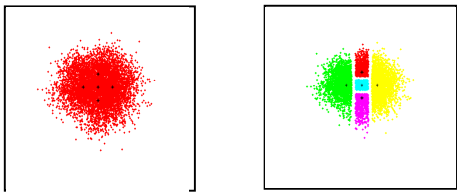


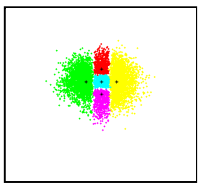
Figure 2: Screen for the interactive driving attribute generation process showing the current parameters of the simulated driving attribute.

Once the parameters have been entered, the program proceeds through an interactive simulation process of attributes. The user is presented with the estimated and model semivariogram and normalized histogram for a driving attribute (Fig. 2). The user can then accept the driving attribute and continue with the next one, or re-generate the current driving attribute if unsatisfied with the generated statistics. After all the driving attributes are generated, the user is prompted for the number of clusters and their seeds. The user interactively controls the clustering process (Fig. 3) by changing the coefficient of proportionality for aggregation and the variance of the attribute "perturbation" noise.



a) Original attributes and cluster seeds

b) Clusters after aggregation



c) Clusters after perturbation noise added

Figure 3: The process of cluster generation in the space of two attributes

In the response variable simulation mode, the user selects the driving attributes upon which the response variable will depend, the number and the type of models to apply (linear, loglinear or plateau). If the user selects more than one model, they must also choose how to assign the different models; models can be assigned to regions in coordinate (x and y) or in attribute space. Next, in an interactive process, the user sets parameters for each model and controls their influence on the simulated response variable. 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 driving attribute $j=1,\ldots,F$, (e.g. nitrogen, potassium…), the user enters parameters $T_{1,m,j}$, $T_{2\ m,j}$, $T_{3,mj}$ (denoted by *T1*, *T2* and *T3* on the screen capture) and $mod_{l,j} = (1-H_{m,j})/(1+H_{m,j})$ (denoted with *mod* on screen) describing the shape of $g_{m,j}$ which is plotted along with the normalized histogram of the driving attribute. By varying G the user can also specify a "ceiling", the maximum allowed value of the response variable.

By inspecting mean, median, standard deviation, minimum and maximum values, the number of outliers (values on which the response variable is not defined or is less than zero), and the normalized histogram for the response variable, the user can tune model parameters to obtain a simulated response matching desired requirements. In the postprocessing mode, unexplained variance and sensor error are added to the response and driving attribute values. 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.

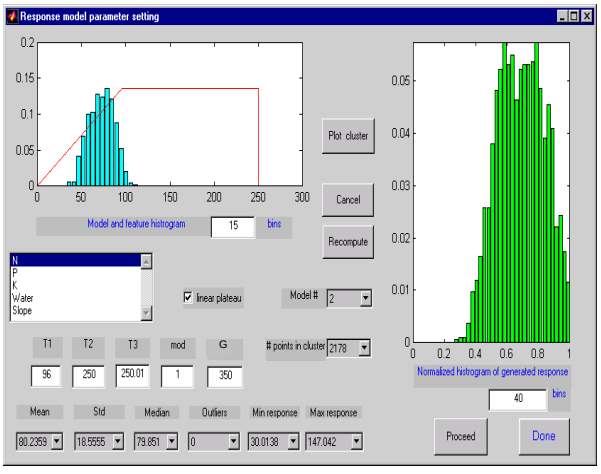


Figure 4: Software screen for tuning plateau model parameters.

When the data simulation process is complete, the simulated driving attributes and response, 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 simulation sessions last 20 minutes to 1 hour. Intermediate results are saved during the simulation process in case program execution is interrupted before the process is complete.

The computational complexity of the data simulation process establishes the practical limits for grid point, seed point, driving attribute, and model numbers. The program has been successfully tested for up to n=15000 grid points, n'=400 seed points, F=10 driving attributes and M=5 models. Since complexity is linear in F,M and n, the critical issue in program execution is the choice of n', since generation of seed points and kriging requires time proportional to the cube of n'. Due to the difficulties of generating driving attributes satisfying specified semivariograms with a small number of points n', a reasonable compromise for the value of n' is necessary.

EXPERIMENTS WITH SIMULATED DATA

Using simulated data, many important spatial data issues in precision agriculture, such as the influence of data parameters (sensor error, unexplained variance, data distribution and heterogeneity) and the type and parameters of regression models on prediction accuracy, can be explored (Pokrajac at al 2000). Here, we will illustrate the application of data generated using the simulator to investigate the interpolation error influence on yield prediction.

We simulated two fields with five driving attributes representing soil and landscape characteristics. Spatial statistic parameters of the driving attributes roughly corresponded to those obtained from a real-world data set (Hess and Hoskinson 1996). Specifically, ranges for the soil fertility driving attributes were all set to 200m. All driving attributes were approximately normally distributed. 256 seed points were used to generate the driving attributes, and the semivariograms were estimated using the robust method and approximated using weighted least squares. All data were generated on a 10*10m^2 grid and the size of each field was 800m*800m. Yield was simulated using a linear plateau homogeneous model, whose parameters were set using fertilizer recommendation guides (Brown 1982) and expert knowledge.

The soil fertility driving attributes were sampled at different densities and then interpolated back to a 10*10m^2 grid. Interpolation accuracy was assessed using coefficient of determination R^2 (Devore 1995). A feed-forward neural network (Haykin 1999) was used to predict yield as the response to all five driving attributes. Yield prediction accuracy was also measured using a R^2 obtained by 2-cross validation (considering each part of a data partition pair both as a training and a test set), and averaging 10 experiment repetitions for each pair of training and test set.

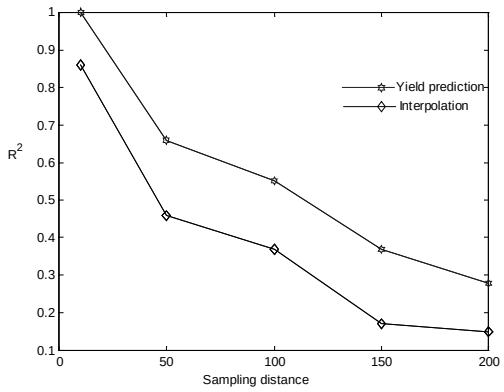


Figure 5. Interpolation and yield prediction accuracy vs. sampling distance of simulated data

When trained on non-interpolated data, the neural network was able to explain 86% of the yield variability in testing fields. However, interpolation errors, even those that occurred when the soil driving attributes were sampled at a spacing equal to ¼ of their geostatistical range (50*50m^2), seriously decreased yield prediction accuracy (Fig. 5). 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 model using interpolated data. Observe also that yield prediction accuracy and interpolation accuracy were highly linearly related (correlation coefficient r= 0.99). This dependence could lead to new methods to predict the accuracy of yield prediction based on the estimation of interpolation accuracy and conversely, to determine the necessary sampling density for obtaining a desired level of yield prediction accuracy.

CONCLUSIONS

A data simulator that can serve as a tool for the development of improved sampling and prediction procedures for spatial data analysis is proposed. Unlike "real" data, using the simulator 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 driving attribute or the response variable) is known at every location. For the simulation of data issues that cannot be currently handled, the modular nature of the simulator will allow new appropriate functions to be easily appended.

REFERENCES

Bay, S.D. 1999. *The UCI KDD Archive [http://kdd.ics.uci.edu]*. Department of Information and Computer Science, University of California, Irvine, CA.

Blake, C.L.; and C.J. Merz, 1998. *UCI Repository of machine learningdatabases[http://www.ics.uci.edu/~mlearn/MLRepository. html]*. Department of Information and Computer Science, University of California, Irvine, CA, 1998.

Brown, B. 1982. *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.

Corá, J.E.; F.J. Pierce; B. Basso; and J.T. Ritchie. 1999. "Simulation of within field variability of corn yield with Ceres-Maize model." In *Proc. of the 4th Int. Conf. on Precision Agriculture*, ASA, CSSA, and SSSA, Madison WI., 1309-1319.

Cressie, N. 1985. "Fitting Variogram models by weighted least squares." *Mathematical geology* 17, no. 5: 563-586.

Cressie, N. 1993. *Statistics for spatial data. Rev. ed*. John Wiley & Sons, inc., New York Chichester Toronto Brisbane Singapore.

Desbrats, J.A. "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, 32-48.

Devore, J.L. 1995. *Probability and statistics for engineering and the sciences. 4th ed*. Int'l Thomson Publishing Company, Belmont.

Dudal R.O., D.G. Stork, P.E. Hart and R.O. Duda. 1999. *Pattern Classification and Scene Analysis: Pattern Classification, 2nd ed.* John Wiley & Sons, New York.

Flury B. *A first course in multivariate statistics*. Springer, New York.

Haykin S. 1999. *Neural Networks: A Comprehensive Foundation, 2nd ed.* Prentice Hall, New Jersey.

Hess, J.R. and R.L. Hoskinson. 1996. "Methods for characterization and analysis of spatial and temporal variability for researching and managing integrated farming systems." In *Precision Agriculture, the Proceedings of the Third International Conference on Precision Agriculture*, ASA, CSSA, SSSA, Madison, WI., 641 – 650.

Pokrajac, D; Z. Obradovic; and T. Fiez. 2000. "Understanding the influence of noise, sampling density and data distribution on spatial prediction quality." In *Proc. of 14. ESM Conference* (in press).