Notes on:

Report on Geostatistics for Spatial Interpolation and Sampling Efficiency in Site Characterisation

William Cadell

What am I writing about?
What is the motivation for this discussion?
What questions am I asking?
Is this process driven or concept driven?
What am I trying to say?
What is the answer?

Introduction

(concept + process)

Geostatistical methods have been accepted as the most representative way of describing the continuous nature of the soil. This report describes the most logical ways of applying the geostatistical methods for common use in studies of the spatial variation of soil properties.

The emphasis will be on applying geostatistics for use with the Land Allocation Decision Support System (LADSS) site characterisation process. Work has been done on this topic previously by Tapping et al *****. They describe a method by which sampling methods can be refined using the Normalised Difference Vegetation Index (NDVI) value derived from multispectral aerial imagery. As part of this work a geostatistical process was also described. The primary need for this is the recognition that within field variability is a major issue in farm based decision support. However, ground based surveys are expensive, so remote sensing has been investigated as a method by which sampling could be made most efficient. This report will focus on the geostatistical methods that will represent within-field variability most efficiently.

This document is structured from both a conceptual and process driven viewpoint. Initially there is a short review of the methods employed in the past with reasons why these are now obsolete. Then there is a conceptual description of the relevant methods in use presently. Finally there is a discussion of the methods directly relevant to the LADSS project.

("Combining Photogrammetric camera and IR videography to define within field soil sampling frameworks" J. Tapping et. al.)
Structure of report

Representation of Spatial variation in Soil

There have been two principle approaches to representing soil variation. Discrete and continuous representations each have their own assumptions, strengths and weaknesses.

According to Heuvelink and Webster (2001) the traditional method of characterising soil properties has been to breakdown the landscape into discrete regions, each to which a class would be assigned. The boundaries would be definite lines across regions where the observations suggested that the greatest change occurred. Inside each region it is assumed that the soil is of a generally homogenous nature. The resultant map would be accompanied by a text describing the classes.

The discrete soil map can be as much to do with the intuition of the soil surveyor as any formal knowledge and understanding of the area. It would involve field observations, lab analysis and aerial interpretation. The observations would be cut to a minimum using the intuition of the surveyors to select representative sites. Sub-class variation would be an acknowledged feature and described qualitatively.

Statistics entered soil classification in the 1960's, this did provide some improvements in the choropleth mapping process, especially if there are close connections between physiography and land cover and aerial photography is available. But on closer inspection this form of soil mapping has some shortcomings such as the delineation of subtle features, which would take a great deal of effort.

Fig Example of Discrete Classification

Laterally the second approach has been to view the soil as an ever-changing medium and represent it as a continuous surface. This method is statistically complicated and computationally very intensive. Kriging has been used for twenty years to describe the continuity in soil characteristics

Fig Example of Continuous Classification

The fundamental idea behind the geostatistical representation of soil is the concept of a *regionalised variable*. Where conventional statistical methods concentrate on the individual, discrete data points. Geostatistical methods look more closely at the differences in value and spatial location of the data points.

It is assumed that, given an adequate population, variables exhibit a degree of continuity within a finite region. It is also a key assumption that the regionalsied variables are subject to a statistically normal distribution.

Geostatistics allows measurements on small volumes of material to represent regions without bias. These samples, called *supports*, can either be small individual sites or they can be enlarged by averaging several measurements together or by taking several small samples and mixing them into a bulked sample.

As mentioned above one of the fundamental aspects of the geostatistical method is the idea that at some scale, the properties in the soil are in some way positively related to each other (autocorrelation). As clarification of this: places close to each other are more similar than places further from each other.

Webster and Oliver 2001 have suggested that geostatistical methods be used for three main reasons: description, interpretation and control.

Description
As well as classical data description methods (medians, variances, histograms, etc) geostatistical data can also be explored from its spatial characteristics. This can be done inside a sample variogram where variance is estimated at increasing distance and in several directions. Further insight into the nature

of the variation can be gained through fitting models to reveal features.

Interpretation
The shape of the points in the variogram can infer much about how the soil's properties change with distance or direction, indicating anisotropy and its nature. It can also indicate the quality of the sampling strategy.

Control
The idea of control in geostatistics suggests that although the spatial nature of the soil characteristics cannot be changed, our response to them can.

Geostatistics in Soil Science

(concept + process)

This chapter will describe the geostatistical concepts, which will be employed in the soil analysis for the LADSS project. The focus will be on specific, implementable methods, however some discussion of major concepts is necessary.

The key concepts and processes behind Kriging are:

- Regionalised Variables,
- Spatial Variability,
- Semi-Variogram Analysis, and
- Semi-Variogram Estimation.

These ideas will be discussed first to provide a conceptual background to the rest of this chapter. Some of the more complex techniques dealing with non-normal sample distributions and anisotropy will then be described.

Regionalised Variables

The fundamental idea behind the geostatistical representation of soil is the concept of a *regionalised variable*. Where conventional statistical methods concentrate on the individual, discrete data points. Geostatistical methods look more closely at the differences in value and spatial location of the data points, and the relationships in these differences.

It is assumed that, given an adequate population, variables will exhibit a degree of continuity within a finite region. It is also a key assumption that the regionalsied variables are subject to a statistically normal distribution.

Spatial Variability

Geostatistical estimation (Kriging) can be tailored to the spatial variability of a regionalised variable. However the absolute value of point can never be known for sure: only an estimate can be made of values outside the dataset. A semi-variogram can provide measures of:

- Sample variability
- Range of influence
- Sample adequacy

Measuring Spatial Variability

The variability of two samples (S_1 and $S_{2)}$ is described by:

$$2\gamma_{12} = (S_1 - S_2)^2$$

Where: S- samples

γ- variability

The distance (h) is the length of the vector between two samples (if their orientation is ignored). The results of the variability can then be averaged over distance intervals (dh):

$$2\gamma_i = 1/n_i \sum (S_j - S_k)^2 \quad \text{for } h < h_{jk} \leq h+dh$$

Where n_l is number of sample pairs in the interval. This is equal to the statistical variance of the sample differences. The semi-variogram is a plot of the semi-variance versus distance for the sample population.

The standard error is then:

$$\sigma_i = \sqrt{(2\gamma_i)}$$

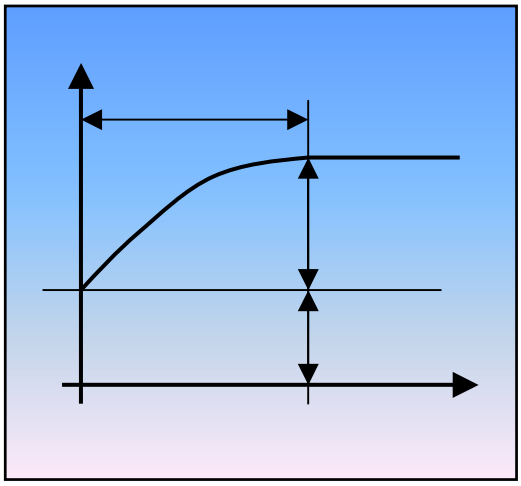
Which then provides a way of establishing uncertainty with the estimation.

The γ relationship should show a general increase within a range then a levelling off or oscillation. To identify a relationship difference distance intervals should be experimented with. Intuatively samples close together have small differences, those far apart have bigger differences and samples at large distances can be expected to be independent of each other.

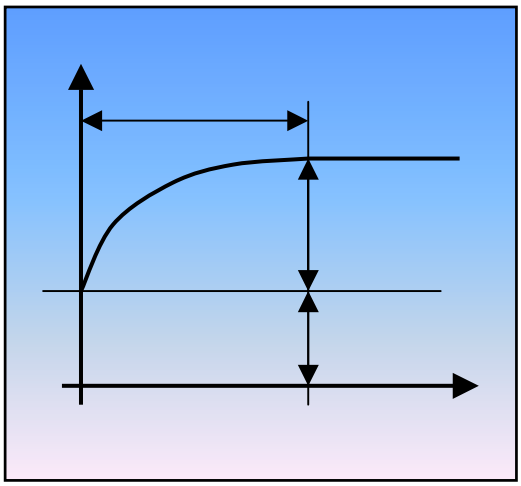
Theoretically γ should approach zero as h approaches zero, however in practice there is a natural randomness and sampling error present. Mathematical expressions can be derived to meet these expectations in representing the spatial variability of a sample.

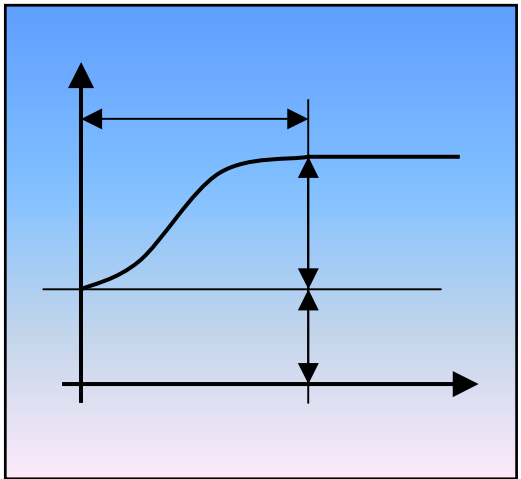
Modelling Spatial Variability

The semi-variogram relates semi-variance to distance. Various models can be used in this process. The most frequently used is the spherical model.



C_o is the inherent random variability of samples at zero distance as described above. This value is known as the *Nugget.* The distance a is known as the *Range of Influence* and is the distance at which samples become independent of each other. The model parameters are determined by interactively fitting the model expression to the results of semi-variogram analysis. There are three other significant model types:

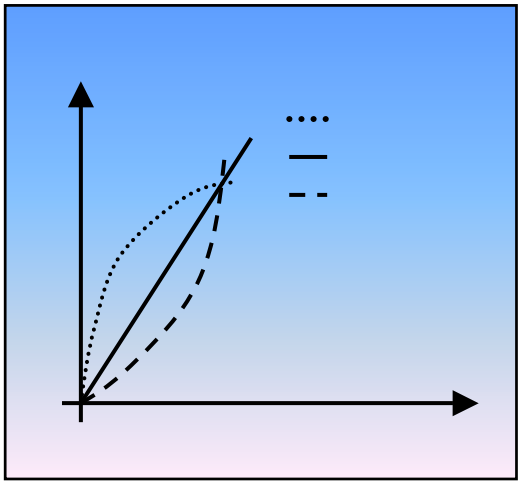




The Exponential Model

The Gaussian Model

The Power Model



The power model has no concept of *range of influence*. In practice it is applied over a limited finite distance. It is described by:

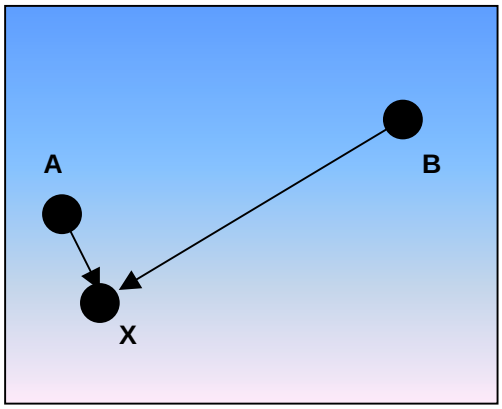
$$\gamma = C_o + \rho S^{\alpha}$$

Where: ρ = Slope

α = Power

It is also possible to have a combination of models, usually incorporating the power model as a multiplier of sorts.

Estimating a Regionalised Variable at a Point



In this diagram, point **A** would provide a better approximation to point **X**'s value than **B**, but **B** is still significant. In practice there will be a large number of samples from which to compare and as such a linear weighted combination of these sample values are used. More weight is given to the closer samples. The weights are assigned to reduce the variance.

This is accomplished by substitution in the expression for statistical variance then differentiating with respect to each of the weights, and solving for minimum variance to obtain the estimation and variance equations:

$\partial / \partial w_i$ (variance) **= 0…** For i = 1,2,3,…,n

On condition that the weights add up to 1, the equations are solved to obtain the sample weights, then back substituted in the linear estimator to obtain an estimate. This process is called *Kriging*.

There can a great abundance of sample points available for use in this process so it is useful to select the most relevant points depending on:

- Their distribution, and
- The N_{max} closest points within range.

From the variance we can derive the standard error, which can be used as a measure of uncertainty

Non Normal Sample Distributions

This describes a frequently occurring situation where the sample data does not fit the criterion of a statistically normal distribution. In cases such as these it may be difficult to discern a measurable relationship in the spatial variability of samples. It may be suitable to apply a transform to the samples to make them more amenable to analysis.

There are four commonly used transforms:

- Log Transforms
- Indicator Transforms
- Rank Order Transforms
- Normal Scores Transforms

Each is described below.

Log Transforms
This converts samples to natural log values, to allow the use of log-normal or exponential distributions. The semi-variogram model derived from transformed samples can be used to estimate variation of logarithmic values of regionalised variables. To obtain variation of real values and anti-log back transform is needed.

Indicator Transform
The indicator transform answers a specific criterion (or query) about the samples and then converts the samples to a 1 or 0 depending on the result. It provides a measure of probability that a point satisfies a specific criterion. This is an effective method for semi-variogram analysis if the population is large enough. Indicators cannot however be back transformed, but can be used to estimate the variation of probability of meeting a specific criterion.

Rank Order Transform
This transform sorts samples into ascending order and converts them into an integer ranking system. This is useful for eliminating the effects of scale in samples, but is only effective for semi-variogram analysis, not estimation.

Normal Scores Transform
The normal scores transform assumes a mean of 0 and standard error of 1 and adjusts each sample by mapping it from the measured culmitive frequency to the standard S-curve of normal distribution. Transformed values have a statistically normal distribution. This method can be

used with semi-variogram analysis and estimation, since values can be back-transformed. It is used in conjunction with simple Kriging for space data sets.

Use of Data Transforms
For semi-variogram analysis, any transform that assists in estimation of the spatial variability of the samples is useful. For estimation there is little point in using a transform which cannot be converted back to useful values.

Semi-variogram Analysis and estimation with a determinate trend

Semi-variogram analysis assumes that the samples of regionalised variables are not subject to any spatially determinate trend. Semi-variogram analysis and estimation is concerned only with the indeterminate component of spatial variability. If a determinate trend exists in the data it may overwhelm the semi-variogram results. Trends can be detected through the use of regression analysis. If trends are located they can be subtracted from the samples and the semi-variogram analysis applied to the residual values.

Removing a determinate trend is similar to using a data transform, which allows back-transformation. The normal estimation procedures are applied to the sample residuals, and the trend component is added back, to obtain the final estimated variation. This is called *Universal Kriging*.

Analysis and estimation with anisotropic spatial variability

Up to this point the assumption has been made that variability is independent of direction, isotropic. However, frequently variability can vary significantly with direction, this is called anisotropy. A directional control can be added to the semi-variogram analysis by:

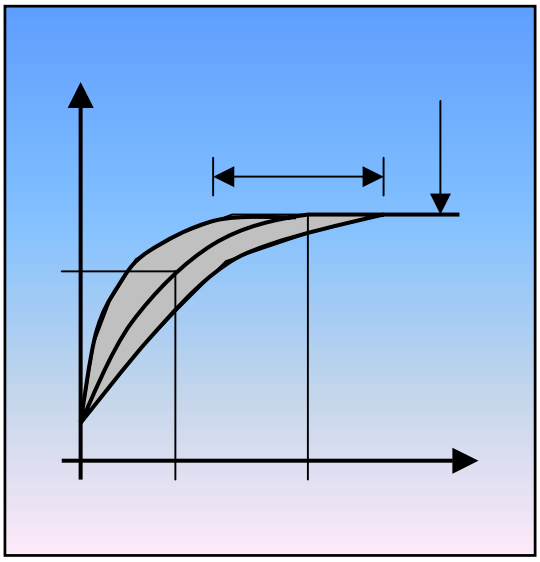
1. determining the orientation of sample pairs,
2. separating them into subsets according to direction, and
3. analysing the subsets independently

The subsets are defined by azimuth and inclination rotations, and angular tolerances. The principle directions of anisotropy, with minimum and maximum ranges of influence, represent the anisotropic variability. The orthogonal ranges of influence define an ellipsoid of influence around a point.

Anisotropy is accounted for in estimation by assuming the range of influence has elliptical variation about a point of estimation.

Range of influence:

$$X^2/a^2X + Y^2/a^2Y + Z^2/a^2Z = 0$$



The semi-variogram model represents this elliptical range of influence by a range of influence envelope, where semi-variance is dependent on direction and distance.

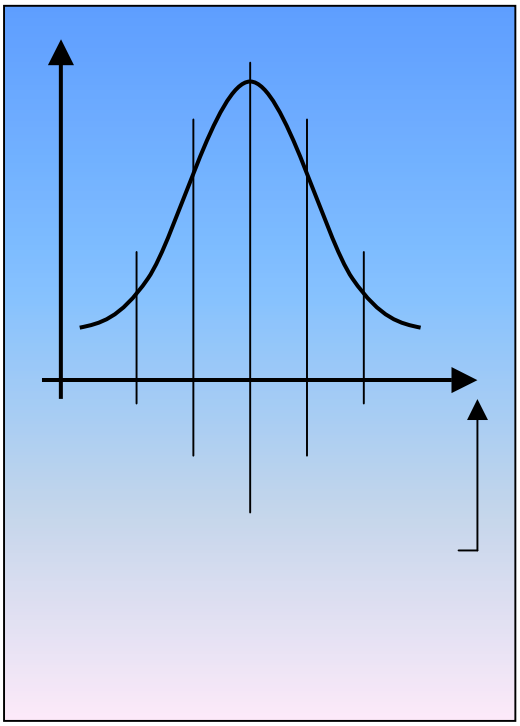
The semi-variance between estimation and sample is obtained from sample orientation, which establishes a range of influence and a scaled value for distance

Estimation with an anisotropic sample set adds an extra dimension to the weighting process. If two samples are equidistant from an estimation point the point closest to the direction of maximum range of influence has less semi-variance than its counterpart, therefore it is assigned a greater weight.

Estimation Uncertainty

A point's estimated value is its most likely value based on sampled and specified parameters. The possible variation in value at a point has a statistically normal distribution of its own. The estimated value is the mean of this distribution, and the estimation variance is the distribution variance.

The standard error associated with the estimated value defines the spread of the normal distribution curve.



Interpolation techniques

External drift

Assessment of uncertainty

Stochastic simulation

(Primary reference "Geostatistics in soil science: state of the art perspectives" Gooverts)

("Geostatistics for environmental scientists" Webster and Oliver)

LADSS implementation

(process)
The LADSS problem
Specific, needs of the project

Software

The Splus statistical package is very well suited to all types of statistical analysis, but it has the added advantage of a specific spatial statistics extension, which provides a platform for serious geostatistical analysis. It provides its own language "S" to automate tasks and conduct more complex analysis. Splus even allows a link to the ArcView GIS package, which together, provide a convenient and very powerful suite of tools for geostatistical analysis and estimation.

This combination allows the user to develop maps and sample sites visually in ArcView, and then analyse them using parameterable statistical solutions inside Splus. Splus can also produce complex graphics for graphs and 3D visualisations, which ArcView can sometime have difficulty with.

Practicalities with interfacing

Splus and Spatial Statistics

This description will outline the field of spatial statistics, discussing the major points, but assuming a level of geostatistical understanding.

Exploritory data Analysis
The initial and essential steps of exploratory data analysis (EDA) are easily conducted in Splus. The basic summary statistics can be accessed using the:

```
> summary (data)
```

command. This will provide the min, mean, median and max etc. In addition to this, the:

```
> hist (data)
```

command will provide a histogram of selected columns. These two commands are not explicitly spatial, and are included in the standard Splus library, but they are important for the EDA of the data set in question. Similarly it is possible to plot the relevant points on a XY grid to visually investigate the spatial distribution of points.

It may also be useful to produce an interpolation to indicate the presence of any 2D trends. This can be done by using the:

```
> interp.data <- Interp
+ (x=dataX, y=dataY, z=dataZ)
> contour (interp.data)
> points(dataX, dataY)
> image(interp.data)
```

From the EDA it should be possible to identify outliers and trends in the data. It will also be possible to investigate the distribution of the points as well as the distribution of the values themselves. Especially as one of the fundamental ideas on which geostatistical analysis and estimation rests is that of a normally distributed data set.

Variogram Estimation
Splus uses an 'empirical variogram' to provide

Variogram Fitting

Ordinary and Universal Kriging

Simulating Geostatistical Data

Practicalities with interfacing

References

Primary reference "Geostatistics in soil science: state of the art perspectives" Gooverts

Heuvelink and Webster (2001)
Webster and Oliver 2001

Further Reading