

CONTENTS

Chapter 2 Statistical Methods	2
2.1 Statistical Notation and Summary	2
2.1.1 Variables	2
2.1.2 Notation	2
2.2 Descriptive statistics	2
2.3 Transformations	3
2.3.1 Logarithmic transformation	3
2.3.2 Square root transform	3
2.4 Exploratory data analysis and display	3
2.4.1 Histograms	3
2.4.2 Box-plots.....	3
2.4.3 Spatial aspects.....	4
2.5 Ordination.....	4
2.5.1 Principal Component Analysis (PCA)	4
2.5.2 Principal Coordinate Analysis (PCO).....	4
2.6 Numerical Multivariate Classification.....	6
2.6.1 Non-hierarchical classification.....	7
2.7 Geostatistics	8
2.7.1 Introduction.....	8
2.7.2 Measuring the Correlation Structure	8
2.7.3 Kriging	12
2.7.4 Co-kriging.....	20
2.8 Geostatistical Simulation.....	23
2.8.1 Sequential Gaussian Simulation.....	25
2.8.2 Turning Bands	25

Figure 2.1 Forms of variograms: (a) unbounded, (b) bounded, (c) pure nugget.

Figure 2.2 Turning bands in two dimensions.

Chapter 2 Statistical Methods

2.1 Statistical Notation and Summary

2.1.1 Variables

The variables recorded in the NSI are of three main kinds: binary (presence, absence) taking values 1 and 0, multistate (classified variables with more than two states, such as soil structure), and fully quantitative variables measured on continuous scales with equal intervals. The latter include the records of the major plant nutrients and trace elements and potentially toxic heavy metals, which are of the greatest interest in this report.

2.1.2 Notation

The data have spatial co-ordinates as well as recorded values, and we distinguish measurement from location. We use the following notation as far as possible throughout. Variables are denoted by italics; an upper case Z for random variables and lower case z for a realization, i.e. the actuality, and also for sample values of the realization. Spatial position in the two dimensions is denoted by bold \mathbf{x} , meaning the vector $\mathbf{x} = \{x_1, x_2\}$. Thus, $Z(\mathbf{x})$ means a random variable Z at place \mathbf{x} , and $z(\mathbf{x})$ is the actual value of Z at \mathbf{x} . In general, we shall use bold lower case letters for vectors and bold capitals for matrices. We also distinguish between variables and parameters by the notation, i.e. constants of populations, and between the parameters themselves and their estimates. For variables, we shall use letters of the Roman alphabet only, and for population parameters, Greek letters. For estimates of these, we shall use either their Roman equivalents or place carats (\wedge) over the Greek: for example, the standard deviation of a population will be σ and its estimate s or $\hat{\sigma}$.

2.2 Descriptive statistics

The following descriptive statistics have been computed on each variable: the arithmetic mean, the median, the variance and its square root, the standard deviation, the coefficient of variation (CV) expressed as a percentage, the coefficient of skewness and of kurtosis.

2.3 Transformations

In many instances the distributions of the measurements are far from normal, and to stabilize variances we have transformed the measured values to new scales on which the distributions are more nearly normal for further analysis.

2.3.1 Logarithmic transformation

The most common departure from normality has been strong positive skewness, i.e. $g > 0$. To correct for this we have transformed the data to logarithms:

$$z' = \log_{10} z \quad \text{or exceptionally} \quad \ln z .$$

The logarithm of 0 is $-\infty$, and to avoid this undesirable transform we have added small values, approximately 0.01 of the range of z , to the data if 0 was recorded. Note that a value of 0 in the dataset represents a value less than the detection limit see McGrath & Loveland, 1992).

2.3.2 Square root transform

For moderately skewed data ($0.5 > g > 1$) we have taken square roots of the data:

$$z' = \sqrt{z}$$

2.4 Exploratory data analysis and display

The data were examined before formal analysis to see the main features, to identify outliers, to detect faults in measurement and recording, and to decide what transformation, if any, might be required. For this we drew histograms and boxplots.

2.4.1 Histograms

For each variable, z , we divided the range into some 25 classes of equal interval, and plotted against the values of z .

2.4.2 Box-plots

We drew box-plots with a box delimiting the interquartile range, the median, and 'whiskers' extending from the limits of the box to the limits of the data.

2.4.3 Spatial aspects

The co-ordinates of the sampling points were plotted on a map, a 'posting', with the boundary of England and Wales superimposed to check that they are reasonable.

2.5 Ordination

We used ordination to explore and describe the relations in the multivariate data by projecting them on to a few new axes that explained much of the variation.

2.5.1 Principal Component Analysis (PCA)

Principal components were computed from various subsets of the data to investigate the multivariate structure and to identify spatial features common to many of the original variables. This was done on variates standardized to unit variance after transformation to logarithms or square roots as above. Components were retained if their *eigenvalues* exceeded 1 (*Kaiser's criterion*, Kaiser, 1958). We interpreted the components by identifying large elements of the *eigenvectors*. We examined the relations between individuals by the *principal component scores*.

2.5.2 Principal Coordinate Analysis (PCO)

PCA adheres strictly to the Euclidean model. The distances between plotted points and the relations between them are approximations to Euclidean distances. However, Euclidean distance is not always the most appropriate measure of the likeness between individuals. An important advance in ordination was Gower's development (Gower, 1966) of PCO in which he was able to find a Euclidean representation from initially non-Euclidean similarities. Principal co-ordinates are calculated from a matrix of distances (dissimilarities) between individuals. A great advantage of the method is that it can be applied to data that are not quantitative. It is also less affected by missing data than PCA. PCO should give the same results as PCA computed from a correlation matrix for quantitative data. The procedure is:

- 1) First calculate a dissimilarity matrix, \mathbf{Q} , between individuals (or convert a matrix of similarities to distances). If there are n individuals, \mathbf{Q} is of order $n \times n$.

- 2) This matrix is adjusted by subtracting the corresponding row and column means from each element, and adding the general mean to give the matrix **F**.
- 3) Latent roots and vectors of **F** are found, and the vectors are arranged as columns in a $n \times n$ matrix **C**.
- 4) The rows of the matrix **C** represent the co-ordinates of the points in relation to the new axes.
- 5) The vectors are now normalised so that the sums of squares of their elements equal their corresponding latent roots. The transformed matrix **C** is the new matrix **G**, with elements:

$$g_{ik} = \sqrt{\frac{c_{ik}^2}{\sum_{i=1}^n c_{ik}^2}} \quad (2.1)$$

That is:

$$\mathbf{G}^T \mathbf{G} = \mathbf{I}, \text{ and } \mathbf{G} \mathbf{G}^T = \mathbf{F}$$

where **D** is the matrix of distances between individuals and **F** contains the scaled distances between individuals. The latent vectors scaled in this way represent exactly the distances between individuals and define their properties relative to principal axes. A difficulty of using PCO for a large data set, such as the NSI, is that the similarity matrix for all sites is too large to be held in computer memory. The solution is to derive a smaller similarity matrix from a representative subset of the data. New points are then added a few at a time, and the distances between the new individuals and the initial individuals are calculated. In this way, the principal co-ordinates can be found for all of the sites in a large data set.

As in PCA, a few roots are usually a good representation of the variation. Although PCO is more flexible than PCA, the latter is usually preferred because the original properties are linearly related to the new axes, which makes interpretation of the new axes more straightforward. A PCO does not allow this because the information about these relations is in the similarity matrix, and this information is lost during the course of the analysis. It is possible to relate the original variables to new axes by computing Snedecor's F-statistic for each variable and then ranking the variables accordingly.

PCA is also more efficient, in computing terms, than PCO if there are many individuals and few variables. The similarity matrix between individuals is much

larger than the correlation matrix between variables, and this has to be kept in memory.

2.6 Numerical Multivariate Classification

Classification has been the conventional way of examining the relations between soil individuals (profiles). The traditional classifications, as developed by the National Soil Surveys, assume that there are natural groups. Clearly the more 'natural' a group, the greater its predictive capabilities. This notion of a natural classification based on overall similarity might not produce strictly natural classes. A general purpose classification based on overall similarity could be arbitrary, and this is often the case for soil. If we think of the classification in this way, the groups can be adjusted according to specific needs, yet still retain some of the benefits of overall similarity and general purpose.

For soil data, it is more realistic in the light of modern knowledge to start with the premise that there are no well-defined natural groups of soil. A specific classification is one with a particular purpose in mind. Classes are differentiated on the basis of a single property or a limited number of properties selected at the outset. Such groups have a limited purpose – that for which they were created. Groups based on a single character have low predictive capability in terms of other properties. However, the NSI data provide the possibility of creating special purpose classifications and general purpose ones as required by the user.

A natural group would appear as a cluster in multivariate property space, i.e. the individuals would be in the more densely occupied part forming a cluster. The clusters would then be separated from other such regions by areas containing a relatively low density of points. Thus, if well defined clusters exist they should be evident in the projection of the points in the plane of the first two principal axes.

With a large data set, such as the NSI, where many properties have been measured, there are so many data that it can be difficult to see any relations, and consequently it is difficult to interpret and to comprehend the information. Classification can help to provide a more simple picture of the relations between individuals as with ordination. It may also help to economise on expensive measurements in the future if clear relations between properties emerge.

There are two approaches to numerical classification: hierarchical and non-hierarchical. Most traditional classifications of the soil have been hierarchical -

following the biological model based on common ancestry. The early numerical classifications were also hierarchical.

2.6.1 Non-hierarchical classification

This is an alternative approach to classification. Oliver & Webster (1987) found that it worked well for soil data where there were no obvious clusters. Non-hierarchical classification is also known as dynamic clustering. The population is subdivided at a single level into as many classes as desired. The approach subdivides a set of individuals into two or more disjoint groups. Each individual belongs to one, *and only one*, group. The general aim is to subdivide the population optimally such that there is minimum variation within the classes and the difference between them is maximised. Rubin (1967) and Friedman & Rubin (1967) described how this could be done, and Crommelin & de Gruitjer (1977), McBratney & Webster (1981) and Oliver & Webster (1987, 1989) have applied it successfully.

A mathematical criterion is chosen as a basis for optimising the subdivision, such as the within groups sums of squares or Wilks' criterion (L), to measure the dispersion within groups, or Trace $W^{-1}B$ to measure the separation between groups. The population can be divided into an arbitrary number of groups at the outset or the number could be based on information from another analysis. The test criterion is calculated, and individuals are then moved from group to group and the criterion recalculated. If the change improves the criterion the move is retained, otherwise it is not. There are different ways of moving individuals from group to group in an iterative way to try to obtain an optimum (Webster & Oliver, 1990). In general, the optimal number of groups to subdivide the population into is unlikely to be known, therefore more groups than are likely should be chosen at first. The number of groups can then be reduced one at a time by fusing the two most similar groups and recalculating the criterion. Individuals are then moved as before.

In general, soil individuals are weakly clustered, and several subdivisions could be reasonable. Non-hierarchical classification has an additional advantage compared with the hierarchical ones, in that once individuals are assigned to groups they can still be moved. This means that the groups change character as individuals are removed or added to optimise the criterion. Once an individual is grouped in a hierarchical method it is irrevocable.

One of the difficulties with non-hierarchical methods is that it is sometimes difficult to decide how many groups is optimal, groups tend to be of a similar size and shape, and weak clusters might not be isolated as distinct classes. There is a solution to this that uses Wilks' criterion, L , and g^2L is plotted against g , where g is the number of classes.

2.7 Geostatistics

2.7.1 Introduction

Geostatistics developed to predict values from more or less sparse sample data of properties that vary in complex ways. We have used geostatistics in this project both to explore the spatial structure and relations in the data as well as for predicting values at unsampled locations, and for designing sparser sampling schemes. The analyses include variography, variogram modelling, ordinary kriging, factorial kriging, disjunctive kriging, co-kriging, and geostatistical simulation.

Underlying geostatistics is a substantial body of theory, regionalized variable theory, that treats variables as the realizations of random processes or random functions (RF). This allows us to model the correlation or spatial dependence (structure) in the processes. Stationary random functions are one kind of RF, and are those usually assumed in geostatistics; they are valuable for dealing with applied problems, in particular. Stationary processes do not undergo any systematic changes; they simply fluctuate about a constant mean in a disordered manner. Using assumptions of stationarity means that we can assume the variation from place to place.

2.7.2 Measuring the Correlation Structure

Stationary random process can be represented by the model:

$$Z(\mathbf{x}) = \mu + \epsilon(\mathbf{x}) \quad (2.2)$$

where the value of the regionalized variable, Z , at \mathbf{x} is the mean of the process, μ , plus a random component $\epsilon(\mathbf{x})$, which has a mean of zero and a covariance function:

$$C(\mathbf{h}) = E[\{\epsilon(\mathbf{x})\}\{\epsilon(\mathbf{x} + \mathbf{h})\}] \quad (2.3)$$

where E is the expectation. Thus, the relation between the values of pairs of points separated by a distance \mathbf{h} (known as the *lag*) can be measured in the same way as the relation between two different properties, by the covariance function:

$$\begin{aligned} \text{cov}[Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})] &= E[\{Z(\mathbf{x}) - \mu\}\{Z(\mathbf{x} + \mathbf{h}) - \mu\}] \\ &= E[\{Z(\mathbf{x})\}\{Z(\mathbf{x} + \mathbf{h})\} - \mu^2] \\ &= C(\mathbf{h}). \end{aligned} \quad (2.4)$$

In general the covariance is converted to the autocorrelation function:

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0}) \quad (2.5)$$

where $C(\mathbf{0})$ is the covariance at lag $\mathbf{0}$, i.e. the variance, σ^2 . If the variance appears to increase as the area of interest increases then the covariance cannot be defined. Matheron (1965) solved this problem by reducing the assumptions of stationarity to those of the intrinsic hypothesis.

2.7.2.1 *Matheron's Intrinsic Hypothesis*: This is based on the expected differences:

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] \quad (2.6)$$

The variance of the differences is based on weaker assumptions than that of second order stationarity. Matheron realised that regionalized variables have an infinite capacity for variation and a constant or finite variance cannot be assumed. The intrinsic model assumes an expectation:

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = 0 \quad (2.7)$$

and a variance $\gamma(\mathbf{h})$:

$$\text{var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] = 2\gamma(\mathbf{h}) \quad (2.8)$$

where $2\gamma(\mathbf{h})$ is the variance of the difference at lag \mathbf{h} , and its half, known as the *semivariance*, is the variance per point when values are considered in pairs. The semivariance depends on \mathbf{h} , and as a function, $\gamma(\mathbf{h})$, is the *variogram* of Z . The advantage of assuming intrinsic variation is that the model has wider generality and application than that of second order stationarity, and the variogram has become the central tool of geostatistics. When a process is second order stationary, the semivariance is related simply to the *covariance*, $C(\mathbf{h})$, and the *autocorrelation*, $\rho(\mathbf{h})$, by:

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}) = C(0)\{1 - \rho(\mathbf{h})\} \quad (2.9)$$

and the variogram, $\gamma(\mathbf{h})$, and the covariance function, $C(\mathbf{h})$, are equivalent for characterizing spatial autocorrelation.

2.7.2.2 Estimating the Sample or Experimental Variogram: The standard equation for computing the experimental semivariance $\hat{\gamma}(\mathbf{h})$ is:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2 \quad (2.10)$$

where $\hat{\gamma}(\mathbf{h})$ is the estimate of $\gamma(\mathbf{h})$, $z(\mathbf{x}_i)$ and $z(\mathbf{x}_i + \mathbf{h})$ are the observed values of Z at (\mathbf{x}_i) and $(\mathbf{x}_i + \mathbf{h})$, respectively, $m(\mathbf{h})$ is the number of paired comparisons at that lag, and the lag, \mathbf{h} , is a vector in both distance and direction. By changing \mathbf{h} , we obtain an ordered set of semivariances, known as the *experimental variogram* or *sample variogram*. The variogram is the function that relates the semivariance to the lag, and is usually presented as a graph of $\hat{\gamma}(\mathbf{h})$ against \mathbf{h} .

2.7.2.3 Variogram Interpretation: The main features of the variogram are:-

- (a) An increasing variance as the lag distance increases from a small value at the shortest lag (Figure 2.1 a). It reflects spatial autocorrelation or dependence in the data, i.e. places near to one another have similar soil, or similar values of the property measured and, as the separation increases, they become increasingly dissimilar on average. Some variograms (Figure 2.1 b) increase indefinitely as the

- lag distance increases: they represent variation that is intrinsic only, i.e. Matheron's Intrinsic Hypothesis holds, but the covariance does not exist.
- (b) The variogram often increases to a maximum at which it remains thereafter. The *sill* variance is an upper bound that the variogram sometimes increases to (Figure 2.1 b). It estimates the *a priori* variance of the random variable and signifies that the variable is second order stationary. The *range* is the lag distance at which the sill is reached: it marks the limit of spatial dependence (Figure 2.1 b). Places separated by distances greater than this are spatially independent. Data points used for interpolating values should be within the range of spatial dependence.
- (c) If the experimental variogram is extrapolated to the ordinate, it often has a positive intercept: the *nugget variance*. This corresponds to the spatially uncorrelated variation. For continuous properties of the soil, such as particle size distribution, pH or concentrations of trace elements, the nugget variance comprises measurement error plus variation that occurs over distances less than the shortest sampling interval. Some variograms appear to be completely flat - *pure nugget*. This means that there is no spatial dependence evident in the data. For continuous variables, this usually arises because the sampling interval is too large: all of the spatially correlated variation is occurring within the smallest sampling interval. More intensive sampling would be needed to identify the spatial structure in the variation. If the variogram is pure nugget, then the data should not be used for any kind of interpolation because there is no spatial relation between the points.
- (d) The shape of the variogram provides insight into the structure of the variation and possible processes that are controlling the variation.

2.7.2.4 *Anisotropy*: Variation can be different in different directions. Therefore, experimental variograms should be computed in at least four directions. If the directional variograms differ substantially from one another then this might signal anisotropy in the random process. Initial gradients or ranges of the directional variograms that are very different suggest that the rate of change in spatial variation and the spatial scale vary with direction. They are often evidence of *geometric anisotropy*, which can be removed by a simple transformation of the spatial coordinates. If the sill variances are different, this suggests that there are different

amounts of variation in different directions. This is a possible indication of *zonal anisotropy*, and is more difficult to deal with. One solution is to stratify the data.

2.7.2.5 Drift and trend: In some instances the variogram approaches the origin with a decreasing gradient: it has a concave upwards form. This can arise from local *trends* or *drift* which are steady progressions in the data. This situation is unlikely to occur with the NSI data because the samples are a large distance apart. However, regional trends might occur giving rise to an upwardly concave section of the variogram after the sill had been reached. In this situation the intrinsic hypothesis no longer holds. A solution for regional trend is to model it by a low order polynomial, and then compute the variogram on the residuals from it. Another is to stratify the data, which is often effective in removing regional trend.

2.7.2.6 Spatial Resolution: The experimental variogram depends to some extent on the scale or resolution of the investigation. It tends to change as the area covered becomes either larger or smaller, because the amount of variation encountered changes accordingly. In general, the sampling interval is increased as the extent of a region becomes larger, with the result that the finer detail is lost; this is embodied in the nugget variance. The variogram also depends on the support of the sample, i.e. its size, shape and orientation. The smaller the support the more variation there is likely to be in the inter-sample area. As the size of the support increases, the more local variation it encompasses. The effect of this on the variogram is to reduce both the nugget and the sill variances.

2.7.3 Kriging

Kriging is the procedure of estimation or prediction embodied in geostatistics. At its simplest, it is a method of local weighted moving averaging of the observed values within a neighbourhood V . Weights, w , are allocated to the sample data within the neighbourhood. They depend on the variogram and on the configuration of the sampling sites, and are allocated to minimize the estimation or kriging variance, and to ensure that the estimates are unbiased. These variances are also estimated. In this sense, kriging is an optimal interpolator. Estimates can be made for points, \mathbf{x}_0 , or over blocks, \mathbf{x}_B . Punctual kriging is an exact interpolator, in that estimates at sampling sites

are the observed values there, and the estimation variance is zero. The latter provide some measure of the reliability of the estimates. These attributes set it apart from all other methods of interpolation.

2.7.3.1 *Ordinary kriging*: As above, an estimate of Z at B , denoted as $\hat{z}(B)$, is a weighted average of the data, $z(\mathbf{x}_1), z(\mathbf{x}_2), \dots, z(\mathbf{x}_N)$:

$$\hat{z}(B) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i) \tag{2.11}$$

where, λ_i are the weights. The kriging variance is given by:

$$\sigma^2(B) = E\{[\hat{z}(B) - z(B)]^2\}$$

$$= 2 \sum_{i=1}^n \lambda_i \mathcal{C}(\mathbf{x}_i, B) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \mathcal{C}(\mathbf{x}_i, \mathbf{x}_j) + \mathcal{C}(B, B), \tag{2.12}$$

where, $\mathcal{C}(\mathbf{x}_i, \mathbf{x}_j)$ is the semivariance of Z between the i th and the j th sampling points, $\mathcal{C}(\mathbf{x}_i, B)$ is the average semivariance between the i th datum and the block for which the estimate is required, and $\mathcal{C}(B, B)$ is the average variance within the block (the within-block variance). The value of $\sigma^2(B)$ is least when:

$$\sum_{i=1}^n \lambda_i \mathcal{C}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^n \lambda_i \mathcal{C}(\mathbf{x}_i, B) \quad \text{for all } j$$

$$\sum_{i=1}^n \lambda_i = 1 \tag{2.13}$$

This is the kriging system to be solved. The Lagrange multiplier, λ , is introduced to achieve minimization. In practice, N is replaced by $n \ll N$, the number of observations near to \mathbf{x}_0 .

2.7.3.2 *Kriging analysis or factorial kriging*

Nested variation: A random process can be a combination of several independent processes, one nested within another and acting at different characteristic spatial scales. These may be explored at individual scales separately by ‘kriging analysis’ or factorial kriging (Matheron, 1982). It treats the variation at each evident scale in turn as the signal, and separates it from variation at all other scales. Where

the variation is nested, the variogram of $Z(\mathbf{x})$ is itself a nested combination of two or more, say S , individual variograms:

$$\gamma(\mathbf{h}) = \gamma^1(\mathbf{h}) + \gamma^2(\mathbf{h}) + \dots + \gamma^s(\mathbf{h}), \tag{2.14}$$

where the superscripts refer to the separate variograms. The formula for the individual variogram is:

$$\text{var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] = 2\gamma(\mathbf{h}). \tag{2.15}$$

If we assume that the processes are uncorrelated then we can represent this by the sum of S basic variograms:

$$\gamma(\mathbf{h}) = \sum_{k=1}^S b^k g^k(\mathbf{h}), \tag{2.16}$$

where $g^k(\mathbf{h})$ is the k th basic variogram function, and b^k is a coefficient that measures the relative contribution of the variances of $g^k(\mathbf{h})$ to the sum. The nested variogram comprises the S variograms with different coefficients, b^k . This is our linear model of regionalization. It represents the real world in which factors such as relief, geology, tree-throw, fauna, and man's divisions into fields and farms, operate on their own characteristic spatial scale(s), and each with its particular form and parameters, b^k , for $k = 1, 2, \dots, S$.

Kriging analysis: In ordinary kriging, $Z(\mathbf{x})$ is estimated in a single operation from the data and the variogram. For kriging analysis, $Z(\mathbf{x})$ itself is regarded as the sum of S orthogonal random functions, corresponding with the components of the variogram, $b^k g^k(\mathbf{h})$, above. Provided $Z(\mathbf{x})$ is second order stationary, this sum can be represented as:

$$Z(\mathbf{x}) = \mu + \sum_{k=1}^S Z^k(\mathbf{x}), \tag{2.17}$$

in which μ is the mean of the process. Each $Z^k(\mathbf{x})$ has expectation 0, and the squared differences are:

$$\frac{1}{2} E[\{Z^k(\mathbf{x}) - Z^k(\mathbf{x} + \mathbf{h})\}\{Z^{k'}(\mathbf{x}) - Z^{k'}(\mathbf{x} + \mathbf{h})\}] = b^k g^k(\mathbf{h}) \text{ if } k = k' \\ = 0 \text{ otherwise.} \tag{2.18}$$

It is possible that the last component, $Z^S(\mathbf{x})$, is intrinsic only, so that $g^S(\mathbf{h})$ in Equation 2.16 is unbounded with gradient b^S . For two components, as for most properties of the NSI data, Equation (2.17) reduces to:

$$Z(\mathbf{x}) = Z^1(\mathbf{x}) + Z^2(\mathbf{x}) + \dots, \tag{2.19}$$

Relation (2.18) expresses the mutual independence of the S random functions $Z^k(\mathbf{x})$. With this assumption, the nested model (2.16) is easily retrieved from relation (2.17). Each spatial component $Z^k(\mathbf{x})$ can be estimated as a linear combination of the observations $z(\mathbf{x}_i)$, $i = 1, 2, \dots, N$. The λ_i^k are the weights assigned to the observations.

$$\hat{Z}^k(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i^k z(\mathbf{x}_i). \tag{2.20}$$

They must sum to 0, not 1 as in ordinary kriging, to ensure that the estimate is unbiased and accord with Equation (2.17). Subject to this condition, they are chosen so that the estimation variance is minimal. This leads to the kriging system:

$$\sum_{j=1}^n \lambda_j^k \gamma^k(\mathbf{x}_i, \mathbf{x}_j) + \lambda^k = b^k g^k(\mathbf{x}_i, \mathbf{x}_0) \text{ for all } i = 1, 2, \dots, n \\ \sum_{j=1}^n \lambda_j^k = 0. \tag{2.21}$$

This system is solved for each spatial component, k , to find the weights, λ_i^k which are then inserted into equation (2.20). The quantity λ^k is the Lagrange multiplier for the k th component. In general, the weights for the individual components will be different, and as result we can extract from the data the different components of the spatial variation identified in the variogram. Estimates are made for each spatial scale, i.e. each k , by solving equations 2.21.

In many instances, data contain long-range trend. This need not complicate the analysis because the kriging is usually done in fairly small moving neighbourhoods centred on \mathbf{x}_0 , as above. From a theoretical point of view it is necessary only that $Z(\mathbf{x})$ is locally stationary, or *quasi-stationary*, thus:

$$Z(\mathbf{x}) = \bar{Z}(\mathbf{x}) + \sum_{k=1}^S Z^k(\mathbf{x}) \quad (2.22)$$

where $\bar{Z}(\mathbf{x})$ is a local mean, which can be considered as a long-range spatial component. To kriging a second order stationary component, we start with the linear combination of the observations $z(\mathbf{x}_j)$:

$$\hat{Z}(\mathbf{x}_0) = \sum_{j=1}^n w_j z(\mathbf{x}_j) \quad (2.23)$$

??

The weights are obtained by solving the kriging system:

$$\begin{aligned} \sum_{j=1}^n w_j C(\mathbf{x}_i, \mathbf{x}_j) - \bar{C}(\mathbf{x}_i) &= 0 \text{ for all } i = 1, 2, \dots, n \\ \sum_{j=1}^n w_j &= 1. \end{aligned} \quad (2.24)$$

Estimation of the long-range component, i.e. the local mean $\bar{Z}(\mathbf{x})$, and the spatial component with the largest range, can be affected by the size of the moving neighbourhood (Galli *et al.*, 1984). To estimate a spatial component with a given range, the distance across the neighbourhood should be at least equal to that range. When there are many data and the range is large, then the effective neighbourhood in the analysis is often much smaller than the one chosen. To overcome this, we have added an estimate of the local mean to the long-range spatial component (Jacquet, 1989).

2.7.3.3 Disjunctive kriging

In addition to estimating $Z(\mathbf{x})$ at unsampled places, we may wish to use such estimates to suggest a course of action, e.g. make a decision. In such situations, we need to estimate the probability that the true value exceeds (or does not exceed) a threshold, z_c . For the NSI, data we are concerned with areas where the concentrations of Cd, Cu, Pb and Zn, in particular, exceed particular thresholds, and where K, P and Mg are less than the recommended limits for crop nutrition.

All estimates are subject to error because we usually have only fragmentary information from which to estimate them. Furthermore, kriging has a smoothing effect on the estimated values, especially where there is a large nugget variance, which adversely affects their value for decision making. In general, decisions are easy where the estimated values are much less than or much greater than a defined threshold, or where the estimation error is small, or both. For instance, if a pollutant in soil far exceeds a critical value then remedial action should be taken immediately. Equally, if the value is much less than the threshold then there is no need for concern. Difficulties arise where the estimated values are close to the threshold, or where the error is large. True values might exceed the threshold. Then there is a risk of making a wrong decision – of doing nothing when we should act and of acting unnecessarily. Disjunctive kriging (DK) solves the problem. It provides a means of assessing the risk taken by accepting the estimate at its face value, i.e. the probability that the true value exceeds or falls short of the threshold, given the estimated value and the data in the neighbourhood. For each estimate, it enables the probability that the true value exceeds (or does not) a threshold to be estimated through non-linear rescaling of the original data. In essence, it is a linear kriging of a non-linear transform. It is a valuable method for dealing with potential excesses or deficiencies in the soil (Rivoirard, 1994).

Indicator Coding: One way of tackling the problem of estimating the probability of a property exceeding a critical threshold, is to transform the data to indicator functions in relation to the threshold. This creates an indicator function, $I(\mathbf{x})$, where the threshold distinguishes between what is, and is not, tolerable. It dissects the scale of Z into two parts: one for which $Z(\mathbf{x}) > z_c$ and the other for which $Z(\mathbf{x}) \leq z_c$, and we can assign the values 1 and 0 to these, respectively. This is known as disjunctive coding. The indicator function, $I(\mathbf{x}) = [Z(\mathbf{x}) > z_c]$, is a random variable, $Y(\mathbf{x})$,

which has a variogram $\gamma_{z_c}(\mathbf{h})$. The most common type of DK, and the one that we have used for the NSI data, is Gaussian disjunctive kriging. The assumptions are:

- a) that $z(\mathbf{x})$ is a realization of a second-order stationary random process $Z(\mathbf{x})$ with a mean, μ , and variance, σ^2 . The variogram must be bounded for this analysis.
- b) that the bivariate distribution for $n + 1$ variates, i.e. for the target site and the sample locations in its neighbourhood, is known, and that it is stable throughout the region. If the distribution of $Z(\mathbf{x})$ is normal and the process second-order stationary, then we can assume that the bivariate distribution for each pair of points is also normal.

Hermite Polynomials: Since most environmental properties are not normal, $Z(\mathbf{x})$, is transformed to a standard normal distribution, $Y(\mathbf{x})$, such that:

$$Z(\mathbf{x}) = \mu + \sigma [Y(\mathbf{x})]. \tag{2.25}$$

This can be achieved using Hermite polynomials, which are defined by Rodrigues’s formula as:

$$H_k(y) = \frac{1}{\sqrt{k!}g(y)} \frac{d^k g(y)}{dy^k}, \tag{2.26}$$

where k is the degree of the polynomial, and $1/\sqrt{k!}$ is a standardizing factor (Matheron, 1976). Since the polynomials are orthogonal, they are independent components of the normal distribution. Almost any function of $Y(\mathbf{x})$ can be represented as the sum of Hermite polynomials:

$$f\{Y(\mathbf{x})\} = f_0 H_0\{Y(\mathbf{x})\} + f_1 H_1\{Y(\mathbf{x})\} + f_2 H_2\{Y(\mathbf{x})\} + \dots, \tag{2.27}$$

and since the Hermite polynomials are orthogonal:

$$\begin{aligned}
 E[f\{Y(\mathbf{x})H_k\{Y(\mathbf{x})\}\}] &= E\left[\sum_{k=0}^{\infty} H_k\{Y(\mathbf{x})\} \sum_{l=0}^{\infty} f_l H_l\{Y(\mathbf{x})\}\right] \\
 &= \sum_{l=0}^{\infty} f_l E[H_l\{Y(\mathbf{x})\} H_k\{Y(\mathbf{x})\}] \\
 &= f_k.
 \end{aligned} \tag{2.28}$$

This enables the coefficients f_k of $[Y(\mathbf{x})]$ to be determined by:

$$\begin{aligned}
 Z(\mathbf{x}) &= [Y(\mathbf{x})] \\
 &= \sum_{k=0}^{\infty} (H_k\{Y(\mathbf{x})\} \sum_{l=0}^{\infty} f_l H_l\{Y(\mathbf{x})\}) \\
 &= \sum_{k=0}^{\infty} f_k H_k\{Y(\mathbf{x})\}.
 \end{aligned} \tag{2.29}$$

This transform is invertible, which means that the results can be expressed in the same units as the original measurements. Any pair of Hermite polynomials is spatially independent, and by kriging them separately the estimates have only to be summed to give the DK estimator:

$$\hat{Z}^{DK}(\mathbf{x}) = \sum_{k=0}^{\infty} f_k \hat{H}_k^K\{Y(\mathbf{x})\} \tag{2.30}$$

If there are n points in the neighbourhood of \mathbf{x}_0 , the target point, the Hermite polynomials are estimated by:

$$\hat{H}_k^K\{Y(\mathbf{x}_0)\} = \sum_{i=1}^n w_{ik} H_k\{Y(\mathbf{x}_i)\}, \tag{2.31}$$

which are then inserted into equation (30). The kriging weights, w_{ik} , are found by solving the equations for simple kriging because we assume that the mean is known:

$$\sum_{i=1}^n w_{ik} Cov[H_k\{Y(\mathbf{x}_j)\}, H_k\{Y(\mathbf{x}_i)\}] = Cov[H_k\{Y(\mathbf{x}_j)\}, H_k\{Y(\mathbf{x}_0)\}] \quad \forall j, \tag{2.32}$$

The procedure enables us to estimate $Z(\mathbf{x}_0)$ by:

$$\hat{Z}(\mathbf{x}_0) = \hat{Y}(\mathbf{x}_0) + \sum_{k=1}^2 [\hat{H}_k^K\{y(\mathbf{x}_0)\}]^2 \sigma_k^2. \quad (2.33)$$

The disjunctive kriging variance of $\hat{Z}(\mathbf{x}_0)$ is:

$$\sigma_{DK}^2(\mathbf{x}_0) = \sum_{k=1}^2 \sigma_k^2. \quad (2.34)$$

Once the Hermite polynomials have been estimated at a target point, the conditional probability that the true value there exceeds the critical value, z_c , is calculated. The transformation $Z(\mathbf{x}) = [Y(\mathbf{x})]$ means that z_c has an equivalent y_c on the standard normal scale. The probability of exceeding the threshold is:

$$\hat{P}^{DK}[Z(\mathbf{x}_0) > z_c] = \hat{P}^{DK}[y(\mathbf{x}_0) > y_c] = G(y_c) + \frac{1}{\sqrt{k}} H_{k-1}(y_c) g(y_c) \hat{H}_k^K\{y(\mathbf{x}_0)\}, \quad (2.35)$$

The probabilities can be mapped in the same way as the estimates and the estimation variances.

2.7.4 Co-kriging

Ordinary co-kriging is the logical extension of ordinary auto-kriging to situations where two or more variables are spatially interdependent or co-regionalized. It needs a model of the co-regionalization, and this must be found first. The two regionalized variables, $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$, denoted by u and v , both have autovariograms and they also have a cross variogram defined as:

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2} E[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})\} \{Z_v(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})\}]. \quad (2.36)$$

This function describes the way in which u is related spatially to v . Provided that there are sites where both properties have been measured $\gamma_{uv}(\mathbf{h})$ can be estimated by:

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z_u(\mathbf{x}_i) - z_u(\mathbf{x}_i + \mathbf{h})\} \{z_v(\mathbf{x}_i) - z_v(\mathbf{x}_i + \mathbf{h})\}, \quad (2.37)$$

which provides the experimental cross variogram for u and v . The cross variogram can be modelled simultaneously with the autovariograms. Each variable is assumed to be a linear sum of orthogonal random variables $Y(\mathbf{x})$:

$$Z_u(\mathbf{x}) = \sum_{k=1}^K \sum_{j=1}^2 a_{uj}^k Y_j^k(\mathbf{x}), \quad (2.38)$$

in which:

$$E[Z_u(\mathbf{x})] = \mu_u$$

and:

$$\frac{1}{2} E[\{Y_j^k(\mathbf{x}) - Y_j^k(\mathbf{x} + \mathbf{h})\} \{Y_{j'}^{k'}(\mathbf{x}) - Y_{j'}^{k'}(\mathbf{x} + \mathbf{h})\}] = g_k(\mathbf{h}), \text{ positive for } k = k' \text{ and } j = j' \\ = 0 \text{ otherwise} \quad (2.39)$$

The variogram for any pair is then:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K \sum_{j=1}^2 a_{uj}^k a_{vj}^k g_k(\mathbf{h}). \quad (2.40)$$

We can replace the products in the second summation by b_{uv}^k to obtain:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K b_{uv}^k g_k(\mathbf{h}) \quad (2.41)$$

The variogram for any pair of variables u and v is:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K \sum_{j=1}^2 a_{uj}^k a_{vj}^k g_k(\mathbf{h}). \quad (2.42)$$

The b_{uv}^k are the nugget and sill variances of the independent components if they are bounded, and for unbounded models they are the nugget variances and gradients. Once the co-regionalization has been modelled it can be used to predict the spatial relations between two or more variables by co-kriging. There are generally two reasons for using co-kriging, as follows:

1. One is where one variable is undersampled compared with another with which it is correlated. The sparsely sampled property can be estimated with greater precision by co-kriging, because the spatial information from the more intensely measured one is used in the estimation. The increase in precision depends on the degree of undersampling and the strength of the co-regionalization.
2. When values of all the variables are known at all sample points, co-kriging can improve the coherence between the estimated values by taking account of the relation between them.

We used co-kriging to try to improve the estimates of sparsely available information using variables that are more readily available. If there are V variables, $l = 1, 2, \dots, V$, and the one to be predicted is u , which in our case has been less densely sampled than the others, then in ordinary co-kriging the estimate is the linear sum:

$$\hat{Z}_u(B) = \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} z_l(\mathbf{x}_i), \tag{2.43}$$

where the subscript l refers to the variables, of which there are V , and the subscript i refers to the sites, of which there are n_l where the variable l has been measured. The λ_{il} are the weights, satisfying:

$$\sum_{i=1}^{n_l} \lambda_{il} = 1, \quad l \neq u; \text{ and } \sum_{i=1}^{n_l} \lambda_{il} = 0, \quad l = u. \tag{2.44}$$

These are the non-bias conditions and, subject to them, the estimation variance of

$\hat{Z}_u(B)$ for a block, B , is minimized by solving the system of equations:

$$\begin{aligned} \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} \gamma_{vl}(\mathbf{x}_i, \mathbf{x}_j) + \sum_{v=1}^V \lambda_v \tau_{uv}(\mathbf{x}_j, B) \\ \sum_{i=1}^{n_l} \lambda_{il} = 1, \quad l = 1, \dots, V \\ \sum_{i=1}^{n_l} \lambda_{il} = 0, \quad l = u, \end{aligned} \tag{2.45}$$

for all $v=1, 2$ to V and all $j=1, 2$ to n_v . The quantity $\gamma_{lv}(\mathbf{x}_i, \mathbf{x}_j)$ is the cross semivariance between variables l and v at sites i and j , separated by the vector $\mathbf{x}_i - \mathbf{x}_j$; $\tau_{uv}(\mathbf{x}_j, B)$ is the average cross semivariance between a site j and the block B , and λ_v is the Lagrange multiplier for the v th variable. The co-kriging variance is obtained from:

$$\sigma_u^2(B) = \sum_{l=1}^V \sum_{i=1}^{n_l} \lambda_{il} \gamma_{ul}(\mathbf{x}_j, B) + \sum_{u=1}^V \lambda_u \tau_{uu}(B, B), \tag{2.46}$$

where $\tau_{uu}(B, B)$ is the integral of $\gamma_{uu}(\mathbf{h})$ over B , i.e. the within-block variance of u .

2.8 Geostatistical Simulation

Another approach to prediction in geostatistics is stochastic simulation, for which there are several different methods (Deutsch & Journel, 1992; Goovaerts, 1997). The simulated values are the outcomes of underlying stochastic processes that are chosen to represent reality. One reason for using simulation is because kriging smooths the variation: variance is lost in kriging. Therefore, if the aim is to retain the variation that is known to be present simulation enables this. The covariance or variogram functions can be used to generate any number of realizations that are as likely to occur as the actuality, and they have the same statistical characteristics. The predicted values are no longer the best estimates, but they retain the variance and provide a view of the spatial variation. Thus simulation differs from kriging because it aims to retain the overall texture of the variation and the statistics of the original data in the simulated values. This takes precedence over the accuracy of the local predictions (Goovaerts, 1997).

There is an initial subdivision of the methods into those in which the predictions are conditioned by the data, conditional simulation, and others that are not, unconditional simulation. For the NSI data we simulated conditionally only. The method uses the variogram model together with the data, and, as with punctual kriging, the generator must return the values at the places where they are known, i.e. the data points. The simulation is conditioned on the n data, and at each sampling point the simulated value must equal the observed one:

$$z_c^*(\mathbf{x}_i) = z(\mathbf{x}_i) \text{ for all } i = 1, 2, \dots, n. \quad (2.47)$$

Elsewhere, the simulated value, $z^*(\mathbf{x})$, should be in accord with the model we have adopted for the spatial dependence. Consider when we kriging Z at \mathbf{x}_0 where we have no observed value; the true value there, $z(\mathbf{x}_0)$, is estimated by $\hat{Z}(\mathbf{x}_0)$ with an error $z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)$ which is unknown:

$$z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0) = \epsilon(\mathbf{x}_0) = z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0). \quad (2.48)$$

A characteristic of kriging is that the error is independent of the estimate, i.e.

$$E[\epsilon(\mathbf{y}) \epsilon(\mathbf{x})] = 0 \text{ for all } \mathbf{x}, \mathbf{y}. \quad (2.49)$$

We make use of this to condition the simulation. We create a simulated field from the same covariance function or variogram as that of the data, but otherwise unconditioned to give values $z_s^*(\mathbf{x}_j)$, $j = 1, 2, \dots, N$, including at the sampling points (\mathbf{x}_i) , $i = 1, 2, \dots, n$. We then kriging at \mathbf{x}_0 from the simulated values at the sampling points to give an estimate $\hat{Z}_s^*(\mathbf{x}_0)$. Its error is $z_s^*(\mathbf{x}_0) - \hat{Z}_s^*(\mathbf{x}_0)$. This error comes from the same distribution as the kriging error in equation (2.48) yet the two are independent. We can use it to replace the kriging error to give the conditionally simulated value as

$$z_c^*(\mathbf{x}_0) = \hat{Z}(\mathbf{x}_0) + \epsilon_s^*(\mathbf{x}_0) = \hat{Z}_s^*(\mathbf{x}_0). \quad (2.50)$$

The outcome has the following properties.

1) The simulated values are realizations of a random process with the same expectation as the original:

$$E[Z_s^*(\mathbf{x})] = E[Z(\mathbf{x})] \quad \text{for all } \mathbf{x}, \tag{2.51}$$

where \bar{z} is the mean.

2) The simulated values have the same variogram as the original.

3) At the data points the kriging errors $z(\mathbf{x}_0) - \hat{z}(\mathbf{x}_0)$ and $z_c^*(\mathbf{x}_0) - \hat{z}_s^*(\mathbf{x}_0)$ are zero, and $z_c^*(\mathbf{x}_0) = z(\mathbf{x}_0)$.

2.8.1 Sequential Gaussian Simulation

A widely used method of conditional simulation is *sequential Gaussian simulation* (Deutsch & Journel, 1992; Goovaerts, 1997). The data are transformed first to a standard normal distribution with a mean of 0 and a variance $\sigma_K^2(\mathbf{x}_0)$. The variogram is computed from the transformed data and modelled. Predictions are then made at each node of a grid as follows:

- 1) Krige to obtain $\hat{z}(\mathbf{x}_i)$ and $\sigma_K^2(\mathbf{x}_i)$.
- 2) Draw a value at random from a normal distribution $N(\hat{z}(\mathbf{x}_i), \sigma_K^2(\mathbf{x}_i))$
- 3) Insert this value into the grid at \mathbf{x}_i , and add it to the sampling data before simulating the next point.
- 4) These steps are repeated for the entire grid.
- 5) Back-transform the simulated values.

2.8.2 Turning Bands

The simulation method of ‘Turning Bands’ due to Matheron (1973) was the earliest for simulating autocorrelated random processes in three dimensions. It is feasible in one and two dimensions, but is more complex for the latter. Isatis has been programmed to enable this. It involves first simulating independent one-dimensional realizations along lines radiating from a central point in the area of interest. Each point in the 2-D space for which a value is required is projected orthogonally on to

every line, and the values at the nearest points to the projections are averaged. The one-dimensional covariance function, $C^1(\mathbf{h})$, corresponding to that in two dimensions, $C^2(\mathbf{h})$, must be known. It is this that is more difficult in two dimensions than it is in three.

2.8.2.1 Generating the turning bands. We define a set of L lines, D_1, D_2, \dots, D_L , $l=1, 2, \dots, L$, in the two dimensional region, R . These radiate from a point at the centre of R and are equally spaced on a circle about it, i.e. their angular separations are constant. Values of the autocorrelated random process are simulated at equal intervals along each line independently with the covariance function $C^1(\mathbf{h})$ appropriate for the two-dimensional covariance, $C^2(\mathbf{h})$.

For any point, \mathbf{x}_0 , in R for which a simulated value is required \mathbf{x}_0 is projected on to line D_l (Figure 2.2). Its position is denoted as x_{0Dl} and a value, $z(x_{0Dl})$, assigned to it at the nearest simulation point on that line. The projections, u_l , are repeated on all lines L and the realization computed by

$$z_s^*(x_0) = \frac{1}{\sqrt{L}} \sum_{l=1}^L z(x_{0Dl}). \quad (2.52)$$

2.8.2.2 Conditioning. The turning bands method creates an unconditional simulation. To condition on the data involves a final stage. Kriging is used with the appropriate autocorrelation function to combine the original and simulated values to produce the conditional simulation. The conditionally simulated values are obtained as:

$$z_c(\mathbf{x}_0) = z_s(\mathbf{x}_0) + \{z_s^*(\mathbf{x}_0) - z^*(\mathbf{x}_0)\}, \quad (2.53)$$

where, $z_s(\mathbf{x})$ is a simulated value at \mathbf{x} from the turning bands, $z_s^*(\mathbf{x})$ is an estimate kriged from the simulated values, and $z^*(\mathbf{x})$ is a kriged estimate from the actual values $z(\mathbf{x}_i)$. This final stage is done automatically in Isatis.

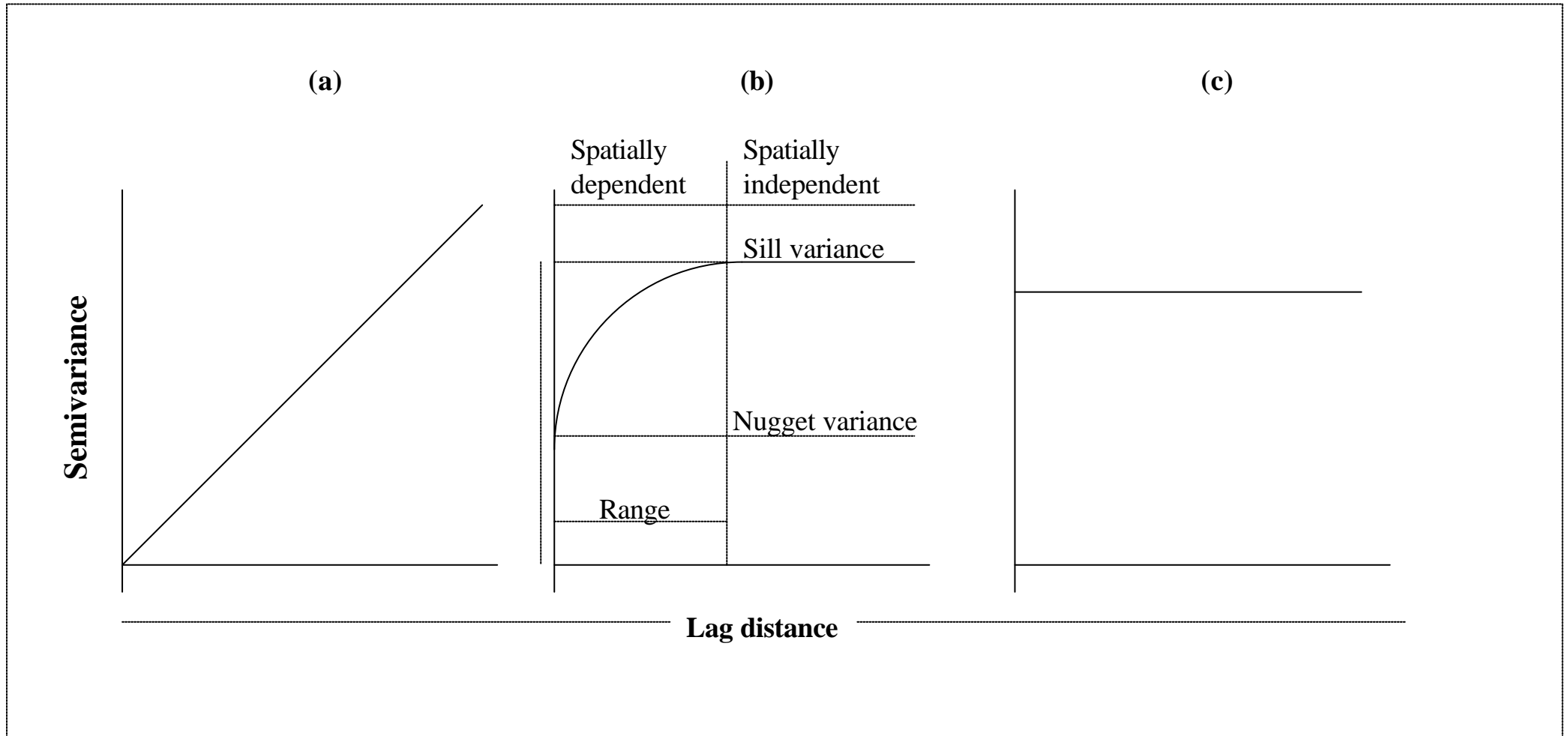


Figure 2.1: Forms of variogram (a) Unbounded (b) Bounded (c) Pure nugget

