SAMPLE DATA AND SURVEY

M. A. Oliver
Department of Soil Science, University of Reading, UK

Keywords: Geostatistics, kriging, nested variation, sampling, survey, variogram

Contents

Summary

The requirements for high quality data are increasing as environmental monitoring and analysis become more sophisticated. For example, geographical information systems (GIS) and geostatistics provide valuable tools for spatial analysis, but the accuracy of the results depends on the reliability of the data that they use. This, in turn, depends on sound survey and sampling. There is a need to decide at the outset what kinds of predictions are required. If global or class means are the chosen predictors, then the sampling scheme should be design-based with randomized sampling locations. For local estimation by kriging or other methods of interpolation the sampling should be model-based. Geostatistical methods also provide a rationale to optimize sampling whether the data are to be analyzed by classical statistics or geostatistics.

The case studies illustrate how different kinds of sampling schemes can be used to determine the spatial scale of variation and hence guide future sampling, how to predict values depending on the nature of the data, and how to design optimal sampling schemes once the variogram is known.

1. Introduction

Features of the environment, such as the soil, atmosphere, landform, oceans, rocks, groundwater, and so on, have properties that vary from place to place and through time. Describing and analyzing variation over the Earth are central to geoinformatics. The emphasis in this Topic-level will be on the spatial component of variation. The variation of interest might be in one or two dimensions, laterally or vertically, or in three dimensions. Here the focus is on lateral changes in one and two dimensions, but the principles can be applied to vertical and three-dimensional variation. The case studies in the second part of this Topic-level are examples from soil science, but the methods can be applied to most branches of environmental science.

Environmental features cover large areas and their spatial variation results in what we recognize as different types of vegetation, rocks, soil, and climate, for example. The transition between different kinds of rock is usually sharp such that discrete units occur with sharp boundaries between. Nevertheless, there is often continuous variation within these units. For the soil and atmosphere, for example, variation is more continuous; the transition between different types of soil and climate can be sharp but it is more likely to be gradual. The steep slope in Figure 1 represents a discontinuity where the value of the property changes rapidly over a short distance. We could regard the two sections of the diagram on either side of the discontinuity as different types of rock or soil, within each of which the property varies continuously.

Environmental properties can also vary at spatial scales that range from microns to hundreds of kilometres, i.e. over several orders of magnitude simultaneously. This is widespread in the environment. For example, seven different scales of spatial variation (nested variation) were identified in the Lorraine iron ore deposit, ranging from 15 μm to several hundred metres. This can arise from the variation of a single process operating at several different scales or from the interaction of several independent processes that operate at different characteristic spatial scales. For example, the observed variation in soil arises from the effects of climate, geology, physiography, hydrology, trees, earthworms, micro biota and so on. As a result patterns in the variation of properties can occur at many levels of spatial scale, one nested within another.

In general, it is impossible to observe and or record environmental properties at all places. Even within a small drainage basin there will be an infinitely large number of possible locations at which observations could be made. Information about environmental properties usually derives from small areas (supports) at selected locations (sampling points) that are separated by much larger areas. Therefore, the detail observed in the variation will depend on the extent to which the intensity of the observations resolves the variation at the level of interest. The latter might be an experimental plot, an agricultural field, a drainage basin, a river system, a mountain range, a country, etc. Since the sample information is generally sparse there are

intervening spaces about which nothing is known and for which predictions are required. The complexity of environmental variation, however, makes prediction difficult.

In spite of the inherent complexity of environmental variation experience has shown that the values of most spatial properties are more similar at places that are close together than those at places that are further apart. In other words there is some spatial correlation in the values at sites that are close together. This relation gives rise to pattern or underlying structure in the variation, which we often want to identify and describe. It can provide clues about the causes of variation, aid prediction and indicate areas that might require different forms of management. It is also important to note that what we observe as structure or spatially correlated variation at one spatial scale or level of resolution can appear as "noise", or uncorrelated variation, at another. Figure 1 shows four possible scales of spatial variation superimposed on one another. That defined by the two classes (on either side of the steep slope) might be over kilometres. Within the classes there are intermediate scales of variation: over hundreds of metres in the right-hand class and over tens of metres in the left-hand one. Superimposed on both of these is the very local variation over distances of less than a few metres. If observations were made at distances greater than the extent of the intermediate scales of variation only the difference between the two classes would be identified. The variation remaining within the classes would be regarded as noise. If the observations were intensive enough to resolve the two intermediate scales of variation, then only the very local variation would appear as noise.

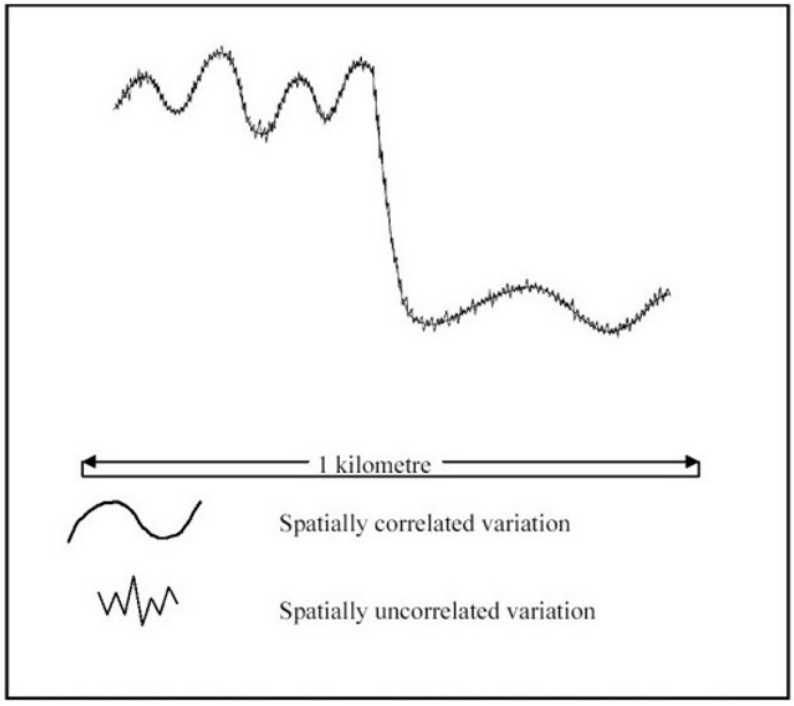

Figure 1: Different scales and types of spatial variation: the horizontal axis represents distance and the vertical one the values of some environmental property.

This Topic includes four Articles: Biogeography, Landform and Earth Surface, Land Hydrology and Geology. This list is necessarily a sample of the wide range of subjects that might be included within the general field of sample data and survey of relevance to geoinformatics and life support systems. The Articles were selected to cover the main domains that affect life support systems (e.g., biosphere, hydrosphere, geosphere). The treatment in each Article is rather different, reflecting each author's preferences. However, each Article covers the basic instrumentation and techniques required for measurement in the field. Such measurement requires some decision over sampling, either implicitly or explicitly (e.g., through sample design). Thus, this Topic-level sets the scene for the Articles that follow in this Topic, with an emphasis on sample data and the theory underlying survey.

2. Survey

Surveys are the framework within which environmental scientists obtain information to describe and analyze the phenomena of interest over a given area. The term survey has two broad meanings in the spatial context. Conventionally, it is the basis for mapping features of the environment in an area, such as the soil, geology, physiography, vegetation, etc. The aim is to define areas that are similar, i.e. they have minimum variation within them, and are delimited by boundaries where there is more marked change in several properties simultaneously or even abrupt change as in Figure 1. Such surveys tend to be general purpose and provide largely qualitative information. Increasingly, surveys are the basis for obtaining and recording georeferenced information about specific features and properties. For example, farmers do specific surveys to obtain information on the nutrient status of their fields for managing fertilizer applications. The resulting information may be quantitative or qualitative, or both. The data from such surveys can then be analyzed in various ways that need not only result in mapping (see *Landforms and Earth Surface*).

Surveys have been associated traditionally with obtaining information from sample locations, but there is a growth in survey information from sensors. These can provide complete cover of information, for example remotely sensed data from satellites or ground based radiometers, electromagnetic scans, crop yield monitors, radar and so on. They are being used to provide information on ground cover, elevation (LiDAR, see *Landforms and Earth Surface*), the soil, crop health, rainfall, and may other components of the environment. Such data are described elsewhere in *Landforms and Earth Surface*. The focus in this chapter will be mainly on surveys to obtain information from sampling.

3. Spatial Sampling

There is no single optimal approach to sampling, but it should be planned with care because sound sampling is central to the accurate estimation or prediction of properties for areas of any size. Poorly designed and inadequate sampling can lead to biased predictions with large errors which, in turn, can have consequences for decision-making. Sampling should also be efficient and this requires information on the variation of the properties. For the classical approach to estimation, knowing the variance or standard deviation of one or more properties can increase the efficiency of sampling.

For the geostatistical approach to prediction, the spatial correlation structures of the properties should be known. The aim of sampling is to reveal information and to enable meaningful statements to be made about the population with confidence.

At the outset before planning sampling, consideration must be given to the following:

1) At what spatial scale should we investigate environmental properties?
2) What is the sample support?
3) What is our population?
4) What kind of sampling scheme should we use?
5) How many samples should we take?
6) What should the interval between sampling locations be?
7) How should we predict values at intervening places?

First, we must define the domain, *D*, of interest so that every point can be assigned to it or not with certainty. This will also define the level of interest and the scale of investigation, such as an experimental plot, a field, or an entire country. The area or volume of material on which observations or measurements are made must also be defined. This is the sample support, which has size, shape and orientation. For example, it is the quadrat in vegetation surveys, the core of soil taken from the ground in soil survey, or the volume of water taken from a river, lake or sea, and so on. In general, the support is very small compared to the region being investigated. The size of the support is often increased by taking several samples from a given area and mixing them to produce a bulked sample. This can diminish the sampling effects that are associated with taking a single sample from each location and it ensures that the information from a site is representative of the surrounding area.

Within *D* are the units that have the dimensions of the supports; this is the population. If only certain areas within *D* are of interest, the units falling within the latter constitute the target population. The domain and population should be defined for both spatial and non-spatial statistical analyses. In general, a subset of the units in the population is selected – this is the sample.

The kind of sampling scheme chosen will depend on the approach chosen to predict values of the properties of interest at unsampled places. For instance, will the mean values of the properties observed for the entire area or for strata within the area be used for estimation? This will require a design-based sampling scheme. Or will the information be used to predict locally, either using mathematical interpolators or geostatistical ones? For this model-based sampling should be used.

3.1. Design-based Sampling Schemes and Estimation

Design-based sampling is essentially the classical statistical approach to sample design, which aims to estimate the population parameters, such as the mean and variance, without bias. The population is the set of all units of interest. The measured values for the property of interest can be written as $z(\mathbf{X}_i)$, $i = 1,2,\ldots, N$, where $\mathbf{X}_i$ represents random locations. The probability of selecting any site is determined by the sampling design. The randomization of the sampling provides a probabilistic basis for inference.

Excluding measurement error, the only variation that plays a role is that resulting from the sampling process. For design-based sampling the number of samples is more important than their geographical location or spacing.

3.1.1. Simple Random Sampling

$$s^2(\mathbf{X}_0) = s^2 + s^2 / N, \tag{1}$$

In simple random sampling the N units are chosen with equal probability from the target population in D. The mean and the variance of the data, z, are unbiased. The global mean, $\bar{z}$, can be used to predict z at a specified location, $\mathbf{X}_0$, or at all places in the region. The estimation variances are different, however. For a point the estimation variance is

and for the mean within D it is

$$s^2(D) = s^2 / N. \tag{2}$$

where s^2 is the sample variance.

The estimation variance and its square root, the standard error, depend on the number of individuals, N, in the sample. For a given error, $s(D)$, that can be tolerated in the estimate from the survey at the 95% confidence level, the size of sample can be calculated from

$$N = (1.96s)^2/s^2(D) . \tag{3}$$

The size of sample is likely to be large and it will increase as the variance increases. The efficiency of sampling can be increased by stratification.

3.1.2. Stratified Random Sampling

The region is divided into strata, D_k, k=1,2, …, K, each of which is represented by a few units chosen at random. If the strata are equal in area and contain the same number of sample points, the mean, $\bar{z}$, of all observations estimates the population mean without bias. If other sizes are chosen then the mean in D is calculated as the weighted average of the individual stratum means, with weights proportional to the $|D_k|$. The estimation variance of stratified sampling depends on the variance within the strata, or the pooled within-stratum variance. It is given by

$$s^2(D)_{\text{stratified}} = \sum_{k=1}^{K} w_k^2 s^2(D_k), \tag{4}$$

where $s^2(D_k)$ is the estimation variance within stratum D_k, and w_k is the weight assigned to the stratum. The weights should sum to 1 to avoid bias.

This scheme can be elaborated depending upon what is known about the region and the variation within it. For example, the strata could have unequal spatial extents and different numbers of individuals per stratum. To combine the data from several strata we assign a weight w_k to each stratum such that

$$w_k = \frac{\text{area of stratum } k}{\text{total area}}. \tag{5}$$

The estimation variance is then given by

$$s^2(D)_{\text{stratified}} = \sum_{k=1}^{K} \frac{w_k^2 s^2(D_k)}{n_k}, \tag{6}$$

where n_k is the number of observations per stratum. Stratification is also the basis of classification, which has been the classical approach to prediction in many environmental sciences. In this case the strata are not arbitrary, but have been identified as different types of rock or soil, for example. The estimator is a simple weighted sum of the n data in a class

$$\overline{z}_k = \frac{1}{n_k} \sum_{k=j}^{n_k} w_j z_j, \tag{7}$$

where w_j is the weight of the jth sampling point in class k. For classification the weights for a class are given by

$$w_j = \frac{1}{n_k} \qquad \text{if } j \in k, \tag{8}$$

otherwise $w_j = 0$.

If the classification accounts for all of the spatially correlated variation then the class mean is the best estimate of the property and one can do no better. However, if there is spatially dependent variation remaining within the classes then local detail is lost (see Figure 1, for example). For any stratification (classification) the precision of the estimates depends on the degree of subdivision in the population.

3.1.3. Systematic Sampling

Sampling is usually most efficient when done on a regular grid. There are two disadvantages, however: it provides no ready estimate of the estimation variance and it can lead to biased estimates of the mean. The first arises because once the origin and orientation of the grid are decided no further randomization is possible. The estimation variances may be approximated, however, by a method such as Yates' balanced differences. Bias can arise where there is trend or periodicity in z in the region. Periodicity is usually evident and an interval and orientation that is "out of tune" with it can be chosen. Alternatively a non-aligned scheme can be used in which each sampling

point on the grid is offset from its node by a random distance along its row and down its column according to a predefined rule.

3.1.4. Nested Sampling

The nested sampling scheme for spatial data is an adaptation of classical multi-stage sampling. The initial aim was to increase the efficiency of replicate sampling based on knowing the separating distance at which most of the variation occurred. It is a design-based approach to sampling, but because it can be used to produce a first approximation to the variogram the method is described in detail later in section 3.3.

-
-
-

Bibliography

Burrough, P. A. (1991). Sampling designs for quantifying map unit composition. In: Spatial Variabilities of Soils and Landforms, SSSA Special Publication Number 28, Madison, Wisconsin: Soil Science Society of America, Inc.. pp. 89-125. [A general paper on the principles of sampling and the analysis of spatial data].

De Gruijter, J. J. and Ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology* **2**, 407-415. [This paper describes the principles of design-based and model-based sampling and estimation].

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, 483 pp. New York: Oxford University Press. [A valuable intermediate text on spatial analysis of environmental properties].

Gower, J. C. (1962). Variance component estimation for unbalanced hierarchical classification. *Biometrics* **18**, 537-542. [This paper gives the method for computing the components of variance for an unbalanced hierarchical sampling design].

Journel, A. G. and Huijbrechts, C. J. (1978). *Mining Geostatistics*, 600 pp. London: Academic Press. [The first major text on the theory and application of regionalized variable theory. It is based on Georges Matheron's seminal work that brought together many isolated ideas on analysing spatial variation into a coherent body of theory].

McBratney, A. B., Webster, R, and Burgess, T. M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables. *Computers and Geosciences* **7**, 331-334. [This text describes in detail how the variogram can be used with the kriging equations to design sampling to meet a specified tolerable error].

Miesch, A. T. (1975). Variograms and variance components in geochemistry and ore evaluation. *Quantitative Studies in the Geological Sciences* (ed. E. H. T. Whitten), pp 333-340. Geological Society of America Memoir 142. [This text shows how the accumulated components from analysis of variance relate to the semivariances of regionalized variable theory].

Oliver, M. A., Frogbrook, Z., Webster, R., Dawson, C. J. (1997). A rational strategy for determining the number of cores for bulked sampling of soil. *Precision Agriculture '97,Proceedings of the 1st European Conference on Precision Agriculture*, J. V. Stafford (Ed.), BIOS Scientific Publishers Ltd., Oxford, pp. 155-162. [Describes a geostatistical approach to determine the number of samples to bulk from for a given sampling support].

Oliver, M. A., Badr, I. (1995) Determining the spatial scale of variation in soil radon concentration. *Mathematical Geology* **27**, 893-922. [This paper shows the difficulties of designing a suitable sampling scheme to describe an environmental property when there is no prior information about the possible scale(s) of spatial variation].

Oliver, M. A., Webster, R. (1987). The elucidation of soil pattern in the Wyre Forest of the West Midlands, England. II. Spatial distribution. *Journal of Soil Science* **38**, 293-307. [A description of the sampling and analysis to explore soil spatial variation using the reconnaissance and conventional variogram].

Oliver, M. A., Webster, R. and Slocum, K. (2000). Filtering SPOT imagery by kriging analysis. *International Journal of remote Sensing* **21**, 735-752. [This paper describes nested variation, the theory of factorial kriging, and shows how imagery can be used to guide sampling for ground surveys with little prior information].

Webster, R., and Oliver, M.A., (1990). *Statistical Methods in Soil and Land Resource Survey*, 316 pp. Oxford: Oxford University Press. [A basic text that describes different sampling designs, and a range of statistical and geostatistical analyses].

Webster, R. and Oliver, M. A. (1992). Sample adequately to estimate the variograms of soil properties. *Journal of Soil Science* **43**, 177-192. [This paper uses a Monte Carlo approach to determine how many samples are needed from which to estimate reliable variograms].

Webster, R., and Oliver, M.A., (2001) *Geostatistics for Environmental Scientists*, 271 pp. Chichester: John Wiley & Sons. [An introductory geostatistics text aimed at environmental scientists].

Youden, W. J., Mehlich, A. (1937). Selection of efficient methods for soil sampling. *Contributions of the Boyce Thompson Institute for Plant Research* **9**, 59-70. [The original application of multi-stage sampling in the spatial context].

Biographical Sketch

Margaret Oliver is Reader in Spatial Analysis in the Department of Soil Science at the University of Reading, UK. She has a strong research team working on the application of geostatistics to remotely sensed data, precision agriculture, soil-landscape variation, the spatial variation in meadow species richness, and malaria in conjunction with soil physical properties. In addition, she teaches modules on pedology, multivariate analysis and geostatistics for soil and environmental scientists.