1 Introduction

1.1 WHY GEOSTATISTICS?

Imagine the situation: a farmer has asked you to survey the soil of his farm. In particular, he wants you to determine the phosphorus content; but he will not be satisfied with the mean value for each field as he would have been a few years ago. He now wants more detail so that he can add fertilizer only where the soil is deficient, not everywhere. The survey involves taking numerous samples of soil, which you must transport to the laboratory for analysis. You dry the samples, crush them, sieve them, extract the phosphorus with some reagent and finally measure it in the extracts. The entire process is both time-consuming and costly. Nevertheless, at the end you have data from all the points from which you took the soil—just what the farmer wants, you might think!

The farmer's disappointment is evident, however. 'Oh', he says, 'this information is for a set of points, but I have to farm continuous tracts of land. I really want to know how much phosphorus the soil contains everywhere. I realize that that is impossible; nevertheless, I should really like some information at places between your sampling points. What can you tell me about those, and how do your small cores of soil relate to the blocks of land over which my machinery can spread fertilizer, that is, in bands 24 m wide?'

This raises further issues that you must now think about. Can you say what values to expect at intervening places between the sample points and over blocks the width of the farmer's fertilizer spreader? And how densely should you sample for such information to be reliable? At all times you must consider the balance between the cost of providing the information and the financial gains that will accrue to the farmer by differential fertilizing. In the wider context there may be an additional gain if you can help to avoid over-fertilizing and thereby protect the environment from pollution by excess phosphorus. Your task, as a surveyor, is to be able to use sparse affordable data to estimate, or predict, the average values of phosphorus in the soil over blocks of land 24 m × 24 m or perhaps longer strips. Can you provide the farmer with spatially referenced values that he can use in his automated fertilizer spreader?

Geostatistics for Environmental Scientists/2nd Edition R. Webster and M.A. Oliver

This is not fanciful. The technologically minded farmer can position his machines accurately to 2 m in the field, he can measure and record the yields of his crops continuously at harvest, he can modulate the amount of fertilizer he adds to match demand; but providing the information on the nutrient status of the soil at an affordable price remains a major challenge in modern precision farming (Lake *et al.*, 1997).

So, how can you achieve this? The answer is to use geostatistics—that is what it is for.

We can change the context to soil salinity, pollution by heavy metals, arsenic in ground water, rainfall, barometric pressure, to mention just a few of the many variables and materials that have been and are of interest to environmental scientists. What is common to them all is that the environment is continuous, but in general we can afford to measure properties at only a finite number of places. Elsewhere the best we can do is to estimate, or predict, in a spatial sense. This is the principal reason for geostatistics—it enables us to do so without bias and with minimum error. It allows us to deal with properties that vary in ways that are far from systematic and at all spatial scales.

We can take the matter a stage further. Alert farmers and land managers will pounce on the word 'error'. 'Your estimates are subject to error', they will say, 'in other words, they are more or less wrong. So there is a good chance that if we take your estimates at face value we shall fertilize or remediate where we need not, and waste money, because you have underestimated, and not fertilize or fail to remediate where we should.' The farmer will see that he might lose yield and profit if he applies too little fertilizer because you overestimate the nutrient content of the soil; the public health authority might take too relaxed an attitude if you underestimate the true value of a pollutant. 'What do you say to that?', they may say.

Geostatistics again has the answer. It can never provide complete information, of course, but, given the data, it can enable you to estimate the probabilities that true values exceed specified thresholds. This means that you can assess the farmer's risks of losing yield by doing nothing where the true values are less than the threshold or of wasting money by fertilizing where they exceed it.

Again, there are analogies in many fields. In some situations the conditional probabilities of exceeding thresholds are as important as the estimates themselves because there are matters of law involved. Examples include limits on the arsenic content of drinking water (what is the probability that a limit is exceeded at an unsampled well?) and heavy metals in soil (what is the probability that there is more cadmium in the soil than the statutory maximum?)

1.1.1 Generalizing

The above is a realistic, if colourful, illustration of a quite general problem. The environment extends more or less continuously in two dimensions. Its properties have arisen as the result of the actions and interactions of many different processes and factors. Each process might itself operate on several scales simultaneously, in a non-linear way, and with local positive feedback. The environment, which is the outcome of these processes varies from place to place with great complexity and at many spatial scales, from micrometres to hundreds of kilometres.

The major changes in the environment are obvious enough, especially when we can see them on aerial photographs and satellite imagery. Others are more subtle, and properties such as the temperature and chemical composition can rarely be seen at all, so that we must rely on measurement and the analysis of samples. By describing the variation at different spatial resolutions we can often gain insight into the processes and factors that cause or control it, and so predict in a spatial sense and manage resources.

As above, measurements are made on small volumes of material or areas a few centimetres to a few metres across, which we may regard as point samples, known technically as *supports*. In some instances we enlarge the supports by taking several small volumes of material and mixing them to produce bulked samples. In others several measurements might be made over larger areas and averaged rather than recorded as single measurements. Even so, these supports are generally very much smaller than the regions themselves and are separated from one another by distances several orders of magnitude larger than their own diameters. Nevertheless, they must represent the regions, preferably without bias.

An additional feature of the environment not mentioned so far is that at some scale the values of its properties are positively related—*autocorrelated*, to give the technical term. Places close to one another tend to have similar values, whereas ones that are farther apart differ more on average. Environmental scientists know this intuitively. Geostatistics expresses this intuitive knowledge quantitatively and then uses it for prediction. There is inevitably error in our estimates, but by quantifying the spatial autocorrelation at the scale of interest we can minimize the errors and estimate them too.

Further, as environmental protection agencies set maximum concentrations, thresholds, for noxious substances in the soil, atmosphere and water supply, we should also like to know the probabilities, given the data, that the true values exceed the thresholds at unsampled places. Farmers and graziers and their advisers are more often concerned with nutrients in the soil and the herbage it grows, and they may wish to know the probabilities of deficiency, i.e. the probabilities that true values are less than certain thresholds. With some elaboration of the basic approach geostatistics can also answer these questions.

The reader may ask in what way geostatistics differs from the classical methods that have been around since the 1930s; what is the effect of taking into account the spatial correlation? At their simplest the classical estimators, based on random sampling, are linear sums of data, all of which carry the same weight. If there is spatial correlation, then by stratifying we can estimate more precisely or sample more efficiently or both. If the strata are of different sizes then we might vary the weights attributable to their data in proportion. The means and their variances provided by the classical methods are regional, i.e. we obtain just one mean for any region of interest, and this is not very useful if we want local estimates. We can combine classical estimation with stratification provided by a classification, such as a map of soil types, and in that way obtain an estimate for each type of class separately. Then the weights for any one estimate would be equal for all sampling points in the class in question and zero in all others. This possibility of local estimation is described in Chapter 3. In linear geostatistics the predictions are also weighted sums of the data, but with variable weights determined by the strength of the spatial correlation and the configuration of the sampling points and the place to be estimated.

Geostatistical prediction differs from classical estimation in one other important respect: it relies on spatial models, whereas classical methods do not. In the latter, survey estimates are put on a probabilistic footing by the design of the sampling into which some element of randomization is built. This ensures unbiasedness, and provides estimates of error if the choice of sampling design is suitable. It requires no assumptions about the nature of the variable itself. Geostatistics, in contrast, requires the assumption that the variable is random, that the actuality on the ground, in the sea or in the air is the outcome of one or more random processes. The models on which predictions are based are of these random processes. They are not of the data, nor even of the actuality that we could observe completely if we had infinite time and patience. Newcomers to the subject usually find this puzzling; we hope that they will no longer do so when they have read Chapter 4, which is devoted to the subject. One consequence of the assumption is that sampling design is less important than in classical survey; we should avoid bias, but otherwise even coverage and sufficient sampling points are the main considerations.

The desire to predict was evident in weather forecasting and soil survey in the early twentieth century, to mention just two branches of environmental science. However, it was in mining and petroleum engineering that such a desire was matched by the financial incentive and resources for research and development. Miners wanted to estimate the amounts of metal in ore bodies and the thicknesses of coal seams, and petroleum engineers wanted to know the positions and volumes of reservoirs. It was these needs that constituted the force originally driving geostatistics because better predictions meant larger profits and smaller risks of loss. The solutions to the problems of spatial estimation are embodied in geostatistics and they are now used widely in many branches of science with spatial information. The origins of the subject have also given it its particular flavour and some of its characteristic terms, such as ‘nugget’ and ‘kriging’.

There are other reasons why we might want geostatistics. The main ones are description, explanation and control, and we deal with them briefly next.

1.1.2 Description

Data from classical surveys are typically summarized by means, medians, modes, variances, skewness, perhaps higher-order moments, and graphs of the cumulative frequency distribution and histograms and perhaps box-plots. We should summarize data from a geostatistical survey similarly. In addition, since geostatistics treats a set of spatial data as a sample from the realization of a random process, our summary must include the spatial correlation. This will usually be the experimental or sample variogram in which the variance is estimated at increasing intervals of distance and several directions. Alternatively, it may be the corresponding set of spatial covariances or autocorrelation coefficients. These terms are described later. We can display the estimated semivariances or covariances plotted against sample spacing as a graph. We may gain further insight into the nature of the variation at this stage by fitting models to reveal the principal features. A large part of this book is devoted to such description.

In addition, we must recognize that spatial positions of the sampling points matter; we should plot the sampling points on a map, sometimes known as a 'posting'. This will show the extent to which the sample fills the region of interest, any clustering (the cause of which should be sought), and any obvious mistakes in recording the positions such as reversed coordinates.

1.1.3 Interpretation

Having obtained the experimental variogram and fitted a model to it, we may wish to interpret them. The shape of the points in the experimental variogram can reveal much at this stage about the way that properties change with distance, and the adequacy of sampling. Variograms computed for different directions can show whether there is anisotropy and what form it takes. The variogram and estimates provide a basis for interpreting the causes of spatial variation and for identifying some of the controlling factors and processes. For example, Chappell and Oliver (1997) distinguished different processes of soil erosion from the spatial resolutions of the same soil properties in two adjacent regions with different physiography. Burrough *et al.* (1985) detected early field drains in a field in the Netherlands, and Webster *et al.* (1994) attempted to distinguish sources of potentially toxic trace metals from their variograms in the Swiss Jura.

1.1.4 Control

The idea of controlling a process is often central in time-series analysis. In it there can be a feedback such that the results of the analysis are used to change the process itself. In spatial analysis the concept of control is different. In many instances we are unlikely to be able to change the spatial characteristics of a process; they are given. But we may modify our response. Miners use the results of analysis to decide whether to send blocks of ore for processing if the estimated metal content is large enough or to waste if not. They may also use the results to plan the siting of shafts and the expansion of mines. The modern precision farmer may use estimates from a spatial analysis to control his fertilizer spreader so that it delivers just the right amount at each point in a field.

1.2 A LITTLE HISTORY

Although mining provided the impetus for geostatistics in the 1960s, the ideas had arisen previously in other fields, more or less in isolation. The first record appears in a paper by Mercer and Hall (1911) who had examined the variation in the yields of crops in numerous small plots at Rothamsted. They showed how the plot-to-plot variance decreased as the size of plot increased up to some limit. ‘Student’, in his appendix to the paper, was even more percipient. He noticed that yields in adjacent plots were more similar than between others, and he proposed two sources of variation, one that was autocorrelated and the other that he thought was completely random. In total, this paper showed several fundamental features of modern geostatistics, namely spatial dependence, correlation range, the support effect, and the nugget, all of which you will find in later chapters. Mercer and Hall’s data provided numerous budding statisticians with material on which to practise, but the ideas had little impact in spatial analysis for two generations.

In 1919 R. A. Fisher began work at Rothamsted. He was concerned primarily to reveal and estimate responses of crops to agronomic practices and differences in the varieties. He recognized spatial variation in the field environment, but for the purposes of his experiments it was a nuisance. His solution to the problems it created was to design his experiments in such a way as to remove the effects of both short-range variation, by using large plots, and long-range variation, by blocking, and he developed his analysis of variance to estimate the effects. This was so successful that later agronomists came to regard spatial variation as of little consequence.

Within 10 years Fisher had revolutionized agricultural statistics to great advantage, and his book (Fisher, 1925) imparted much of his development of the subject. He might also be said to have hidden the spatial effects and therefore to have held back our appreciation of them. But two agronomists, Youden and Mehlich (1937), saw in the analysis of variance a tool for revealing and estimating spatial variation. Their contribution was to adapt Fisher’s concepts so as to analyse the spatial scale of variation, to estimate the variation from different distances, and then to plan further sampling in the light of the knowledge gained. Perhaps they did not appreciate the significance of their research, for they published it in the house journal of their institute, where their paper lay dormant for many years. The technique had to be rediscovered not once but several times by, for example, Krumbein and Slack (1956) in geology, and Hammond *et al.* (1958) and Webster and Butler (1976) in soil science. We describe it in Chapter 6.

We next turn to Russia. In the 1930s A. N. Kolmogorov was studying turbulence in the air and the weather. He wanted to describe the variation and to predict. He recognized the complexity of the systems with which he was dealing and found a mathematical description beyond reach. Nowadays we might call it chaos (Gleick, 1988). However, he also recognized spatial correlation, and he devised his 'structure function' to represent it. Further, he worked out how to use the function plus data to interpolate optimally, i.e. without bias and with minimum variance (Kolmogorov, 1941); see also Gandin (1965). Unfortunately, he was unable to use the method for want of a computer in those days. We now know Kolmogorov's structure function as the variogram and his technique for interpolation as kriging. We deal with them in Chapters 4 and 8, respectively.

The 1930s saw major advances in the theory of sampling, and most of the methods of design-based estimation that we use today were worked out then and later presented in standard texts such as Cochran's *Sampling Techniques*, of which the third edition (Cochran, 1977) is the most recent, and that by Yates, which appeared in its fourth edition as Yates (1981). Yates's (1948) investigation of systematic sampling introduced the semivariance into field survey. Von Neumann (1941) had by then already proposed a test for dependence in time series based on the mean squares of successive differences, which was later elaborated by Durbin and Watson (1950) to become the Durbin–Watson statistic. Neither of these leads were followed up in any concerted way for spatial analysis, however.

Matérn (1960), a Swedish forester, was also concerned with efficient sampling. He recognized the consequences of spatial correlation. He derived theoretically from random point processes several of the now familiar functions for describing spatial covariance, and he showed the effects of these on global estimates. He acknowledged that these were equivalent to Jowett's (1955) 'serial variation function', which we now know as the variogram, and mentioned in passing that Langsaetter (1926) had much earlier used the same way of expressing spatial variation in Swedish forest surveys.

The 1960s bring us back to mining, and to two men in particular. D. G. Krige, an engineer in the South African goldfields, had observed that he could improve his estimates of ore grades in mining blocks if he took into account the grades in neighbouring blocks. There was an autocorrelation, and he worked out empirically how to use it to advantage. It became practice in the gold mines. At the same time G. Matheron, a mathematician in the French mining schools, had the same concern to provide the best possible estimates of mineral grades from autocorrelated sample data. He derived solutions to the problem of estimation from the fundamental theory of random processes, which in the context he called the theory of regionalized variables. His doctoral thesis (Matheron, 1965) was a *tour de force*.

From mining, geostatistics has spread into several fields of application, first into petroleum engineering, and then into subjects as diverse as hydrogeology, meteorology, soil science, agriculture, fisheries, pollution, and environmental protection. There have been numerous developments in technique, but Matheron's thesis remains the theoretical basis of most present-day practice.

1.3 FINDING YOUR WAY

We are soil scientists, and the content of our book is inevitably coloured by our experience. Nevertheless, in choosing what to include we have been strongly influenced by the questions that our students, colleagues and associates have asked us and not just those techniques that we have found useful in our own research. We assume that our readers are numerate and familiar with mathematical notation, but not that they have studied mathematics to an advanced level or have more than a rudimentary understanding of statistics.

We have structured the book largely in the sequence that a practitioner would follow in a geostatistical project. We start by assuming that the data are already available. The first task is to summarize them, and Chapter 2 defines the basic statistical quantities such as mean, variance and skewness. It describes frequency distributions, the normal distribution and transformations to stabilize the variance. It also introduces the chi-square distribution for variances. Since sampling design is less important for geostatistical prediction than it is in classical estimation, we give it less emphasis than in our earlier *Statistical Methods* (Webster and Oliver, 1990). Nevertheless, the simpler designs for sampling in a two-dimensional space are described so that the parameters of the population in that space can be estimated without bias and with known variance and confidence. The basic formulae for the estimators, their variances and confidence limits are given.

The practitioner who knows that he or she will need to compute variograms or their equivalents, fit models to them, and then use the models to krige can go straight to Chapters 4, 5, 6 and 8. Then, depending on the circumstances, the practitioner may go on to kriging in the presence of trend and factorial kriging (Chapter 9), or to cokriging in which additional variables are brought into play (Chapter 10). Chapter 11 deals with disjunctive kriging for estimating the probabilities of exceeding thresholds.

Before that, however, newcomers to the subject are likely to have come across various methods of spatial interpolation already and to wonder whether these will serve their purpose. Chapter 3 describes briefly some of the more popular methods that have been proposed and are still used frequently for prediction, concentrating on those that can be represented as linear sums of data. It makes plain the shortcomings of these methods. Soil scientists are generally accustomed to soil classification, and they are shown how it can be combined with classical estimation for prediction. It has the merit of being the only means of statistical prediction offered by classical theory. The chapter also draws attention to its deficiencies, namely the quality of the classification and its inability to do more than predict at points and estimate for whole classes.

The need for a different approach from those described in Chapter 3, and the logic that underpins it, are explained in Chapter 4. Next, we give a brief description of regionalized variable theory or the theory of spatial random processes upon which geostatistics is based. This is followed by descriptions of how to estimate the variogram from data. The usual computing formula for the sample variogram, usually attributed to Matheron (1965), is given and also that to estimate the covariance.

The sample variogram must then be modelled by the choice of a mathematical function that seems to have the right form and then fitting of that function to the observed values. There is probably not a more contentious topic in practical geostatistics than this. The common simple models are listed and illustrated in Chapter 5. The legitimate ones are few because a model variogram must be such that it cannot lead to negative variances. Greater complexity can be modelled by a combination of simple models. We recommend that you fit apparently plausible models by weighted least-squares approximation, graph the results, and compare them by statistical criteria.

Chapter 6 is in part new. It deals with several matters that affect the reliability of estimated variograms. It examines the effects of asymmetrically distributed data and outliers on experimental variograms and recommends ways of dealing with such situations. The robust variogram estimators of Cressie and Hawkins (1980), Dowd (1984) and Genton (1998) are compared and recommended for data with outliers. The reliability of variograms is also affected by sample size, and confidence intervals on estimates are wider than many practitioners like to think. We show that at least 100–150 sampling points are needed, distributed fairly evenly over the region of interest. The distances between sampling points are also important, and the chapter describes how to design nested surveys to discover economically the spatial scales of variation in the absence of any prior information. Residual maximum likelihood (REML) is introduced to analyse the components of variance for unbalanced designs, and we compare the results with the usual least-squares approach.

For data that appear periodic the covariance analysis may be taken a step further by computation of power spectra. This detour into the spectral domain is the topic of Chapter 7.

The reader will now be ready for geostatistical prediction, i.e. kriging. Chapter 8 gives the equations and their solutions, and guides the reader in programming them. The equations show how the semivariances from the modelled variogram are used in geostatistical estimation (kriging). This chapter shows how the kriging weights depend on the variogram and the sampling configuration in relation to the target point or block, how in general only the nearest data carry significant weight, and the practical consequences that this has for the actual analysis.

A new Chapter 9 pursues two themes. The first part describes kriging in the presence of trend. Means of dealing with this difficulty are becoming more accessible, although still not readily so. The means essentially involve the use of REML to estimate both the trend and the parameters of the variogram model of the residuals from the trend. This model is then used for estimation, either where there is trend in the variable of interest (universal kriging) or where the variable of interest is correlated with that in an external variable in which there is trend (kriging with external drift). These can be put into practice by the empirical best linear unbiased predictor.

Chapter 10 describes how to calculate and model the combined spatial variation in two or more variables simultaneously and to use the model to predict one of the variables from it, and others with which it is cross-correlated, by cokriging.

Chapter 11 tackles another difficult subject, namely disjunctive kriging. The aim of this method is to estimate the probabilities, given the data, that true values of a variable at unsampled places exceed specified thresholds.

Finally, a completely new Chapter 12 describes the most common methods of stochastic simulation. Simulation is widely used by some environmental scientists to examine potential scenarios of spatial variation with or without conditioning data. It is also a way of determining the likely error on predictions independently of the effects of the sampling scheme and of the variogram, both of which underpin the kriging variances.

In each chapter we have tried to provide sufficient theory to complement the mechanics of the methods. We then give the formulae, from which you should be able to program the methods (except for the variogram modelling in Chapter 5). Then we illustrate the results of applying the methods with examples from our own experience.