Chapter 5
Sampling

Abstract Reliable analysis and kriging demand sound sampling, which must be sufficient and have an acceptable configuration. Sampling to estimate the variogram is problematic because the spatial scale of variation is often unknown, yet there must be numerous pairs of sampling points within the correlation range, if it exists. One might determine the spatial scale from visible features such as landforms and vegetation on the ground or from remote sensing. If that is not possible a nested survey and hierarchical analysis, by either analysis of variance or residual maximum likelihood (REML), can provide a first approximation to the variogram and a guide for subsequent sampling. Variograms from previous surveys or from ancillary data, in particular aerial image data, may also be used to guide sampling. Once a variogram with known parameters is available sampling for kriging can be optimized so that some tolerable kriging error is met but never exceeded. Alternatively, if the budget for sampling is set the kriging equations can be solved to determine the kriging errors everywhere within the region of interest and in particular the maximum absolute error.

Keywords Model-based sampling · Spatial scale · Ancillary information · Nested sampling · Hierarchical analysis of variance · REML · Kriging variance · Optimal sampling · Tolerable error · Mapping

In Chap. 1 we introduced the need for sampling of the environment because of the extent of area usually covered and because the variation is usually continuous. We mentioned design-based and model-based approaches, and here we focus on the latter where our principal concern will be to sample adequately and without bias to enable us to predict accurately throughout the region. This requires sample data that are suitable to estimate both the variogram and to krige. If one knows the variogram of a variable for a particular region and can specify the maximum tolerable error in predictions using it then one can optimize one's sampling scheme (see Sect. 5.2). In most instances, however, one must first estimate the variogram, and we therefore describe the associated problems and the way to tackle them before dealing with the kriging.

M.A. Oliver and R. Webster, *Basic Steps in Geostatistics: The Variogram and Kriging*, SpringerBriefs in Agriculture, DOI 10.1007/978-3-319-15865-5_5

5.1 Sampling for the Variogram

Sampling to estimate the variogram is one of the most problematic tasks in geostatistics. It receives too little attention among both research workers and practitioners with the result that in many instances the data are too few or the spacings are unsuitable for reliable estimates of the variogram. There have been several attempts to optimize sampling for variograms, but without knowing the true variogram one cannot succeed. Lark (2002) and Webster and Lark (2013) show that without prior information on a variogram's likely form and model parameters designing a sampling scheme is little better than guesswork. In particular, one must guess the limit of spatial dependence, if such exists in the region. In Oliver's (Oliver and Webster 1987) initial survey of the soil the Wyre Forest in England, the sampling with even coverage was too sparse; the distances between neighbouring sampling points exceeded the range of spatial correlation in the soil variables.

We return to our search for that range below. Before that we state some general principles.

1. The maximum lag to which you compute the variogram should exceed the correlation range, and if it exists the sampling plan should ensure that.
2. The steps by which the lag is incremented should be small enough and the number of lags large enough for the experimental estimates to reveal the functional form of the variogram. Ideally you should aim for about six estimates within the correlation range, if it exists, and another four beyond, and sampling should be designed to provide them.
3. The size of sample should be large enough to place the estimates of the semivariances within acceptable confidence limits. A good working rule is to aim for at least 100–150 sampling points.

You might be able to judge the first from your understanding of the environment and from visible features of the landscape; physiography is a good guide. Alternatively, or in addition, you might already have or know of empirical variograms for similar land nearby. Item 2 depends to some extent on item 1, because only if you know the correlation range can you decide the interval between estimates and the sampling intervals on the ground to provide them. If you want the variogram solely for kriging then you should have one that is well estimated at short lag distances, and you should design a scheme that includes many pairs of points separated by short distances. Therefore, for a grid survey sample more intensively from randomly selected nodes to provide such pairs of points (Fig. 5.5).

Item 3 is widely misunderstood. You cannot apply the classical formula based on χ^2 to obtain confidence intervals on the experimental variogram calculated by the method of moments, Eq. (3.1), because the same data are used many times over and successive estimates are correlated. The advice in several texts to aim for 30–50 pairs of comparisons in each estimate, $m(\mathbf{h})$ in Eq. (3.1), is seriously misleading. It implies fewer than 50 points for a grid in two dimensions, and we know from empirical studies (Webster and Oliver 1992) that it leads almost inevitably to poor estimates and to erratic variograms.

We have already drawn attention to this shortcoming in Chap. 3, and we reinforce the matter in Fig. 3.5. That figure shows confidence intervals on experimental variograms computed from samples of four sizes. The upper two, Fig. 3.5a, b, in the figure for samples of size 49 and 81 are wide at all lags. As we have stated before, you should aim to sample at 100–150 points to obtain a reliable variogram.

5.1.1 Nested Sampling

Surveyors often have little or no idea of the range of spatial dependence or of the form of the variogram within its range. This is especially true when they begin investigations in unfamiliar regions. Guesswork can be expensive, either because the sampling is too sparse resulting in a variogram that is all nugget and is useless for kriging or because it is unnecessarily dense. In these circumstances sampling can be staged, with the first stage one of nested sampling followed by hierarchical analysis of variance (ANOVA) or its equivalent by REML.

The aim of such a scheme is to estimate efficiently the contribution made to the variation over scales ranging widely from fine to coarse in the region. The general principle was first proposed by Youden and Mehlich (1937) for sampling soil. Although the authors' original paper lay buried for a long time the technique was rediscovered and is now well documented in texts by Webster and Oliver (2007) and Webster and Lark (2013). The latter includes several novel options in an attempt to optimize the approach. Here we concentrate on the basic features of the strategy.

Stages are defined in terms of spacings between sampling points. At the lowest stage pairs or triplets of points are separated by the shortest distance of interest. At the highest stage, stage 1, pairs or triplets of groups are separated by the largest distance of interest. In between are several stages with points separated by intermediate distances. The distances progress in geometric sequence such that at any stage above the lowest the distance is at least 3 times that of the one below. The separating distances are fixed, but the orientations of the separations are chosen at random. The effects of distance are assumed to be random, and so the appropriate model for the analysis of variance is Model II of Marcuse (1949).

For a design with p stages the model of variation is

$$Z_{ijk\ldots m} = \mu + A_i + B_{ij} + C_{ijk} + \cdots + \varepsilon_{ijk}. \tag{5.1}$$

The quantity μ is the mean, and the A_i, B_{ij}, C_{ijk}, …, $\varepsilon_{ijk\ldots m}$ are independent random variables associated with stages 1, 2, 3, …, p, respectively, with means of zero and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_p^2$. These latter are the components of variance for the p stages, and each one is a measure of the variation attributable to that stage, i.e. to that separating distance. Together they sum to the total variance:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_p^2. \tag{5.2}$$

Miesch (1975) pointed out that if estimates of these components are accumulated, starting with that at the smallest spacing, they form a first approximation to the experimental variogram, thus:

$$\begin{aligned} \widehat{\sigma}_p^2 &= \widehat{\gamma}(h_p) \\ \widehat{\sigma}_{p-1}^2 + \widehat{\sigma}_p^2 &= \widehat{\gamma}(h_{p-1}) \\ \widehat{\sigma}_{p-2}^2 + \widehat{\sigma}_{p-1}^2 + \widehat{\sigma}_p^2 &= \widehat{\gamma}(h_{p-2}), \end{aligned} \tag{5.3}$$

and so on, where the h_p, h_{p-1}, h_{p-2}, …, h_1 are separating distances equivalent to the lag distances in geostatistical convention.

The analysis of variance for Model II above can be set out as in Table 5.1 in which there are four stages and N data, each of which belongs to one and only one group in each stage.

The table is quite general. It can be extended for more than four stages, and it can be simplified for fully balanced designs in which the same number of divisions is made at any particular stage into groups at the stage below. Balanced designs are attractive statistically because they lead to a straightforward analysis, and the variance components are readily calculated from the table because, for example, $u_{3,3} = u_{2,3} = u_{1,3}$ and u$_{2,2}$ = $u_{1,2}$. Their big disadvantage is that the number of sampling points increases exponentially, at least two-fold for each additional stage, as the number of stages increases and soon becomes unaffordable.

Balance is not necessary, however, because one does not need the very many degrees of freedom in the low stages to obtain reliable estimates of the components. Unbalanced designs can still be analysed by ANOVA, but calculating the components of variance is more complex because their coefficients, the u in the table, change from stage to stage. Gower (1962) devised formulae for calculating the coefficients, and a worked example appears in the 6th edition of *Statistical Methods* of Snedecor and Cochran (1967), but not in later editions. For theoretical reasons we now prefer to estimate the components by residual maximum likelihood (REML) as described by Webster et al. (2006).

For balanced designs the results are the same, but for unbalanced ones they generally differ somewhat.

Table 5.1 Hierarchical analysis of variance

Stage	Degrees of freedom	Parameters estimated by mean squares
Stage 1	$f_1 - 1$	$u_{1,1}\sigma_1^2 + u_{1,2}\sigma_2^2 + u_{1,3}\sigma_3^2 + \sigma_4^2$
Stage 2	$f_2 - f_1$	$u_{2,2}\sigma_2^2 + u_{2,3}\sigma_3^2 + \sigma_4^2$
Stage 3	$f_3 - f_2$	$u_{3,3}\sigma_3^2 + \sigma_4^2$
Residual (stage 4)	$N - f_3$	σ_4^2
Total	$N - 1$	

5.1.1.1 Illustrative Example: Nested Sampling in the Wyre Forest

Following the initial survey of the soil of the Wyre Forest Oliver (Oliver and Webster 1987) planned a second one to discover the scale(s) of variation in the soil. The sampling comprised nine principal nodes on a grid at intervals of 600 m; this was stage 1. The points for stage 2 were selected 190 m from each node in a random direction. From each point in stage 2 a point was selected 60 m away to form stage 3, and from each of those points another was chosen 19 m away (stage 4). Finally, from half of the stage 4 points, points were chosen 6 m away to form the fifth stage. This gave 9 × 2 × 2 × 2 = 72 sampling points in the first four stages plus a further 36 in the fifth stage, giving 108 points in all. The structure of the scheme is shown as a topological tree in Fig. 5.1. The hierarchy is unbalanced in that at stage 4 only half of the sampling points have pairs in stage 5. Figure 5.2 shows the sampling configuration on the ground for one node.

The design might not have been optimal, but it was almost certainly a better use of resources than a balanced design, and, perhaps surprisingly, better than a design that distributes the degrees of freedom equally among the stages (Webster and Lark 2013).

Oliver and Webster originally estimated the components of variance by Gower's method, but later they re-analysed their data by REML (Webster et al. 2006), Table 5.2 lists the resulting components for three depths.

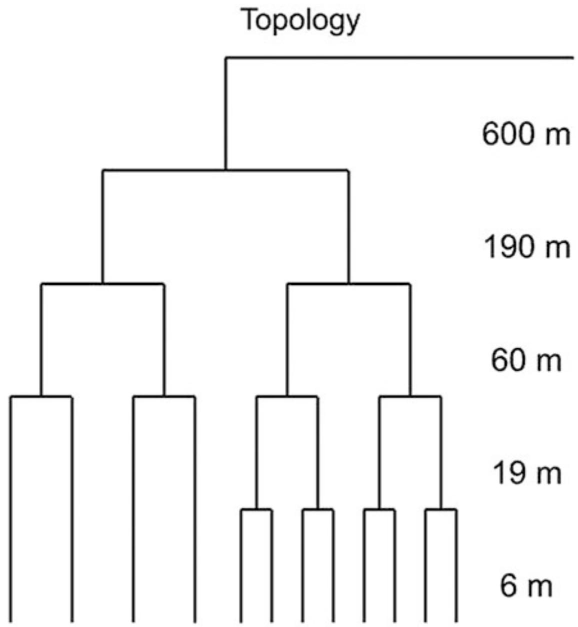


Fig. 5.1 Topology of one branch of the nested sampling scheme by Oliver (see Oliver and Webster 1987) to sample the soil of the Wyre Forest. Notice that only half of the branches at Stage 4 (19 m) are divided in the unbalanced design

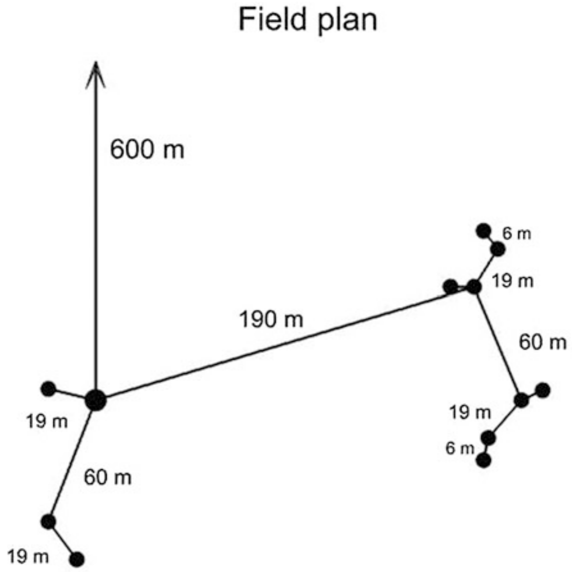


Fig. 5.2 Sampling plan of sites for one of the main branches from a grid node in the Wyre Forest with distances 190, 60, 19 and 6 m (Oliver and Webster 1987)

Table 5.2 Components of variance of percentage of sand in the soil of the Wyre Forest estimated by REML (from Webster et al. 2006)

Source (stage)	Distance/m	Components of variance		
		Depth/cm		
		0–5	25–30	50–55
1	600	38.12	16.68	33.75
2	190	–58.03	–90.02	–100.19
3	60	102.50	198.51	314.81
4	19	131.50	131.96	–38.89
5 (residual)	6	54.9	108.56	303.26

By accumulating the components from the bottom of the table upwards, as in Eq. (5.3), we obtain the variograms shown in Fig. 5.3. The variograms are erratic, but all three have maxima at 60 m.

Evidently, the range is roughly half of the distances between neighbouring points in the first survey. The figure also shows that for the first and second depths, 0–5 cm and 25–30 cm, a large proportion of the variance is between 6 and 60 m. Oliver (Oliver and Webster 1987) went on to sample the region at 5-m intervals on transects at various orientations and obtained accurate variograms by the method of moments and modelled them for kriging from data on a grid with nodes at 20-m intervals.

The above shows something of what can be achieved by splitting survey into distinct stages. Marchant and Lark (2006, 2007) developed this line of investigation, combining estimation of the variogram and kriging in stages such that the information gained in one stage is used to adapt the sampling in the next, and so on with the hope that eventually one would be able to predict and map a variable with acceptable confidence within specified budgets, starting, as it were, with a blank

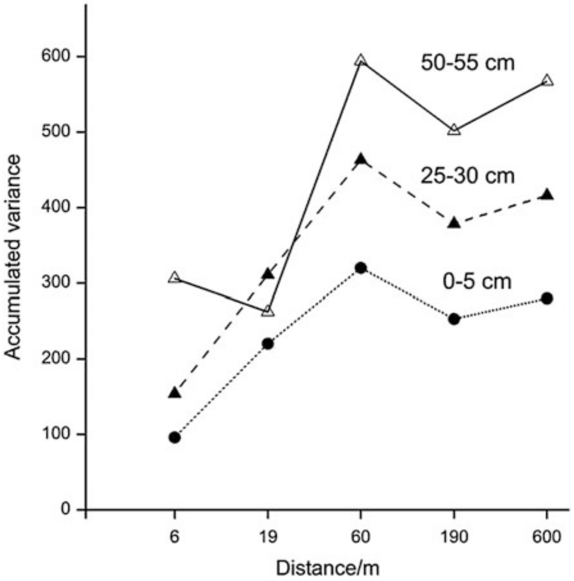


Fig. 5.3 Approximate variograms of percentage sand at three depths from the nested survey of the Wyre Forest obtained by accumulating the components of variance estimated by REML

sheet of paper. We leave the reader to pursue their strategy in the papers mentioned and in the book chapter by Marchant and Lark (2010).

What should you do if you must do the field work in a single stage? This is often the case, perhaps because of logistic difficulties and costs of getting to remote regions, perhaps because clients want quick assessments, perhaps because money is available for only a single season in the field. In these situations surveyors find that they must sample in such a way as to estimate the variogram and model it and krige from the same set of data. They cannot expect to optimize any of the steps. Pragmatically, a surveyor must start somewhere. One starting point, mentioned already, is prior knowledge of the region, especially of the landscape and physiography if one is dealing with attributes of the soil or land more generally. That should enable one to decide sampling intervals on transects for estimating the variogram and perhaps wider ones on a grid for the kriging. One will not know what the maximum errors are until one has finished, and that is a hazard.

The example below shows how variograms of ancillary data from aerial photographs, sensors and yield monitors, and existing variograms of the properties of interest can be used to guide sampling for future surveys. The data are from a 23-ha field on the Yattendon Estate, Berkshire, England (Oliver and Carroll 2004). A colour aerial photograph for 1991 was digitized and the variogram computed from the digital numbers for the red waveband. Figure 5.4a shows the experimental variogram and the fitted nested spherical model, Eq. (3.12), and Table 5.3 lists the model parameters. The yield of wheat was recorded in the field in 1995 and the variogram was computed and modelled. Figure 5.4b shows the experimental values and the fitted nested spherical function, and Table 5.3 lists the parameters of that model. The topsoil (0–15 cm) was sampled on a 30-m grid with additional samples at

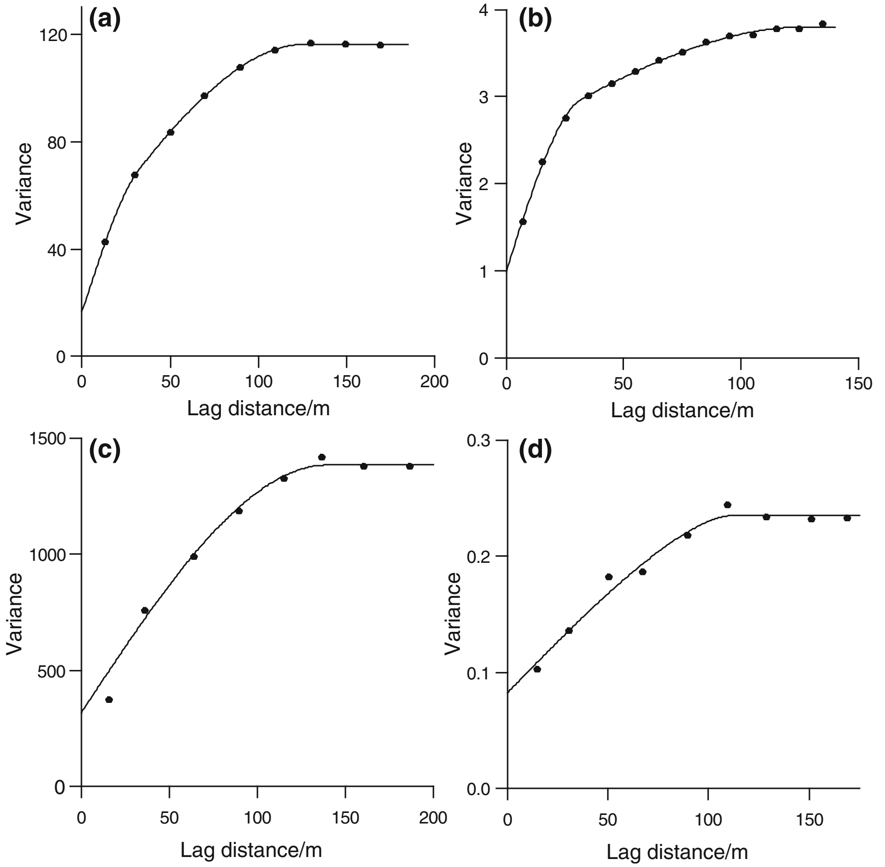


Fig. 5.4 Experimental variograms and fitted models of: **a** red waveband of a digitized colour aerial photograph taken in 1991, **b** wheat yield recorded in 1995, **c** potassium of the topsoil (0–15 cm) and **d** subsoil pH (30–60 cm) for a field on the Yattendon Estate, Berkshire, UK

Table 5.3 Model parameters of soil and ancillary data for the Yattendon Estate

Variable	Model type	Estimates of parameters				
		c_0	c_1	c_2	a_1/m	a_2/m
Soil						
Potassium—0–30 cm	Spherical	318.4	1065.0		140.1	
pH—30–60 cm	Circular	0.0824	0.152		109.8	
Ancillary						
Aerial image 1991—red waveband	Double spherical	16.86	24.91	74.52	32.66	126.8
Yield—1995	Double spherical	0.995	1.494	1.311	32.37	127.6

randomly selected grid nodes 15 m apart, and the subsoil (30–60 cm) was sampled on a 60-m grid with additional samples at selected grid nodes 15 m and 30 m apart. The experimental variogram and fitted spherical function, Eq. (3.10), of topsoil available potassium are shown in Fig. 5.4c, and the model parameters are listed in Table 5.3. Figure 5.4d shows the experimental variogram of subsoil pH with a circular function fitted, Eq. (4.16); the model parameters are listed in Table 5.3. Note that the variogram ranges of the longer structure for the aerial photograph and yield, and the ranges for potassium and pH are similar.

Kerry et al. (2010) suggested after repeated sampling of a large set of simulated values that sampling at 0.33 or less of the variogram range would provide an adequate basic grid. The average range of the variograms examined in the above example is about 126 m, and sampling at 0.33 times the range of the variogram would give an interval of 42 m for the grid. However, we recommend strongly that additional samples are taken at intervening intervals as above for the field at Yattendon to ensure that the variogram is estimated well near to the origin.

Aerial photographs are an excellent source of information for environmental surveys where the patterns of variation they show are linked with those of the variables of concern. Variograms can be computed from the digitized values prior to field work and used to guide the sampling. Milne et al. (2010) made good use of them in their analysis of gilgai patterns in Australia.

An alternative starting point is the budget; that will determine the total number of sampling points. If all the points are placed on a grid then the interval might be too large to estimate the variogram; there might be no comparisons from which to estimate the semivariances at short enough lags.

Atteia et al. (1994) planned their survey, which had to be done in a single season, with random nested sampling around 23 of their grid nodes. More often practitioners place their additional sampling points on some of the grid lines joining the nodes, as in Fig. 5.5. In Fig. 5.5a the additional points are 1.1 and 1.3 units

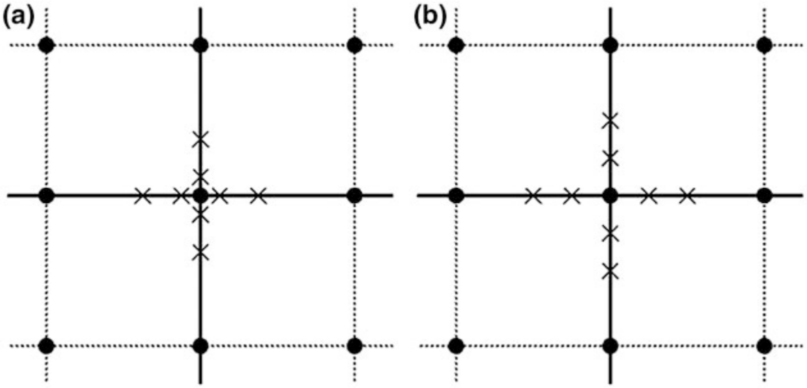


Fig. 5.5 Configurations for additional sampling at the node of a square sampling grid. The supplementary sampling points are shown as *crosses* at distances of: **a** 0.1 and 0.3 times the grid interval from the central node marked by a *circle* and **b** 0.2 and 0.4 times the interval

away from the central node. This would allow one to compute semivariances, $\hat{\gamma}(h)$, at lag distances 0.1, 0.2, 0.3, 0.4, 0.6, 0.7 and 0.9 units on the principal axes. In Fig. 5.5b the additional points are placed 0.2 and 0.4 units away from the central node, and that allows one to compute $\hat{\gamma}(h)$ at lag distances of 0.2, 0.4, 0.6 and 0.8 units. These schemes are not optimal, but both are better than a strict grid in that they enable one to compute and model the variogram over lags distances shorter than the grid interval and which one needs for predicting values between the nodes.

5.2 Sampling Plans for Mapping

The prediction of variables at unvisited places without bias is a central aim of geostatistics, and in Chap. 4 we presented the kriging equations to achieve that. The kriging equations also minimize the variance of any prediction, and their solution leads to an estimate of the kriging variance or error. In addition to being able to map a variable at a fine resolution from sample data we can also map the kriging variance or its square root, the kriging error. Such a map might show where extra sampling is needed to diminish the error and increase confidence. We can also use the kriging equations to plan sampling to map within some tolerable error—provided we have an accurate model of the variogram.

You can see that Eqs. (4.3) and (4.5) contain only semivariances, which derive from the variogram and the configuration of the sampling points in relation to the target point or block. They do not depend on the observed values at the sampling points. If you know the variogram then you can add points to the kriging systems where data seem to be too sparse and calculate what the kriging variances would be if you sampled at those points. To some extent choosing the additional sampling points is a matter of trial and error. You add a point where the existing kriging variance is greatest and solve the new kriging system, and you repeat the procedure until the kriging error is small enough everywhere.

If you know the variogram beforehand you can plan a sampling that is nearly optimal in that it will minimize the maximum kriging variance for a given cost. In general, the further a target point is from data the larger is the kriging variance. You can minimize the maximum distance between target and data by sampling on a grid; in those circumstances the maximum distance is from the centre of a grid cell to the nearest grid nodes. For punctual kriging the kriging variance is greatest there.

These maximum distances are minimized for a given sampling density with triangular configurations, and the maximum kriging variance is least. Figure 5.6 shows the situation. For a square grid the maximum distance is $1/\sqrt{2} \approx 0.7071$ units, whereas for an equilateral triangular grid with the same density the maximum distance is 0.6204 units. Square grids are more convenient, however, and as the maximum distance between a target point and the distance to the nearest sampling

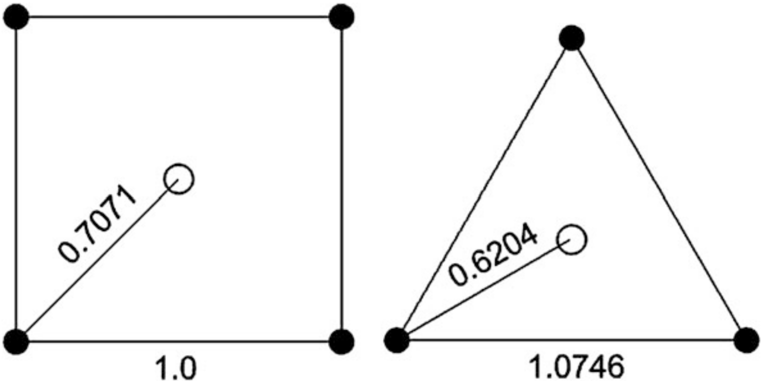


Fig. 5.6 Distances between the centres of grid cells and nearest sampling points for square and equilateral triangular grids with the same sampling density of one point per unit area

points is little more than for triangular grids of the same density and there are four near points instead of three the maximum kriging variance is only slightly larger: see Fig. 8.23a in Webster and Oliver (2007) and Fig. 9.7a in Webster and Lark (2013).

Note, however, that kriging variances tend to increase as the margins of the region are approached and that for irregularly shaped regions a regular grid should be modified to achieve best results.

The following procedure, proposed by Burgess et al. (1981) and reiterated by Webster and Lark (2013), will enable you to plan a grid.

1. Set up the kriging equations for a square configuration of sampling points with the target point or block at its centre.
2. Solve the equations for a small sampling interval, the smallest that is likely to be feasible, and compute the kriging variance.
3. Increase the sampling interval in steps and repeat the calculations in 2 above at each step.
4. Draw a graph of kriging variance (or its square root, the kriging error) against the sampling interval and link the points by a smooth curve.
5. Draw a horizontal line on this graph at your chosen maximum variance or error to cut the curve, and drop a perpendicular from the intersection to the abscissa.

That perpendicular gives the required sampling interval, from which you can determine the number of sampling points for mapping and hence the budget. Alternatively, if the budget for survey is fixed then that will determine the sampling interval, and you follow step 5 in reverse. You draw a perpendicular from the abscissa to cut the curve and read the corresponding maximum kriging variance or error on the ordinate.

5.2.1 Illustrative Example: Sampling to Map Chromium in the Swiss Jura

Atteia et al. (1994) sampled the topsoil of part of the Swiss Jura in a survey of potential toxicity caused by heavy metals, among which was included chromium (Cr). From 366 measurements they obtained the omnidirectional experimental variogram shown by the points plotted in Fig. 5.7 and to which they fitted an exponential model with equation

$$\gamma(h) = 19.98 + 98.34 \times \left\{ 1 - \exp\left(-\frac{h}{174} \right) \right\}. \tag{5.4}$$

Here h is the lag distance, and the distance parameter, a = 174, is in metres.

Using this variogram we can calculate the maximum kriging variances or errors for points or blocks of any reasonable size against sample spacing by following steps 1 to 4 above. Usually we shall be interested in blocks, and in Fig. 5.8 we show the maximum kriging errors for two sizes of block, 50 m × 50 m (= 0.25 ha) and 100 m × 100 m (= 1 ha), as the curves.

If the grid interval is much shorter than the side of the block the maximum kriging variance can occur for blocks centred on grid nodes (Burgess et al. 1981; Webster and Lark 2013), but the differences between it and that from cell-centred blocks are small and of little practical significance.

Let us suppose that in some future survey the maximum kriging error is to be no more than 10 % of the tolerable maximum concentration. The threshold for Cr set in the VSBo of 1986 for Switzerland (FOEFL 1987) is 75 mg kg^{-1} of soil. That leads

Fig. 5.7 Variogram of chromium in the topsoil in the Swiss Jura. The *line* is the fitted exponential model with the parameters shown on the graph

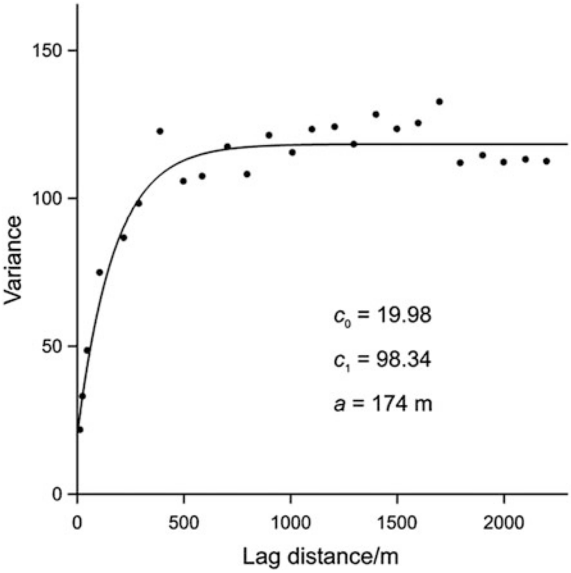
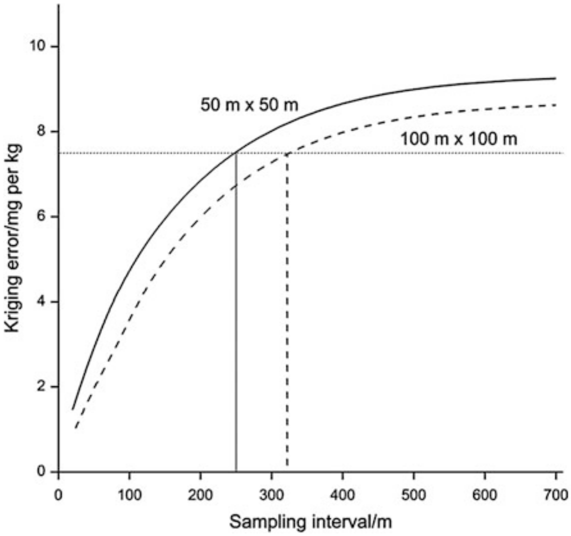


Fig. 5.8 Maximum kriging errors for chromium in the topsoil in the Swiss Jura for square blocks of 0.25 and 1 ha. The *horizontal line* is drawn at concentration 7.5 mg kg^{-1}, which is 10 % of the tolerable maximum set in the VSBo (FOEFL 1987)



to a maximum tolerable kriging error of 7.5 mg kg^{-1}. So we draw a horizontal line at that value to cut the curves and drop the perpendiculars shown in the figure. For 0.25-ha blocks the spacing is 245 m, and for the 1-ha blocks it is 322 m.

5.3 Summary

We can provide guidelines for sampling for geostatistical interpolation and mapping *if you have a satisfactory model for the variogram.* The best advice is to sample on a grid, for which either the survey budget or the maximum tolerance on a prediction determines the grid interval. If you have to estimate the variogram first and have little idea of its form then the best approach is to survey in stages, beginning with a nested scheme with analysis by REML to estimate the spatial components of variance, followed by systematic sampling to estimate the variogram and model it, and finally a grid for the mapping. If the survey cannot be staged then your best approach is to survey on a grid with its interval determined by whatever information you can glean from existing sources and an understanding of the landscape—or by the budget if that is fixed—and augment the grid with additional sampling points between the grid nodes.