Chapter 2
Sampling in Precision Agriculture

R. Kerry, M.A. Oliver and Z.L. Frogbrook

Abstract  This chapter considers the importance of spatial scale in sampling and investigates various methods by which the variogram can be used to determine an appropriate sampling scheme or interval for grid sampling. When no prior information is available on the scale of variation, and the variable of interest is unlikely to be strongly correlated to available ancillary data, a nested survey and analysis provides a first approximation to the variogram and the approximate spatial scale. If the variable of interest appears related to ancillary data such as aerial photographs or elevation, variograms of these data can provide an indication of the likely scale of variation in the soil or crop. Existing variograms of soil or crop properties can be used to determine how many cores of soil or samples from plants should be taken to form a composite (bulked) sample to reduce the local noise. Such variograms can also be used with the kriging equations to determine a grid sampling interval with a specific tolerable error, or an interval of less than half the variogram range can be used to ensure a spatially dependent sample. Finally, if the scale of variation is large in relation to the field size, a variogram estimated by residual maximum likelihood (REML) or standardized variograms from ancillary data can be used to krig data from a small, but spatially dependent sample. Each of the methods investigated is illustrated with a case study.

Keywords  Sampling · Method of moments variogram (MoM) · Residual maximum likelihood (REML) · Nested sample design · Bulking strategy · Soil data · Ancillary data

R. Kerry (✉)  Department of Geography, Brigham Young University, 690 SWKT, Provo, UT 84602, USA  e-mail: ruth_kerry@byu.edu

M.A. Oliver  Department of Soil Science, The University of Reading, Whiteknights, Reading RG6 6DW, United Kingdom  e-mail: m.a.oliver@reading.ac.uk

Z.L. Frogbrook  Environment Agency Wales, Ty Cambria, 29 Newport Road, Cardiff CF24 0TP, United Kingdom  e-mail: zoe.frogbrook@environment-agency.wales.gov.uk

2.1 Introduction

Accurate information about the variation in soil and crop attributes within fields is crucial for precise management in agriculture. This is the essence of precision agriculture (PA) which began in the early 1990s, (see papers in Robert et al. 1995, 1996; Schueller 1997; Stafford 1997, 1999). The value of geostatistics to predict accurately for digital mapping of soil and crop attributes was recognized at a similar time (Blackmore 1994; Whelan et al. 1996; Oliver and Frogbrook 1998). The accuracy of such predictions, however, depends on the quality of information on the soil and crops. Many soil and crop attributes have to be determined physically from samples in the field, which can be time-consuming and expensive. The geostatistical method of local prediction, kriging, depends on having accurate variograms and spatially dependent data from which to predict. All methods of interpolation assume implicitly that data are spatially dependent, which means that sampling should be at an interval that is well within the correlation range of spatial variation. Despite the popularity of geostatistics in PA, we have found that the methods have been, and continue to be, applied to unsuitable data. Therefore, the aim of this chapter is to guide potential users of geostatistics in precision agriculture on the sampling requirements and cost effective approaches to sampling. Chapter 3 continues the theme of sampling with examples of more elaborate methods.

Planning the sampling for surveys appears to be the ‘Cinderella’ of many environmental studies; the temptation is to rush into the field without adequate preparation. However, sampling underpins the quality and accuracy of subsequent analyses and decision making because these depend on the data being suitable for the purposes intended. It is notable that there are few papers on sampling in the proceedings of either the International or European Conferences on Precision Agriculture and in the Journal of Precision Agriculture. We need a more objective approach to sampling in PA to provide the quantitative information that is required. McKenzie et al. (2008) reinforce the need for a clear and consistent conceptual framework for sampling. Although much of the information used in agriculture in the past was quantitative, it was based on field averages which were only adequate for uniform applications and management at the field level. Management of the variation within fields, however, requires more detailed knowledge, which can be obtained only by intensive sampling.

The variation in soil and crop properties within fields comprises variation over short distances of a few metres and that over longer distances of tens or hundreds of metres. It is this last component of variation that the precision farmer wants to resolve for management, and we can regard the short-range variation as ‘noise’ or a sampling effect. Sampling for traditional farm surveys in the United Kingdom and many other countries was at about one sample per hectare and sometimes it was even more sparse. This approach has also become widely used in precision farming surveys (Godwin and Miller 2003) because it is considered a sampling density that the farmer can afford. However, sampling at this density takes no account of either the scale of variation or of how many sampling points might be needed for further analyses. Geostatistics, and in particular the variogram, can be used to guide
Sampling. The importance of spatial scale in sampling is considered in Section 2.1.1. For some attributes, especially of the soil where variation is complex and cannot be observed at the surface, for example nematodes (see Chapter 9), and there is no prior information, a nested survey and analysis can provide a first approximation to the variogram. This approximation can then be used to determine the approximate scale(s) of variation and a suitable sampling interval for a more detailed survey (see Section 2.2). For other attributes, such as weeds, plant diseases, some soil properties, etc., the variation might be evident visually in the field or in remote and proximally sensed imagery. The latter types of data are often referred to as ancillary data and variation in these is often linked to properties of the soil and of the crop. Variograms computed from ancillary data can be used to determine the approximate scale of spatial variation, and so can be used to guide sampling based on the ‘rule of thumb’ of sampling at less than half the variogram range (see Section 2.2.2).

Although we more or less dismiss the local sample to sample variation above as ‘noise’, we cannot disregard it because it can affect the accuracy of any future predictions. Local variation can be smoothed by mixing together several small cores of soil or, for example, by taking several cotton bolls from one or more plants, etc. over a given sample support to create a bulked or composite sample. The number of samples to take for bulking can be optimized using the variogram (see Section 2.3).

Sampling should match the objectives of the survey; for precision farming this is often to produce accurate digital maps. For this, a grid survey has been the usual choice (Viscarra-Rossel and McBratney 1998) because of its efficiency for sample collection in the field, prediction and mapping. If variograms of soil and crop properties from previous surveys exist for an area with a similar soil parent material and similar crop, they can be used with the kriging equations to determine an optimal grid interval.

If the sampling intervals recommended by the above methods are large, there may be too few data from which to compute a reliable variogram by the usual method of moments estimator. Kerry and Oliver (2007) showed that a variogram estimated by residual maximum likelihood (REML) could provide more accurate predictions with fewer data than one estimated by the conventional method of moments (see Section 2.5.1). A standardized variogram based on ancillary data or existing variograms of soil properties can also be used to krige spatially dependent data that are too sparse from which to compute a variogram (see Section 2.5.2).

### 2.1.1 The Importance of Spatial Scale for Sampling

Soil and crop properties can vary at markedly different spatial scales both within and between fields. Therefore, it is essential when designing a sampling scheme that the spatial scales of variation in the properties of most importance for PA management are used to guide sampling. Figure 2.1a,b illustrates the effect of spatial scale in the pixel maps of two random processes simulated with a spherical variogram. Consider that each pixel represents information on a 5-m sampling grid and the area is a 25-ha field. The variation in Fig. 2.1a occurs over short distances, whereas that in Fig. 2.1b
occurs over much longer distances. The variation represented in Fig. 2.1c is ‘white noise’. To sample to provide spatially dependent data would require quite different sampling intensities for the fields in Fig. 2.1a,b, whereas for Fig. 2.1c the regional mean only could be estimated. If these random processes were superimposed on one another they would result in nested variation as described for yield in Chapter 1. Therefore, we should need to choose which of the processes illustrated in Fig. 2.1a or b was the more important to resolve, and how many samples to bulk from to remove the effect of the ‘noise’ in Fig. 2.1c. Figure 2.1d–f shows the variograms used to generate the simulated fields; for the short-range variation the range was 50 m and for the long-range variation it was 125 m. Figure 2.1c was generated by a pure nugget variogram (Fig. 2.1f).

### 2.1.2 How Can Geostatistics Help?

Geostatistics embodies techniques to describe spatial autocorrelation of a regionalized variable, $Z$, and that use this information for local prediction by kriging. Kriging requires a model of the spatial correlation structure derived from either the
covariance function or more usually the variogram. These functions are not known a priori, and they must be estimated from sample data. Sampling for subsequent geostatistical analysis must serve two purposes: first estimation and modelling of the variogram and second local prediction by kriging. To satisfy the first purpose, sampling must be sufficient to estimate the semivariances precisely.

### 2.1.3 How can the Variogram be Used to Guide Sampling?

Webster and Oliver (1992) showed that at least 100 samples are required to estimate the variogram reliably, but for some agricultural fields this number might be too few to resolve the variation present if the scale is short, whereas it might be too many where the scale is long. In some situations there are no visible signs of variation because of the nature of the property of interest, and there are no clues to the approximate scale of variation. So how can we deal with these situations?

1. In the absence of any prior information about the spatial scale of variation and no visible signs of the variation in the property of interest, a nested survey and analysis can provide a first approximation to the variogram. This approximation can then be used to determine a suitable sampling interval for a more detailed survey. This is described in Section 2.2.1 together with an example.

2. In the absence of existing variograms of soil or crop properties and the property of interest appears related to ancillary data, such as those from remote and proximal sensing, digital elevation models, etc., they could be used to compute variograms from which to judge the approximate spatial scale of variation and a suitable sampling interval.

3. If variograms of the soil or crop properties are available, they can be used to determine how many cores of soil or samples from plants should be taken over a given support to form composite samples for laboratory analyses. The aim is to reduce local ‘noise’.

4. If there are variograms of the soil or crop properties from previous surveys, they can be used with the kriging equations to determine an optimal sampling interval, for a future grid survey with a specified tolerable error to avoid over- or under-sampling. Existing variograms can also be used with the ‘half variogram range rule of thumb’ to ensure that survey data will be spatially dependent.

5. If the scale of variation is large and there are 50–100 samples, a variogram can be estimated by residual maximum likelihood (REML) and could be used to krig data from a small, but spatially dependent sample. Alternatively, existing variograms of ancillary data, soil or crop properties could be used to compute standardized average variograms to krige sparse data that have been standardized.
2.2 Variograms to Guide Sampling

2.2.1 Nested Survey and Analysis: Reconnaissance Variogram

A nested survey and analysis is advantageous in precision agriculture when there are few or no clues as to the spatial scale of variation in the property of interest and the costs of management are large. The nested survey enables several orders of magnitude of spatial scale to be examined in a single analysis (Fig. 2.2a) to determine the approximate scale of variation with no more than about 108 samples. Webster and Oliver (2007) describe nested analysis in detail, and Section 9.3.2 provides some of the theoretical background to the method. The idea underlying the model of nested variation is that a population can be divided into classes at two or more categoric levels or stages in a hierarchy. The population can then be sampled with a nested scheme to estimate the variance at each stage, i.e. the components of variance. The individual component for a given stage measures the variation attributable to that stage, and the components sum to the total variance. Miesch (1975) showed that there is a link between the results of a hierarchical analysis of variance and the semivariances of geostatistics. If the components of variance are accumulated, starting with the smallest spacing, they are equivalent over the same range of distances to the semivariances of geostatistics.

The nested analysis provides a first approximation to the variogram (Oliver and Webster 1986; Webster and Oliver 2007). It can indicate the range of spatial scales over which most of the variation occurs making it a valuable reconnaissance tool. It indicates the spatial scale at which most of the variation occurs and this information can be used to guide sampling for an overall survey or to obtain a more accurate variogram. For properties for which we have no clues about their scale of variation, such as nematode infestations, a reconnaissance variogram could avoid wasted sampling effort and costs of analysis by indicating a suitable sampling interval for a more detailed survey. The samples from the nested survey could be integrated with the later samples provided the time interval was not too great for temporally variable properties such as nematodes. For more permanent properties of the soil, the temporal aspect would not be an issue.

2.2.1.1 Unequal Sampling

Youden and Mehlich’s (1937) sampling design was fully balanced with replication at each stage; the sample size doubled for each additional stage. Oliver and Webster (1986) showed that full replication at each stage is not necessary because the mean squares for the lower stages are estimated more precisely than those of the higher ones. Economy can be achieved by replication of only a proportion of the sampling centres in the lower stages. Such a scheme is then unbalanced, which makes estimating the components more complex (see Gower 1962). Webster et al. (2006) have shown that computing the hierarchical analysis of variance by residual maxi-
mum likelihood (REML), which is a general method of model fitting, is preferable for unbalanced nested surveys because it gives a unique result, whereas there are several methods for finding the components by the analysis of variance (ANOVA). In general, the unbalanced approach would be preferable in PA because more stages can be used with no more sampling effort.

Components of Variance by Residual Maximum Likelihood

For balanced designs ANOVA and REML give the same results, but for unbalanced ones they do not in general (Pettitt and McBratney (1993). If the random effects are assumed to be normally distributed, maximum likelihood estimates of the variance components can be calculated from Eq. 9.6. Patterson and Thompson (1971) developed the method of residual maximum likelihood (REML) to adjust for the fixed degrees of freedom before estimating the variance components. In the context here there is only one fixed effect, the grand mean, \( \mu \); therefore the differences between the estimates from REML and ANOVA are small. Readers are referred to Webster et al. (2006) for a full description of the method.

Case Study

Oliver and Webster (1987) used an unbalanced nested sampling strategy to determine the spatial variation in clay content of the soil in part of the Wyre Forest (1440 ha) of central England. Their scheme had five stages covering sampling intervals from 6 to 600 m increasing in a geometrical progression of approximately threefold increments. The survey had nine main centres on a \( 3 \times 3 \) square grid with a spacing of 600 m; the grid was orientated randomly over the region. All other sampling positions were located from these grid nodes on random orientations as follows (see Fig. 2.2a for the plan at one main centre). From each main centre, a second site was chosen 190 m away to provide the second stage. From each of the now 18 points another site was chosen 60 m away (stage 3) and the procedure was repeated at stage 4 to locate points 19 m away. At stage 5 only half of the stage 4 sites were replicated by sampling 6 m away. The result was a sample size of 108 rather than 144 for a balanced design. Since this survey Oliver and Badr (1995) applied a nested design to a survey of soil radon concentration and Webster and Boag (1992) to a survey of nematodes; both had seven stages with only 108 sampling sites.

We estimated the variance components by REML (Webster et al. 2006). Table 2.1 gives the accumulated components of variance and also the percentage variance explained by each stage. Figure 2.2b shows the accumulated components of variance for clay content at each of four sampling depths plotted against distance on a logarithmic scale to provide a first approximation to the variogram.

Figure 2.2 shows that about 80% of the variance occurs within 60 m, and that stages 1 and 2, i.e. distances of 190–600 and 60–190 m, respectively, account for less than 20%. For these variables and many others at this site, the components
Fig. 2.2 Nested sampling scheme in the Wyre Forest, England: (a) sampling plan for one of the main centres and (b) accumulated components of variance estimated by REML for clay content in the soil at four depths in the profile, with the lag distance on a logarithmic scale.

Table 2.1 Accumulated components of variance estimated by REML for clay content in the soil of the Wyre Forest at four depths in the profile, and the percentage variance accounted for by each stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Depth 0–5 cm</th>
<th>Depth 10–15 cm</th>
<th>Depth 25–30 cm</th>
<th>Depth 50–55 cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (600 m)</td>
<td>63.82 (17.02)</td>
<td>74.73 (6.25)</td>
<td>209.3 (7.72)</td>
<td>365.1 (10.57)</td>
</tr>
<tr>
<td>2 (190 m)</td>
<td>52.96 (−20.01)</td>
<td>70.06 (−16.82)</td>
<td>193.1 (−7.18)</td>
<td>326.5 (−14.93)</td>
</tr>
<tr>
<td>3 (60 m)</td>
<td>65.73 (23.38)</td>
<td>82.63 (34.59)</td>
<td>208.1 (14.56)</td>
<td>381.0 (58.97)</td>
</tr>
<tr>
<td>4 (19 m)</td>
<td>50.81 (46.44)</td>
<td>56.78 (41.69)</td>
<td>177.6 (42.63)</td>
<td>165.7 (4.60)</td>
</tr>
<tr>
<td>5 (6 m)</td>
<td>21.17 (33.17)</td>
<td>25.92 (34.69)</td>
<td>88.41 (42.24)</td>
<td>148.9 (40.78)</td>
</tr>
</tbody>
</table>

for stage 2 are negative. These suggest that either there is some repetition in the variation of the property at that distance or there is no contribution to the variance at this stage. At the lowest stage, there is a considerable contribution to the variance, especially for clay at depths of 25–30 and 50–55 cm. This represents the unresolved variation within 6 m plus any errors of measurement. This is similar to the nugget variance in geostatistics (see Chapter 1). From this information we could design a survey to estimate the variogram more precisely by sampling along transects or we could design an overall survey with a maximum sampling interval of less than half the correlation range of 60 m identified (see Oliver and Webster (1987) for a full account of these results). Although this example is not agricultural, the principles are the same and the study area here would be equivalent to doing a nested survey of a whole farm. A case study described in Chapter 9 of this book shows an application of nested sampling and analysis to determine a suitable grid sampling interval to estimate the pattern of variation in the cereal cyst nematode, *Heterodera avenae*.
2.2.2 Variograms from Ancillary Data

Many of the more permanent soil properties, such as soil texture, appear to co-vary spatially with ancillary data, which are usually intensive and cheaper to obtain. Ancillary data include digitized aerial photographs, electrical conductivity (ECₐ), yield, remotely and proximally sensed data, and digital elevation models. Each type of ancillary data is likely to be more strongly correlated with some soil properties than with others. In other words they are coregionalized which suggests that we can use these data to indicate the approximate spatial scale of variation in more expensive variables. Variograms computed from intensive ancillary data can then be used to guide sampling of the variables of interest. Chapter 7 describes how inexpensive coregionalized secondary information can be used to improve the accuracy of predictions of the primary variable by cokriging and other multivariate geostatistical methods. These approaches are likely to be of value to precision farmers because they often have suitable ancillary data or they can obtain it with little additional expense.

Experimental variograms computed from ancillary data and modelled can be used to guide sampling with an often used ‘rule of thumb’ of sampling at less than half of the variogram range (see Section 2.4.2). This use of ancillary data should avoid over- or under-sampling, both of which will result in wasted effort. The case study described below shows that the variogram ranges of the more permanent soil properties and ancillary data are reasonably consistent.

2.2.2.1 Case Study

Data from a field at Wallingford with soil developed on the plateau gravels of the Thames valley near to Oxford, England illustrate the above approach. The site was sampled and observations were made in the winter of 2000 on a 30-m grid. Six samples of the topsoil (0–15 cm) were taken from a 1 m² support and bulked at each grid node. Several soil properties of the air-dried <2 mm soil fraction were measured using standard methods. The property of interest here is loss on ignition (LOI). Ancillary data were obtained as follows. Aerial photographs of the bare soil from aerial surveys were digitized to give a ground pixel size of 3.5 × 3.5 m. Apparent soil electrical conductivity (ECₐ) was recorded with a Geonics EM38 instrument in the vertical position about every 1 m along transects about 20-m apart. Elevation data were obtained from the z coordinate of differential global positioning system (DGPS) surveys of the fields, and yield data from several years were obtained with the Massey Ferguson Fieldstar system (www.masseyferguson.com).

Omnidirectional variograms were computed for each of the soil and ancillary variables at the site. Multivariate variograms (Bourgault and Marcotte 1991) were computed from the aerial photograph data by Eq. 2.2. The multivariate variogram is defined by:

\[ \gamma(h) = \frac{1}{2} \mathbb{E}[(Z(x) - Z(x + h))^T M[Z(x) - Z(x + h)]], \]  

(2.1)
where \( E \) is the expectation, \( Z(x) \) and \( Z(x + h) \) are vectors of random variables at positions \( x \) and \( x + h \) separated by the lag, \( h \), \( T \) is the transpose, and \( M \) is a \( p \times p \) positive-definite symmetric matrix defining the relations between the variables. The experimental variogram was calculated by the standard formula adapted for the multivariate case:

\[
\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{z(x_i) - z(x_i + h)\}^T M \{z(x_i) - z(x_i + h)\},
\]  

(2.2)

where \( z(x_i) \) and \( z(x_i + h) \) are the vectors of observations at \( x_i \) and \( x_i + h \).

Table 2.2 gives the parameters of the models fitted to each experimental variogram. All soil properties and most ancillary variables have variogram ranges of similar magnitudes, except for yield. Yield shows more complex variation with two scales of variation for most years. The average variogram range for the soil properties is 238 m and for the ancillary data excluding yield it is 221 m. The difference in the variogram ranges for yield can be explained by the fact that several factors other than the soil, such as disease, weeds, pests, management practices, weather, etc., affect the yield in a given year. At Wallingford the physiography is fairly complex, and the soil properties appear to vary in harmony with it. Our analyses at other sites showed that in general aerial photograph and EC<sub>a</sub> data had variogram ranges that were more similar to those of the soil properties than elevation and yield data. Therefore, we do not recommend yield and elevation data to guide sampling unless there is evidence that these data are consistently related to patterns of soil variation in the field of interest.

Figure 2.3a, c shows the aerial photograph and EC<sub>a</sub> data for Wallingford, respectively. The patterns of variation for both sets of data are similar. The paler areas have smaller conductivities and vice versa; they indicate places where the soil is particularly gravelly, or where the gravel is largely calcareous. The experimental variograms and the fitted models for the aerial photograph and EC<sub>a</sub> data (Fig. 2.3) have similar ranges. There are also some similarities in the patterns observed in the aerial photograph, EC<sub>a</sub> and the soil data (compare Figs. 2.3a, c and 2.4a).

Based on the variogram results in Table 2.2 and sampling at an interval of less than half the range of a variogram from appropriate ancillary data, grid intervals of 100–120-m should suffice at the Wallingford site. Data on the 30-m grid were sub-sampled to give a range of coarser sampling intervals: 60-m (70 sites), 90-m (36 sites), 120-m (23 sites), 150-m (14 sites) and 120 + 60-m (50 sites). The sub-sampled data were then used with the model parameters of the variogram computed from the 30-m data for kriging. The variogram for the 30-m data was used because it was computed from >100 data, whereas the sub-samples would provide too few data from which to compute accurate variograms. Figure 2.4 shows the kriged maps of loss on ignition (LOI) for the full and sub-sampled data. The main features of the variation are preserved for sampling intervals of 120-m and less, although there is some loss of detail for all sub-samples. The main features of the variation are no longer evident when the sampling interval is 150-m (Fig. 2.4e). The results suggest
<table>
<thead>
<tr>
<th>Variable</th>
<th>Model</th>
<th>(c_0)</th>
<th>(c_1)</th>
<th>(c_2)</th>
<th>(\alpha)</th>
<th>(a_1) (m)</th>
<th>(r_1) (m) (3(r_1) (m))</th>
<th>(a_2) (m)</th>
<th>(r_2) (m) (3(r_2) (m))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay</td>
<td>Spherical</td>
<td>0.04556</td>
<td>28.87</td>
<td></td>
<td>134.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth</td>
<td>Exponential</td>
<td>310.00</td>
<td>271.30</td>
<td></td>
<td></td>
<td></td>
<td>77.46 (232.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOI</td>
<td>Pentaspherical</td>
<td>0.0535</td>
<td>0.4610</td>
<td></td>
<td>226.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCF</td>
<td>Circular</td>
<td>0.03091</td>
<td>0.3151</td>
<td></td>
<td>185.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Munsell value</td>
<td>Pentaspherical</td>
<td>0.1273</td>
<td>0.3967</td>
<td></td>
<td>218.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>Pentaspherical</td>
<td>5.894</td>
<td>81.39</td>
<td></td>
<td>231.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stoniness</td>
<td>Exponential</td>
<td>11.73</td>
<td>205.8</td>
<td></td>
<td></td>
<td></td>
<td>89.05 (267.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VWC</td>
<td>Pentaspherical</td>
<td>1.891</td>
<td>4.770</td>
<td></td>
<td>202.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerial(^{66})</td>
<td>Stable exponential</td>
<td>0</td>
<td>538.3</td>
<td>1.459</td>
<td>68.20 (204.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerial(^{77})</td>
<td>Exponential</td>
<td>1.861</td>
<td>1.180</td>
<td></td>
<td>82.57 (247.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECa</td>
<td>Pentaspherical</td>
<td>3.525</td>
<td>43.19</td>
<td></td>
<td>201.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elevation</td>
<td>Circular</td>
<td>0.2309</td>
<td>10.03</td>
<td></td>
<td>228.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yield(^{96})</td>
<td>Double spherical</td>
<td>0.6027</td>
<td>0.4363</td>
<td>2.281</td>
<td>62.62</td>
<td>56.60 (169.8)</td>
<td>650.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yield(^{97})</td>
<td>Exponential</td>
<td>0.2082</td>
<td>0.3993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yield(^{98})</td>
<td>Double spherical</td>
<td>0</td>
<td>0.216</td>
<td>0.074</td>
<td>22.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yield(^{99})</td>
<td>Double spherical</td>
<td>0</td>
<td>0.7842</td>
<td>0.361</td>
<td>14.31</td>
<td></td>
<td>316.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yield(^{00})</td>
<td>Double exponential</td>
<td>0</td>
<td>0.9201</td>
<td>0.776</td>
<td>11.37</td>
<td></td>
<td>89.40</td>
<td>31.13 (93.39)</td>
<td></td>
</tr>
</tbody>
</table>

LOI is loss on ignition; MCF is moisture correction factor and VWC is volumetric water content.
The parameters are: \(c_0\), the nugget variance; \(c\), the sill of the autocorrelated variance; \(a\), the range of the spatial dependence; for the exponential and stable exponential functions, \(r\) is the distance parameter of the model and because the sill is asymptotic an approximate range is determined as 3\(r\) (value in parentheses).
that a sampling interval of 100–120-m would be adequate for precise management in PA, which corresponds with less than half the range of all ancillary variables apart from yield (Table 2.2). Figure 2.4f suggests that additional sampling at half the grid interval for randomly selected nodes has increased the observed detail. It would also improve the accuracy of the variogram near to the origin.

As the pattern of variation is generally unknown at the sampling stage the importance of any given sample location cannot be known. Taking additional samples at shorter intervals reduces the potential loss of information and reduces the nugget component. Additional samples can be targeted according to the variation in a key ancillary variable to avoid missing important features in the soil variation. The further samples are from one another, the larger the nugget is likely to be due to unresolved variation at distances less than the sampling interval. Accurate kriged predictions depend on accurate estimation of the nugget variance. The nugget variance can also be reduced by bulking soil or plant samples to reduce the effects of variation at distances much less than the sampling interval.
2.3 Use of the Variogram to Guide Sampling for Bulking

Soil and crop attributes can vary considerably over short distances. This local fluctuation in values between points might mask the variation over the tens or hundreds of metres that is of greatest interest to the precision farmer. The material taken or the observations made at a sampling point are intended to represent the property of interest reliably over the surrounding area, the size of which will depend on the degree of spatial variation. If there is sufficient information from several cores of soil or several plants then this short-range fluctuation can be smoothed by local averages. However, obtaining so many individual measurements would be too costly in practice. The alternative is to bulk several cores of soil from a small area (Oliver et al. 1997) and mix them thoroughly for analysis or bulk material from several plants (Willers et al. 2009). The concentration measured in a bulked sample should equal the arithmetic mean of the individual cores or plants contributing to it, unless some chemical reaction takes place within the sample.

The number of soil cores or plants to be bulked depends on the local variation and the error that can be tolerated. Burgess and Webster (1984) described how to
determine the number of cores or plants for bulking if the variogram of the property is known. The variogram model is used to determine the variance for various combinations of support and configuration of sampling points within the blocks. The estimation variances (or errors) are calculated for a range of sample sizes and configurations, and the combination that meets the tolerance can be determined.

For a small area \( B \), we can estimate the average of property \( Z \), \( \mu(B) \), in \( B \) from \( n \) known values, \( z(x_i) \), at positions \( x_i, i = 1, 2, \ldots, n \) within \( B \) by

\[
\hat{\mu}(B) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i z(x_i), \tag{2.3}
\]

where \( \lambda_i \) are the weights associated with positions \( x_i \) (Burgess and Webster 1984).

The estimation variance of \( \hat{\mu}(B) \) is

\[
\sigma^2(B) = 2 \sum_{i=1}^{n} \lambda_i \gamma(x_i, B) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i, x_j) - \bar{\gamma}(B, B), \tag{2.4}
\]

where \( \gamma(x_i, x_j) \) is the semivariance between points \( x_i \) and \( x_j \), \( \bar{\gamma}(x_i, B) \) is the average semivariance between data point \( x_i \) and the block \( B \), and \( \bar{\gamma}(B, B) \) is the within-block variance.

In principle, bulking is equivalent to averaging the values at the positions from which the individual samples of soil or plants are taken in \( B \). Every sample is the same size and shape, so the weights, \( \lambda_i \), are equal. The number and positions of the cores can be varied to achieve a particular precision expressed in terms of the estimation variance. For practical purposes the estimation variance is minimized for a given \( n \) when the sampling points are on a centred regular grid (Burgess and Webster 1984). With the parameters of an existing variogram, Eq. 2.4 can be solved for a range of \( n \) on a square grid. To determine the number of cores required for a bulked sample, the calculated estimation variance or error is plotted against \( n \). The smallest value of \( n \) can be determined from the graph to satisfy a predefined tolerable error. Tolerable errors can be determined with Eq. 2.5 as in Section 2.3.1 if none is known beforehand.

### 2.3.1 Case Study

An arable field on a chalk plateau on the Yattendon Estate, Berkshire, England, was sampled (Oliver et al. 1997). The sampling scheme comprised six nodes 100-m apart. At each node there were two orthogonal transects, the mid-points of which coincided with the node. The transects were 14 m long, and the soil was sampled at 1-m intervals to a depth of 15 cm. Twenty nine samples were taken at each grid node, giving 174 samples in total. The samples were air-dried and sieved, and the
Table 2.3  Variogram model parameters for the soil survey on the Yattendon Estate, Berkshire, England

<table>
<thead>
<tr>
<th>Property</th>
<th>Model</th>
<th>( c_0 )</th>
<th>( c )</th>
<th>( a ) (m)</th>
<th>( w )</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>Spherical</td>
<td>159.10</td>
<td>298.00</td>
<td>3.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Power</td>
<td>9.82</td>
<td>1.228</td>
<td>1.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mg</td>
<td>Circular</td>
<td>5.24</td>
<td>6.43</td>
<td>4.38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The parameters are: \( c_0 \), the nugget variance; \( c \), the sill of the autocorrelated variance; \( a \), the range of the spatial dependence; \( w \) and \( \alpha \) are the intensity and exponent, respectively, of the power function. The equation for the spherical function is given in Section 1.3.2, Eq. 1.27. The equations for the power and circular functions are:

\[
\gamma(h) = c_0 + wh^\alpha \\
\gamma(h) = c_0 + c \left[ 1 - 2/\pi \cos^{-1}(h/a) + 2h/\pi a \sqrt{1 - h^2/a^2} \right],
\]

respectively.

Available potassium (K), phosphorus (P) and magnesium (Mg) were determined by the standard methods of MAFF (1986). Experimental variograms were computed for the variables by the usual method of moments estimator, and models were fitted in GenStat (Payne 2008).

The variogram model parameters for each property in Table 2.3 were used with Eq. 2.4 to determine the estimation errors for three block sizes (of side 2-, 5- and 10-m) and sample sizes \( n = 4, 9, 16, 25, 36, 49 \). The errors plotted against \( n \) for each block size are shown in Fig. 2.5. All the blocks were square, and the results are for square sampling configurations. There are no intermediate values of \( n \) with these configurations; the lines are drawn to guide the eye and aid interpretation. In all cases the estimation errors decrease roughly in inverse proportion to the sample size, and as there is more variation in the larger blocks their errors are always somewhat larger than those of the smaller blocks.

To determine the optimal number of cores for bulking, an acceptable margin of error is needed. If this value is known, or has been determined by Eq. 2.5, it can be represented as a horizontal line drawn across the graph at the specified tolerance (Fig. 2.5). The intersections of the line with the graphs indicate the number of cores from which to bulk. Since the block is square the next largest square number from the intersection is used.

Recommended tolerance values were used for this study. For K, the value was \( 7 \text{ mg}\text{l}^{-1} \) and the line intersects the graph at \( n \approx 4 \) for the 2 × 2 m block. Bulking from this number of cores in a centrally placed square configuration will ensure that the tolerable error is not exceeded. For the 5 × 5 m block the line intersects the graph at \( n \approx 6 \) and the next largest square number, nine, would be used. The 10 × 10 m block would also require a bulked sample of nine cores. For P the tolerance was \( 1 \text{ mg}\text{l}^{-1} \), and Fig. 2.5 shows that a sample of sixteen cores is needed for all block sizes. A tolerance of \( 3 \text{ mg}\text{l}^{-1} \) was used for Mg (the line for this is not shown because it is much larger than the largest error), and Fig. 2.5 shows that a bulked sample of four cores would be sufficient.
If there are no recommended values for the tolerance then we suggest that the departure from the mean for a 95% level of confidence is calculated by

\[ L = \frac{1.96\sigma}{\sqrt{n}}. \]  

(2.5)

where \( L \) is the tolerance, \( \sigma \) is the standard deviation of the variate and \( n \) is the sample number (Webster and Oliver 1990).

Different nutrients require different sampling schemes; therefore, the number of samples should be based on the nutrient that has most effect on the crop to be grown. In many instances, phosphorus may be limiting and, therefore, a bulked sample of 16 should be used. Although this number results in over sampling for potassium and magnesium, it is clear from Fig. 2.5 that a sample of 16 reduces the variance considerably and will improve the estimates of all these nutrients.

The results from other fields with different soil types showed that the sample size needed for bulking also varies with soil type. On more variable soil, such as the alluvium that Oliver et al. (1997) examined, the optimal number of cores was
much larger. In the absence of prior knowledge, and no variogram from which to compute the estimation variances, a reasonable ‘rule of thumb’ would be to bulk from 16 cores.

### 2.4 The Variogram to Guide Grid-Based Sampling

Sampling on a grid is favoured in geostatistics and PA because it provides an even cover of values and minimizes the maximum estimation variance (or error) for a given grid interval. It is also easy to implement in the field. However, practitioners must still decide how many sample sites are needed and what the interval between them should be. The variogram can provide the answer to the latter. Two approaches for identifying a suitable interval for grid-based soil sampling are considered here and are illustrated with case studies.

#### 2.4.1 The Variogram and Kriging Equations

McBratney et al. (1981) and McBratney and Webster (1981) showed how the variogram and kriging equations could be used to determine an optimal sampling interval for prediction by kriging before obtaining new data from a survey. The basis of their approach is that the kriging weights, and also the kriging variances (see Chapter 1) depend on the configuration of the sampling points in relation to the target point or block and on the variogram. They do not depend on the observed values at these points. If we have a variogram function from a previous survey of a field then we can determine the kriging errors for any grid size before sampling. It is possible to optimize the sampling by designing a scheme to meet a specified tolerance or precision. Although a triangular grid is the most efficient, square or rectangular grids are preferred because they are easier to implement, and there is little difference in precision between them in practice.

Sampling can be optimized for punctual or block kriging (see Eqs. 1.8–1.16). Webster and Oliver (2007) provide more detail on this; here we consider only block kriging. Precision farmers want to manage areas (blocks) of their fields that relate to their machinery. For block kriging practitioners must decide where to determine the kriging variances, i.e., whether for blocks centred on grid cells or ones centred on grid nodes. This is because the position at which the kriging variance is greatest varies according to block size, and it is the largest variances that should be used.

#### 2.4.1.1 Case Study

The topsoil (0–15 cm) of the same field on the Yattendon Estate in Section 2.3.1 (Frogbrook 2000) was sampled in December 1997 and 1998, and available K,
Mg and P were determined. There was no anisotropy in the variation, and omnidirectional experimental variograms were computed from the measurements on a bulked sample of 10 cores of soil from a support of 2 × 5 m taken at the nodes of a 20-m grid. The experimental values were modelled, and the model parameters are given in Table 2.4. The parameter values are similar for both years for each nutrient, which suggests that a tolerance can be fixed in the absence of any marked temporal variation.

The model parameters for K, Mg and P (Table 2.4) were used with the kriging equations to provide estimates of the kriging standard errors associated with sampling intervals on a square grid between 2- and 200-m and blocks of side 20-, 40- and 60-m. These block sizes are ones that farmers might use to manage inputs and are associated with the size of areas over which farm machinery operates. Figure 2.6 shows that for each nutrient and block size, as the grid spacing increases, the kriging error increases until it reaches a maximum. For all examples in Fig. 2.6, when the sampling interval is very short the maximum estimation variances decrease from some small value and then increase again. This is an artefact that arises because only observations nearest the centres of the blocks have been used for estimation.

Tolerable kriging errors of 2, 5 and 10 mg kg⁻¹ for P, Mg and K, respectively, were used to determine sampling intervals for these nutrients. These concentrations were considered to be limiting to cereal yield for each nutrient. Table 2.5 summarizes the results for the various block sizes and tolerance values used in Fig. 2.6. The sampling intervals indicated for K and P are consistent between years, they are greater for K than for P and they increase as block size increases. Thus a variogram from a survey done in a previous year could be used to improve sampling at a later stage. For Mg, the largest interval is for the 40-m block in 1997; the tolerance exceeds the estimation error for the 60-m block in 1997 and for all blocks for 1998. The variation of Mg in this field is small compared to the tolerable error and a large sampling interval would achieve the specified error.

It is possible to take anisotropy into account by adjusting the grid spacing so that sampling is more intense in the direction of maximum variation (see Webster and Oliver 2007, for more detail). The intervals can be determined as above with the model parameters of an anisotropic model.
Fig. 2.6  Graphs of kriging error against grid spacing for: (a) K$_{97}$, (b) K$_{98}$, (c) Mg$_{97}$, (d) Mg$_{98}$, (e) P$_{97}$ and (f) P$_{98}$ on the Yattendon Estate, Berkshire
Table 2.5  Optimal grid spacing (m) for three different block sizes at the Yattendon Estate

<table>
<thead>
<tr>
<th>Property and tolerance (mg kg(^{-1}))</th>
<th>20-m block</th>
<th>40-m block</th>
<th>60-m block</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>± 10</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>Mg</td>
<td>± 5</td>
<td>43</td>
<td>*</td>
</tr>
<tr>
<td>P</td>
<td>± 2</td>
<td>10</td>
<td>12</td>
</tr>
</tbody>
</table>

*Tolerance exceeds the estimation error or the variation in the field is smaller than the tolerable error so a large sampling interval would suffice.

2.4.2  Half the Variogram Range ‘Rule of Thumb’ as a Guide to Sampling Interval

The case study on ancillary data in Section 2.2.2.1 suggests that a sampling interval of less than half the range of a variogram from such data can be used to guide sampling. We tested this ‘rule of thumb’ on large fields simulated by the turning bands method (Journel and Huijbregts 1978). Following the approach of Webster and Oliver (1992) the fields were sub-sampled randomly 100 times for a given sampling interval based on the variogram range. The sub-samples were designed to use a datum no more than once and to cover the same area. The minimum sample size was 144 so that we could discount the effects on the accuracy of too small a sample to compute a reliable MoM variogram (Webster and Oliver 1992). Variograms were computed for each random sub-sample, and confidence limits were computed based on the 100 experimental variograms. These limits narrowed as the sampling interval decreased (Fig. 2.7). The variograms computed on data with an interval of 0.66 of the range appeared to be almost all pure nugget (Fig. 2.7a), and the confidence intervals were almost twice as wide as those for the other sampling intervals.

Standard deviations of the observed semivariances (not shown) were similar for sampling intervals of 0.5 and 0.4 of the variogram range when there was no nugget component. However, they were larger for 0.5 of the range when there was a nugget component. Nevertheless, they were still markedly less than those for an interval of 0.66 of the range. These results show that a sampling interval of 0.66 of the range is too large and will result in large errors in the predictions and probably a pure nugget variogram. An interval of 0.5 of the range would give acceptably accurate predictions for site-specific management. The results for intervals of 0.4 and 0.33 of the range, which are similar, would give more accurate predictions, especially when there is a nugget component. Therefore, sampling at an interval of 0.4 of the variogram range or less should be adopted if possible.
Fig. 2.7 Observed semivariances and the 90% confidence limits (dashed lines) for various sampling intervals for a spherical field with a range of 50 m and zero nugget: (a) 0.66 of range (33 m), (b) 0.5 of range (25 m), (c) 0.4 of range (20 m), and (d) 0.33 of range (17 m) (the grey line is the function used for simulation and the black line is the model fitted to the exhaustive variogram of the sub-sample)

2.5 Variograms to Improve Predictions from Sparse Sampling

2.5.1 Residual Maximum Likelihood (REML) Variogram Estimator

We describe above how a suitable sampling interval can be determined from variograms of ancillary data or existing variograms of relevant soil properties. However, if this interval is large compared with the extent of the field, then there will be too
few data from which to compute Matheron’s (1965) method of moments (MoM) variogram reliably (Webster and Oliver 1992). Some potential solutions to this situation are illustrated below with a case study.

2.5.1.1 Case Study

For the Wallingford site (see Section 2.2.2.1), variograms from ancillary data suggested that an interval of 120-m would resolve the spatial variation in this 43.5 ha field. This sampling interval results in only 23 sampling points, which is too few data from which to compute a reliable MoM variogram. Pardo-Igúzquiza (1998a) suggested that a reliable variogram could be computed from a ‘few dozen’ data by maximum likelihood or residual maximum likelihood (REML). Lark (2000) and Kerry and Oliver (2007) examined this idea further and suggested that 50–60 data might suffice. Based on a simulation study, Kerry et al. (2008) showed that the accuracy of a variogram estimated from 100 data by MoM was similar to one estimated from 50 data by REML. Nevertheless, the variogram estimated from 100 data by REML was more reliable.

Figure 2.8 shows the variograms computed by MoM and REML from data with sampling intervals between 30- and 120-m at Wallingford, and Table 2.6 gives the model parameters. All variograms have been standardized to a sill of 1 so that the parameters can be compared with the approach in Section 2.5.2. The MoM variogram for the 90-m grid is unbounded, and that for the 120-m grid appears as pure nugget, whereas the variograms estimated by REML for these intervals are similar to that for the 30-m grid. Table 2.6 shows that when data on the 120-m grid are supplemented with additional points at 60-m to give 50 data, the model parameters for the MoM and REML estimators are closer to those for the 30-m data. Although a plausible model was fitted to the experimental MoM variogram computed from these data, Fig. 2.8e shows that the semivariances are erratic and the model is a poor fit.

Table 2.7 gives the results of cross-validation for the standardized LOI data. Data on the 30-m grid were used with the model parameters of variograms estimated from the subsets to determine how appropriate these models were for representing the variation in LOI. For the MoM variogram there is a marked increase in mean squared error (MSE), Eq. 1.29, as the sampling interval increases and the mean squared deviation ratio (MSDR), Eq. 1.30, is markedly less than 1. For the variogram estimated by REML the increase in MSE with increasing sampling interval is less marked. The MSDRs for some sub-samples and the REML estimator are around 2, which show that the kriging variance is underestimated by the relevant model. The MSEs, however, indicate that sample size affects variograms estimated by REML less and as a consequence there is less loss of accuracy in the predictions. This result has important implications for PA because farmers often cannot afford to sample intensively. Pardo-Igúzquiza (1998b) provides a published program to compute the variogram by REML.

Loss on ignition was kriged using data from the various sub-samples and their associated MoM and REML variograms, however, for pure nugget variograms
Fig. 2.8 Variograms for percentage loss on ignition (LOI) at Wallingford estimated by the method of moments (MoM) and residual maximum likelihood (REML) for: (a) 30-m grid (296 data), (b) 60-m grid (70 data), (c) 90-m grid (36 data), (d) 120-m grid (23 data) and (e) 120-m + 60-m grid (50 data)
Table 2.6 Variogram model parameters for residual maximum likelihood (REML), method of moments (MoM) and standardized (Std) variograms of percentage loss on ignition (LOI) at Wallingford

<table>
<thead>
<tr>
<th>Grid interval (number of data)</th>
<th>Method</th>
<th>Model</th>
<th>(c_0)</th>
<th>(c_1)</th>
<th>(w)</th>
<th>(a) (m)</th>
<th>(r) (m), (3(r) (m))</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-m (296)</td>
<td>REML</td>
<td>Exp.</td>
<td>0.04</td>
<td>0.96</td>
<td></td>
<td>50.07</td>
<td>(159.2)</td>
</tr>
<tr>
<td>60-m (70)</td>
<td>REML</td>
<td>Spher.</td>
<td>0.41</td>
<td>0.59</td>
<td></td>
<td>227.8</td>
<td></td>
</tr>
<tr>
<td>90-m (36)</td>
<td>REML</td>
<td>Spher.</td>
<td>0</td>
<td>1</td>
<td></td>
<td>163.1</td>
<td></td>
</tr>
<tr>
<td>120-m + 60-m (50)</td>
<td>REML</td>
<td>Spher.</td>
<td>0.36</td>
<td>0.64</td>
<td></td>
<td>205.9</td>
<td></td>
</tr>
<tr>
<td>120-m + HML Std (38)</td>
<td>REML</td>
<td>Spher.</td>
<td>0</td>
<td>1</td>
<td></td>
<td>206.8</td>
<td></td>
</tr>
<tr>
<td>30-m (296)</td>
<td>MoM</td>
<td>Penta.</td>
<td>0.10</td>
<td>0.89</td>
<td></td>
<td>226.1</td>
<td></td>
</tr>
<tr>
<td>60-m (70)</td>
<td>MoM</td>
<td>Penta.</td>
<td>0.27</td>
<td>0.73</td>
<td></td>
<td>344.4</td>
<td></td>
</tr>
<tr>
<td>90-m (36)</td>
<td>MoM</td>
<td>Linear</td>
<td>0.71</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120-m (23)</td>
<td>MoM</td>
<td>Nugg.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120-m + 60-m (50)</td>
<td>MoM</td>
<td>Penta.</td>
<td>0.17</td>
<td>0.83</td>
<td></td>
<td>291.3</td>
<td></td>
</tr>
<tr>
<td>120-m + LMS Std (38)</td>
<td>Std.</td>
<td>Exp.</td>
<td>0</td>
<td>1</td>
<td></td>
<td>68.2</td>
<td>(204.6)</td>
</tr>
</tbody>
</table>

The models are: Exp. exponential, Spher. spherical, Penta. pentaspherical and Nugg. pure nugget. The model parameters are: \(c_0\) the nugget variance; \(c\) the sill of the autocorrelated variance; \(a\) the range of the spatial dependence and \(w\) the intensity of the linear function. For the exponential function, the sill is asymptotic and an approximate range is determined as \(3r\) (values in parentheses), where \(r\) is the distance parameter of the model, and LMS is large, medium or small digital numbers (DNs).

Table 2.7 Cross-validation results for Wallingford for loss on ignition (LOI) data with model parameters estimated by residual maximum likelihood (REML), the method of moments (MoM) and standardized (Std) variograms

<table>
<thead>
<tr>
<th>Variogram estimator</th>
<th>Number of data</th>
<th>MSE</th>
<th>MSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-m REML</td>
<td>296</td>
<td>0.471</td>
<td>0.970</td>
</tr>
<tr>
<td>60-m REML</td>
<td>70</td>
<td>0.505</td>
<td>0.887</td>
</tr>
<tr>
<td>90-m REML</td>
<td>36</td>
<td>0.478</td>
<td>2.091</td>
</tr>
<tr>
<td>120-m REML</td>
<td>23</td>
<td>0.497</td>
<td>0.929</td>
</tr>
<tr>
<td>120-m + 60-m REML</td>
<td>50</td>
<td>0.477</td>
<td>2.188</td>
</tr>
<tr>
<td>120-m + LMS REML</td>
<td>38</td>
<td>0.477</td>
<td>2.657</td>
</tr>
<tr>
<td>30-m MoM</td>
<td>296</td>
<td>0.477</td>
<td>1.504</td>
</tr>
<tr>
<td>60-m MoM</td>
<td>70</td>
<td>0.495</td>
<td>1.191</td>
</tr>
<tr>
<td>90-m MoM</td>
<td>36</td>
<td>0.543</td>
<td>0.661</td>
</tr>
<tr>
<td>120-m MoM</td>
<td>23</td>
<td>0.574</td>
<td>0.521</td>
</tr>
<tr>
<td>120-m + 60-m MoM</td>
<td>50</td>
<td>0.482</td>
<td>1.444</td>
</tr>
<tr>
<td>120 m + LMS Std</td>
<td>38</td>
<td>0.471</td>
<td>1.367</td>
</tr>
</tbody>
</table>

bilinear interpolation was used; Figs. 2.9 and 2.10, respectively, show the mapped predictions. These maps contrast with those in Fig. 2.4 where the model for the 30-m data was used to krige the sub-sampled data. Figure 2.9 shows that there is considerable loss of detail in the maps based on the sub-sampled data and that this is much greater than in Fig. 2.4. The variation in Fig. 2.9e is the least degraded, which
Fig. 2.9  Kriged maps of LOI for Wallingford for sub-samples with grid spacings of: (a) 30-m, (b) 60-m, (c) 90-m, (d) 120-m and (e) 120-m + 60-m data and associated MoM variograms

indicates the advantage of the additional sampling at half the grid interval (Fig. 2.4f) when sampling is sparse. Figure 2.10 shows that the patterns of variation are far less degraded for the small sample sizes when the variogram estimated by REML is used.

2.5.2 Standardized Variograms

As an alternative to estimating the variogram by REML, Kerry and Oliver (2008) suggested using standardized variograms from ancillary data with the sill scaled to 1 and nugget:sill ratios that were appropriate for different soil properties. These authors found that the most reliable method of determining the nugget:sill ratio was to compute a variogram by REML from the soil data, together with data from 15 additional sampling sites targeted to areas in an aerial photograph with large, medium and small (LMS) digital numbers.

Figure 2.11a shows samples selected from the 30-m grid based on large, medium or small (LMS) digital numbers (DNs) in Fig. 2.3a. Five samples were selected for
Fig. 2.10  Kriged maps of LOI for Wallingford for sub-samples with grid spacings of: (a) 30-m, (b) 60-m, (c) 90-m, (d) 120-m and (e) 120-m + 60-m grid data and associated variograms estimated by REML

Table 2.6 gives the model parameters of this variogram and those of the standardized variogram (last line). Table 2.7 gives the associated cross-validation results for comparison with the other sub-samples. The MSE is small for the variogram estimated by REML from data on the 120-m grid and targeted LMS data, but the MSDR is large. By contrast, the MSE is slightly smaller and MSDR is markedly smaller (Table 2.7) using the range and model type of the standardized variogram determined from ancillary data and the nugget:sill ratio from the variogram estimated by REML from just 38 data. The results for the standardized variogram are comparable to those for the 30-m data and better than those for each of the other sub-samples. There was also a strong correlation between the predictions based on the standardized variogram model and those from the 30-m model. Figure 2.11b shows the variation in LOI using the standardized variogram model and the 38 data for kriging. The map from the 30-m data (Fig. 2.9a) and that for the 38 data based on the standardized variogram (Fig. 2.11b)
2.6 Conclusions

This chapter has shown the importance of planning the sampling and analyses to be done before going into the field and the problems of applying a prescriptive approach to sampling based on economic considerations (i.e. one sample per hectare). This book focuses on geostatistical applications in PA and this chapter on sampling for them. Readers should not forget, however, that the message is the same for other types of data analysis although the requirements will be different. Nested sampling and analysis provide a first approximation to the variogram when little or nothing is known about the spatial scale of variation. If there is a relation between soil or crop properties and ancillary data, variograms of the latter can indicate the approximate scale of variation in the former. Existing variograms of soil or crop properties can be used to determine how many samples to bulk from to reduce sampling effects on predictions and also to design optimal sampling schemes on a grid for kriging. The half variogram range ‘rule of thumb’ can guide sampling to provide predictions that are accurate enough for management in PA. If the sampling intervals indicated are large, there might be too few samples to estimate the variogram accurately by MoM. Variograms estimated by REML and standardized variograms from ancillary data, together with a few judiciously located additional samples, can improve the accuracy of predictions. These approaches provide an interim solution until more intensive information about soil and crop properties can be obtained more cheaply. A general message that emerges from the case studies is that when sampling for eventual geostatistical analysis there is a need for balance between overall cover of
the field, for which a grid survey at an appropriate interval is suitable, and resolution of localized variation which requires some degree of nesting in the sample configuration. This is discussed further in Chapter 3. When there are fewer than 100 soil or crop data, the variogram should be estimated by REML rather than MoM.

References


Geostatistical Applications for Precision Agriculture
Oliver, M.A. (Ed.)
2010, XIV, 331 p., Hardcover