6

Introduction to Sampling in Space

As regards 'space', it is assumed that in the structured approach to designing survey and monitoring schemes (Chap. 3.2), the design information specifies that the universe of interest is purely spatial, i.e., no time dimension is involved. This part therefore deals with the situation in which a once-only survey can deliver the required information. Of course, the methods presented here can be applied more than once in the same area. That would, however, constitute a form of monitoring, the implications of which are dealt with in Part IV 'Sampling in Space–Time'.

Sampling for survey of natural resources can be done in 1D, 2D or 3D space. Although the spatial universe of interest is often a three-dimensional body, sampling is mostly carried out in the horizontal plane, i.e., in 2D space, so that the sampling locations have only two coordinates. Therefore we present the methods in terms of 2D sampling; for instance, we will use the term 'area' rather than 'spatial universe'. Sampling in 1D or 3D space is discussed separately, in Sect. 7.2.16.

In Sect. 4.1 we already stressed the importance of the choice between design-based or model-based inference because design-based inference requires probability sampling, whereas for model-based inference non-probability sampling is most appropriate. We shall discuss the pros and cons of these two approaches to sampling and inference now in more detail in the context of spatial survey. Broadly speaking, the suitability of design-based methods, relative to model-based methods, is greatest for global quantities such as the spatial mean, and diminishes as one moves to smaller and smaller sub-areas, and finally to estimation at specific points. Neither of the two approaches has a monopoly, not even at the extremes of the spatial resolution continuum, viz. the area as a whole and individual point locations. This broad picture is illustrated in Fig. 6.1. It should be noted that the relative suitability functions depicted in this figure only reflect our global expectations of suitabilities, 'averaged' over a broad class of different cases that could be encountered in practice.

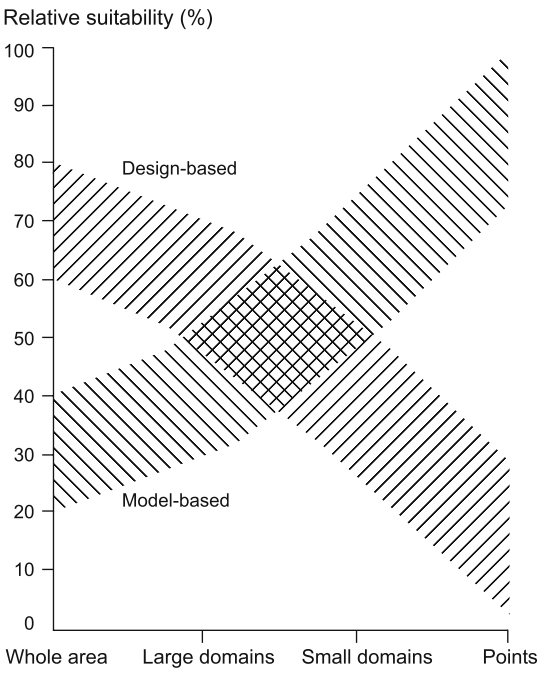

Fig. 6.1. Relative suitability of the design-based and model-based approaches to sampling, as a function of the spatial resolution at which estimates are required

In reality there are more factors that determine the suitability than spatial resolution alone. These factors relate to the following questions.

1. Should the estimation or the test of the global quantity be 'design-unbiased', i.e., correct on average over repetitions of the sampling process using the same sampling design? Design-unbiasedness can be regarded as a strict guarantee against bias in sampling, such as may arise in convenience or purposive sampling. If this guarantee is required, then a design-based method is the only option.
2. Should the accuracy of the estimate or the power of the test be quantified objectively, i.e., without recourse to assumptions on the spatial variation? A positive answer to this question rules out model-based methods.
3. Is random sampling in the field practically feasible? If not, then some form of convenience or purposive sampling combined with model-based inference is the obvious choice.
4. Is a reliable model of the spatial variation available? Only if this is the case, can the model-based approach be sensibly applied.

5. Do substantial spatial autocorrelations exist between sampling locations and prediction locations? If not, then the computational effort involved in model-based inference will be fruitless.
6. Is composite sampling acceptable and would it reduce costs significantly? The conditions under which composite sampling is acceptable are discussed in Sect. 4.3. In principle, costs can be significantly reduced by compositing if laboratory analyses of individual aliquots would consume a considerable portion of the total budget. If composite sampling is an attractive option, this could be a reason to prefer a design-based method over a model-based one. The reason is that design-based methods allow compositing of aliquots taken at large mutual distances, possibly across the entire area, whereas with model-based methods, compositing is in practice always limited to aliquots from within small neighbourhoods. In general, compositing of aliquots that are wider apart yields a greater reduction of sampling variances, hence greater precision of the final estimates.
7. Are multiple realizations of a random field needed for the inference about the target quantity? Such realizations are to generated by simulation with a stochastic model of the variation, hence a model-based method must be used. A condition that makes simulation inevitable is when the target quantity is a nonlinear function of multiple values of the target variable. This is the case, for instance, with detection problems (the target quantity being the maximum of a 0/1 indicator variable), and with target quantities defined by neighbourhood operations, such as the (surface) area of a watershed.

As there are several misconceptions in the literature on this issue, we repeat from Sect. 4.1 that the design-based methods presented in Sect. 7.2 are valid, regardless of the structure of the spatial variation, because they do not make any assumption about that structure.

A typical application of design-based sampling strategies is to estimate the areal mean of a directly measured quantitative variable. However, the scope of these strategies is much wider than this, and can be expanded in three directions: derived variables, other parameters and smaller areas or sub-areas.

First, the target variable need neither be quantitative, nor directly measured. If the target variable is measured on a nominal or ordinal scale, the sample data consist of class labels, and these can be analyzed statistically by first transforming them into 0/1 indicator variables. The presence and absence of a given class are thereby re-coded as 1 and 0, respectively. Of course, if there are k mutually exclusive classes, only $k-1$ indicator variables are needed. The mean of an indicator variable can be interpreted as the fraction of the area in which the class occurs. Transformation into indicator variables can also be applied to quantitative variables in order to estimate the areal fraction in which the variable exceeds a given threshold. This technique can be extended to estimate the entire Spatial Cumulative Distribution Function

(SCDF) of a quantitative variable. In that case, areal fractions are estimated for a series of threshold values.

Apart from the simple 0/1 transformations, the target variable may be the output of a more or less complicated model for which the input data are collected at the sampling locations. Another important case of indirect determination is in validation studies, where the target variable represents an error, i.e., the difference between a measured value and a value predicted by a process model or a spatial distribution model, such as a thematic map. A common example is the error resulting from a classification algorithm applied to remotely sensed images. The errors determined at the sampling locations can be used to estimate their spatial mean (which equals the bias), the mean absolute error, the mean squared error or the entire SCDF of the errors.

Second, the target parameter does not need be the spatial mean. For instance, it may also be a quantile, such as the 90^{th} percentile, the spatial variance, a tolerance interval or a parameter of a model relating one or more predictor variables to a variable of interest. See Krishnaiah and Rao (1988) and Patil and Rao (1994) for design-based statistical inference on these and other target parameters.

Third, the region for which estimation or hypothesis testing is required need not be the entire area sampled; interest may also focus on one or more sub-areas, or in estimation at points. This subject is dealt with in Sect. 8.2.

Traditionally, the design-based method focused on discrete populations, and therefore representation of the universe is discrete in this approach. For instance, the mean is defined as an average over all N population elements. In this book we adhere to this usage in most cases, even when the universe is continuous. The continuous universe is first discretized by a fine grid of which the nodes represent the possible sampling locations. These methods are thus presented in a finite population mode, whereby the size of the universe is a dimensionless quantity (the number of nodes). In the model-based approach, on the other hand, the universe consists of an infinite number of possible sampling locations, and its size is measured in units of length, (surface) area or volume.

6.1 Contents

This part is divided into three chapters, according to what the aim of the sampling is: sampling for *global* quantities in space (Chap. 7), for *local* quantities in space (Chap. 8), or for *variograms* to model the spatial variation (Chap. 9). The chapters 7 and 8 form the main body of this part. They are each divided into a section on design-based methods and a section on model-based methods.

The section on *design-based* methods for *global* quantities (7.2) is the largest section of this part. It contains not only subsections on basic and advanced types of sampling designs and on how to choose from them (7.2.2–7.2.8), but also subsections on special sampling techniques like Probabilities-

Proportional-to-Size Sampling (7.2.9), Sequential Random Sampling (7.2.10), Line-Transect Random Sampling (7.2.13), and Line-Intercept Random Sampling (7.2.14). Two subsections deal explicitly with the use of ancillary information in sampling and in inference from sample data (7.2.11 and 7.2.12). Finally, there is a special subsection on model-based optimization of sample sizes for design-based sampling (7.2.15), and one on sampling in 1D or 3D space (7.2.16).

The section on *model-based* methods for *global* quantities (7.3) treats Centred Grid Sampling (7.3.2) and Geostatistical Sampling (7.3.3), i.e., optimizing the sampling pattern with the aid of a geostatistical model. The section on *model-based* methods for *local* quantities (8.3) in addition contains a subsection on Spatial Coverage Sampling (8.3.3). Both sections (7.3) and (8.3) contain a separate subsection on sampling of hot spots. Sampling for answering the question 'Is there a hot spot?' is treated in Sect. (7.3.4), while the question 'Where is the critical threshold exceeded?' is dealt with in Sect. 8.3.5).

The section on *design-based* methods for *local* quantities (8.2) deals with probability sampling for quantities defined on sub-areas (8.2.2) and for estimation of values at points (8.2.3).

Finally, Chap. 9 presents sampling and inference methods for variogram estimation. The sampling methods entail regular patterns (9.2) and optimized patterns (9.3). The inference methods are the method-of-moments (9.4.1) and maximum likelihood estimation (9.4.2).