5

Optimization of Sample Selection

Decisions on sample selection are often taken on the basis of experiences in comparable cases, practical considerations, prescribed protocols or convention. This may be inevitable, and it may work satisfactorily. However, in principle one would like to optimize the selection of the sample in terms of costs and quality, employing the available information about the universe in question. This is especially true for large projects requiring relatively large investments, and when good prior information is available.

In this chapter we discuss how sample selection can be optimized in terms of costs and the quality of the result. Our discussion focuses on point sampling from a continuous universe and on two modes of optimization:

1. *Quality maximization*, under the constraint that the costs must not exceed a given budget;
2. *Cost minimization*, under the constraint that the quality of the result must meet a given minimum requirement.

As we have seen in previous chapters, the sample is not the only scheme item that determines costs and quality. The assessment method, sample support and bulking of aliquots are other items that affect costs and quality. Hence, the sample selection can only be optimized after decisions have been made (at least provisionally) on these other issues.

In the following sections we discuss various options with respect to quality measures as the basis of optimization, and approaches to the optimization process itself.

5.1 Quality Measures

Defining a quality measure is a prerequisite for ex-ante and ex-post evaluation of the quality of the result, as well as for optimization of the sample selection. As indicated in Sect. 3.2, this definition is usually the result of a 'translation' of an initial, more general formulation of the aim of the survey or monitoring

project. This translation generally leaves room for choice, and the designer should be aware of this, because this choice has direct consequences for the design process. We distinguish three types of quality measures.

Utility Measures

These are functions of the error distribution, which specify the expected economic losses due to given errors in the result. An error in the resulting data may cause a loss when it causes the user to make a sub-optimal or wrong decision. Hence, functions of this kind are derived from a detailed quantitative analysis of the consequences of errors. This involves analysing how the error in the results propagate to the outcome of the economic analysis. See Bie and Ulph (1972) for an early example of this approach in the context of landuse planning on the basis of an imperfect soilmap. In many cases such an uncertainty analysis is unfeasible, but if a realistic utility function can be defined, then this is to be preferred over statistical or geometric functions, because it offers a more accurate representation of the quality. This type of measure requires the availability of a stochastic model of the variation as prior information.

Statistical Measures

These are also functions of the error distribution, but they are generic and do not specify expected economic losses. Instead, they reflect the accuracy, precision or reliability of the result. Common measures for estimates and predictions are the standard error, the error variance and the half–width of confidence or prediction intervals. These measures are appropriate for more-or-less symmetrically distributed errors. For non-symmetrically distributed errors, and in case one is able to estimate the conditional probability density of these errors one could choose to minimize the entropy (Bueso et al., 1998).

For qualitative results, as produced by hypothesis testing or classification, the power of the test and error rates are common measures. Statistical measures are easier to apply than utility measures, but less closely related to the actual use of the result. As with utility measures, statistical measures need a stochastic model of the variation, either in the form of prior estimates of one or more variance components, or in the form of a geostatistical or time-series model.

Geometric Measures

These measures can be viewed as substitutes for statistical and utility measures. They can be used to facilitate the optimization of spatial sampling locations when the application of a statistical or utility measure is unfeasible, e.g., because a realistic model of the variation is not available. An example is the Mean Squared Shortest Distance, a measure used to create ‘spatial coverage’ samples (Sect. 8.3.3).

5.2 Approaches to Optimization

The aim of this section is to give a broad overview of approaches to optimization of sample selection. More details are given in various sections on sampling methods. Section 5.2.1 discusses optimization for a single target quantity. It appears that optimization in the design-based approach is quite different from optimization in the model-based approach. Section 5.2.2 discusses the more complicated case of optimization for multiple target quantities. Here, the choice of an optimization method will be largely governed by the reason why there are multiple target quantities.

5.2.1 Optimization for a Single Target Quantity

As explained below, optimization of sample selection is different in the design-based approach and the model-based approach (Sect. 2.2.1). Optimization in model-based sampling tries to find the best sampling pattern within the target universe. This works roughly as follows.

If the optimization mode is 'quality maximization', the affordable sample size is first determined from the given budget. An iterative search algorithm is then used to find the optimal sample of that size. To keep computation time within reasonable limits, the search is typically confined to some subset of possible positions, e.g., selections from a user-specified discretization grid. During the search, the chosen quality measure is evaluated for a large number of candidate samples, but the algorithm generally does not guarantee that a global optimum will be reached. Therefore the search should be repeated with different initial solutions.

If the optimization mode is 'costs minimization', the problem is more complex because now the best combination of both sample size *and* sampling pattern has to be found. A practical approach is to conduct a 'quality maximization' for each of a series of eligible sample sizes, and to retain the combination with the smallest sample size that still meets the given quality requirement.

Optimization in design-based sampling is different from that in model-based sampling, because design-based inference implies that a probability sample is drawn, i.e., the sampling pattern is stochastic and cannot be optimized as such. However, the randomized selection takes place within certain randomization restrictions, which characterize the sampling design and vary in nature and complexity. Thus, optimization in design-based sampling tries to find the best sampling design rather than the best sampling pattern.

In order to explain the optimization of sampling designs, we must have a closer look at the randomization restrictions that they impose. A very basic kind of restriction is imposed by the design 'Simple Random Sampling (Sect. 7.2.3) with sample size n', which is that all samples of a size other than n are excluded from selection. A slightly more complex restriction is imposed by the design 'Stratified Simple Random Sampling (Sect. 7.2.4) with sam-

ple sizes $n_1, \ldots, n_k$ in the strata $\mathcal{U}_1, \ldots, \mathcal{U}_k$', allowing only samples with the pre-specified sample sizes in the strata.

An even more complex combination of restrictions is imposed, for instance, by 'Cluster Random Sampling (Sect. 7.2.6) by n random transects in a north-south direction and equidistant sampling locations d metres apart'. While these restrictions determine the number and direction of the transects and the inter-point distance, the only randomness left is in the starting points of the transects. These examples illustrate that randomization restrictions can be quantitative (e.g., numbers, distances) or qualitative (e.g., type of cluster, shape of strata).

Optimization of qualitative randomization restrictions is computationally more cumbersome in general, so in practice optimization will often be limited to quantitative restrictions, given a provisional decision on possible qualitative restrictions. The quantitative restrictions are related to total sample size, sample size per stratum, number of primary and secondary units (as in Two-Stage Random Sampling (Sect. 7.2.5)) and number of clusters.

Optimization methods for these parameters work with a stochastic model of the variation, as with optimization in model-based sampling, according to a utility or statistical measure. The use of a model in the design-based approach may be confusing at first sight. Keep in mind, however, that the model is only used to optimize the sampling design, not for inference from the sample data.

What was said above about costs minimization versus quality maximization in model-based sampling applies to design-based sampling as well. Specific details on optimization in design-based sampling in space are presented in Sect. 7.2, along with the various types of random sampling designs.

5.2.2 Optimization for Multiple Target Quantities

A predominant factor which complicates optimization is that survey and monitoring often have more than one purpose, in the sense that more than one target quantity is defined. This may be because there is more than one domain, or more than one target variable, or more than one target parameter. Application of a single quality measure to multiple target quantities yields multiple qualities (or a multivariate quality), and the outline given above is then no longer directly applicable.

The choice of a suitable approach to optimization with multiple qualities depends on the reason why there are multiple qualities. Therefore we discuss these situations separately.

More Than One Domain

We discuss two cases, one where the domains have extensions in space and/or time, and one where the domains are prediction points.

Non-Point Domains

In this case, the universe is divided into a number of disjoint parts for which separate results are required.

First let us consider costs minimization. This optimization mode assumes that a quality requirement is given for each domain individually, e.g., the error variance for each domain must be smaller than a user-specified maximum. If the overall costs can be reasonably approximated by the sum of the costs per domain, and if sampling and inference is done in such a way that the qualities of the results for the domains are approximately independent of each other, then *overall* costs minimization can simply be done by costs minimization *per domain*, along the line indicated for the case of a single target quantity.

In the design-based approach, the condition of quality independence between domains can be met by using the domains as strata and sampling them independently (Sect. 7.2.4). In the model-based approach, this independence can be introduced by assuming – if only to simplify the optimization – that the inference will be limited to the sample data from within each domain. This may not be unreasonable, because it leads to conservative predictions of the qualities if, in reality, data from other domains are also used.

For quality maximization, given a total budget for all domains together, there are two options. The first option is to assign weights to the qualities in the domains, and then to define a single overall quality measure as the *weighted average* of the qualities per domain (see Sect. 8.2.2). The other option is to define the overall quality measure as the *minimum* of the qualities per domain. Whatever the definition, this single overall quality can in principle be maximized through a search algorithm dividing the total budget among the domains.

Point Domains

A common instance is a regular grid of prediction points created for the construction of a map by means of model-based inference (e.g., kriging, Sect. 8.3).

Quality maximization, given a total budget and a maximum sample size derived from it, also offers the same two options as in the case of domains with an extension. One option is to assign weights to the qualities at the prediction points, and then to define a single overall quality measure as the *weighted average* of the qualities per point. The other option is to define the overall quality measure as the *minimum* of the qualities per point. However it is defined, this single overall quality can be maximized through a search algorithm such as simulated annealing (Sect. 8.3.4).

Costs minimization assumes that a quality requirement is given for each prediction point, and must be implemented over the combination of sample size and sampling pattern. The difference with costs minimization in the case of domains with an extension is that here the assumption of quality independence is untenable; a change in sample size or pattern will affect the qualities

at various prediction points simultaneously. This makes optimization more complicated.

Similar to the case with a single target quantity (Sect. 5.2.1), a practical approach would be to optimize[1] the pattern for each of a series of eligible sample sizes, and to retain the combination with the smallest sample size that still meets the given quality requirements for the prediction points.

More than one target variable

This situation occurs when multiple target variables are to be determined at the same sampling events. We briefly discuss three typical cases.

Spatial Cumulative Distribution Function (SCDF)

In order to estimate or predict the SCDF of some variable, a series of indicator variables is defined that correspond with increasing threshold values of the variable. The indicator variables are treated as the target variables. The spatial means of the indicators are interpreted as fractions, which are the target quantities.

In principle, there are as many qualities as there are threshold values, but in this case it may be appropriate to reduce these to a single quality on prior grounds. For instance, the fraction closest to 0 or 1 will have the smallest relative precision, and this may be selected as the most *critical* single quality, on the basis of which either quality maximization or costs minimization is performed.

Spatial Coverages of Plant Species

In the context of survey of vegetation, information may be required about the spatial coverages of different plant species. The presence or absence of the species is represented by indicator variables which, once again, are treated as the target variables, and their spatial means as target quantities.

Just as in the previous case, there are as many qualities as there are plant species, and to simplify the optimization one might select a single quality on prior grounds. In this case, however, it may be more appropriate to select the quality for the ecologically most *relevant* species as the one for which either quality maximization or costs minimization is performed.

The underlying assumption of this approach is that, while the sample selection is optimized for a single species, the results for the other species will still be satisfactory. If this assumption is deemed unrealistic, it may be necessary to have recourse to the approach outlined for the case of prediction points as domains.

[1] The pattern could be optimized by simulated annealing (Sect. 8.3.4 and Appendix A) with a penalty function accounting for differences between the expected qualities and the required qualities

Concentrations of Pollutants

In the context of environmental monitoring, information is often required about the concentrations of a set of pollutants. The concentration of each pollutant is treated as a target variable, and the spatio-temporal means of these concentrations are the target quantities. This case resembles the previous one as far as optimization is concerned, and sample selection may be optimized in the same way.

There may be one difference, however, namely when the option is available not to measure all concentrations at all sampling events. This is the case, for instance, if the aliquots taken on the sampling events can be analyzed for varying sets of pollutants. It may then be cost-efficient to limit the more expensive analyses to a subsample and to use possible correlations in the inference. This can be done by Two-Phase Random Sampling and regression estimators (Sect. 7.2.11).

More than one target parameter

A typical example is when an estimate is required of both the mean and the standard deviation of a quantitative target variable. Exact optimization in such a case would generally require stochastic simulation.

A rough approach to costs minimization, given a quality requirement for each parameter, would be to assume a parametric distribution on the basis of the available prior information, e.g., a normal distribution with a prior estimate of the standard deviation. Assuming Simple Random Sampling and using standard formulae one can predict the precision of estimates of both parameters for each of a series of eligible sample sizes. The smallest sample size that still satisfies the quality requirements is then chosen. This leaves a safety margin if a more efficient sampling design is applied, as will normally be the case.