

Water Quality Assessments - A Guide to Use of Biota, Sediments and Water in Environmental Monitoring - Second Edition

Edited by Deborah Chapman

© 1992, 1996 UNESCO/WHO/UNEP

ISBN 0 419 21590 5 (HB) 0 419 21600 6 (PB)

Appendix 10.1 Basic design for sampling programmes

General guidance on sampling design (location, frequency, sampling equipment, number of samples) has been given in the relevant chapters for different levels of assessment in rivers, lakes, reservoirs and groundwaters. However, in many water bodies (e.g., complex reservoirs) the issues of where to sample and how many samples to take in order to give representative results for the whole water body, need more than an intuitive choice of sampling design. Nevertheless, the task of deciding the optimum number of samples to take and the most suitable locations in a water body in order to characterise its water quality in a meaningful way, and with the most economic use of resources, can be quite daunting. Statistically based methods of sampling design can help this task and also ensure that the data collected are appropriate for later statistical analysis and interpretation. A full coverage of this subject is beyond the scope of this guidebook but some basic principles are described briefly in this appendix. A full description of the sampling process is given in Keith (1988). Gilbert (1987) also gives a comprehensive account of sampling theory, with many excellent examples based on environmental studies. Much of the source material is based on standard texts such as Cochran (1963) and Snedecor and Cochran (1980).

Basic sampling design naturally falls into seven aspects: (i) reasons to sample, (ii) what to sample, (iii) how to sample, (iv) when to sample, (v) where to sample, (vi) how many samples to take, and (vii) sampling evaluation. The issues of what, where, when and how to sample are defined by the assessment programme objectives and are discussed in detail in relation to rivers, lakes, reservoirs and groundwaters in Chapters 6, 7, 8 and 9. Some guidance for obtaining statistically valid measurements is given below.

Where to sample

Although random sampling is often suggested as the basis of a sampling design, it is rarely the most appropriate approach, particularly when there is already some knowledge of the nature of the water body and the characteristics of the variable being monitored. The choice of sample site is, therefore, usually based on:

- *Judgement sampling*: Sample site location is based on experience of the type of water body and the variables, together with their anticipated distribution pattern. Although judgement sampling can provide good estimates of many variable types, it is difficult to provide a quantified assessment of the applicability of the estimates so derived to the whole water body.

• *Probability sampling*: Sample sites and numbers are based on statistical probability theory, and so are ostensibly based on an objective methodology. However, as will be seen later, in practice many subjective aspects have to be involved. Pure, sampling theory may also lead to logistically impractical schemes. The main options are:

- *Simple random*: Sites are chosen at random across the surface of the water body, and within the water mass.

- *Stratified random*: Sample sites are selected randomly within areas chosen relative to the more homogeneous components of an otherwise heterogeneous variable. For example, randomly sampling within the epi-, meta- and hypolimnion of a stratified reservoir or within multiple basins of a dendritic water body.

- *Systematic sampling*: This is usually the method of choice for practical purposes. Sample sites are chosen at appropriate locations. These locations are chosen on the basis of experience and/or judgement. Samples are usually drawn at prescribed, not necessarily regular, points along horizontal and vertical transects.

- *Cluster sampling*: Specialised sampling, e.g. for fish or water weeds. Usually involves taking all specimens within an isolated sampling area.

- *Double sampling*: Used where a more simple variable, which bears a known relationship to a more complex one, may be sampled more intensively to provide an estimated distribution of the lesser sampled, more complex variable, e.g. measuring conductivity instead of comprehensive anion/cation analysis.

How many samples to take

A. Random sampling

In schemes which have to use discrete sample sites or volumes, the numbers of samples required to attain various monitoring objectives may be basically derived from the relevant probability distributions and error statistics (see sections 10.4.6, 10.4.7,

10.4.8). These generally derive: $(\bar{X} - \mu)/(s/\sqrt{n}) = t$, where t is the Student's t probability distribution function, \bar{X} is the sample mean, μ is the "true" mean, s is the sample standard deviation, and n is the number of samples.

Bearing in mind that such formulations apply to at least near-normally distributed and independent samples (see section 10.4.4), by simple rearrangement: $n = (t*s/(\bar{X} - \mu))^2$.

If $d = \bar{X} - \mu$, then $n = (t*s/d)^2$.

In order to use such a formulation, it is necessary (because of the terms it contains) to cast the sampling objective in the form of: How many samples n are necessary from a population with variance s^2 to provide an estimate of the population mean within $\pm d$ units of the true mean, with a probability of p per cent? (90 or 95 per cent confidence levels are commonly used). Even this specification involves two difficulties. The main difficulty lies in having already available an acceptable estimate of the variance of the population

about to be sampled! The secondary problem is that t is also dependent on the number of samples involved n , so it is only possible to solve the equation by successive trials.

Example 1

Suppose the sampling and measurement variability of a mixed reservoir population of algae produced a standard deviation of 3 mg m⁻³ chlorophyll a. Assuming this value remains true, how many samples would be necessary to ensure that the measured mean was within 2 mg m⁻³ chlorophyll a of the actual population mean, with 95% confidence?

Thus, $s = 3$, $d = 2$ and $p = 95$. The probability of obtaining a difference greater than d is $\alpha = (100 - p)/100 = 0.05$ (this change is convenient for the manner in which t is usually tabulated). As there is no initial estimate of sample numbers, the formula containing f can only be used with an estimated value for the degrees of freedom df , which along with α determines t . For present purposes, $df = n - 1$. Due to the form of the f distribution (it is nearly normal, and relatively constant for values of df greater than about 30), it is convenient to start our trial solutions with $df = 30$.

Thus:

$t_{0.05[30]} = 2.042$. Then: $n_1 = (2.042 \cdot 3/2)^2 = 9.4 \Rightarrow 9$
 $t_{0.05[8]} = 2.306$. Then: $n_2 = (2.306 \cdot 3/2)^2 = 12.0 \Rightarrow 12$
 $t_{0.05[11]} = 2.201$. Then: $n_3 = (2.201 \cdot 3/2)^2 = 10.9 \Rightarrow 11$
 $t_{0.05[10]} = 2.228$. Then: $n_4 = (2.228 \cdot 3/2)^2 = 11.2 \Rightarrow 11$
As $n_4 = n_3$, the solution has been found: 11 samples.

If the solutions oscillate between two figures, then the higher figure should be chosen for the basis of the sample collection as these formulations give theoretical, minimum sample numbers. Any spatial correlation amongst samples, for example, will usually require greater numbers of samples for similar statistical accuracy (see below).

The ease of the example computation above rested almost entirely on the pre-knowledge of the variable standard deviation. If that value is not known or cannot be reasonably estimated, then the computation cannot be made. As having a reasonable estimate of the sample variance is so essential, efforts are usually made to provide a reasoned value. The most likely approaches include: preliminary survey, previous monitoring data, expert judgement, and reasoned assessment.

Even the first two of these approaches also require a subjective judgement that the previous survey results are still applicable, particularly with qualities which are potentially highly variable in horizontal and vertical extent, and time. It is also usually true that many of the variables of interest in some water bodies do not have similar relative variability, or that their relative variabilities change differently through the year. At any sampling instant, therefore, the relatively most variable of the qualities of interest must be used to set the

sample number calculation. If these differ widely, then a sub-sampling approach may be used without loss of efficiency.

An example of a reasoned assessment when there is no direct information available on quality variability can be based on a diatom bloom in a lake or reservoir. In such cases the variable of interest (i.e., chlorophyll *a*) often has a more readily apparent maximum and minimum value at any particular sampling time. For example, a Spring dominance by a diatom in a temperate reservoir might produce possible biomass measurements up to about 100 mg m⁻³ chlorophyll *a*. If the reservoir is reasonably mixed at this time, it might be assumed that it is unlikely that the minimum value sampled will be much less than 50 mg m⁻³. The sample range *R* is, therefore, 100 - 50 = 50 mg m⁻³ chlorophyll *a*. Assuming distribution of the results to be somewhat non-normal, then $s = 0.3 * R$ (see section 10.4.2). Thus the estimate of *s* is 15 mg m⁻³ chlorophyll *a* which can be used in the sample number calculation.

The assessment of sample numbers may also be simplified by considering the problem in more relative terms. For example: How many samples are necessary for a variable with a coefficient of variation (see section 10.4.3) of 40 per cent so that, with 90 per cent confidence, the mean is estimated to within 10 per cent of its true value? Casting the problem in this way will always help for those water quality measurements in which sample variability tends to be related to sample concentration (i.e., high variability is associated with high values and low variability with low values). It is then easier to assume that previous results, for example, will still hold for a subsequent sampling occasion.

Example 2

Using the values from the previous paragraph: $cv = s/\mu = 0.4$; $\delta = d/\mu = 0.1$, and $\alpha = 0.1$; then $n = (t^* cv/\delta)^2$. Starting once again with $df = 30$:

$t_{0.1[30]} = 1.697$. Then: $n_1 = (1.697 * 0.4 / 0.1)^2 = 46.1 = 47$

$t_{0.1[46]} = 1.681$. Then: $n_2 = (1.681 * 0.4 / 0.1)^2 = 45.2 = 46$

$t_{0.1[45]} = 1.680$. Then: $n_3 = (1.680 * 0.4 / 0.1)^2 = 45.2 = 46$

Thus, at least 46 samples would be required to meet the specified sampling objective.

The examples show how the numbers of samples required rapidly mount up if overly stringent accuracy and associated confidence in the results are applied to inherently variable sites or circumstances. Once again, considerable judgement may be necessary in balancing the assessment needs and the sampling resources available.

One usual response to the need for large sample numbers is to spread the sampling over a number of sample sites located within the water body. If the sites are vertically mixed, and samples from them are not spatially correlated, then spreading the sample numbers over a series of sampling sites is a natural development. However, such a simple approach is not appropriate if the sample values are not independent but are

correlated, as will frequently be the case if many sample sites are used. For example, if there is a tendency for a high result for a given variable at a particular site to be accompanied by another high result for the same variable at a neighbouring site (or a low value with a low value) then the true variability will be greater than that simply measured. If the lesser variance is then used in the formulae given above it will result in an under-estimate of the numbers of samples required to meet the sampling objective.

If experience or previous results allow some estimate of a uniform spatial correlation coefficient r amongst sampling sites, then:

$$n \cdot n_s = (ts/d)^2 \cdot (1+r(n_s - 1)), \text{ where } n_s \text{ is the number of sampling sites.}$$

With positive correlation $r > 0$, so a factor $(1 + r(n_s - 1))$ as many samples will be required. If, for example, $r = 0.1$ and $n_s = 10$ virtually twice as many samples relative to the basic number would be required. Clearly, with only a single sampling site, or no correlation, the formula becomes the same as that used before. Moreover, serial correlation may also exist, for example amongst sample sites spaced along a transect or within samples taken relatively close together in time, and further allowance may have to be made for that situation. Furthermore, if the means at each sample site differ, but are presumed to be measurements of the same global mean, then: $s = \sqrt{(s^2 + n \cdot s_s^2)}$ where s_s^2 is the inter-site means variance. Where such complexity occurs, it is highly desirable that expert assistance is involved in the sampling design.

Example 3

Nitrate samples taken at monthly intervals from three, neighbouring groundwater sampling wells, show a standard deviation of 60 mg l⁻¹ and a correlation coefficient of 0.45, but independence from month to month. Could those samplings give an estimate of the annual mean nitrate value to within 20 mg l⁻¹, with 90% confidence?

From the formula above, $n \cdot 3 = (t_{0.1[3 \cdot n_s - 1]} \cdot 60/20)^2 \cdot (1 + 0.45(3 - 1))$. Iterative solution of this equation leads to $n = 16$ samples at each site. Therefore, the monthly sampling is insufficient. Assuming that adding a few more sampling stations would reduce the site-to-site correlation coefficient to about 0.4, but all else remained the same; how many stations are necessary to achieve the original sampling aim?

$$\text{Then, } 12 \cdot n_s = (t_{0.1[12 \cdot n_s - 1]} \cdot 60/20)^2 \cdot (1 + 0.4(n_s - 1)).$$

Re-arranging: $n_s = (1 - 0.4) / (12 / t_{0.1[12 \cdot n_s - 1]} \cdot 3)^2 - 0.4$ and solving by successive approximation gives $n_s = 7$; so a further four sites would be necessary. If, more realistically, there was also correlation between sample estimates at succeeding time intervals, then further modification is necessary. A reasonable approximation is:

$$n \cdot n_s = (ts/d)^2 \cdot (1 + r(n_s - 1)) \cdot (1 + 2 \sum r_k), \text{ where } r_k \text{ are the correlations between samples } i, i+1; i, i+2; i, i+3; \text{ etc.; and the sum is taken to } n - 1.$$

If, for this example, it is assumed that $r_1 = 0.6$; $r_2 = 0.3$; $r_3 = 0.1$, and all other $r_k = 0$; then $(1+2*(0.6+0.3+0.1)) = 3$ times as many samples would be required. Correlations clearly reduce the amount of information gained from each individual sample.

B. Stratified random sampling

In reservoirs and lakes the phenomenon of summer thermal stratification is a further source of complexity in sampling design because of its potential support of vertical water quality differences (see Chapters 7 and 8). The most usual basis for design in such circumstances is stratified random sampling. In this method, a series of non-overlapping vertical strata are pre-chosen, so that the quality to be sampled may be presumed to be very much less variable within any stratum than between the strata. The strata are often chosen on the basis of criteria such as vertical temperature or oxygen gradients. The epilimnion, metalimnion and hypolimnion may, therefore, form appropriate strata. However, for some organisms such as buoyant blue-green algae or zooplankton, further stratification may be necessary. Significant bottom depressions or general bathymetric complexity may also form the basis for stratum choice. The weight, w_i to be given to any stratum is the proportion of the total sampling units which may be sampled that is contained within the stratum. For example, for the epilimnion w_{epi} , could represent the volume of the epilimnion as a proportion of the total reservoir volume, i.e., $w_{\text{epi}} = \text{Vol}_{\text{epi}}/\text{Vol}_{\text{total}}$.

Samples are drawn from random depths within these pre-defined strata (note: random depths are not haphazard depths, but are chosen according to an appropriate randomising scheme). Within any stratum, the numbers of samples to be taken relative to the total are usually set by one of two procedures: proportional allocation, or optimised allocation.

In proportional allocation, the number of samples in the i th layer is $n_i = n * w_i$. The numbers of samples are simply allocated on the basis of the relative weights w_i of the various strata. The proportional allocation scheme then leads to $n = t^2 (\sum w_i s_i^2) / d^2$ as an estimate of the total number of samples to be taken at the site, where s_i^2 is the quality variance in the i th layer.

In optimised allocation, the numbers of samples taken from any layer reflects both the variability of the quality within that layer, and the difficulty (cost) of sampling from that layer. Then the optimum number of stratum samples would be $n_i = n * (w_i s_i / \sqrt{c_i}) / \sum (w_i s_i / \sqrt{c_i})$, where c_i is the cost of sampling in the i th layer. This formula suggests taking more samples where the stratum is larger, the stratum quality is more variable and the sampling cost is least. If the costs are similar for each layer, then: $n_i = n * w_i s_i / \sum (w_i s_i)$. Optimised allocation then leads to $n = t^2 (\sum w_i s_i)^2 / d^2$ as an estimate of the total numbers of samples to be taken. It is usually the case that optimised allocation requires fewer samples for a given precision and confidence of estimate.

Example 4

Consider a lake, with soluble reactive phosphorus (SRP) concentrations and distribution similar to those illustrated in Figure 7.19 for Heilingensee in late August, 1987. How many samples are required to provide an estimate of the mean amount of phosphorus in the lake to within 10 mg P m^{-3} , with 90 percent confidence, if other sample variances relative to the concentration ranges hold?

| i | Depth range (m) | w_i | Concn. Range (mg P m^{-3}) | Prev. Data s_i | $w_i s_i$ | Optimal allocation n_i | $w_i s_i^2$ | Propn. allocation n_i |
|----------|-----------------|-------|---------------------------------------|------------------|-----------|--------------------------|-------------|-------------------------|
| 1 | 0-5 | 0.7 | 30-50 | 6.4 | 4.5 | 8 | 28.7 | 241 |
| 2 | 5-7 | 0.2 | 50-1,000 | 171.6 | 34.3 | 61 | 5,889.4 | 69 |
| 3 | 7-10 | 0.1 | 1,000-2,000 | 260.7 | 26.1 | 47 | 6,796.4 | 34 |
| Σ | | | | | 64.9 | 116 | 12,714.5 | 344 |

Optimal: $n = t_{0.1[115]}^2 (\Sigma w_i s_i)^2 / d^2 = 116$ samples

Proportional: $n = t_{0.1[\infty]}^2 (\Sigma w_i s_i)^2 / d^2 = 344$ samples

Samples would then be drawn from random sites on the lakes, and at random depths

within the strata. The overall mean would then be $\bar{X}_t = \Sigma w_i \bar{X}_i$, where \bar{X}_i is the i th stratum mean. The total SRP in the lake would also be estimated as $\text{SRP}_{\text{total}} = \text{Vol}_{\text{total}} \bar{X}_t$. It will be noticed how substantially the optimal allocation strategy has reduced the sampling load necessary for the specified accuracy and confidence.

If no previous variability statistics exist, then a possible approach to provide rough estimations of the stratum standard deviations and sample numbers is as follows. Calculate s_i for each stratum by $(\text{SRP}_{\text{max}} - \text{SRP}_{\text{min}})_i / 4$; the divisor being chosen relative to about 30 samples in the absence of any other data (see section 10.4.2). This leads to: $s_i = 5; 238; 250$ for the three strata. Then $n_i = 7; 99; 52$; (total =158). Therefore, even this simple procedure gives a sample programme quite close to that of the more specific optimal allocation strategy previously considered. Since for stratum 1 only 7 samples are suggested, a more appropriate divisor of, for example, 3 ($\cong 7 - 10$ samples) might be used instead. Similarly for stratum 2, a divisor of 5 ($\cong 50 - 100$ samples) could be used. Recalculation on these factors leads to $n = 126$, with $n_i = 9; 71; 46$ respectively. These approximations show that very useful sampling guidance can be derived from reasoned estimation, even when accurate, critical data are not available.

Example 4 might also be considered on the basis of whether random sampling is the best approach. An alternative procedure would be to use systematic sampling, where sites are chosen to give appropriate coverage of the water-mass, and then sampled at regular depths so as to allow reasonable depth profiles to be drawn. These could, for example, then be planimetrically integrated. A possible problem with this approach is the

difficulty of ascribing appropriate confidence levels to the result. Nevertheless, it illustrates the need always to consider alternative sampling approaches as part of the initial sample planning stage.

Advanced statistical analysis may also be used to identify spatial and temporal patterns amongst sampling results drawn from special surveys. These patterns may then be used as the basis for more general sampling site allocation. This is a particularly important aspect for dealing with the variability associated with reservoirs (see section 8.2.1). A good example of such an approach is given in Thornton *et al.* (1982) (see section 8.5.2) Essentially similar approaches to sampling number estimation may be made in rivers (Montgomery and Hart, 1974).

Sampling evaluation

A very important part of any sampling exercise is to review the extent to which the programme has achieved the desired objectives. Once the samples have been taken, contemporary information is available as to distributions and variability. If the required precision has not been achieved, these new data may be used in many of the foregoing equations to establish how many extra samples are required, and how the sampling strategy may be further optimised. If necessary, any extra sampling may then be immediately undertaken.

If the water quality variable under investigation is not likely to be normally distributed (e.g., count data), or is found on the basis of first sampling to be distinctly non-normal, then it would usually be necessary to apply an appropriate transformation (see section 10.4.5) so that these normal distribution based formulae may be applied.

References

Cochran, W.G. 1963 *Sampling Techniques*. Second edition. Wiley and Sons, New York, 413 pp.

Gilbert, R.O. 1987 *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold Company, New York, 320 pp.

Keith, L.H. [Ed.] 1988 *Principles of Environmental Sampling*. American Chemical Society. 458 pp.

Montgomery, H.A.C. and Hart, I.C. 1974 The design of sampling programmes for rivers and effluents. *Wat. Pollut. Control*, **73**, 77-101.

Snedecor, G.W. and Cochran, W.G. 1980 *Statistical Methods*. 7th edition. Iowa State University Press, Ames, 507 pp.

Thornton, K.W., Kennedy, R.H., Magoun, A.D. and Saul, G.E. 1982 Reservoir water quality sampling design. *Water Resources Bull.*, **18**, 471-480.
