

ESTIMATING A GLOBAL MEAN

A guide for data analysts and interpreters on the estimation of an average contaminant concentration over a large area or volume.

This guidance document is one of a series that outlines important basic statistical concepts and procedures that are useful in contaminated sites studies. BC Environment recommends that these suggestions be followed where applicable, but is open to other techniques provided that these alternatives are technically sound. Before a different methodology is adopted it should be discussed with BC Environment.

April 2001

THE GENERAL IDEA

At various stages in the study of a contaminated site, estimates are required of the average value of the contaminant concentration over a large area or volume. If the available samples fairly represent the underlying population, then the arithmetic average of the samples serves as an unbiased estimate of the mean of the underlying population. Furthermore, the level of uncertainty can easily be quantified. Unfortunately, the available samples are often clustered, with "hot spots" being preferentially sampled once they are encountered. In such situations, the mean of the sample values is a biased estimate of the true average since the more highly contaminated areas are over-represented in the sample data base.

This guidance document addresses the problem of estimating the global mean and quantifying the uncertainty in this estimate. It begins with the most tractable and convenient situation, in which the samples fairly represent the underlying population and can all be given equal weight. It then considers the more common practical situation, where the available samples are preferentially clustered in certain areas and do not fairly represent the underlying distribution. It specifically discusses two approaches to the problem of preferential clustering in the sample locations: cell declustering and polygons of influence. Though there are other ways of producing unbiased estimates of the global mean from spatially clustered samples, these two methods are among the most common and will provide a good indication of the sensitivity of the estimate of the global mean to preferential sampling.

EQUALLY WEIGHTED AVERAGES

The most straightforward procedure for estimating the average value over a large area or volume is to use as an estimate the arithmetic average of the available samples:

$$\text{Estimate of global mean} = \frac{1}{n} \sum_{i=1}^n v_i$$

v_1, \dots, v_n are the n available sample values, each one of which receives the same weight in the estimation of the global mean. This equal weighting of the available samples is reasonable in situations where any sample is as representative of the underlying population as any other sample. As discussed in the document entitled *SAMPLING PLANS*, however, the available samples from a contaminated site are usually not equally representative of the underlying population. The more common situation is that the samples have been preferentially located in certain regions, either based on visual observation or on high sample

values from earlier sampling campaigns. Later in this document, we will present procedures for dealing with preferentially clustered samples.

One situation in which the available samples can be regarded as equally representative is where they are located on a regular grid. Another is where the sample locations are randomly selected with no thought being given to visual criteria or earlier sample information. Though these situations are rare in most contaminated site statistical applications, they may occur when material has been stockpiled and samples are being collected to allow an estimation of the average contaminant concentration of the entire stockpile.

QUANTIFYING UNCERTAINTY

In addition to estimating the mean of the underlying population, we also often need to quantify the uncertainty on such an estimate. The uncertainty on a global mean is usually expressed by a quantity known as the "standard error", which is calculated as follows:

$$\text{Standard error of global mean} = \sigma_{\text{global mean}} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the available samples and n is the number of available samples. The standard error can be thought of as the standard deviation of the distribution of the underlying true global mean. Though there is only one true global mean, we don't know what it is and our uncertainty entails that there is some range of possible values; the standard error describes the breadth of this range. If the standard error is very high, then the range of possible values is very broad and we don't know very much about the true underlying mean; this can be caused either by having a large value of s (which means that the available sample values are very erratic) or by having a small value of n (which means that we have only a very few samples). If s is small or if n is large, then the standard error will be small, which signifies that the true underlying mean must fall within a narrow range of possible values.

For many classification problems, we need to make sure that the average concentration of the material being classified is almost certainly below a specified threshold. Rather than comparing the arithmetic average of the sample values to the specified threshold, we need to choose a pessimistically high estimate of the underlying mean and make sure that even if the true average concentration of the material is as high as this pessimistic estimate, it would still fall below the threshold. The pessimistic estimate of the global mean needed for this kind of classification problem is usually calculated by taking the arithmetic average of the sample values and adding twice the standard error. As

an example of this procedure, suppose that we are trying to check whether the average arsenic concentration in a stockpile is below 100 $\mu\text{g/g}$, and that we have 25 randomly selected samples whose arithmetic average is 87 $\mu\text{g/g}$ and whose standard deviation is 45 $\mu\text{g/g}$. In this example, the available samples are all equally representative of the stockpile since all sample locations were randomly selected, and the arithmetic average serves as an unbiased estimate of the true mean concentration of the stockpile. The standard error on the global mean is $45 \div \sqrt{25} = 9 \mu\text{g/g}$. A pessimistically high estimate of the mean concentration of the stockpile would be $87 + 2 \times 9 = 105 \mu\text{g/g}$. For this particular example, there is enough uncertainty about the global mean that it is not safe to assume the mean arsenic concentration of the entire stockpile is below 100 $\mu\text{g/g}$ even though the average of the available samples is only 87 $\mu\text{g/g}$.

If we have more than 20 samples that are statistically independent from one another, we can assume the probability distribution of the unknown global mean is a normal distribution. Under this assumption, there is a 95% chance that the unknown global mean will be within two standard errors of the arithmetic average of the available data values. The pessimistic estimate described in the previous paragraph is often referred to as the "upper 95% confidence limit of the global mean".

BIAS CAUSED BY PREFERENTIAL SAMPLING

Figure 1 shows an example of mercury measurements taken from a contaminated site in three sampling campaigns. The first 7 samples were taken haphazardly throughout the site since no coherent sampling plan had yet been developed. The second group of 14 samples covers the area with a regular grid and the third group of 12 samples provides additional detail in the areas with the highest mercury concentrations.

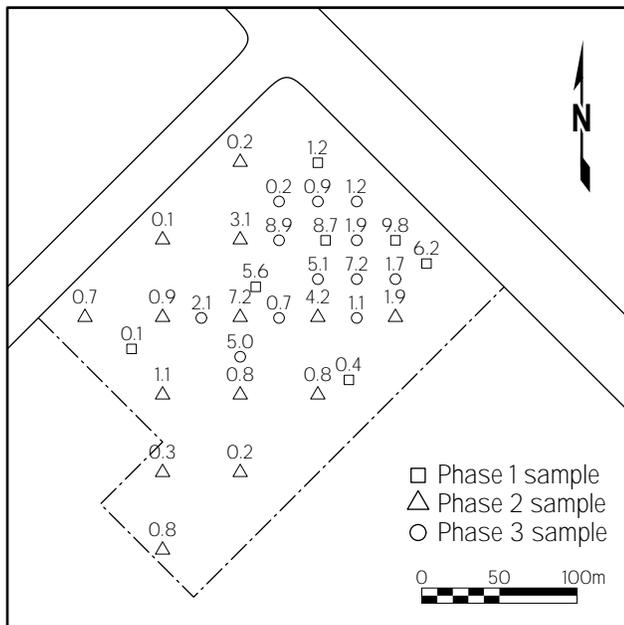


Figure 1 Mercury samples from three sampling campaigns.

Table 1 shows how the mean of the sample values varies in each of the three sampling campaigns. In the initial group of 7 samples, the average mercury concentration was 4.57 $\mu\text{g/g}$.

In the second group of 14 samples, the average dropped to 1.59 $\mu\text{g/g}$. In the third group of 12 samples, the average increased to 3.00 $\mu\text{g/g}$. It is clear that the preferential sampling of the most highly contaminated regions has caused the higher mercury values to be over-represented with the result that the naive sample mean of 2.74 $\mu\text{g/g}$ based on all 33 samples is likely an overestimate of the actual average mercury concentration over the entire site.

Table 1 Sample means by campaign.

	Number of Samples	Average Hg Concentration
Phase 1	7	4.57
Phase 2	14	1.59
Phase 3	12	3.00

WEIGHTED AVERAGES

Estimates of the global mean from spatially clustered data can be produced by using a weighted average of the data values rather than the equally-weighted average discussed earlier. A weighted average can be written as

$$\text{Weighted average} = \sum_{i=1}^n w_i \cdot v_i$$

where v_1, \dots, v_n are the n available sample values, and where w_1, \dots, w_n are the corresponding weights that sum to 1.

The weight given to each sample reflects its importance to the global mean. To mitigate the influence of preferential sampling on an estimate of the global mean, we need to give lower weights to the values from densely sampled areas and higher weights to the values from sparsely sampled areas. Though this general principle is used by many different declustering methods, they differ in the details of the calculations and in the exact weight assigned to each sample.

Cell declustering

One of the simplest procedures for choosing declustering weights is to overlay a grid of cells, as shown in Figure 2, and to make the weight of each sample inversely proportional to the number of samples in the same cell. This is equivalent to calculating the average value within each cell and then averaging the cell averages. Figure 3 shows the cell averages for the 17 cells shown in Figure 2; the average of these cell averages, 1.97 ppm, serves as a declustered estimate of the global mean.

With cell declustering, the main stumbling block in practice is the choice of the cell size. In Figure 2, we chose to use 50x50 m cells; but why didn't we choose 100x100 m, or 25x25 m, or even 100x50 m? Each of these cell sizes would result in a different estimate of the global mean. The recommended practice with cell declustering is to choose a cell that matches the spacing of the most regularly spaced subset of the samples. In our example with the mercury contamination, the second phase of samples was on a 50x50 m grid. If there is no quasi-regular subset of samples, then the common practice is to try many different cell sizes, from very small ones to very large ones, and then select the one that minimizes the global mean.

The selection of the minimum estimate of the global mean is predicated on the assumption that all of the clustered samples are in areas with high values. If this is not the case — if some of the clustered samples are in areas with moderate or low values — then it is difficult to justify any particular choice of cell size.

polygon of influence. Figure 4 shows the polygons of influence for the mercury samples used in the earlier examples. The edges of these polygons are the perpendicular bisectors between the pairs of samples; all locations within any polygon are closer to the central sample than to any other sample. In densely sampled areas, the polygons will tend to be smaller and this makes the area of the polygon of influence a natural candidate for a declustering weight.

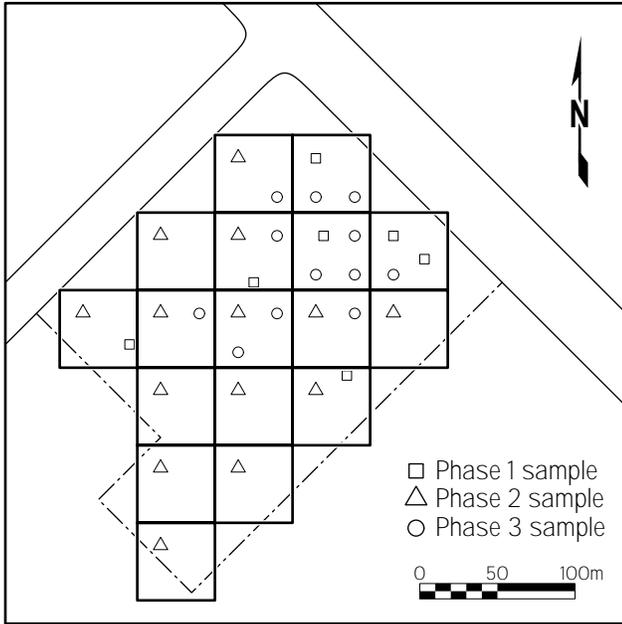


Figure 2 50x50 m cells over the site.

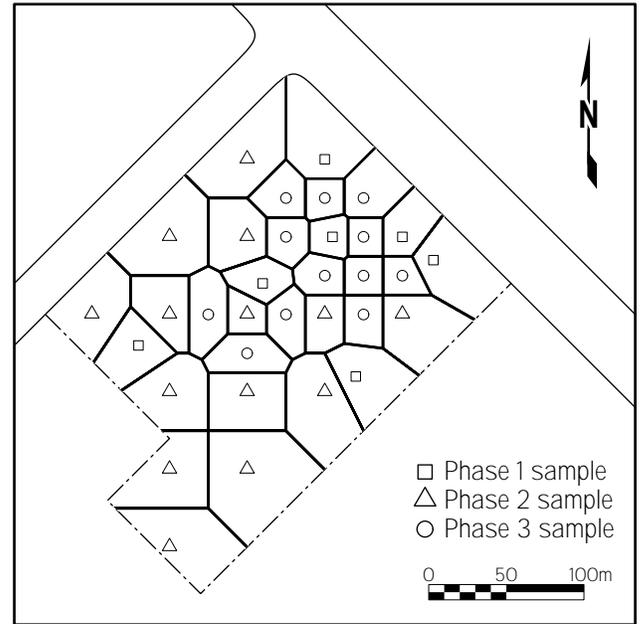


Figure 4 Polygons of influence.

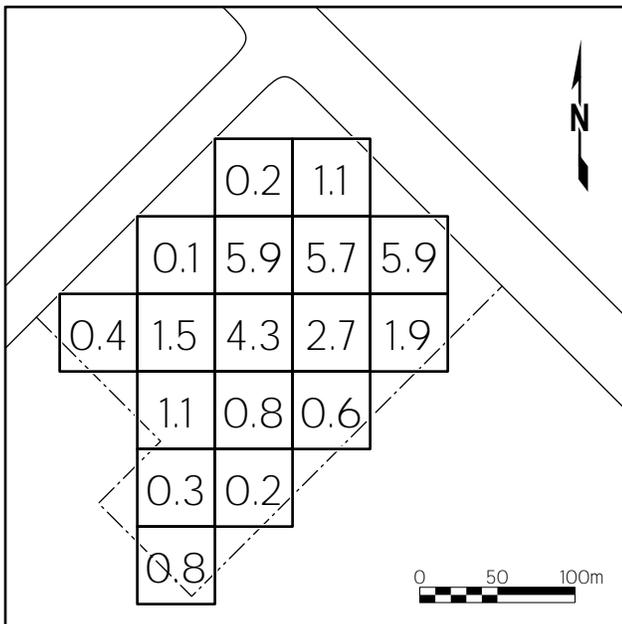


Figure 3 Average Hg concentration in each cell.

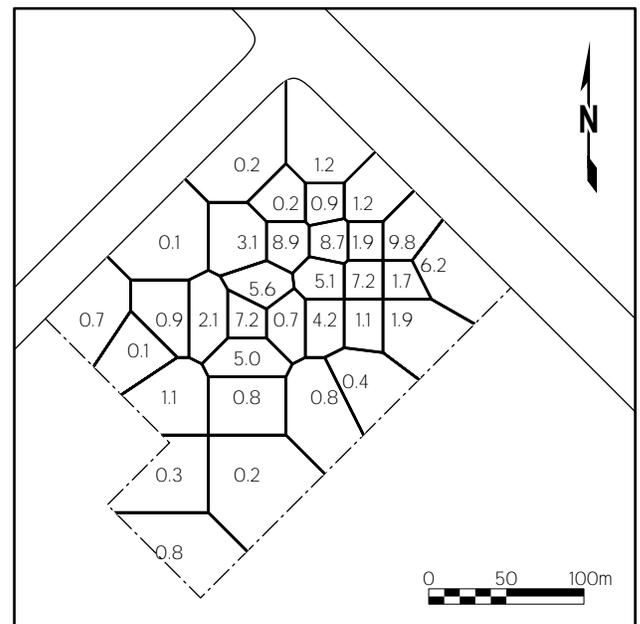


Figure 5 Sample values for polygons of influence.

Further detail on the cell declustering method can be found in Isaaks and Srivastava (1989); Deutsch (1989) provides a computer program that implements this approach.

Polygons of influence

One of the oldest methods for spatial declustering is to give each sample a weight that is proportional to the area of its

Figure 5 shows the sample values assigned to each of the polygons from Figure 4. When these sample values are weighted by the areas of their respective polygons, the resulting estimate of the global mean is 1.98 $\mu\text{g/g}$. The remarkable agreement

between this polygonal estimate of the global mean and the cell declustering estimate of 1.97 $\mu\text{g/g}$ obtained earlier is purely fortuitous in this case. For most site studies, the estimates obtained by the two approaches are more different.

With the polygonal approach, the main source of arbitrariness is the decision about how to close the outer polygons that are not naturally bounded by other sample values. In the example shown in Figure 4, the property boundary was used to limit the areal extent of the edge polygons. Unfortunately, in practice the areal extent of the polygons often does not have a clearly defined limit. When no such boundary suggests itself, the common practice is either to limit the size of the polygons to the average spacing between the available samples or to the range of correlation as defined through geostatistical analysis of the spatial variation. The guidance document entitled *SAMPLING PLANS* provides a brief introduction to the range of correlation and the analysis of spatial variation.

Further detail on the use of polygons of influence can be found in Isaaks and Srivastava (1989); Hayes and Koch (1984) provide a computer program that implements this approach.

UNCERTAINTY FOR WEIGHTED AVERAGES

When the global mean is estimated from a weighted average, the uncertainty in the estimate can be quantified using the following equation:

$$\sigma_{\text{global mean}} = s_{\text{weighted}} \times \sqrt{\sum_{i=1}^n w_i^2}$$

where w_i is the weight given to the i -th sample value, and s_{weighted} is the standard deviation of the samples calculated using the same set of weights:

$$s_{\text{weighted}} = \sqrt{\sum_{i=1}^n w_i \cdot (v_i - \text{weighted mean})^2}$$

RECOMMENDED PRACTICE

1. When the available samples are located on a regular grid or are the result of a formal randomization of sample locations, then the arithmetic average of the sample values is an appropriate estimate of the underlying global mean.
2. When the available samples have been preferentially located in certain areas and not in others, then a weighted average of the available sample values should be used to estimate the underlying global mean.
3. Since the cell declustering and the polygonal method both have a certain arbitrariness, the recommended practice is to try both procedures rather than to rely exclusively on one or the other. If both approaches result in an estimate that is markedly different from the equally weighted sample mean, then the spatial clustering of the data does have a pronounced effect on sample statistics and this should be taken into account whenever the study calls for an estimate of a global statistic from sample data.

If the results of cell declustering and the polygonal method are very similar, as in the mercury example shown in this guidance document, then either estimate is acceptable. If the two values are quite different, then the cell declustering estimate should be accepted if a subset of the available samples is on a quasi-regular grid. If no such regular grid exists then the estimate from the polygons of influence is preferable.

4. When ever the cell declustering approach has been adopted, the report should contain a clear discussion of the choice of cell size. Whenever the polygonal method has been adopted, the report should contain a clear discussion of the choice of the boundary that limits the edge polygons.

REFERENCES AND FURTHER READING

In addition to the other guidance documents in this series, the following references provide useful supplementary material:

- Deutsch, C.V., "DECLUS: A Fortran 77 program for determining optimum spatial declustering weights," *Computers and Geosciences*, v. 15, p. 325-332, 1989.
- Hayes, W., and Koch, G., "Constructing and analyzing area-of-influence polygons by computer," *Computers and Geosciences*, v. 10, p. 411-431, 1984.
- Isaaks, E.H., and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.