BRITISH
COLUMBIA
Ministry of Environment

12

TECHNICAL GUIDANCE
ON CONTAMINATED SITES

October 2005

Statistics for Contaminated Sites

The series of documents listed here provides information and guidance on various aspects of the application of statistics to contaminated sites studies. These publications have been developed by FSS International for the ministry.

Summary of current FSS documents

There are currently 16 guidance documents in this series.

12-1, "Univariate Description" discusses statistical concepts and presents tools for describing the statistical characteristics of a single variable.

12-2, "Bivariate Description" addresses the statistical analysis of pairs of variables and presents tools for describing the relationship between variables.

12-3, "Spatial Description" presents various tools for describing and analyzing data in their spatial context.

12-4, "Distribution Models" presents basic information on some of the statistical distribution models that are commonly used in contaminated site studies.

12-5, "Non-Parametric Statistics" discusses statistical procedures that do not depend on distribution models.

12-6, "Choosing a Distribution" presents advice on how to decide which distribution model is most appropriate.

12-7, "Identifying Populations" presents tools that can be used to help with the decision of whether to treat the data as one population or to split them into two or more subpopulations.

12-8, "Outliers" discusses the evaluation and treatment of unexpected and erratic high values.

12-9, "Estimating a Global Mean" addresses the estimation of an average value over a large area and the quantification of the uncertainty on such estimates.

12-10, "Composite Samples" presents advice on the interpretation of analytical values from composite samples that have been created from two or more discrete samples.

12-11, "Statistical QA/QC" discusses issues related to the monitoring, documentation, and control of the reliability and repeatability of sample information.

12-12, "Sampling Plans" addresses the design of appropriate sampling plans for various purposes throughout the life a contaminated site project.

12–13, "Classification" provides information on how to classify contaminated material into an appropriate regulatory category.

12–14, "Stockpiling" discusses the appropriate sampling and classification of stockpiled material.

12–15, "Reporting" provides general advice on the content of a report of a statistical study for a contaminated site.

12–16, "Randomization" presents procedures for randomly selecting samples from larger batches and for randomly selecting sample locations.

Alternatives to this technical guidance

All of these FSS statistical guidance documents have the following brief explanatory note in the header on their front page:

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

This makes it clear that the ministry views the information in these guidance documents as an appropriate starting point for statistical studies of contaminated sites. It recognizes that prescribing a rigid procedure is not appropriate for all contaminated sites, and it is willing to consider site-specific alternatives that are technically defensible. We recommend, however, that such alternatives be discussed with the ministry so that consensus can be reached on the appropriateness of different approaches.

Also important to note is that these 16 documents were written before the Contaminated Sites Regulation came into effect. Therefore terms such as "industrial waste," "residential waste," and "total PAH" still appear in some of the documents even though they are not used in the new Regulation.

For more information, contact the Environmental Management Branch at site@gov.bc.ca .

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-1

UNIVARIATE DESCRIPTION

A guide for report writers, reviewers, data analysts and interpreters on exploratory data analysis for one variable

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

The application of statistics to contaminated site studies requires a clear and coherent understanding of the available data. For those directly involved in statistical analysis and interpretation, a clear and coherent understanding of the data will help them to select appropriate statistical tools and to make critical assumptions about statistical populations. For those who prepare statistical reports, it is important that their reports convey a clear and coherent understanding of the data to their audience; the readers of a report will not be able to form an opinion about the validity of the study's conclusions without a good understanding of the data on which it is based.

This guidance document discusses tools for exploratory data analysis, a statistical study's first step in which we investigate the available data, form tentative opinions and modify these opinions as our understanding of the data improves and evolves. The same tools that help us explore and interpret the available data are also ideal for presenting and summarizing our understanding of the data to those not directly involved in the study. This guidance document should therefore be of assistance not only to those who actually do the statistical analysis and interpretation, but also to those who are responsible for writing reports. This document is not intended to provide a rigid prescription for how to perform and present an exploratory data analysis; indeed, as noted in the final section of this document, such a rigid prescription would not permit us to exercise the curiosity that is one of the cornerstones of thorough exploratory data analysis. This document does intend, however, to encourage some much needed consistency in the performance and presentation of statistical studies by providing a simple and straightforward approach to exploratory data analysis.

This guidance document focuses on the exploratory data analysis of a single variable, such as the concentration of a single contaminant. Two other documents in this series focus on other aspects of exploratory data analysis. *BIVARIATE DESCRIPTION* focuses on tools for analyzing the relationship between pairs of variables; *SPATIAL DESCRIPTION* focuses on tools for analyzing the data in their spatial context.

PROVIDING DETAIL & CONVEYING INFORMATION

With all statistical presentations there is a tradeoff between the level of detail in the presentation and the amount of information that it conveys. Table 1 and Figure 1 demonstrate this tradeoff using data from a site contaminated with mercury. Table 1 provides the most detailed and complete information about the available data values and yet it does not immediately convey much information. By sacrificing some of the detail, the histogram shown in Figure 1 more immediately conveys useful information about the available data by giving us a quick appreciation of the fact that there are many low values around 1 ug/g and only a few erratic high ones above 10 ug/g. Though this fact can also be extracted from Table 1, the histogram makes it more readily apparent.

Table 1 Hg measurements (in ug/g) from a contaminated site.

1.08	1.10	7.27	0.30	5.01	1.97	1.58	0.67	0.06	0.22
0.28	3.22	0.46	1.13	0.78	7.44	0.08	0.45	14.91	0.26
0.32	11.47	8.93	0.93	1.81	9.40	0.52	12.89	5.40	1.52
1.25	15.87	2.31	0.70	11.31	4.64	0.69	0.06	0.45	3.59
0.54	0.76	4.80	7.92	8.46	0.90	0.71	8.94	2.01	9.83

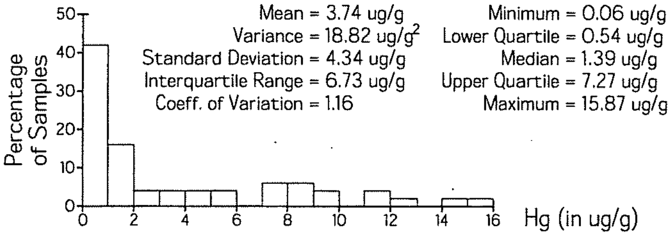


Figure 1 A histogram of the mercury data in Table 1.

As we compile a statistical description, we should keep this tradeoff in mind and should select graphical and numerical summaries that convey useful and salient information about the data. The various components of our statistical description will often be somewhat redundant; some of the statistics in Figure 1, for example, convey similar information as others. Such redundancy is not a flaw, however, as long as each component successfully conveys useful additional information. The guiding principle should be clarity of understanding — a good statistical presentation is one that enables others who are unfamiliar with the site to share our understanding of the data.

SUMMARY STATISTICS

Measures of center

The statistic most commonly used to summarize where the center of a distribution lies is the mean, which is simply the arithmetic average of the data values:

$$\text{Mean} = m = \frac{1}{n}\sum_{i=1}^{n} v_i$$

Though the mean is the traditional measure of the center of a distribution, it is strongly influenced by erratic high values and may not correspond to our intuitive sense of the middle of the distribution. For the mercury data shown in Figure 1, for example, more than two-thirds of the values are smaller than the mean value of 3.74 ug/g, so it is not clear why this qualifies as a "central" value. For contaminated site data, which often span several orders of magnitude, it is common to find that the vast majority of the data values fall below the mean. Had the largest value in Table 1 been an order of magnitude higher, at 158.7 ug/g rather than 15.87 ug/g, the mean would nearly double to 6.60 ug/g higher than 75% of the data.

For data that span several orders of magnitude, the median is less sensitive to extreme values and provides a stable statistic that corresponds more closely to our intuitive sense of the center of the distribution. The median is the number that appears halfway down the list of values when they are sorted from smallest to largest; for an even number of data, the median is the average of the middle two values. Since it depends only on the ordering of the data, the median would not be changed if the largest value was an order of magnitude higher.

Taken together, the mean and the median provide an indication of the influence of extreme values in a data set. If the two measures of the center are close to each other, then extreme values do not play much of a role. This is not typically the case with contaminated site data. It is not unusual to find that there are some influential extreme values that cause the mean of the data to be more than twice the median.

Measures of location

The statistics that are used to describe the location of other parts of the distribution can all be calculated easily from a sorted list of the data values. The minimum is the first value on the sorted list and the maximum is the last value on the sorted list. The "quartiles" provide two other useful measures of location. In the same way that the median splits the data set into halves, the quartiles split it into quarters. 25% of the data values are below the lower (or first) quartile and 25% of them are above the upper (or third) quartile.

In most contaminated site studies, the lowest values are below the detection limit. Rather than reporting these values as exactly half the detection limit, as is often done, it is more useful in a statistical summary to state the detection limit and to report how many values fall below it.

Measures of spread

In addition to describing where the center of the distribution lies, a complete statistical description should also report how the data values are spread around the center — are they all tightly grouped close to the center or are they scattered far away from the center? The statistics that are commonly used to describe the spread of the distribution are the variance, s^2, and the standard deviation, s. The sample variance is the average squared difference of the data values from their mean:

$$\text{Variance} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(v_i - m)^2$$

The standard deviation is the square root of the variance. Like the mean, and all other statistics that involve an averaging of the data, the variance and standard deviation are both sensitive to extreme values. Had the largest value in Table 1 been an order of magnitude higher, at 158.7 ug/g rather than 15.87 ug/g, the variance would soar from less than 20 to nearly 500 ug/g^2! With a single extreme value having such a profound influence, the variance and standard deviation are often difficult to interpret. For most contaminated site data, the interquartile range (IQR) is a more stable and interpretable alternative. The IQR is the difference between the upper quartile and lower quartile and provides a direct measurement of the spread of the middle 50% of the data values. Since it depends only on the quartiles, the IQR is insensitive to the exact values of the most extreme data.

Several of the guidance documents in this series warn that certain procedures should not be used if the data are not from a homogeneous population. Though it is difficult to give exact specifications for "homogeneity", an appropriate starting point is the measure of spread called the coefficient of variation, which is the ratio of the standard deviation to the mean: CV = s ÷ m. This measure of relative variation is often expressed in percent, rather than as a ratio. For data whose CV is 1 (or 100%), their standard deviation is as big as their mean. If the data are to be considered "homogeneous", their CV should be smaller than 1. By itself, this is not a guarantee that the data do come from a common population; qualitative information, such as the site history and the provenance of the data, also needs to be taken into account. If the CV is larger than 1, however, it is unlikely that the samples are from a single population; the large spread in the data values is likely a warning that different samples have been affected by different physical and chemical processes.

Measures of shape

The final summary statistic that is often included in a statistical description is a measure of the shape or symmetry of the distribution. The symmetry of a distribution can be described by comparing the mean to the median, and can also be captured in a statistic called the skewness. Though the skewness does have a specific formula, we rarely need to know its precise value and usually report only its sign, either positive or negative.

Positively skewed distributions have a lot of low values and a decreasing proportion of high values; the histogram of positively skewed data is asymmetric with a tail to the right, like the one shown in Figure 1. Negatively skewed distributions, which are rare in contaminated site studies, have a lot of high values and a decreasing proportion of low values; their histogram has a tail to the left. Occasionally we encounter contaminated site data whose skewness is very minor and whose histogram appears symmetric. The guidance document *CHOOSING A DISTRIBUTION* offers a rule of thumb that can be used to decide if the distribution can be deemed symmetric.

GRAPHICAL TOOLS

By themselves, summary statistics do not always convey all of the important information about a data set. Graphical presentations of the data provide valuable visual support to readers who are trying to follow the details of a statistical study. A combination of graphical displays and numerical summaries is the most effective vehicle for conveying our understanding of the data to readers who are not familiar with the project.

Histograms

The most common graphical presentation of data is a histogram that shows how many samples fall in different categories. Using the mercury data presented on Table 1 on the previous page, Table 2 records how many samples fall within each of sixteen classes, from 0–1 ug/g to 15–16 ug/g. Figure 1 presents this information as a histogram on which the height of a bar is equal to the percentage of samples in that class.

Table 2 Frequency table of mercury data in Table 1.

Class (in ug/g)	No. of samples	% of total	Class (in ug/g)	No. of samples	% of total
0–1	21	42	8–9	3	6
1–2	8	16	9–10	2	4
2–3	2	4	10–11	0	0
3–4	2	4	11–12	2	4
4–5	2	4	12–13	1	2
5–6	2	4	13–14	0	0
6–7	0	0	14–15	1	2
7–8	3	6	15–16	1	2

It is often awkward to select a class width for histograms of contaminated site data since the data span several orders of magnitude. If a large class width is used in an attempt to display the entire range of the data values, then the first class on the histogram often gets the lion's share of the samples and the display does not provide much detail on the distribution of the lower values. If a small class width is used in an attempt to show more of the detail of the distribution of the lower values, then the number of classes needed to span the entire range becomes unmanageable. One solution to this common problem is to show two histograms, one that spans the entire range with wide classes and another that shows the details of the low end of the distribution with smaller classes.

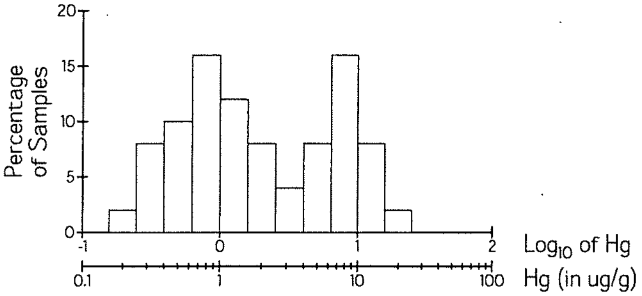


Figure 2 A logarithmic histogram of the data in Table 1.

Another way of dealing with data that span several orders of magnitude is to choose classes that have equal widths on a logarithmic scale. Figure 2 shows the mercury data from Table 1 plotted as a histogram on a logarithmic scale. One of the advantages of a logarithmic histogram is that it often makes different populations more apparent. On Figure 2, for example, the low background mercury values form one clear bump or "mode" around 1 ug/g, while the high contaminated values show another mode around 10 ug/g.

Cumulative plots and probablity plots

Using the mercury data presented on Table 1 on the previous page, Table 3 records the number and percentage of samples that fall below the sixteen thresholds from 1 to 16 ug/g; in the cumulative plot shown in Figure 3, the threshold values in the first column of Table 3 are used as the x-coordinates and the cumulative percentages in the last column are used as the y-coordinates. The cumulative plot in Figure 3 has a logarithmic x-axis to accommodate the skewness in the data; if the distribution had been more symmetric, a linear x-axis would have been more appropriate.

Table 3 Cumulative frequency table of mercury data in Table 1.

Threshold (in ug/g)	No. of samples below	% of total	Threshold (in ug/g)	No. of samples below	% of total
1	21	42	9	43	86
2	29	58	10	45	90
3	31	62	11	45	90
4	33	66	12	47	94
5	35	70	13	48	96
6	37	74	14	48	96
7	37	74	15	49	98
8	40	80	16	50	100

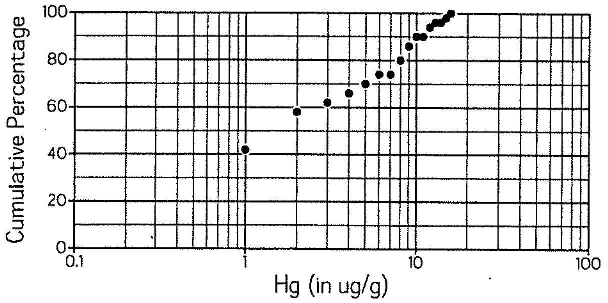


Figure 3 A cumulative plot of the mercury data in Table 1.

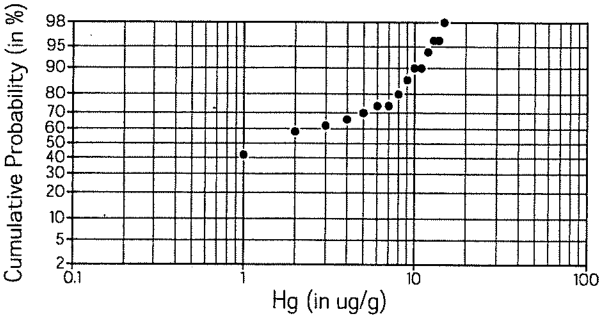


Figure 4 A probability plot of the mercury data in Table 1.

When cumulative plots are presented on special graph paper called "probability paper" they are usually called "probability plots". Figure 4 presents the data from Table 3 as a probability plot. The probability axis on this kind of plot is squashed in the middle and stretched at the top and bottom. The reason for plotting cumulative curves on this distorted grid is that it simplifies the checking of whether the distribution of the data values is close to that of a mathematical model called the "normal" or "gaussian" distribution. The histogram of normally distributed data will be shaped like a bell. Rather than checking how bell-like the histogram is, it is easier to check how straight the probability plot is. The distorted grid of probability paper is designed in such a way that the cumulative curve of normally distributed data will plot as a straight line.

Boxplots

Figure 5 shows another graphical tool that is becoming popular in exploratory data analysis. The box goes from the lower quartile to the upper quartile and therefore spans the middle 50% of the data values. The bar in the middle of the box shows the median and the dot shows the mean. The arms that stick out of the box go to the minimum and maximum. As with other graphical presentations, logarithmic scaling often makes the boxplot more informative if the distribution of data values is skewed. The simple boxplot captures most of the critical information about a distribution — its center, its spread and its skewness — in a format that is more compact than a histogram.

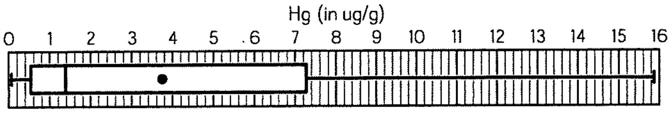


Figure 5 A boxplot of the mercury data in Table 1.

DATA BASE COMPILATION AND VERIFICATION

An exploratory data analysis is only as good as the data on which it is based; when there are errors in the data base, our exploratory data analysis is often useless and misleading. With the data from contaminated site studies often having to be transcribed, keypunched or electronically merged from some other source, there are ample opportunities for human error. Before attempting an exploratory data analysis, we need to know how the data base was created. If the data base is not accompanied by a clear audit trail that explains all of the steps involved in its creation, then it should be verified against original records wherever possible.

One of the best ways to verify a data base is by using teams of two people to proofread the data. One person reads out loud the data values from a hardcopy of the data base while the other checks each value against *original* records, such as laboratory reports or surveyor's notes. Though two person proofing is a tedious exercise, it is a very important one if the integrity of the data base is uncertain. Experience has shown that it provides a much more complete and exhaustive verification of a data base than automated or computer-based techniques, which can do no more than check one electronic version of the data base against another. If the original data were not recorded electronically, but were recorded manually and later transcribed, then verifying one electronic version against another cannot catch mistakes that crept into the data before or during the creation of the first electronic version.

Once the integrity of the data base is well documented, every effort should be made to maintain this integrity. In many contaminated site studies, where there are several phases of data collection, the integrity of the early data base is lost as various people merge new data and modify old data to suit their individual purposes. Concentrations may need to be converted from one unit of measurement to another, or the coordinate system used by one group may need to be transformed to the coordinate system used by another group. With every such modification of the data there are opportunities for error. Such opportunities increase with every new person who has access to the data base and is able to make modifications. Once the integrity of a data base is lost, restoring it will either require considerable effort or will be completely impossible.

On large projects, where there are more than 100 samples or where several groups have been collecting data, one person should have the responsibility for maintaining the authoritative and verified data base. Others may obtain copies for their own work but none of their individual changes should be accepted in the single authoritative version until the data base coordinator approves the change and prepares documentation that explains exactly what changes were made and why.

RECOMMENDED PRACTICE

It is not possible to give a rigid prescription for exploratory data analysis since a thorough understanding of the data requires both creativity and curiosity. The sequence of steps that worked on one project will not always work on another one. The following general guidelines, however, should improve any exploratory data analysis for contaminated site studies:

1. Before exploratory data analysis, the integrity of the data should be documented, either by reference to a report on procedures used to compile and verify the data or by a complete check of all data against original records.
2. Complete listings of all data used in statistical studies should be included as appendices to reports; these do not, however, constitute an appropriate statistical summary. Statistical summaries of univariate data should include:
   (a) Graphical presentations of the data, such as histograms, probability plots or boxplots. If the data are skewed, then logarithmic scaling will often make such graphical presentations more informative.
   (b) Summary statistics that describe the center, location, spread and shape of the distribution of data values. If the data are skewed, then the mean and standard deviation should not be used alone to summarize the data but should be accompanied by other measures, such as the median and interquartile range, that are not so sensitive to extreme values.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Understanding Robust and Exploratory Data Analysis, (Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds.), John Wiley & Sons, New York, 1983.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Moore, D.S., *Statistics: Concepts and Controversies*, W.H. Freeman and Company, New York, 1985.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-2

BIVARIATE DESCRIPTION

A guide for report writers, reviewers, data analysts and interpreters on exploratory data analysis for two variables

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

The application of statistics to contaminated site studies requires a clear and coherent understanding of the available data. For those directly involved in statistical analysis and interpretation, a clear and coherent understanding of the data will help them to select appropriate statistical tools and to make critical assumptions about statistical populations. For those who prepare statistical reports, it is important that their reports convey a clear and coherent understanding of the data to their audience; the readers of a report will not be able to form an opinion about the validity of the study's conclusions without a good understanding of the data on which it is based.

This guidance document discusses tools for exploratory data analysis, a statistical study's first step in which we investigate the available data, form tentative opinions and modify these opinions as our understanding of the data improves and evolves. The same tools that help us explore and interpret the available data are also ideal for presenting and summarizing our understanding of the data to those not directly involved in the study. This guidance document should therefore be of assistance not only to those who actually do the statistical analysis and interpretation, but also to those who are responsible for writing reports. This document is not intended to provide a rigid prescription for how to perform and present an exploratory data analysis; indeed, as noted in the final section of this document, such a rigid prescription would not permit us to exercise the curiosity that is one of the cornerstones of thorough exploratory data analysis. This document does intend, however, to encourage some much needed consistency in the performance and presentation of statistical studies by providing a simple and straightforward approach to exploratory data analysis.

This guidance document focuses on the exploratory data analysis of the relationship between pairs of variables. Two other documents in this series focus on other aspects of exploratory data analysis. *UNIVARIATE DESCRIPTION* focuses on tools for analyzing a single variable; it also addresses the important first step of verifying the data base. *SPATIAL DESCRIPTION* focuses on tools for analyzing the data in their spatial context.

PROVIDING DETAIL & CONVEYING INFORMATION

With all statistical presentations there is a tradeoff between the level of detail in the presentation and the amount of information that it conveys. Table 1 and Figure 1 demonstrate this tradeoff using data from a site contaminated with PCBs. Table 1 provides the most detailed and complete information about the available data values and yet it does not immediately convey much information. The scatterplot shown in Figure 1 does not show us the precise values of all the data, and is therefore slightly less detailed than the complete listing. By sacrificing some of the detail, however, the scatterplot more immediately conveys useful information about the available data by giving us a quick appreciation of the fact that there is a strong relationship between clay content and PCB concentration — high PCB values tend to be associated with soil that has a high clay content. Though this fact could also have been extracted from Table 1, the scatterplot makes it more readily apparent.

Table 1 Measurements of clay content (in %) and PCB concentration (in ug/g) from a contaminated site.

Clay	PCB	Clay	PCB	Clay	PCB
80	0.9	90	7.6	14	1.1
90	9.4	010	121.8	41	6.2
100	59.7	90	4.5	52	9.3
80	3.7	90	35.4	49	11.2
90	11.8	100	95.6	54	3.2
90	16.2	100	27.2	58	19.1
80	0.6	80	2.1	21	5.6
80	5.4	80	0.8	93	184.0
80	1.3	90	105.6	39	3.5
80	3.2	80	0.8	67	49.5
80	1.5	90	54.5	54	8.5
90	86.8	100	53.3	55	11.2
80	1.5	100	32.8	31	3.0
80	3.4	90	12.4	94	124.2
80	1.3	100	59.5	74	44.9
90	26.3	100	105.5	32	3.8
90	8.8	90	28.0		

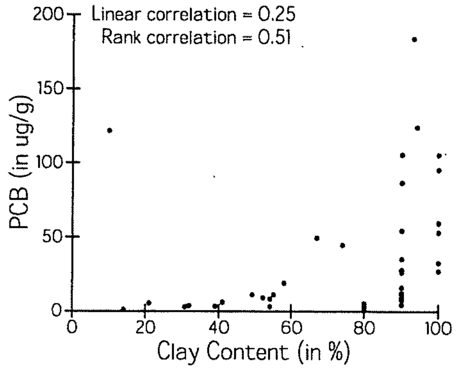


Figure 1 A scatterplot of the clay – PCB data in Table 1.

As we compile a bivariate statistical description, we should keep this tradeoff in mind and should select graphical and numerical summaries that convey useful and salient information about the relationship between the two variables. The various components of our statistical description will often be somewhat redundant; the two statistics given in Figure 1, for example, convey similar information. Such redundancy is not a flaw, however, as long as each component successfully conveys useful additional information. The guiding principle should be clarity of understanding — a good statistical p resentation is one that enables others who are unfamiliar with the site to share our understanding of the data.

SUMMARY STATISTICS

Linear correlation coefficient

The correlation coefficient that is commonly used to summarize the relationship between two variables is calculated as follows:

$$\text{Linear correlation} = r = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i\right) - m_x \cdot m_y}{s_x \cdot s_y}$$

The term in the brackets is the average of the products between each pair of data values; using the example from Figure 1, the x_i's would be the clay content values and the y_i's would be the PCB concentrations. m_x is the mean of the x values and s_x is their standard deviation; m_y is the mean of the y values and s_y is their standard deviation.

The correlation coefficient is always between −1 and +1. When the two variables are perfectly linearly related and increase together, then their correlation coefficient will be +1. If the two variables are perfectly linearly related but one increases when the other one decreases, then the correlation coefficient will be −1. Figure 2 shows examples of scatterplots with correlation coefficients ranging from −0.8 to +0.8.

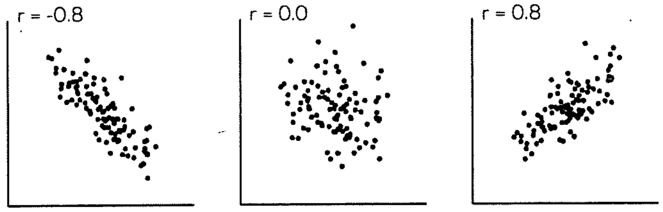


Figure 2 Examples of different correlation coefficients.

A correlation coefficient close to +1 or −1 usually indicates a strong relationship between two variables. It should be noted, however, that a strong correlation does not necessarily imply a causal relationship between the two variables.

Shortcomings of the linear correlation coefficient

Though the linear correlation coefficient is the most common summary of the relationship between two variables, it has some practical shortcomings for contaminated site studies. Like other statistics that involve an averaging of the data values, such as the mean and variance, the linear correlation coefficient is strongly influenced by extreme values.

An example of this sensitivity to extreme values can be seen in Figure 1. Though there is some visible relationship between the two variables — high PCB values tend to be associated with high clay content — the linear correlation coefficient is only 0.25, a value so low that we might mistakenly believe the two variables to be unrelated. The cause of this low correlation is a single aberrant sample that has a low clay content but a high PCB concentration. A quick check of the original data in Table 1 strongly suggests that the clay content for this sample is erroneous, and should have been recorded as 100 rather than 010. If we remove this questionable sample from the data set, and calculate the correlation coefficient on the remaining 49 samples, we find that the correlation coefficient rises to 0.43.

Extreme values do not always cause the correlation coefficient to deteriorate; they can also enhance the low correlation of a weak relationship. The linear correlation coefficient essentially measures how close the paired values come to plotting on a straight line. As shown in Figure 3, a single very extreme sample can cause the linear correlation coefficient to be high not because there is a strong relationship between the variables but rather because a straight line can be fit through the aberrant sample and the cloud formed by the rest of the data.

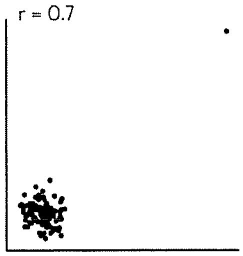


Figure 3 Example of aberrant sample enhancing an otherwise poor correlation.

Rank correlation

The rank correlation is an alternative to the traditional linear correlation that is not so sensitive to extreme or aberrant values; it is calculated by assigning ranks to the data and then calculating the traditional linear correlation on these ranks.

Table 2 Data from Table 1 along with their ranks.

Clay		PCB		Clay		PCB	
%/Rank		ug/g/Rank		%/Rank		ug/g/Rank	
80	28	0.9	4	90	38	105.6	47
90	32	9.4	26	80	27	0.8	3
100	49	59.7	43	90	36	54.5	41
80	25	3.7	16	100	45	53.3	40
90	40	11.8	29	100	50	32.8	36
90	30	16.2	31	90	33	12.4	30
80	24	0.6	1	100	46	59.5	42
80	21	5.4	19	100	47	105.5	46
80	18	1.3	6	90	29	28.0	35
80	17	3.2	12	14	2	1.1	5
80	22	1.5	8	41	7	6.2	21
90	41	86.8	44	52	9	9.3	25
80	23	1.5	9	49	8	11.2	27
80	20	3.4	14	54	10	3.2	13
80	19	1.3	7	58	13	19.1	32
90	31	26.3	33	21	3	5.6	20
90	39	8.8	24	93	42	184.0	50
90	37	7.6	22	39	6	3.5	15
010	1	121.8	48	67	14	49.5	39
90	35	4.5	18	54	11	8.5	23
90	34	35.4	37	55	12	11.2	28
100	48	95.6	45	31	4	3.0	11
100	44	27.2	34	94	43	124.2	49
80	16	2.1	10	74	15	44.9	38
80	26	0.8	2	32	5	3.8	17

Table 2 gives the ranks for the data shown earlier in Table 1. The ranks, which are values from 1 to the number of samples, identify where the original data values would appear on a sorted list. For example, the smallest PCB value in Table 1 is 0.6 ug/g; this PCB value gets a rank of 1. The largest PCB value is 184.0 ug/g; this PCB value gets a rank of 50. The ranking of the clay content measurements is a little bit tricky since there are many values that are identical; for example, there are seven samples whose clay content is reported as 100%. One common way of breaking these ties is simply to assign the ranks randomly within each group of tied values. In Table 2, the highest seven ranks, from 44 through 50, are assigned randomly to the highest seven clay content values.

To calculate the rank correlation coefficient for the clay – PCB data we use the equation shown earlier for the correlation coefficient, but rather than plugging in the actual data values, we use their ranks instead. The x_i's would be the ranks of the clay content, m_x would be the mean of these ranks and s_x would be their standard deviation; the y_i's would be the ranks of the PCB measurements, m_y would be the mean of these ranks and s_y would be their standard deviation.

As can be seen from the statistics reported along with the scatterplot in Figure 1, the rank correlation coefficient for the clay – PCB data is noticeably higher than the linear correlation. This is due to the fact that the rank correlation is not as sensitive to the aberrant (and probably erroneous) data value. Earlier, when we removed this single aberrant value, the linear correlation coefficient climbed from 0.25 to 0.43. The removal of this same dubious sample causes the rank correlation to change from 0.51 to 0.60; while the rank correlation is definitely affected by the aberrant sample, it is not as sensitive to this aberrant sample as is the traditional linear correlation.

The rank correlation coefficient will not always be higher than the linear correlation cofficient. With the example shown in Figure 3, where a single aberrant sample was enhancing an otherwise poor correlation, the rank correlation coefficient would be virtually 0, much lower value than the linear correlation coefficient of 0.7.

The main advantage of the rank correlation coefficient is that it provides a useful supplement to the traditional linear correlation coefficient. In the same way that the difference between the mean and the median can provide insight into the skewness of a distribution, the difference between the rank and linear correlation coefficients can provide insight into the nature of the relationship between two variables. If the rank correlation is lower than the linear correlation, then the relationship between the two variables might not be as good as the linear correlation suggests since aberrant samples could be enhancing an otherwise poor correlation. If the rank correlation is higher than the linear correlation, then the relationship between the two variables might not be as bad as the linear correlation suggests since aberrant samples could be ruining an otherwise good correlation. If the rank and linear correlation coefficients are about the same, as they would be for the three examples shown in Figure 2, then aberrant samples likely have little effect and either statistic provides an appropriate summary of the strength of the relationship.

GRAPHICAL TOOLS

By themselves, rank and linear correlation may not convey all of the important information about the relationship between two variables. Graphical displays provide valuable visual support to those who are trying to follow the details of a statistical study. A combination of graphical displays and numerical summaries is the most effective vehicle for conveying our understanding of the data to those who are not familiar with the project.

Scatterplots

The common graphical display for paired data is a scatterplot or x-y plot like the one shown in Figure 1. The values of one variable serve as the x coordinates for the plot and the values of the other variable serve as the y coordinates. In the example shown in Figure 1, each sample listed in Table 1 is shown as a dot, with the clay content serving as the x coordinate and the PCB concentration serving as the y coordinate.

In addition to their value as graphical summaries, scatterplots are often very useful for detecting errors in the data base. With the data in Table 1, for example, the sample with a clay content of 10% might not attract much attention during univariate analysis; even though it is the smallest clay content in the data base, there are some other low values in the 10–20% range, so this particular sample would not stand out on a histogram or cause any of our univariate statistical summaries to attract attention. When we plot the clay content against the PCB measurements on a scatterplot, however, this particular sample does provoke our curiosity because it does not follow the general trend of the rest of the sample data.

When a scatterplot reveals aberrant samples, these should not be discarded without a complete examination of the reasons for these aberrations. The guidance document entitled *OUTLIERS* provides advice on recognizing, interpreting and dealing with aberrant samples.

With skewed data that span several orders of magnitude, a conventional scatterplot may not be very revealing or informative since much of the data will be squashed along the axes. In Figure 1, for example, even though we can see that there is a tendency for high PCB values to be associated with high clay content, we don't really get a good look at what is happening with the half of the data for which the PCB value is below 10 ug/g. In such situations, logarithmic scaling of one or both of the axes may bring useful additional insight into the data.

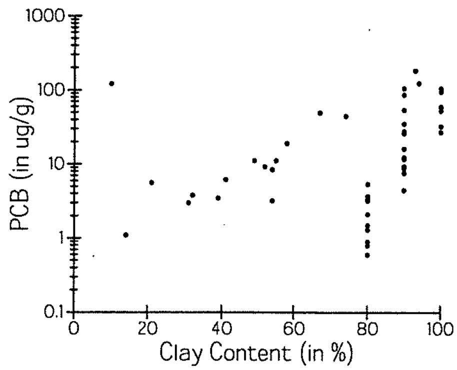


Figure 4 A scatterplot of the clay – PCB data in Table 1 with the y-axis logarithmically scaled.

Figure 4 shows a scatterplot of the clay – PCB data from Table 1 with the y-axis logarithmically scaled. The reason that we have not also used logarithmic scaling on the x-axis is that the clay content measurements are not skewed nor do they span more than one order of magnitude. Logarithmic scaling was used on the y-axis because the PCB values are positively skewed and go from less than 1 ug/g to more than 100 ug/g.

With the logarithmic scaling, the scatterplot better reveals one of the unusual characteristics of this particular data set. There appears to be two groups of data, both of which show a tendency for PCB content to increase with clay content. One of the groups spans a broad range of clay content, from about 15% to 95% while the other one spans a much narrower range, from 80% to 100%, and only includes exact multiples of 10. For this particular data set, the reason for this odd relationship is that the data were collected by two different groups; one group made visual estimates of clay content while the other one made direct measurements of clay content as part of their laboratory procedure. The group that made visual estimates used only exact multiples of 10 and tended to overestimate the clay content.

RECOMMENDED PRACTICE

With sites that have several different contaminants, the number of pairings of all the different variables may be very large; with 20 contaminants, for example, there are nearly 200 different pairings of the different variables. It is not necessary to explore the relationship and provide a complete bivariate description for all possible pairs. The analysis of the relationship between variables should focus on those relationships that are deemed to be important for the study. Using the earlier example of the clay – PCB data, the fact that the PCBs are concentrated in layers with a high clay content may lead to a remediation strategy that treats only the clay layers; in such a situation, documentation of the relationship between clay content and PCB concentration is critical. Other common situations in which the relationship between variables should be documented include the following:

- One contaminant is often selected as the principal contaminant from a group of several known contaminants with an assumption that the remediation of this principal contaminant will also entail the remediation of all the minor contaminants. With heavy metals contamination problems, for example, lead is often identified as the primary contaminant and becomes the focus of the study even though other metals may also occur in sufficient quantities to require remediation. In such situations, scatterplots of lead versus each of the other possible contaminants will provide good documentation of whether the remediation of lead will also address the concerns about minor contaminants.
- Some remediation strategies target only a portion of the soil on the contaminated site. Soil washing, for example, may be used to remediate the medium and fine grain sizes if the coarser material is thought to be uncontaminated. In such situations, a scatterplot of contaminant concentration versus grain size will provide good documentation of the appropriateness of such a remediation strategy.

In general, for any study in which information about one variable is being used as the basis for making assumptions about the behaviour of another variable, then bivariate exploratory data analysis should be performed and summarized in the study report.

It is not possible to give a rigid prescription for exploratory data analysis since a thorough understanding of the data requires both creativity and curiosity. The sequence of steps that worked on one project will not always work on another one. The following general guidelines, however, should improve the exploratory data analysis of the relationship between variables for any contaminated site study:

1. Before exploring the relationship between pairs of variables, the integrity of the data should be documented, either by reference to a report on procedures used to compile and verify the data or by a complete check of all data against original records; see the guidance document entitled *UNIVARIATE DESCRIPTION* for further advice on the issue of data base compilation and verification.
2. Complete listings of all data used in statistical studies should be included as appendices to reports; these do not, however, constitute an appropriate statistical summary. Statistical summaries of bivariate data should include:
   (a) Scatterplots that display the relationship between pairs of variables; if the data are skewed, then logarithmically scaled scatterplots should also be included.
   (b) Linear and rank correlation coefficients that summarize the strength of the relationship.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Understanding Robust and Exploratory Data Analysis, (Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds.), John Wiley & Sons, New York, 1983.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Moore, D.S., *Statistics: Concepts and Controversies*, W.H. Freeman and Company, New York, 1985.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-3

SPATIAL DESCRIPTION

A guide for report writers, reviewers, data analysts and interpreters on exploratory data analysis for spatial data

April 2001

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.

THE GENERAL IDEA

The application of statistics to contaminated site studies requires a clear and coherent understanding of the available data. For those directly involved in statistical analysis and interpretation, a clear and coherent understanding of the data will help them to select appropriate statistical tools and to make critical assumptions about statistical populations. For those who prepare statistical reports, it is important that their reports convey a clear and coherent understanding of the data; the readers of a report will not be able to form an opinion about the validity of the study's conclusions without a good understanding of the data on which it is based.

This guidance document discusses tools for exploratory data analysis, a statistical study's first step in which we investigate the data, form tentative opinions and modify these opinions as our understanding of the data improves and evolves. The same tools that help us explore and interpret the data are also ideal for presenting and summarizing our understanding of the data to those not directly involved in the study. This guidance document should therefore be of assistance not only to those who actually do the statistical analysis and interpretation, but also to those who are responsible for writing reports. This document is not intended to provide a rigid prescription for how to perform and present an exploratory data analysis; This document does intend, however, to encourage some much needed consistency in the performance and presentation of statistical studies by providing a simple and straightforward approach to exploratory data analysis.

This guidance document focuses on tools for analyzing data in their spatial context. Two other documents in this series focus on other aspects of exploratory data analysis. *UNIVARIATE DESCRIPTION* focuses on tools for analyzing a single variable; it also addresses the important first step of verifying the data base. *BIVARIATE DESCRIPTION* focuses on tools for analyzing the relationship between pairs of variables.

THE IMPORTANCE OF SPATIAL CONTEXT

One of the aspects of statistical studies of contaminated sites that distinguishes them from many other statistical studies is that the data have a spatial context. This spatial context helps us decide how to group the available data into statistical populations; it can also help us catch errors in calculation and interpretation. Another reason for analyzing the data spatially is that the key to successful remediation often lies in qualitative information, such as surficial geology or the history and usage of the site, that can be integrated with a statistical understanding only through visual displays such as maps and cross-sections.

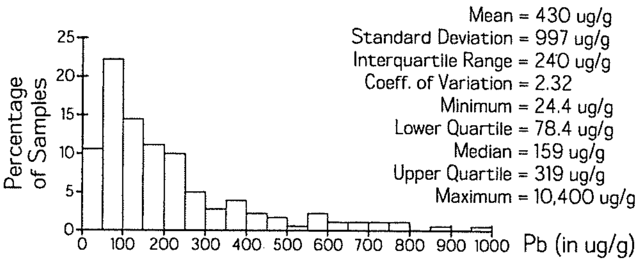


Figure 1 A histogram of 180 lead samples.

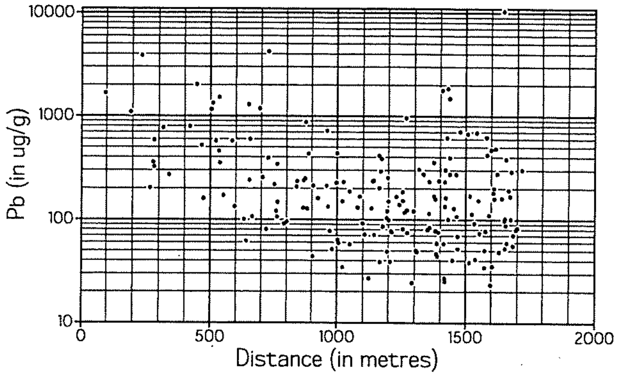


Figure 2 A scatterplot of lead versus distance from the smelter.

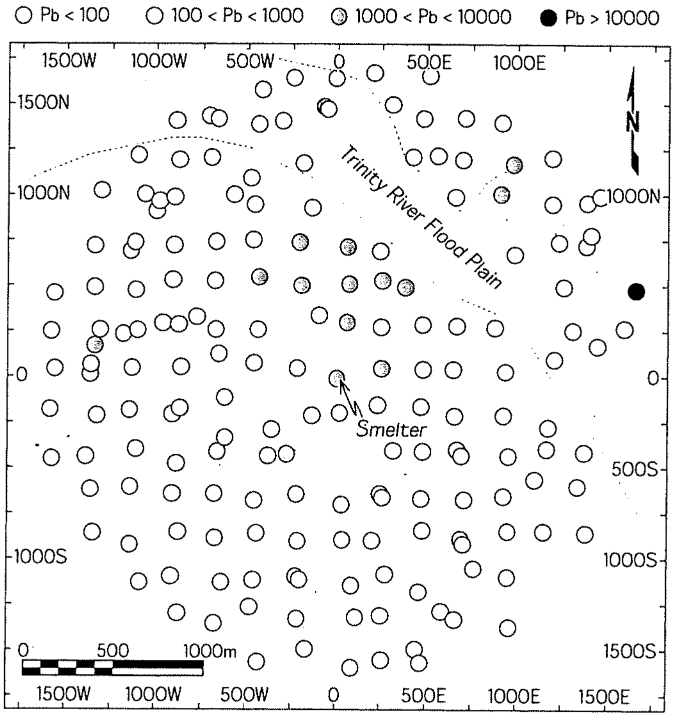


Figure 3 A greyscale map of the 180 lead samples.

Figures 1 and 2 summarize 180 lead samples from the soil near a smelter. The histogram and summary statistics in Figure 1 show that the available data span several orders of magnitude, from roughly 20 to 10,000 ug/g. As with most contaminated site studies, statistical analysis and interpretation of these lead data may need to recognize two separate populations: a "background" population, with concentrations around 100 ug/g, and a "contaminated" population with concentrations around 500 ug/g or greater. A scatterplot of lead concentration versus distance from the smelter (Figure 2) shows that lead concentrations tend to be higher close to the smelter.

The histogram and the scatterplot both help to document the understanding that the soil has been contaminated by lead from the smelter. Though these conventional statistical summaries certainly help to document the effect of the smelter, a simple map, such as the one shown in Figure 3, is usually much more direct and obvious. The map in Figure 3 has lost some of the detail in the data by coding the lead values according to their order of magnitude rather than presenting the exact value. By sacrificing this detail, however, the visual display is a more effective vehicle for communicating an understanding of the effect of the smelter.

Figure 3 is also a rich source of information on other aspects of the contamination. It shows us that the high lead values tend to be located north and northeast of the smelter; had we not already recognized the importance of wind direction, the north-northeasterly spread of the contamination plume might prompt us to find out more about local meteorological conditions. We might also need to learn more about the effect of the river's floodplain on lead in the soil, since Figure 3 shows a band of low values that cut across the plume in the northeast quadrant of the map area. The map also alerts us to short scale variability at several locations where the sample values change by an order of magnitude over short distances.

From the factors that control the broad scale features to those that create short scale variability, all of this information is important to a thorough study of a contaminated site. An understanding of the broad scale controls is critical for characterizing the site, for estimating the amount of soil that will need to be remediated and for identifying specific areas that require remediation. An understanding of the short scale variability is essential for planning a remediation strategy that can deal with the "hot spots" that commonly occur on contaminated sites.

Though conventional statistical analysis can assist with developing and documenting our understanding of the data, exploratory data analysis is much more effective when it incorporates displays of data in their spatial context. As with the example shown in Figure 3, maps and cross-sections often alert us to other information that will help us to predict contaminant concentrations and to plan an appropriate remediation strategy. Even though this additional information is often qualitative — "the predominant wind direction is from the southwest" or "the river tends to wash out lead" — and not in the form of hard quantitative data, it needs to be taken into account. Such qualitative information will often be helpful in making decisions about whether to divide the data into separate populations and can also assist with the identification and treatment of outliers.

DATA POSTINGS

One of the simplest and most common ways to display data in their spatial context is to post each data value beside its corresponding sample location. Figure 4 shows an example of a data posting for the 180 lead samples discussed earlier.

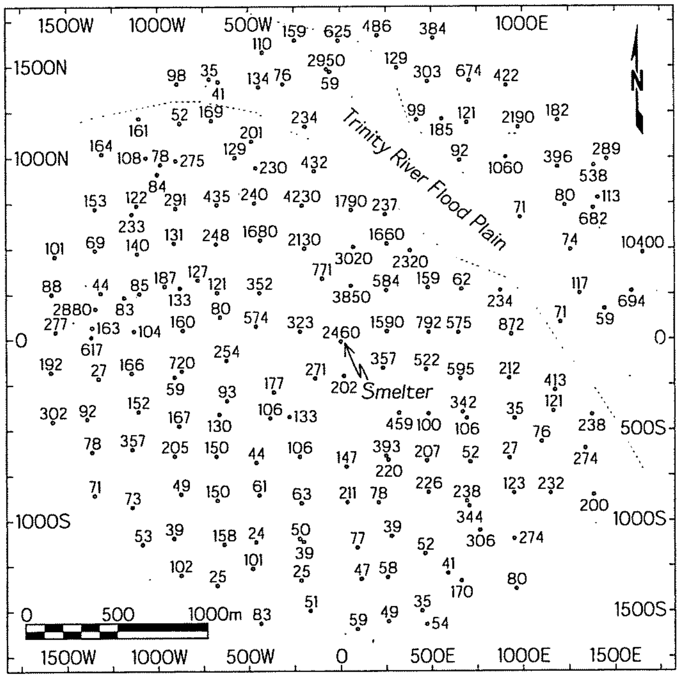


Figure 4 A data posting of the 180 lead samples.

One advantage of such a display is that it provides a detailed look at each and every sample location and therefore serves as a good basis for checking whether any samples are mislocated. When data are merged from different sources, errors can easily creep into the coordinate information. Coordinates can be inadvertently reversed if a data base that lists latitude before longitude is merged with one that lists east before north or x before y. Coordinates can also become confused when each organization that collected samples has used their own version of a local coordinate system. With these and other sources of location errors, a data posting is often the first warning of problems with the coordinate information.

Data postings are often necessary for recognizing and interpreting aberrant sample values or outliers. As discussed in the document entitled *OUTLIERS*, a sample value may be regarded as an outlier if it is inconsistent with all of the other nearby sample values. A data posting also serves as a good basis for detecting errors in numerical computations. If a contour map is used as the basis for calculating remediable volumes, for example, the interpreted contour lines should be checked for consistency with the original data. Between human error and software bugs, it is possible that computer-generated contour lines might not correctly honour the data. A data posting provides a straightforward way of checking whether software is producing sensible numerical interpretations.

The distinct disadvantage of data postings is that they usually present so much detail that they do not give a quick visual appreciation of where the contamination lies. By sacrificing some of the detail in the display, and colour coding the data values into several different categories, we can make the map more effective as a vehicle for communicating our understanding of the data. Though Figure 4 is more detailed than Figure 3, it is less effective as graphical support for the point of view that the contamination extends downwind from the smelter.

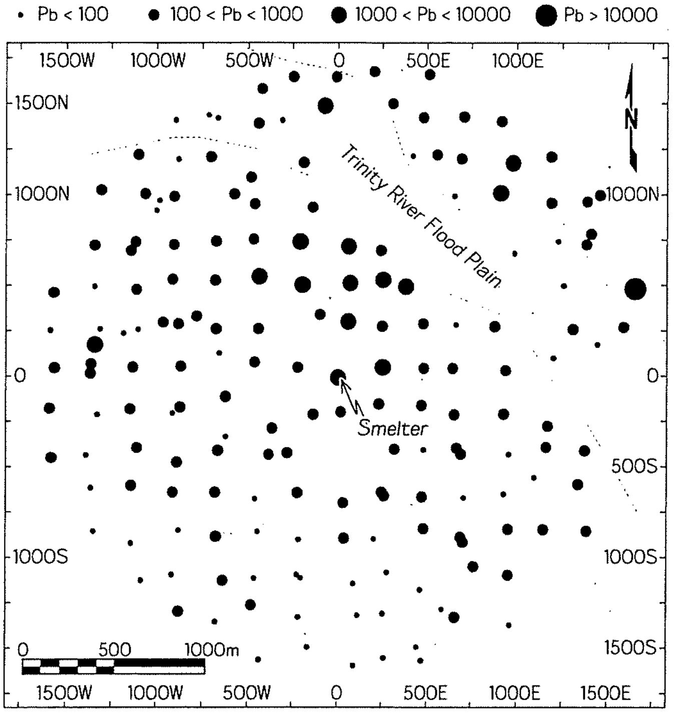


Figure 5 A map of the lead sample values on which the symbol size is proportional to magnitude of the lead concentration.

If the creation of colour or greyscale maps and cross-sections is difficult, due to computer hardware or software limitations, a similar visual effect can often be accomplished by using symbols of different sizes to code the data values into different categories. Figure 5 shows the 180 lead samples with the size of each sample dot scaled to the magnitude of the lead concentration. As with Figure 3, this style of presentation does not carry all of the detail of a data posting but presents a more immediate visual sense for where the contamination is highest.

CONTOUR MAPS

Perhaps the most traditional format for displaying earth science information is a contour map. On this type of display, locations of equal value are connected to form contour lines (or "isopleths"). For those who are familiar with this type of display, these contour lines communicate useful information about the spatial arrangement of the data values. Figure 6 shows a contour map of the lead data used in earlier examples.

One of the problems with contouring contaminated site data is that the skewness of the data makes it hard to choose a single appropriate contour interval. Attempts to use a common contour interval for the entire map usually result in some regions of the map being cluttered with too many contour lines and other regions being empty. With skewed data, it is often necessary to show two contour maps or, as in the example in Figure 6, to take some liberty with the conventional format by using two different contour intervals. In Figure 6, the contour interval for the thinner lines is 100 ug/g and 500 ug/g for the thicker lines.

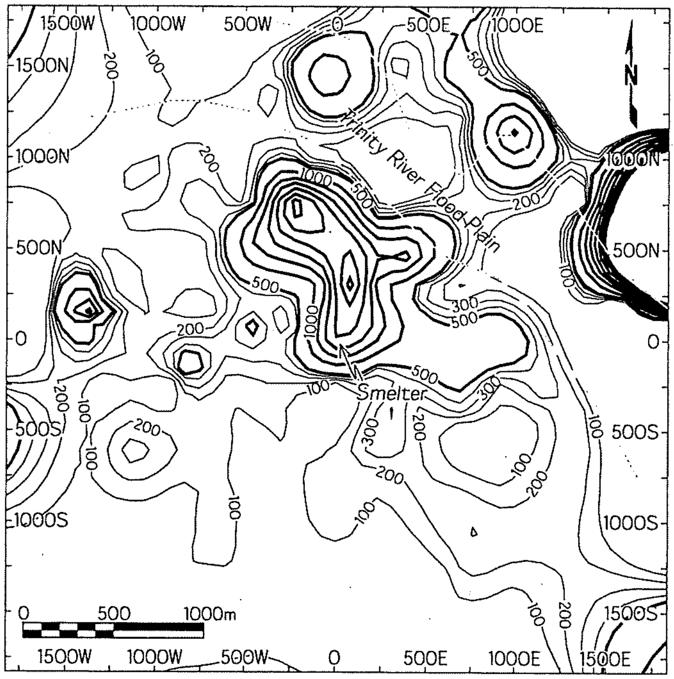


Figure 6 A contour map based on the lead data in Figure 4.

Though contour maps are a familiar display format for most earth scientists, they are not always ideal for exploratory data analysis since they do not present the raw data in their original form but present instead an interpretation that involves numerical processing of the original data. Contouring is not a unique exercise; whether it is done manually or on a computer, different people (or different programs) can produce different contour maps from the same set of original data.

Different contour maps of the same data reflect different approaches to various arbitrary choices that need to be made. One of the most critical of these is the choice of a method for interpolating between the available sample data; another is the choice of a method for tracing curved lines through a series of control points. In the most popular and commercially successful contouring software packages, a lot of emphasis is placed on aesthetics — a contour ma p that shows smooth lines and gentle undulations is preferred over one that has jagged contour lines and a lot of short scale variation. Though aesthetics do play an important role in the visual display of quantitative information, we should not turn a blind eye to other issues. There is an implicit tradeoff in making aesthetics our first priority: smooth and gentle undulations usually come at the price of ignoring short scale variation. The data posting in Figure 4 shows that the actual data fluctuate much more than the contour map in Figure 6 suggests. For example, due north of the smelter in the floodplain of the river is a pair of samples that are very close to one another; one has a lead concentration of 2,950 ug/g and the other has a lead concentration of only 59 ug/g. In this same area, the contour map in Figure 6 does not show this sudden short scale variation.

For contaminated sites where "hot spots" are a major concern, smooth contour maps can instill a complacent belief that the contamination is well behaved and easily mapped. When short scale variability is not properly recognized in remediation planning, the resulting remediation exercise often experiences large overruns as unanticipated "hot spots" trigger additional remediation that was not evident on the original contour maps.

Due to their tendency to smooth away short scale variations, contour maps should not be the sole graphical display of the spatial distribution of the available data. The impact of the smoothing that is fundamental to contour maps can be assessed if the contour map is accompanied by other displays that present the available data with little or no numerical processing, such as the data postings, greyscale and symbol maps in Figures 3 through 5.

For many audiences, particularly those who do not have a technical background, colour or greyscale postings of the data are much more comprehensible and effective than contour maps. As a vehicle for communicating our understanding of the spatial context of the data, a contour map is best suited to technical audiences who are already familiar with the conventions of contouring. Even when the intended audience is familiar with contour maps, this type of display should be used only for communicating broad features of the spatial distribution since the smoothing inherent in contouring causes large scale features to be emphasized at the expense of small scale ones.

LOCAL STATISTICS

The issue of statistical populations is a recurring theme in statistical studies of contaminated sites; though it is often convenient and tempting to lump all of the data into a single statistical population, it is usually more appropriate to split the data into two or more separate populations. A simple procedure that provides useful insight into the lumping-or-splitting decision is to calculate local statistics within sub-areas. If the available data have similar statistical characteristics in all the sub-areas, then it is appropriate to treat them as a single population. The more common situation is that the statistical characteristics of the available data are markedly different in some regions. In such situations, the data should either be separated into different populations or, if no clean separation is possible, the trends in the data should be analyzed and accommodated in subsequent statistical analysis.

As an example of the calculation and use of local statistics, Table 1 presents a few summary statistics for the lead data in each of the main quadrants of the map area in Figure 4. These local statistics show notable changes in the statistical characteristics across the map area. The lead values tend to be much higher in the northeast quadrant than in the southwest quadrant; in addition to being higher, the available data in the northeast quadrant also tend to be more erratic. These statistical observations should not be used as support for carving the site neatly into four quadrants; instead, they should be regarded as a first step in developing an appropriate treatment of the data. When integrated with our earlier remark on the wind direction, this preliminary set of local statistics could lead to a more detailed examination of directional trends in the lead concentrations. When integrated with the earlier remark on the rivers floodplain, these local statistics could lead to further examination of whether the samples from the floodplain should be treated as a separate population.

Table 1 Summary statistics for each quadrant.

Quadrant	N	Mean	s	CV	Median	IQR
Northeast	40	940	1740	1.85	384	675
Northwest	56	428	794	1.85	159	179
Southwest	40	131	124	0.95	92.2	108
Southeast	44	240	367	1.53	170	216

RECOMMENDED PRACTICE

In addition to the following guidelines, the documents entitled *UNIVARIATE DESCRIPTION* and *BIVARIATE DESCRIPTION* also contain guidance that is relevant to spatial description.

1. Reports on statistical studies of contaminated sites should contain graphical displays that present the available data in their spatial context.
2. Data should be posted on maps or cross-sections that show the location of each sample along with the corresponding sample value.
3. Data postings should be simplified and summarized through the use of colour, greyscale or symbol size to highlight the locations of the highest sample values.
4. Contour maps should be used to show the broad features of the spatial distribution.
5. Local statistics should be presented to assist the reader in understanding and evaluating decisions about statistical populations and trends.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Jones, T., Hamilton, D. and Johnson, C., *Contouring of Geological Surfaces with the Computer*, Van Nostrand Reinhold, New York, 1986.

Tufte, E.R., *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-4

DISTRIBUTION MODELS

A guide for reviewers, data analysts and interpreters on the statistical properties of common distribution models

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

Statistical applications for contaminated site studies commonly make use of theoretical distributions such as the normal and lognormal distribution. This document presents some of the common choices for distribution models and discusses their properties; it is intended to serve two audiences:

- reviewers who need an overview of the common distribution models and their characteristics, and
- data analysts and interpreters who need information on how to calculate percentiles and other summary statistics for some of the common distributions.

This guidance document discusses the normal, lognormal and exponential distributions, three of the more common and useful distribution models for statistical studies of data from contaminated sites. Before discussing the characteristics of these particular distributions, there are two things that should be made clear. First, it may not be necessary to choose a distribution model at all; for many of the statistical problems commonly encountered in contaminated site studies, a distribution model is not necessary. Second, in addition to the three distribution models discussed here, there are many others that might also be useful for particular problems at specific sites; Johnson and Kotz (1970) provide details on a wide variety of alternatives.

Other guidance documents in this series discuss these two closely related topics. The document entitled *NONPARAMETRIC METHODS* discusses whether or not a distribution model is necessary and presents some statistical methods that do not require any distribution assumption. If it is necessary to choose a distribution model, the document entitled *CHOOSING A DISTRIBUTION* provides advice on how to select an appropriate model and how to document the reasons for this choice. Readers are strongly encouraged to read these other two documents so that they have a more complete appreciation of the various issues surrounding the selection of a distribution model.

The distribution models discussed in this document have only one mode; histograms of actual data from contaminated sites sometimes show two or more modes. Such multimodal behaviour is usually due to a mixture of two or more populations. It is common to find with heavy metals, for example, that the data represent a mixture of two distributions, one that reflects the naturally occurring background concentrations and another that reflects the concentrations of material affected by industrial and other anthropogenic contamination. The document entitled *IDENTIFYING POPULATIONS* discusses the issue of separating data into different subpopulations.

THE NORMAL DISTRIBUTION

Overview

The most commonly used (and chronically misused) distribution model in statistics is the normal distribution. Data values from a normally distributed population have a histogram that looks like the one in Figure 1: fairly symmetric with the most common values representing the middle of the distribution and with extremely low or high values being equally uncommon.

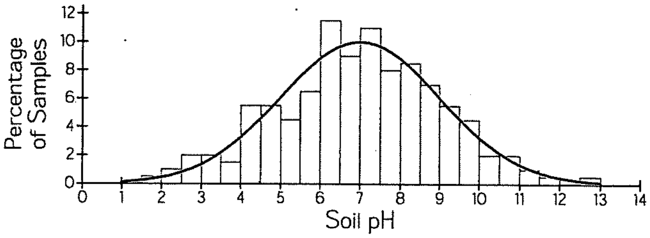


Figure 1 A histogram of normally distributed data.

The use of the normal distribution is often defended by invoking the Central Limit Theorem, which states that any distribution will tend to look more and more like a normal distribution as we average values together. While this is an interesting statistical fact, it has some important restrictions that limit its practical relevance for contaminated site studies. The most important of the assumptions that underlie the Central Limit Theorem is the assumption that the values being averaged together are independent. Violation of the independence assumption is, apart from systematic errors, the most critical violation of the common statistical assumptions. Soil or water contamination is not the result of the averaging of several independent events. Sample values from contaminated site studies rarely have the pleasing symmetry of the normal distribution; it is much more "normal" to see a lot of low values and a decreasing proportion of erratic high ones.

Despite the fact that the normal distribution usually does a poor job of modelling the distribution of contaminant concentrations, it does have a useful role to play in some specific applications. Certain variables, such as porosity in many groundwater studies or the pH of soil or water, have distributions that are fairly symmetric and that rarely have the kind of extreme values that are characteristic of contaminant concentrations.

Classifying stockpiled material based on the average concentration of the stockpile is the other common situation in which the normal distribution may be an appropriate model. The distribution of the average concentration of a contaminant over large homogeneous volumes of material will definitely be more symmetric than the distribution of the contaminant concentration from discrete samples. The average concentrations of several stockpiles may, in some situations, be viewed as averages of many independent values from a common population. They can be viewed as averages since the true average grade of any individual stockpile is the average of the vast numbers of discrete samples that make up that stockpile; they can be viewed as coming from a common population as long as all of the stockpiled material is drawn from a homogeneous area; and they can often be viewed as independent because the stockpiling process usually removes the spatial correlation that might have existed in the *in situ* material. For these reasons, we may be justified in assuming that average concentrations of stockpiled material will follow a normal distribution.

Calculation of percentiles

The smooth symmetric curve in Figure 1 shows the relative frequencies that the normal distribution model predicts. One of the attractions of the normal distribution is that this theoretical curve of relative proportions is fully determined by the mean and standard deviation of the distribution.

The "standard" normal distribution is one with a mean of 0 and a standard deviation of 1; many statistical textbooks contain tables of the percentiles of the standard normal distribution. The percentiles of any other normal distribution can be calculated by first calculating the corresponding percentile of the standard normal distribution, and then multiplying the result by the standard deviation and adding the mean. For example, suppose we need to calculate the 90th percentile of a normal distribution whose mean is 50 ug/g and whose standard deviation is 10 ug/g. From a table that gives the percentiles of the standard normal distribution, such as Table 26.1 in Abramowitz and Stegun (1970), we know that 1.28 is the 90th percentile of the standard normal distribution. The 90th percentile of our normal distribution is therefore

$$\begin{aligned} \text{90th percentile} &= 1.28 \times \text{Standard deviation} + \text{Mean} \\ &= 1.28 \times 10 + 50 = 62.8 \text{ ug/g} \end{aligned}$$

In addition to the tables provided in many books, there are also some approximations that can be implemented on a programmable calculator or a computer; Kennedy and Gentle (1980) provide a good discussion on the numerical approximations for the percentiles of the standard normal distribution.

68% and 95% confidence intervals

Many statistical procedures make use of the fact that a value drawn randomly from a normal distribution has a 68% chance of falling within one standard deviation of the mean, a 95% chance of falling within two standard deviations of the mean and a 99% chance of falling within three standard deviations from the mean (see Figure 2). Wherever we see a statistical statement involving 68% or 95% "confidence intervals", we can be fairly sure that an assumption of normality has been made. Not all statistical statements involving 68% and 95% depend on an assumption of normality, but the vast majority do.

An example of a remediation decision for which the normal distribution is commonly assumed is the classification of stockpiled material. We don't know the true average contaminant concentration of the stockpile, but we may choose to view this unknown average concentration as a value drawn randomly from a normal distribution. Having decided to model the unknown average concentration in this way, we now need to choose the parameters for this distribution. Using samples from the stockpile, and the methods outlined in the guidance document entitled *ESTIMATING A GLOBAL MEAN*, we may decide that our normal distribution has a mean of 80 ug/g and a standard deviation of 10 ug/g. Having selected the parameters for our normal distribution, we are now able to make predictions about the chance that the unknown average concentration will exceed various thresholds. With a mean of 80 ug/g and a standard deviation of 10 ug/g, there is a 95% chance that the unknown average concentration will be between 60 ug/g and 100 ug/g (two standard deviations from the mean). If we are concerned only with the chance that the unknown average will exceed a regulatory limit of 100 ug/g, then the symmetry of the normal distribution entails that there is only a 2.5% chance that the unknown average concentration will exceed 100 ug/g.

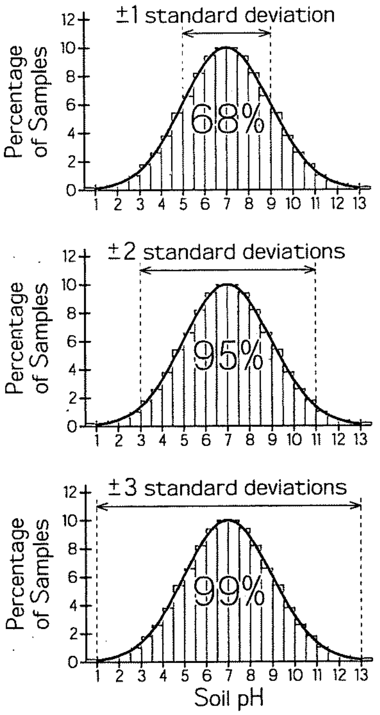


Figure 2 Probability of a value drawn at random from a normal distribution falling within one, two and three standard deviations of the mean.

THE LOGNORMAL DISTRIBUTION

Overview

Distributions of contaminant concentrations rarely have the kind of symmetry that makes the normal distribution an appropriate model; it is common to find that data from contaminated site studies contain many low values and a decreasing proportion of high values. A distribution model that captures this kind of asymmetry is the lognormal distribution. Figure 3 shows an example of the histogram of data drawn from a lognormal distribution; it has a lot of low values and a steadily decreasing proportion of high ones.

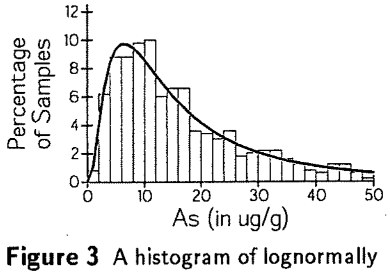


Figure 3 A histogram of lognormally distributed data.

The lognormal distribution, as the name implies, is one for which the logarithms of the data values are normally distributed. It is used for many earth science problems in which the data values span several orders of magnitude and have an asymmetric distribution. In the same way that the use of the normal distribution is often defended by arguing that the data are the result of a large number of independent additive events, the lognormal distribution can be defended by arguing that the data are the result of a large number of independent multiplicative events. There are a few papers in the technical literature that use arguments based on reaction rates and chemical reactions to support the point of view that the genesis of certain kinds of contamination does, indeed, involve a series of independent multiplicative events. Despite such observations, it is fair to say that the success of the lognormal model is not really due to independent multiplicative events, but due to the fact that by offering us an asymmetric distribution, the lognormal model corrects the major practical deficiency of the normal distribution.

Calculation of percentiles

The lognormal distribution shares with the normal distribution the fact that it is completely determined by the mean and standard deviation; it does not, unfortunately, share its computational convenience. The calculation of a percentile is typically accomplished by first calculating the percentile in terms of the logarithm and then exponentiating the result. In order to calculate the percentile in terms of the logarithm, we first need to know the mean and standard deviation of the logarithms. The following equations describe how the mean, m, and the standard deviation, s, of lognormally distributed values are related to the mean, α, and the standard deviation, β, of their logarithms:

$$m = \exp\left(\alpha + \frac{\beta^2}{2}\right) \qquad s = m\sqrt{\exp\left(\beta^2\right) - 1}$$

$$\alpha = \log(m) - \frac{\beta^2}{2} \qquad \beta = \sqrt{\log\left[1 + \left(\frac{s}{m}\right)^2\right]}$$

where all of the logarithms are natural (base e) logarithms.

As an example of how to calculate a percentile for a lognormal distribution, we can take the lognormally distributed data shown in Figure 3 and find their 90th percentile. The mean of the arsenic values shown in Figure 3 is 18.9 ug/g and their standard deviation is 20.1 ug/g. Using the equation given above for β, we can calculate that the standard deviation of their logarithms should be 0.87; and, using the equation for α, their mean should be 2.62. As discussed earlier in the section on calculating percentiles for a normal distribution, tables from reference books tell us that the 90th percentile of a normal distribution is 1.28 standard deviations above the mean. So, in terms of the logarithms, the 90th percentile would be

$$\begin{aligned}\text{90th percentile of logs} &= 1.28 \times \text{Standard deviation} + \text{Mean} \\ &= 1.28 \times 0.87 + 2.62 = 3.73\end{aligned}$$

This result needs to be exponentiated to get the 90th percentile of our original arsenic values:

$$\text{90th percentile of original values} = \exp(3.73) = 41.8 \text{ ug/g}$$

In addition to the exact calculations that can be done with the equations given above, there are some rules of thumb that may be useful to get a quick idea of where various high percentiles of a lognormal distribution lie. Most of these back-of-the-envelope calculations make use of the coefficient of variation (CV), which is the ratio of the standard deviation to the mean, and express the percentile as a multiple of the median (not the mean). For a lognormal distribution with a CV of 1 (the standard deviation is equal to the mean), the 90th percentile is roughly three times the median, the 95th percentile is nearly four times the median and the 99th percentile is nearly seven times the median. If the CV climbs to 2 (the standard deviation is twice the mean), then the 90th percentile is five times the median, the 95th percentile is eight times the median and the 99th percentile is almost twenty times the median.

The arsenic data shown in Figure 3 have a median of 13.7 ug/g, and their coefficient of variation is close to 1. We can use the rules of thumb given above to conclude that the 90th percentile will be fairly close to three times the median, or roughly 41 ug/g — a quick, but still very good, approximation to the exact value of 41.8 ug/g that we calculated earlier.

THE EXPONENTIAL DISTRIBUTION

Overview

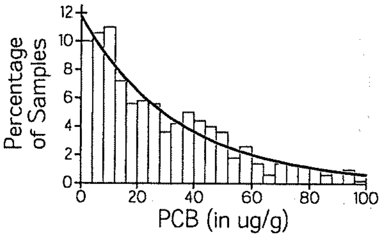


Figure 4 A histogram of exponentially distributed data.

The lognormal distribution is not the only distribution that allows us to capture the fact that low values are more common than high ones. One of the other common distributions that has the same kind of asymmetry as the lognormal distribution is the exponential distribution. Figure 4 shows an example of the histogram of data drawn from an exponential distribution. Like the lognormal distribution, it has a lot of low values and a steadily decreasing proportion of high ones. It differs from the lognormal distribution in the behaviour of the very lowest values. For the exponential distribution, lower values are always more common than higher ones; for the lognormal distribution, the very lowest values are actually not quite as common as some of the slightly higher values. In Figure 4, the tallest bar on the histogram is the first one; on Figure 3, however, the first bar is not the tallest one. The curve drawn with the heavier line in Figure 4 shows the relative frequencies that the exponential distribution model predicts.

While normal and lognormal distributions need two parameters, the mean and the standard deviation, the exponential distribution is completely determined by its mean. The standard deviation of an exponential distribution happens to be equal to the mean, so this distribution may be appropriate for data whose coefficient of variation is close to 1. The calculation of percentiles for an exponential distribution is more straightforward than for the lognormal distribution since the mean is the only parameter involved. Percentiles for an exponential distribution can be calculated using the following equation:

$$\text{p-th percentile} = -m \times \log\left[1 - \frac{p}{100}\right]$$

where m is the mean and the logarithm is the natural (base e) logarithm. Using the PCB data shown in Figure 4 as an example, their mean is 34.2 ug/g, so their 90th percentile is calculated as follows:

$$\begin{aligned} \text{90th percentile} &= -34.2 \times \log \left[1 - \frac{90}{100}\right] \\ &= 78.7 \text{ ug/g} \end{aligned}$$

In addition to the exact calculation given above, there are some rules of thumb that can be used to get a quick approximation of some of the high percentiles. For an exponential distribution, the 95th percentile is three times the mean and the 99th percentile is almost five times the mean. The 90th percentile is not very close to being a simple multiple of the mean; it happens to be 2.3 times the mean.

ASYMMETRIC CONFIDENCE INTERVALS

When discussing the normal distribution, we pointed out that 68% of the values fall within one standard deviation of the mean and 95% fall within two standard deviations. This property of the normal distribution is often used as the basis for making statements about the "confidence intervals" for an estimate. The typical assumption is that the quantity we are trying to estimate can be modelled by a normal distribution, that our estimate represents the mean of this distribution, and that we have somehow been able to express the uncertainty in our estimate as a standard deviation. The guidance document entitled *ESTIMATING A GLOBAL MEAN* gives an example of this type of procedure where, under an assumption of independence, the mean of a statistical population can be estimated by m, the mean of the available samples, and the standard deviation of this estimate is $\sigma_m = s \div \sqrt{N}$ where s is the standard deviation of the individual samples and N is the number of available samples. Though this approach is valid for quantifying the uncertainty on the mean of any distribution of values, whether normal or not, an assumption of normality is made as soon as we use this information to report $m \pm \sigma_m$ as our "68% confidence interval" or $m \pm 2\sigma_m$ as our "95% confidence interval".

Table 1 Lead concentrations (in ug/g).

12	191	872	13	52	92	43	17	5	59

With data that are clearly skewed (as are most data from contaminated site studies), the uncertainty about the mean is not likely to follow a normal distribution, especially if there are only a few samples available for estimation. Table 1 shows an example of 10 samples of lead concentrations in the soil from a contaminated site. Using these ten data, and assuming that they are independent, we can estimate that the mean of the population from which they came is 135.6 ug/g and that the standard deviation of this estimate is 83.7 ug/g. Up to this point, we have made no assumption about the underlying distribution, we have simply applied the equation given above. Given the very evident skewness of these data, it makes little sense to assume that the uncertainty on our estimate is going to follow a normal distribution. The normal 95% confidence interval, for example, would be 135.6±167.4 ug/g; a dose of common sense tells us that there's not a lot of meaning in a confidence interval that goes down to -31.8 ug/g on the low side.

When the quantity we are trying to estimate is better modelled by a skewed distribution, it is more useful to calculate confidence intervals directly from the percentiles than to use the classical $\pm\sigma$ and $\pm 2\sigma$ intervals. Regardless of the distribution, there is a 95% chance that a value will fall between the 2.5th percentile and the 97.5th percentile, so we can use these percentiles directly to report a 95% confidence interval. This approach works for any distribution, even a normal one, and is the only sensible way to report confidence intervals where the distribution is not normal. To continue with the example of the data in Table 1, it is more appropriate to assume that their unknown true mean follows a lognormal distribution with the mean and standard deviation reported above. Using the method outlined earlier, in which we calculated the percentile in terms of the logarithm and then exponentiated the result, the 2.5th percentile of our unknown mean is 37.9 ug/g and the 97.5th percentile is 351.4 ug/g. Using this information, we can report an asymmetric 95% confidence interval of 37.9 – 351.4 ug/g for our estimate of the mean.

RECOMMENDED PRACTICE

1. If a distribution model is necessary for some calculation, and if a normal, lognormal or exponential model has been chosen, the equations given in this guidance document can be used to calculate percentiles. It is recommended that the exact equations be used wherever possible and that the rules of thumb be used only for rough calculations. In all cases where a percentile or confidence interval is calculated from some distribution model, the type of model should be reported along with its parameters.
2. If a mean and standard deviation are being used to calculate confidence intervals, the classical $\pm\sigma$ 68% confidence interval and $\pm 2\sigma$ 95% confidence interval should not be used unless there is good reason to believe that the quantity being estimated follows a normal distribution. If a skewed distribution is more appropriate, the 95% confidence interval should be reported as the range from the 2.5th percentile to the 97.5th percentile.

REFERENCES AND FURTHER READING

This guidance document does not provide specific guidance on when to choose a distribution model or how to choose an appropriate distribution model. These issues are addressed in the guidance documents entitled *NONPARAMETRIC METHODS* and *CHOOSING A DISTRIBUTION*. In addition to the other guidance documents in this series, the following references provide useful supplementary material.

Abramowitz, M. and Stegun, I.A., (eds.), *Handbook of Mathematical Functions*, Dover, New York, 1970.

Blake, I.F., *An Introduction to Applied Probability*, John Wiley & Sons, New York, 1979.

Johnson, N.L. and Kotz, S., *Distributions in Statistics — Continuous Univariate Distributions, Volume 1*, Houghton Mifflin, Boston, 1970.

Kennedy, W.J. and Gentle, J.E., *Statistical Computing*, Marcel Dekker, New York, 1980.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-5

NONPARAMETRIC METHODS

A guide for data analysts and interpreters on statistical methods that do not require a distribution model

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

The validity of statistical statements can easily be challenged by questioning any distribution assumption. For example, we might be tempted to take data from a contaminated site, calculate that their mean is 50 ug/g and their standard deviation is 10 ug/g, and then use this information to predict that there is a less than a 1% chance that samples from the same population will exceed a threshold of 80 ug/g. This statement is defensible only if we can also defend the implicit assumption that the data values follow the classical bell-shaped normal distribution. The type of contaminant concentration data that we typically collect from contaminated sites very rarely follow a normal distribution, however, and any predictions that follow from this initial assumption are difficult to defend.

Though we try to make sure that our assumptions about underlying distributions are appropriate — choosing skewed distributions, for example, to model contaminant concentrations — we always run the risk that regardless of the distribution we choose, someone is going to challenge our predictions based on the fact that we *assumed* a particular distribution that we can never prove is correct. Fortunately, for many of the statistical problems that arise in contaminated site studies, there are methods that allow us to solve the problem without making any assumption about the underlying distribution. The predictions that we get from such *nonparametric* procedures will be defensible regardless of the assumption that anyone wants to make about the underlying distribution.

This guidance document presents some of the more common and practically useful nonparametric methods. In addition to demonstrating how they can be used in practice, this document also discusses the advantages and disadvantages of these nonparametric methods. There are two other documents in this series, *DISTRIBUTION MODELS* and *CHOOSING A DISTRIBUTION*, that discuss related issues.

ADVANTAGES OF NONPARAMETRIC METHODS

Inappropriateness of the normal distribution

The main advantage of nonparametric methods is that they do not require us to assume that data are normally distributed. Even though an assumption of normality underlies the vast majority of statistical procedures that are in common use, it is a very questionable assumption in contaminated site studies. Figure 1 shows a typical example of a histogram of sample values from a contaminated site along with some of the common summary statistics. These data have a mean that is much larger than their median; they show a large proportion of low values and a decreasing proportion of high ones. A normal distribution would show none of these characteristics; its mean and median would be very similar and its histogram would look symmetric, with similar proportions of low and high values.

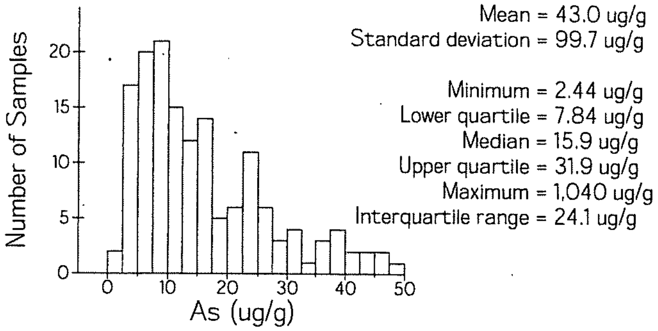


Figure 1 A histogram for measurements of the arsenic concentration in the soil from a contaminated landfill site.

Typical of the kinds of statistical predictions that depend on a prior assumption of normality is the use of the mean and standard deviation to build confidence intervals. Wherever w e see $m\pm\sigma$ being used as a 68% confidence interval, or $m\pm2\sigma$ as a 95% confidence interval, we are seeing a result that depends on an assumption of normality. If the unknown values that we are trying to predict do follow a normal distribution, then 68% of the values will fall within one standard deviation of the mean and 95% of them will fall within two standard deviations. If, however, the values do not follow a normal distribution (and this is more commonly the case in practice), then the traditional confidence intervals are meaningless.

In a nonparametric approach we make no assumption about the underlying distribution. This makes our predictions more robust in the sense that they do not depend on whether or not the underlying distribution is normal.

No need for any distribution model

Nonparametric methods are particularly useful in the early stages of a contaminated site study, where there are typically very few data yet available and, even if we intend ultimately to use a parametric technique that assumes some distribution model, we do not yet have enough data to allow us to choose an appropriate distribution model. Table 1 shows an example of a few measurements of the PCB concentration in the first ten samples collected from a contaminated site. Suppose that at this very early stage in the study we wanted to make some statement about whether the median for the entire population could be 10 ug/g. Any parametric technique would require us to first make an assumption about the underlying distribution from which these ten values come. With so few data at our disposal, it is very difficult to decide what kind of distribution might be an appropriate model for the PCB values. As discussed in greater detail later in this guidance document, this question about the median can be answered with a nonparametric technique that does not require us to assume anything about the underlying distribution.

Table 1 PCB values (in ug/g).

<1	51.2	17.9	34.6	<1	22.4	11.5	48.2	7.8	31.4

Ease of calculation and interpretation

The common nonparametric methods are very simple to apply. They usually work with the ranks of the data or with simple counts of values above and below the median and are therefore easy to calculate manually or to implement on a computer.

Another advantage of nonparametric statistics is that they are often easier for non-statisticians to understand and interpret. As discussed later, some of the graphical displays that are based on nonparametric statistics, such as the percentiles of the distribution, are more straightforward than other more traditional displays and still convey as much useful information.

Ability to work with no-detects

One of the other advantages of many nonparametric techniques is that they can accommodate values below detection limit without assigning such samples some arbitrary value (such as half the detection limit). As we will see later, we can make statistical tests with data such as those shown in Table 1 even though we do not know the exact value of every sample.

DISADVANTAGES OF NONPARAMETRIC METHODS

Not as efficient

The principal limitation of nonparametric methods is that they are not as efficient or powerful as parametric methods that are based on a known underlying distribution. For example, if we are trying to use statistics to document that two groups of data should be treated as separate populations, and if we already know that it is reasonable to assume that the data values in both groups are normally distributed, then a parametric test, such as the t-test, will be able to discriminate more effectively between the means of the two groups than would the corresponding nonparametric test described below.

Unable to extrapolate beyond data

The assumption of a specific distribution model is very powerful and buys us a lot of predictive power. Once we claim to know the distribution that the data represent, and we have chosen the parameters of our assumed distribution (such as the mean and the standard deviation for the normal distribution), we are then able to leverage our assumption and predict the behaviour of the entire distribution. For example, a parametric approach gives us the ability to predict the 99th percentile even if we haven't actually got a sample value that high yet. With its fundamental philosophy of avoiding unnecessary distribution models and letting the data speak for themselves, a nonparametric approach has no additional information to leverage beyond the data themselves; if we have only ten sample values, it will not be possible to predict the 99th percentile with a nonparametric approach.

Need more data

By letting data speak for themselves rather than letting a distribution model do the speaking for them, nonparametric methods provide statistical predictions that are not compromised by unnecessary distribution assumptions. The price for this strict adherence to data, however, is that nonparametric methods cannot make strong statistical statements with few data.

NONPARAMETRIC DATA ANALYSIS

Percentile-based statistics

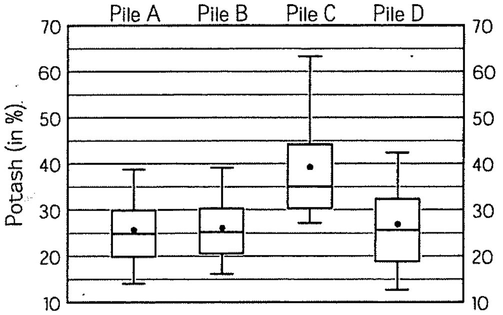
The two statistics that are most commonly used to describe a distribution are the mean and standard deviation. The first of these is a measure of the center of the distribution, the second is a measure of the spread of the distribution. The popularity of these two particular statistics is due, in large part, to the fact that they are the common parameters for the normal distribution. Though they are commonly used, these two statistics are often of little value for exploratory data analysis since they are both strongly influenced by extreme values. With the arsenic data shown in Figure 1, for example, it is questionable whether the mean of 43.0 ug/g is really describing the center of the distribution, or whether the standard deviation of 99.7 ug/g is telling us anything useful about the spread of the values. In this particular example, as in many other actual data sets from contaminated site studies, a few extremely high values have a profound influence on these two statistics.

Nonparametric methods rely on "rank" or "order" statistics that are simply the percentiles of a distribution. Rather than use the mean to describe the center of the distribution, nonparametric approaches more commonly use the median or 50th percentile. The difference between the upper quartile (75th percentile) and the lower quartile (25th percentile) is called the "interquartile range" and is the nonparametric alternative to the standard deviation for describing the spread.

For most people, the median corresponds more closely to their visual sense of where the center of the histogram lies than does the mean. Similarly, their visual sense for the spread of the distribution is closer to the interquartile range than to the standard deviation. Most of us, statisticians and non-statisticians alike, have a stronger intuitive feel for what the interquartile range is measuring — the spread of the middle half of the data — than we have for whatever it is that the standard deviation is measuring — the square root of the average squared deviations from the mean?! For the purposes of communicating statistical information to a non-technical audience, nonparametric statistics are therefore an excellent supplement to the more conventional mean and standard deviation.

Boxplots

A boxplot provides a concise graphical format for displaying the key nonparametric statistics. Figure 2 shows an example of a set of boxplots for the K_2O (potash) concentrations from discrete samples taken from four different stockpiles of cement kiln dust. The box in the middle of a boxplot extends from the lower quartile to the upper quartile; the bar in the middle of the box shows where the median lies. There are a couple of different conventions for how to draw the arms that stick out of the box; the one used in Figure 2 shows them extending all the way to the minimum on the low side and to the maximum on the high side. The other common convention is to draw the arms only part way to the extremes and to plot a star at each of the very extreme values. Boxplots commonly also pay homage to the fact that the mean is by far the most common summary statistic of all and, even though it is not a percentile-based statistic, it is usually shown with some special symbol — a black dot in the examples shown in Figure 2.



Number of data	39	31	63	85	Number of data
Mean	25.7	26.1	39.3	26.9	Mean
Maximum	38.8	39.1	63.3	42.4	Maximum
Upper quartile	29.9	30.3	44.2	32.3	Upper quartile
Median	24.8	25.2	35.0	25.6	Median
Lower quartile	19.8	20.5	30.3	18.8	Lower quartile
Minimum	14.0	16.1	27.1	12.6	Minimum

Figure 2 Side-by-side boxplots.

A boxplot presents most of the relevant univariate information that we need from an exploratory data analysis. It gives us a sense for where the middle of the distribution lies, how spread out it is and whether or not it is symmetric. The boxplot therefore offers most of the useful information that a histogram contains, but in a more compact form that is more amenable to side-by-side comparisons between different groups of data.

NONPARAMETRIC TESTS

Chebyshev's inequality for confidence intervals

Earlier, we pointed out that the use of $m \pm \sigma$ for calculating 68% confidence intervals is fine for the normal distribution but does not work for other distributions. There is a century-old non-parametric result known as "Chebyshev's inequality" that allows us to build confidence intervals using the mean and standard deviation even if we don't know the underlying distribution. Chebyshev's inequality says that for any constant k, the proportion of data that are within k standard deviations from the mean cannot be less than $1 - (1 \div k)^2$. If we take k=2, for example, this inequality tells us that at least 75% of the distribution must be within two standard deviations of the mean; for k=10, at least 99% of the distribution must be within ten standard deviations of the mean.

Compared to confidence intervals predicted from any distribution model, those predicted using Chebyshev's inequality are broader. For example, the opening example on the first page of this document involved a distribution with a mean of 50 ug/g and a standard deviation of 10 ug/g; with these statistical parameters, an assumption of normality leads to the conclusion that less than 1% of the data should exceed 80 ug/g. For this same threshold, which happens to be three standard deviations above the mean, Chebyshev's inequality states that any possible distribution must have at least 89% of the data within three standard deviations of the mean; no more than 11% could possibly be more than three standard deviations from the mean. This gives us a pessimistic upper bound on how much of the distribution *might* exceed 80 ug/g if our assumption of normality is inappropriate: for any distribution whatsoever, it is not possible to get more than 11% of the values to be greater than three standard deviations above the mean.

The sign test for the median

Earlier in Table 1 we showed ten PCB values and asked if the median could be as low as 10 ug/g. The "sign test" is a non-parametric procedure in which all data values above the proposed median are given + signs and all others are given − signs. We can test whether the median could be as low as some specified threshold, T, by noting that if T is, indeed, the median, then regardless of the shape of the distribution, each data value has the same probability of getting a + sign as a − sign:

$$p_+ = p_- = \frac{1}{2}$$

In a sample of size N, the number of observations with a + sign, N_+, will follow a binomial distribution. The probability of getting more than n + signs is:

$$\mathrm{Prob}[N_+ \geq n] = \left[\frac{1}{2}\right]^N \times \sum_{i=n}^{N} \frac{N!}{(N-i)! \times i!}$$

These binomial probabilities are tabulated in most reference and textbooks on probability and statistics. For large values of N, most introductory probability books, such as Blake (1979), discuss good approximations to these binomial probabilities.

Using the data from Table 1 and a proposed median of 10 ppm, seven of the values would get + signs. The no-detect samples do not create any difficulty; even though we do not know exactly the PCB concentration of these samples, we can still assign them − signs since they are definitely below 10 ppm. The probability of getting seven or more + signs out of a total of ten tries is:

$$\left[\frac{1}{2}\right]^{10} \times \left[\frac{10!}{3! \times 7!} + \frac{10!}{2! \times 8!} + \frac{10!}{1! \times 9!} + \frac{10!}{0! \times 10!}\right] = 0.172$$

Regardless of the underlying distribution, the chance that its median is 10 ug/g or lower given the ten observed values shown in Table 1 is about 17%.

The sign test can be adapted to test for any percentile by changing the equation given above to accommodate the fact that p_+ and p_- are no longer the same:

$$\mathrm{Prob}[N_+ \geq n] = \sum_{i=n}^{N} \frac{N!}{(N-i)! \times i!} \times {p_+}^{i} \times {p_-}^{(N-i)}$$

The Wilcoxon rank-sum test

Nonparametric methods for testing the difference between two groups of data usually deal with the ranks of the data. In a group of N data, the ranks are simply numbers from 1 to N that order the data from smallest to largest: the smallest data value has a rank of 1, the second smallest has a rank of 2 and so on up to the largest data value, which has a rank of N.

With two groups of data, the first containing N_1 samples and the second containing N_2 samples, the Wilcoxon rank-sum statistic, W, is created as follows:

1. Combine both groups of data, creating a large group with N samples.
2. Assign ranks to the data.
3. Let W be the sum of the ranks of all the data that came from the first group.

To test whether the two groups are significantly different, the Wilcoxon rank-sum test compares W against tabulated values of critical values. These tables are given for various values of N_1 and N_2. They show the range of values that W can have if the two groups of data actually come from the same population. If the observed value of W falls outside the range given in such tables, we accept this as evidence that the differences between the data values in the two groups are too large to be explained by chance alone; a more plausible explanation than mere chance is that the data values in each group were drawn from different populations.

If the values of N_1 and N_2 are larger than those that appear in reference tables, there is an another way to check if W is too extreme. We calculate the following test statistic

$$z = \frac{W - \frac{N_1 \cdot (N_1+N_2+1)}{2}}{\sqrt{\frac{N_1 \cdot N_2 \cdot (N_1+N_2+1)}{12}}}$$

and check to see if $|z|$ is greater than 3. If it is, then the chance that the differences between the two groups are due to chance alone is less than 1%, so values of z outside the range -3 to +3 are accepted as evidence that there are significant statistical differences between the two groups.

As an example of the application of the Wilcoxon rank-sum test, consider the problem of checking whether the following four PCB values might belong in the same group as the ten shown earlier in Table 1: <1, 5.2, 9.2 and 1.9 ug/g. These four values seem to be low compared to those seen earlier, but could this just be chance?

When the four new values are combined with the other ten to make a group of 14 samples, the three lowest values are all no-detects. Since we can't sort out the order of these three and don't know which should get the rank of 1, which should get the rank of 2 and which should get the rank of 3, we assign the average rank of 2 to each of these three tied values. The four new values therefore get ranks of 2, 4, 5 and 7; the sum of these ranks is 18. Tabulated values of the Wilcoxon rank-sum statistic (Finkelstein and Levin, p. 563 – 564) show that with a group of 4 samples being compared to a group of 10 samples, there is a 90% chance that W will be between 16 and 44. So although the new values tend to be on the low side, we cannot reject the possibility that they could actually be from the same population as the original ten values shown earlier.

RECOMMENDED PRACTICE

1. When presenting a statistical summary of data collected from a contaminated site, use percentile-based statistics, such as the quartiles and the median to supplement the more traditional mean and standard deviation.
2. Use boxplots as an alternative to histograms for graphical display purposes, especially when documenting a comparison between two or more groups of data.
3. Wherever a statistical prediction calls for a prior assumption about the underlying distribution, use a nonparametric alternative as a way of checking the sensitivity of the conclusion to the distribution assumption.

REFERENCES AND FURTHER READING

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

Blake, I.F., *An Introduction to Applied Probability*, John Wiley & Sons, 1979.

Conover, W., *Nonparametric Statistics*, John Wiley & Sons, 1980.

Finkelstein, M.O., and Levin, B., *Statistics for Lawyers*, Springer-Verlag, 1990.

Gibbons, J.D., *Nonparametric Methods for Quantitative Analysis*, 2nd ed., American Sciences Press, 1985.

Gibbons, R.D., "General statistical procedure for ground water detection monitoring at waste disposal facilities," *Ground Water*, v. 28, p. 235 – 248, 1990.

Kendall, M.G., *Rank Correlation Methods*, 4th ed., Griffin, 1970.

Lehmann, E.L., *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, 1975.

Millard, S.P. and Deverel, S.J., "Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits," *Water Resources Research*, v. 24, n. 12, p. 2087 – 2098, 1988.

Mosteller, F., and Rourke, R.E.K., *Sturdy Statistics: Nonparametric and Order Statistics*, Addison-Wesley, 1973.

United States Environmental Protection Agency, "40 CFR Part 264: Statistical methods for evaluating ground-water monitoring from hazardous waste facilities; final rule," *Federal Register*, v. 53, n. 196, p. 39720 – 39731, U.S. Government Printing Office, 1988.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-6

CHOOSING A DISTRIBUTION

A guide for data analysts and interpreters on how to select an appropriate distribution model and document the choice

April 2001

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.

THE GENERAL IDEA

Many statistical procedures used in contaminated site studies involve assumptions about the underlying distribution of data values. If these assumptions are poorly founded, our statistical interpretations may be very misleading; we must be clear about our assumptions in order not to waste time on meaningless calculations. Even if our assumptions are well founded, a failure to state them clearly or to justify them in a report may leave doubt in a reviewer's mind about the validity of our conclusions; we owe it to those who eventually review our work to provide a clear statement of what has been assumed and why.

This document discusses the choice of a distribution model and recommends procedures for providing supporting documentation. It begins with an example that demonstrates how different assumptions about the underlying distribution can lead to very different remediation decisions. It then presents three common distribution models and describes statistical tools and procedures that can help us make an appropriate selection. It closes with a section that discusses whether or not a distribution model is necessary; many of the issues that we confront in statistical applications on contaminated site studies can be addressed adequately without assuming any distribution model.

There are other documents in this series that the reader should also examine. *DISTRIBUTION MODELS* provides an introduction to the distribution models discussed here and provides more detail on their statistical characteristics than is covered in this document. *NONPARAMETRIC METHODS* discusses statistical methods that do not require any distribution assumption. *IDENTIFYING POPULATIONS* addresses the problem of separating data into subpopulations, a common concern when the distribution of the available data appears multimodal and may reflect a mixture of two or more distributions.

INTRODUCTORY EXAMPLE

Even with exactly the same sample data, different choices of distribution models can lead to different remediation decisions. To take a simple example, consider the data shown in Table 1; these are measurements of the total volatile petroleum hydrocarbon (VPH) concentration in ten discrete samples taken at an early stage from a site containing roughly 50,000 cubic metres of soil, some of which may be contaminated.

Table 1 VPH values (in ug/g).

11	41	3	196	52	7	107	81	22	16

As a preliminary step, we might need to get a ballpark estimate of how much soil is considered industrial quality according to BC Environment regulations and how much has to be regarded as waste and removed from the site. We therefore decide to use statistics to help us estimate the proportion of material that exceeds the BC Environment industrial soil quality threshold of 200 ug/g VPH. Three of many possible distribution models we could adopt are:

- The data are from a normal distribution.
- The data are from an exponential distribution.
- The distribution of total VPH values over the entire area is exactly the same as that currently shown by the ten available samples (i.e. the highest value is exactly 196 ug/g).

If we assume that the sample values given in Table 1 are independent, we can use them to estimate a mean of 53.6 ug/g and a standard deviation of 60.6 ug/g for the underlying population. Using methods described in *DISTRIBUTION MODELS*, we can calculate that if the data are normally distributed, then 0.8% of the underlying distribution would exceed 200 ug/g This corresponds to about 400 cubic metres of soil that could not be considered industrial quality and would have to be removed from the site. Under the second assumption, we can use the same information to calculate that if the data are exponentially distributed, then 2.4% of the underlying distribution would exceed 20 ug/g. This corresponds to about 1,200 cubic metres of soil that could not be considered industrial quality. Finally, if we adopt the third assumption, then all of the soil could be considered industrial quality and none would have to be removed from the site.

The three assumptions lead to very different predictions about how much material will need to be remediated, with the difference between their corresponding costs amounting to several hundreds of thousands of dollars.

As this example shows, the choice of an appropriate distribution model may be critical, especially when it is based on few samples and is used to predict the probability of extreme events.

SOME COMMON DISTRIBUTION MODELS

This guidance document discusses the three distribution models shown in Figure 1: the normal, lognormal and exponential distributions; these are the distribution models whose properties and statistical characteristics are described in *DISTRIBUTION MODELS*. While these are three of the more common and useful distribution models for statistical studies of contaminated sites, there are many other distribution models that may also be useful for particular problems at specific sites; Johnson and Kotz (1970) provide details on a wide variety of alternatives.

HOW TO CHOOSE A DISTRIBUTION MODEL

When confronted with the need to document why we have chosen a particular distribution model, there are several different types of arguments that we can present. Some of these are specific quantitative calculations, others are more qualitative. Ideally, we should be able to use both kinds of arguments.

Symmetric or not?

One calculation that can help us decide whether to choose a symmetric distribution, such as the normal. distribution, or an asymmetric one, such as the lognormal or exponential distributions, is to compare the mean to the median. With symmetric distributions the two should be just about the same; with the asymmetric distributions commonly encountered in contaminated site studies, the mean is often much larger than the median. A boxplot provides a quick graphical check of whether the mean is close enough to the median to warrant an assumption that the underlying distribution is symmetric; if the mean plots outside the box (i.e. above the 75th percentile) then a symmetric distribution is not an appropriate model.

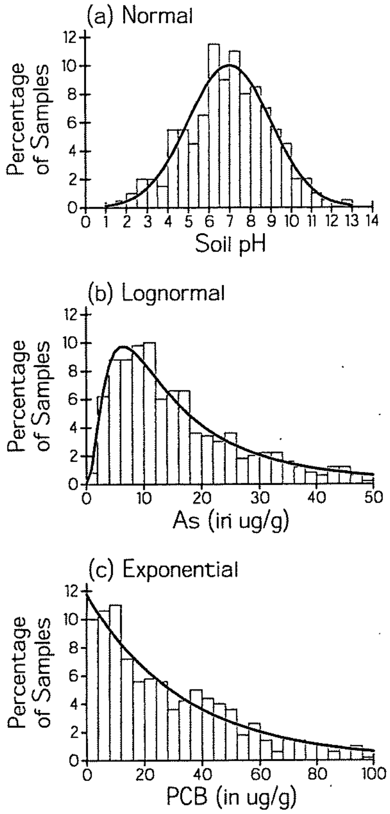


Figure 1 Examples of (a) normally distributed data, (b) lognormally distributed data and (c) exponentially distributed data.

Even if the mean is below the 75th percentile, this does not mean that a symmetric distribution is appropriate; with large data sets, we should expect much closer agreement between the mean and the median if we are going to assume a symmetric distribution. For a data set containing N values, a symmetric distribution is not an appropriate model if the difference between the mean and the median is larger than the standard deviation divided by $\sqrt{N}$.

Histograms, cumulative plots and probability plots

Figure 1 shows a graphical presentation that helps document the rationale for a distribution model: a plot of the relative frequencies predicted by the model along with the histogram of the data. Though this style of presentation may help to sort out hopelessly inappropriate models, it can also be somewhat deceptive. Figure 2 shows a histogram of arsenic measurements from a contaminated site. Superimposed on this histogram is the relative frequency curve predicted by a lognormal distribution with a mean of 14.1 ug/g and a standard deviation of 10.0 ug/g Though this looks like a good fit, it is very misleading to claim that the data come from a distribution whose mean is 14.1 ug/g because the actual mean of the data is 43.0 ug/g more than three times that of our theoretical model!

Figure 3 shows a cumulative plot of the arsenic data along with the cumulative curve predicted by the same theoretical lognormal model we considered earlier. From this plot it is clear that although the lognormal does a reasonably good job with the lower values, it does a very poor job with the high values. The plot shown in Figure 2 is deceptive because it doesn't show how badly we do at the very high end. If we are trying to give convincing graphical support for our distribution model, we should show how the cumulative plot of the actual data compares to the corresponding theoretical curve predicted by our distribution model.

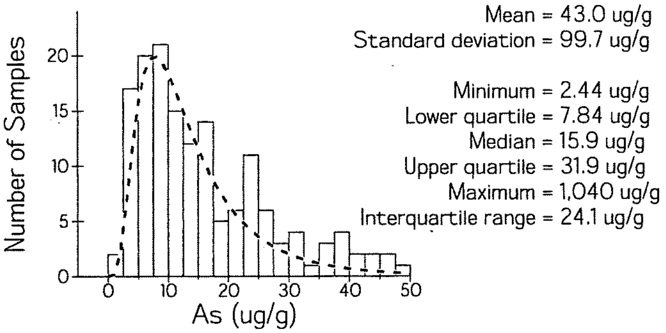


Figure 2 A histogram of arsenic concentrations along with a lognormal distribution model.

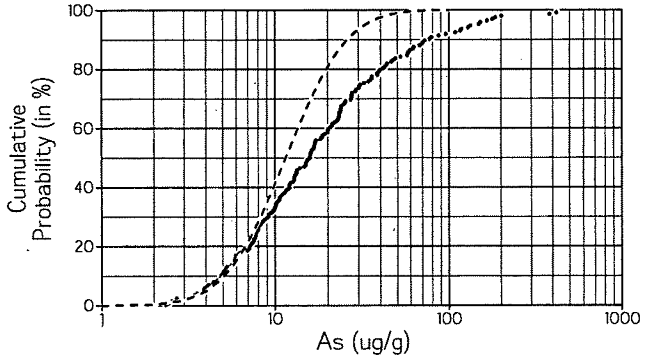


Figure 3 Cumulative plot of arsenic concentrations from Figure 2 along with the same lognormal distribution model.

Probability paper

With certain distribution models, such as the normal and lognormal ones, there is a convenient way to check if the cumulative plot of the actual data is close to what the theoretical model predicts. Rather than plot the actual and theoretical curves, as we did in Figure 3, we can use special probability paper whose axes have been scaled in such a way that the cumulative probabilities of the actual data will plot on a straight line if the data do, in fact, represent the distribution model we have chosen. With normal probability paper, the cumulative probability axis is squashed in the middle and stretched at the ends so that a cumulative normal distribution, which plots as an S-curve on an arithmetic scale, will plot as a straight line. If the cumulative probabilities of the data plot as a straight line on normal probability paper, then we have some justification for choosing a normal distribution as a model.

In addition to distorting the cumulative probability axis, lognormal probability paper also uses a logarithmic scale on the data value axis. If the cumulative probabilities of the data plot as a straight line on lognormal probability paper, then we have some justification for choosing a lognormal distribution as a model.

Statistical tests

For any distribution model that we might choose, there will always be some differences between the proportion of the actual data values that fall within a particular class on our histogram and the proportion that should fall within that class according to our theoretical model. The chi-square test (Bratley et al., 1983) provides a way of testing whether these differences between the actual and theoretical proportions are significant enough that we should abandon our distribution model. Unfortunately, this test is very permissive; in those few cases where it rejects our distribution assumption, we would likely reach the same conclusion through a comparison of the actual and theoretical cumulative probability plots. Other disadvantages of the chi-square test are that it needs more data to work well than are commonly available in contaminated site studies, and that independence is a necessary assumption.

Is there a good default model?

It is tempting to tackle the problem of choosing a distribution model by taking the point of view that one particular model is the best choice barring any strong evidence to the contrary. Unfortunately, there is no distribution that can serve as a good default for the wide variety of statistical problems that arise in contaminated site studies since the shape of the distribution depends on the volume of material in question. A histogram of discrete sample values from a contaminated site will look much more skewed than the histogram of the average concentrations of large stockpiles. Since one of the main reasons for choosing one distribution model over another is its skewness, it is important when choosing a preliminary distribution model to be clear on what volume the data are based.

When dealing with values defined on a relatively small volume, such as concentrations of discrete or composite samples, we should expect the distribution of data values to be skewed. Until we have enough data to confirm or refute a specific distribution, data values based on small volumes should be assumed to follow an asymmetric distribution, such as the lognormal or exponential distribution; the normal distribution is *not* a good default choice for such data.

A normal distribution is a good default choice only if we know that we are dealing with values defined on a large and homogeneous volume, such as the average concentration of an entire stockpile. It is important that the volume be both large and homogeneous. A common yardstick for measuring homogeneity is the coefficient of variation, which is discussed in *UNIVARIATE DESCRIPTION*; if the coefficient of variation is greater than 1, then it is not appropriate to assume homogeneity. If the material is not homogeneous, the contaminant concentrations typically span several orders of magnitude, and the distribution of the average concentration of entire stockpiles, even large ones, may still be noticeably skewed.

As an example of the linkage between the choice of a distribution model and the volume of material under consideration, consider the problem of interpreting the results of a composite sample taken from a stockpile. If we are trying to address the issue of whether any single discrete sample value in the composite might have exceeded some threshold then we are concerned with data values that are based on a very small volume of material: a single discrete sample. For this issue, our preliminary assumption should be that the data values follow a skewed distribution. With the same sample information, however, we might be trying to address the issue of whether the average concentration of the entire stockpile is above some threshold; we are now interested in data defined on a much larger volume: an entire stockpile. For this issue, we could adopt a normal distribution as our preliminary model if we can assume that the stockpiled material is homogeneous; as discussed in the document entitled *STOCKPILING*, such an assumption of homogeneity is best supported by a careful and thorough *in situ* characterization study.

WHY CHOOSE A DISTRIBUTION?

Before choosing a distribution model, it is worth considering whether we really need one. How is our work made any easier or better by assuming a distribution model?

A historical perspective

The reason why statisticians seem to spend so much time worrying about an appropriate distribution model lies in the early part of this century, when data were scarce and computers nonexistent. In an era without computers or calculators, the initial focus of a statistical study was on finding a tractable, well understood mathematical model that described the distribution of the data values. Though the data set in a typical statistical study from the early part of this century would now be considered quite a small data set, it was still difficult to deal with the raw data. Even with as little as 20 or 30 data values, simple mathematical calculations, such as the mean or the standard deviation, are tedious when they have to be done manually.

The life of statisticians in the early part of this century was made much easier by the pioneering work of mathematical statisticians like Sir Robert Fisher, who added a great deal to the knowledge of how certain distributions behave. With a large and growing literature on the properties of various distribution models, statistical studies were considerably simplified if the actual data were replaced by a model. It is possible to make much quicker progress with a normal distribution model, for example, than to struggle through manual calculations with actual data. Once the parameters of the distribution model have been chosen, typically the mean and standard deviation, it is possible to make many different kinds of predictions about the behaviour of the entire population. We could calculate its percentiles, for example, its skewness, its peakedness, its mode, its median, and so on — all without having to grind the actual data through another set of calculations.

With the advent of modern computers, however, the need for a distribution model becomes questionable. With computers able to rapidly sort data, even if there are thousands of values, and able to calculate even the most complicated statistics in a few seconds, why should a tractable and well-studied mathematical model be of much interest?

Advantages of distribution models

Even though their computational convenience is now largely a matter of historical curiosity, distribution models possess other advantages.

Some kind of model is necessary if we are trying to make predictions about events that are so rare that they are never (or hardly ever) observed. Those weird and wonderful statistics about how much more likely it is that we'll get hit by a meteorite than suffer a fatal accident related to nuclear reactors, are all based on distribution models for low probability events. Statistical predictions of this type are very sensitive to the way that the distribution model behaves for extreme values. As we saw in the introductory example with VPH concentrations, there can be considerable variability in predictions about the chance of exceeding a threshold that no data value has yet exceeded. If our distribution model predicts a rapid decrease in the occurrence of extreme values (like the normal distribution does), then we're not going to calculate a very high chance of exceeding the threshold; if, on the other hand, the model predicts a slower decrease in the occurrence of extreme values (like the exponential distribution does), then we're going to calculate a higher chance of exceeding the threshold.

Another advantage of choosing a distribution model is that it gives us a very compact way of describing a data set. Where a need exists to communicate the essential features of a data set to other people, it is often easier to say something like "the VPH concentrations follow an exponential distribution with a mean of 54 ug/g" than to list all of the available data. When used in this way, the distribution model is useful only if our audience is already familiar with its shape and statistical parameters.

The use of distribution models as a kind of shorthand notation for describing data takes on a sharper focus when, in certain fields of study, the use of specific distributions is so common that workers in the same discipline can use the parameters of the distribution as diagnostic features. For example, although the parameters of the Weibull distribution, commonly called λ and α, are not likely to be familiar to most people, they are so commonly understood by many of the researchers who study the failure rates of communication systems that experimental data sets from this area of application are often summarized with these two parameters alone.

The final advantage of some distribution models is that they simplify certain inferences and predictions. The most notably convenient and computationally simple distribution model is the normal distribution. In order to quantify the uncertainty on an estimate, our job is made much easier if we assume that the errors we might make with our estimate are normally distributed. Having made this assumption, all that is needed to develop confidence intervals is an estimate of the standard deviation of the estimation errors. Once this is available, the 95% confidence interval goes from two standard deviations below to two standard deviations above the estimate. When used in this way, the distribution model is not something that we choose after thoughtful consideration of our data, it is something that we *hope* is appropriate because it makes our calculations easier. When this hope has no justification,eagerness for a simple and tractable calculation usually leads to misapplication of a distribution model.

Disadvantages of distribution models

The main disadvantage of using a distribution model is that it may not be appropriate for a particular set of data. While one of the commonly used distributions can usually do a good job of fitting most of the data, few of them do a good job for all of the data values. Typically, there are departures between what a distribution model predicts for the occurrence of extreme values and what the data actually show. With our interest in contaminated site studies often focused on the high values, the good fit of a model over the lower 90% of the data may be useless if it does a poor job of fitting the critical upper 10%.

RECOMMENDED PRACTICE

1. All distribution assumptions should be made explicit in reports.
2. If the difference between the mean and the median of a data set containing N samples is greater than the standard deviation divided by $\sqrt{N}$, then it cannot be assumed that the data come from a symmetric distribution, such as the normal distribution.
3. The appropriateness of a distribution model should be documented graphically by comparing the cumulative probabilities of actual data to the cumulative probabilities predicted by the theoretical model. Such a comparison should be done on probability paper, if available.
4. If there are too few data to adequately support or refute a distribution model, then discrete samples should be assumed to follow an asymmetric distribution. Average values over large volumes, such as stockpiles, may be assumed to follow a symmetric distribution, such as the normal distribution, if *in situ* characterization has demonstrated the material to be homogeneous.

REFERENCES AND FURTHER READING

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

Bratley, P., Fox, B.L. and Schrage, L.E., *A Guide to Simulation*, Springer-Verlag, 1983.

Jaeger, R.M., *Statistics — A Spectator Sport*, Sage Publications, 1990.

Johnson, N.L. and Kotz, S., *Distributions in Statistics — Continuous Univariate Distributions, Volume 1*, Houghton Mifflin, 1970.

Moore, D.S., *Statistics — Concepts and Controversies* W.H. Freeman and Company, 1985.

Size, W.B., (ed.), *Use and Abuse of Statistical Methods in the Earth Sciences*, IAMG Studies in Mathematical Geology, Volume 1, Oxford University Press, 1987.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-7

IDENTIFYING POPULATIONS

A guide for data analysts and interpreters on
the identification of statistical populations

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

The essence of statistical inference is the borrowing of information from a group of data to make predictions about how a particular population behaves. The grouping of data and the definition of the population(s) of interest are fundamental and recurring problems in the application of statistical methods to contaminated site studies. On one hand, the uniqueness of each and every sample encourages us to split the data into smaller and smaller groups and to recognize multiple populations in our data. On the other hand, the need for a sufficient number of data to support statistical calculations encourages us to group data together into larger groups and to work with fewer populations, each of which contains more data.

As an example of this problem, consider the twenty sample values listed in Table 1 and shown as a histogram in Figure 1. Are there two populations here, a "background" of low values in the 0 to 10 ug/g range and another population of higher "contaminated" values? Or is there just a single population that happens to be highly skewed with a lot of low values and a decreasing proportion of higher ones?

Table 1

44	140	6.3	76	6.5
2.7	89	86	2.6	14
6.3	67	49	97	5.9
4.3	52	4.0	32	72
4.8	8.6	39	61	89
1.8	120	4.4	63	9.7
7.8	100	20	3.2	8.5
14	5.4	7.2	80	6.2
1.1	5.4	2.3	16	3.8
81	94	3.1	6.4	110

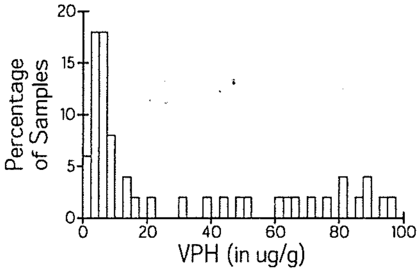


Figure 1 VPH concentrations in soil samples from a contaminated site.

If these data are viewed as a single population, then an appropriate distribution model would need to be asymmetric with a long tail; a lognormal distribution, for example, could do the job. On the other hand, if these data are viewed as a mixture of two populations, then it might not be necessary to use skewed distribution models; a combination of two normal distributions might be more appropriate. These two different approaches to data analysis and interpretation will lead to quite different predictions, particularly when trying to estimate the probability of extreme events.

Unfortunately, there is no statistical test that unambiguously proves that data belong in a single population or that they need to be split into separate populations. Trying to test for whether the data should be grouped or split is a chicken-and-egg problem. Until we assume some underlying population, we have no point of reference against which we can compare our actual data. Developing such a point of reference requires some data (or some bold assumptions) and, depending on which data we choose (or which assumptions we choose to make), we will either conclude that the data should be grouped or we will conclude that they should be split.

Despite the awkwardness of documenting that a particular grouping or splitting decision is appropriate, the issue of identifying the population(s) in a data set is critical. Without a clear definition of the population(s), other important issues, such as the evaluation of outlier data, can not be resolved.

This document presents guidance on identifying statistical populations. It begins with a discussion of graphical tools and then addresses statistical tests that can help with the decision of whether to treat the data as a single population or as several separate populations. There are other guidance documents in this series that contain related material. The one entitled *OUTLIERS* provides additional insight into methods for evaluating whether or not a particular sample should be treated as part of the population; *NONPARAMETRIC METHODS* provides alternatives to the statistical tests outlined here.

QUALITATIVE INFORMATION

The decision to group data into a single population or into several separate populations should, wherever possible, take into account qualitative information. An understanding of the historical use(s) of a site is invaluable in developing an appropriate statistical treatment of the available data. Field notes that describe local conditions in the immediate vicinity of each sample location are also very useful since these often provide critical clues to the physical, chemical and geological conditions that influence contaminant concentrations. A clear understanding of the goal of the study is also necessary in making appropriate decisions about the statistical treatment of the data; though it may be appropriate to group all of the data into a single population for assessing the total volume of soil that requires remediation at a contaminated site, it may be necessary to split the data into several populations if the goal of the study is detailed local mapping of contaminant concentrations.

There are several graphical tools and statistical tests that can be used to support decisions about grouping data together or splitting them into several separate populations. These should not be used by themselves, however, to justify a decision regarding statistical populations. If a probability plot, for example, suggests that there may be a mixture of two populations at the site, then this observation, which is based purely on a quantitative consideration of the data, should be reconciled with the qualitative information about the site. Do the two populations reflect natural background and industrial contamination? Are the magnitudes of the values in the population being designated as "background" consistent with other infarmation about what the natural background levels should be? Could the two populations both be due to industrial contamination but from different sources? If so, are both sources part of the focus of the study? All of these questions, and dozens of similar ones, can not be answered with the data values themselves and need supporting historical and geological information.

GRAPHICAL TOOLS

Probability plots

One of the most useful graphical displays for exploratory data analysis is a probability plot, an example of which is shown in Figure 2. On such a graph, data values are plotted on the x-axis against the cumulative probability on the y-axis. Using Figure 2 as an example, about 35% of the data values are below 100 ug/g and slightly more than 90% are below 1000 ug/g

The scaling on the axes of a probability plot is often confusing to non-statisticians. The y-axis is squashed in the middle and stretched at the ends; the distance between 50% and 60%, for example, is smaller than the distance between 80% and 90%. This kind of y-axis scale is used on probability plots because it makes it easier to tell whether the data are close to being normally distributed. If both the x and y axes have a conventional linear scale, then cumulative curves of normally distributed data will plot as an S-curve. It is difficult to tell if an S-shaped curve is close to the kind of S-shape that normal data would produce; it is much easier to tell if the data are plotting on a straight line. By stretching out the y-axis for the very low and very high values, and squashing it for the middle ones, we end up with distorted graph paper on which cumulative curves of normally distributed data will plot as a straight line.

If the distribution of data is skewed, with many low values and a decreasing proportion of high ones, it is common to use a logarithmic scale on the x-axis; this style of log-probability plot is the one that has been used in Figure 2. With logarithmic scaling on the x-axis and the distorted probability scale on the y-axis, cumulative data will plot as a straight line if the data are lognormally distributed.

Whether the data o r their logarithms are normally distributed or whether they follow some other distribution, a probability plot is useful for exploring possible sub-populations. If the probability plot has kinks, with some of the values not following the trend of the others, this is often taken as evidence that the data should be separated into different groups. With the example in Figure 3, the mercury data values show a consistent trend up to about the 90th percentile; the highest 10% of the data, however, do not follow the same trend as the lowest 90%. This behaviour indicates that the highest 10% of the values could be treated as a separate population. As discussed earlier, however, the decision to treat the highest 10% as a different population should be supported by qualitative information regarding the history of the site and the source of contamination.

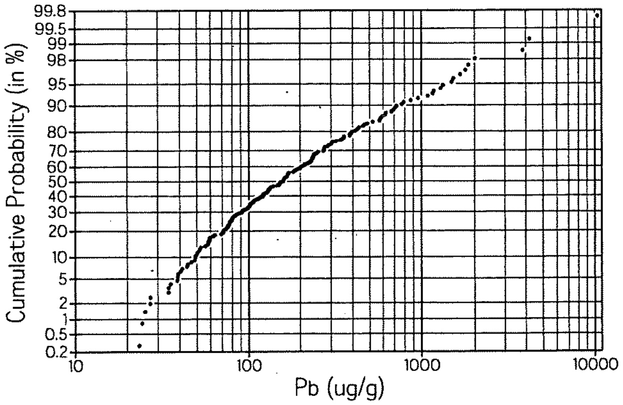


Figure 2 A log-probability plot for lead concentration measurements from the soil in the vicinity of a smelter.

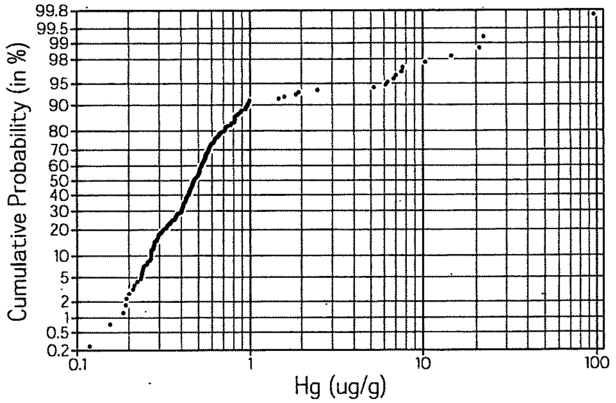


Figure 3 A log-probability plot showing multiple populations.

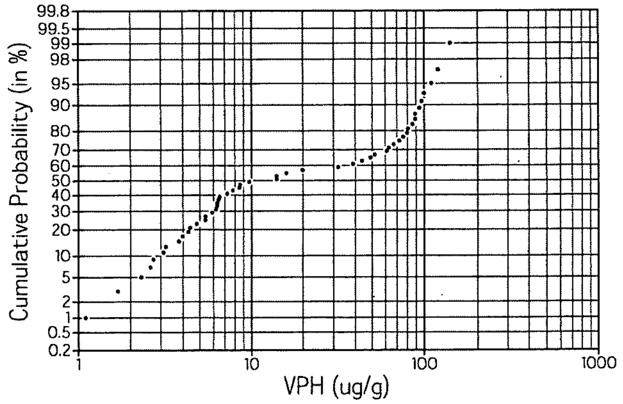
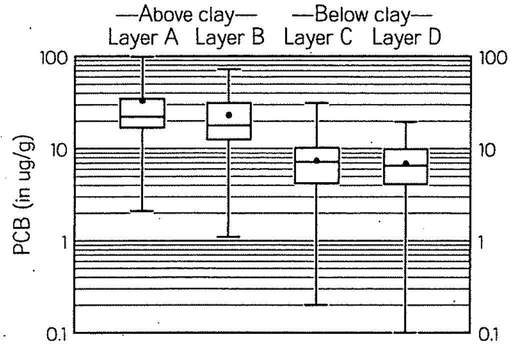


Figure 4 A log-probability plot that shows a mixture of "background" and "contaminated" populations.

Figure 4, which uses the VPH data shown in Table 1, shows another type of behaviour that probability plots often exhibit: the data values seem to plot on two different trends with a gradual transition between the two. Probability plots of this type are usually due to overlapping mixtures of several populations. Such mixtures are common in contaminated sites where the lower end of industrial contamination overlaps with the higher end of natural background contamination. Sinclair (1974) discusses how the information from such probability plots can be used to develop distribution models for each of the mixed populations as well as to calculate the proportion of each population in the mixture.

Side-by-side boxplots

In many contaminated sites there is qualitative infamation about the site or the soil conditions that allows us, if we deem it necessary, to subdivide the data into different groups. When considering whether such a splitting is appropriate or not, it is often useful to be able to compare the distributions in the different groups under consideration. Fa example, we might expect the level of PCB contamination at depth to indicate whether or not the surface contamination has passed through a clay layer. This situation differs somewhat from those that we considered in the previous section. With all of the probability plot examples shown earlier we started with the data in a single group and used the probability plot as a tool to study whether it might be more appropriate to subdivide them. With the PCB example given in this section, we already have a grouping in mind — depth from surface and position relative to the clay layer — and we are trying to decide whether this distinction is important or whether no subdivision is necessary.



No. of data	69	58	39	31	No. of data
Mean	33.2	23.1	7.4	6.9	Mean
Stand. Dev.	65.2	47.1	20.4	19.2	Stand. Dev.
Maximum	98.8	72.1	31.3	19.4	Maximum
Upper quartile	34.6	31.3	10.2	9.9	Upper quartile
Median	22.2	17.9	7.2	6.5	Median
Lower quartile	16.8	12.5	4.2	4.1	Lower quartile
Minimum	2.1	1.1	0.2	0.1	Minimum

Figure 5 Side-by-side boxplots of PCB contamination.

One way of comparing distributions from several groups of data is to plot their histograms and tabulate some key summary statistics. Side-by-side boxplots, such as those shown in Figure 5, provide a more useful graphical comparison. The box in the middle of each individual boxplot extends from the lower quartile to the upper quartile of the distribution; the bar in the middle of the box is the median and the "arms" define the range (minimum to maximum). The black dot shows the mean of the distribution.

A boxplot presents most of the relevant univariate information that we need from an exploratory data analysis. It gives us a sense for where the middle of the distribution lies, how spread out it is and whether it is symmetric or not. The boxplot therefore offers most of the useful information that a histogram contains, but in a more compact format that is more amenable to side-by-side comparisons between different groups of data.

Where side-by-side boxplots fa two groups of data show that their boxes do not overlap — the central 50% of one group does not overlap with the central 50% of the other group — this is evidence that supports treating the two groups separately. As with the other graphical and statistical tools for examining populations, a decision to split data into separate populations should not be based on boxplots alone but should also be supported with qualitative information that explains why the distributions are different.

Scatterplots

Where probability plots suggest a mixture of two overlapping populations, it is often difficult to identify the population to which each sample belongs since intermediate values could be high values of one population or low values of the other. Most contaminated site studies involve a suite of possible contaminants and scatterplots can often be the key to sorting out which samples belong to which populations.

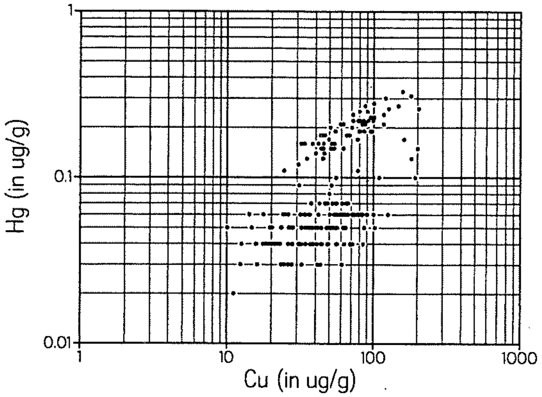


Figure 6 Separate populations on a scatterplot.

Figure 6 shows an example in which the copper contamination consists of two populations that overlap. Using only the copper concentration it is not possible to assign each sample to the background or contaminated population since the low end of the contaminated distribution overlaps with the high end of the background distribution. At this site, mercury concentrations have also been measured. When the mercury and copper concentrations are plotted together on a scatterplot, the separate populations become clearer. With the use of both variables the separation of the data into two separate populations is much more straightforward than when the copper or mercury concentrations are used separately.

STATISTICAL TESTS

In addition to the graphical tools that can be used to support a decision to treat two groups of data separately, there are some statistical tests that can also provide support for such a decision. Statistical tests are often used when the decision to recognize separate populations is not immediately obvious. With the data shown in Figure 5, for example, it is clea that the PCB concentrations above the clay layer are different from those below the clay layer. For the A and B soil layers that are above the clay layer, however, it is not as obvious whether or not these should be treated as different populations. Though their means are different, this could either be due to chance a could also be due to the underlying populations being different in the two layers; even if the two sets of data were drawn from the same underlying population, we would still expect to see some differences in their means. There are statistical tests that help us to decide if a difference between the statistics of two groups of data is due to chance alone or if it is more likely that the two groups were drawn from different populations.

The t-test

The t-test is used to determine if the difference between the means of two populations is "statistically significant". This test begins by assuming that the two groups of data are from the same population and then tries to refute this assumption.

If sets of N data are drawn from a common population, the means of these different sets will fluctuate around the mean of the parent population. How much fluctuation we should expect depends on the value of N; if N is large then the mean of the actual data will be closer to the mean of the parent population than if N is small. If the N values in each set are independent, then σ_m, the standard deviation of the means of the different sets, is related to σ, the standard deviation of the parent population, by the following equation: $\sigma_m = \sigma \div \sqrt{N}$.

In addition to using this equation that describes how much the sample means can fluctuate from the mean of their underlying population, the t-test assumes that the means will be normally distributed. Under all of these assumptions — that the means are, in fact, the same, that the samples are all independent and that the means are normally distributed — the following statistic should follow a standard normal distribution:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

N_1, m_1 and σ_1 are the number of samples in the first group, their mean and their standard deviation; N_2, m_2 and σ_2 are the corresponding values for the second group. If the value of t calculated from the equation above is well within the range of values expected for a standard normal distribution, -3 to +3, then the difference between the means can well be explained by the random fluctuations that we expect between groups of samples drawn from the same distribution. If t is less than -3 or greater than +3 then it is very unlikely that random fluctuations alone are causing the difference. The conventional approach is to interpret extreme values of t as evidence that the two groups come from different populations.

As an example of the use of the t-test, let us determine whether the Layer A and Layer B samples from Figure 5 have significantly different means. The values that we need to substitute in the equation for the t-statistic are the following:

$$N_1 = 69 \qquad m_1 = 33.2 \qquad \sigma_1 = 65.2$$

$$N_2 = 58 \qquad m_2 = 23.1 \qquad \sigma_2 = 47.1$$

With these values, the calculated value of the t-statistic is 1.01, well within the -3 to +3 range that we expect for a standard normal distribution. This tells us that the difference between the mean value of the 69 Layer A samples and the 58 Layer B samples is not large enough to lead us to believe that the underlying populations are different.

The t-test given above is what is called a "two-sided" t-test since it takes into account that neither of the sample means is exactly the same as that of the underlying population. In situations where the mean of the underlying population is known (not a very common situation in contaminated site studies), the equation given above can be turned into a "one-sided" t-test by setting m_2 to the true mean and σ_2 to zero. An example of a situation in which we may prefer to do a one-sided t-test is the comparison of laboratory measurements of a reference standard to the accepted value of the standard. In this case, the true mean is the accepted value of the standard and we are interested in whether the mean of repeated measurments is significantly different from this accepted reference value.

The philosophy of the t-test is to make an assumption, namely that the data are from the same population, and then use extreme values of the t statistic to argue that this assumption is not very plausible. It should be noted, however, that in believing that the calculated t-statistic should come from a standard normal distribution, we are making several other assumptions. It is possible that the assumption of a common population is correct and that it is one of the other ones — independence of the samples or normality of the means — that is incorrect.

There are some statistical tests for differences between populations that do not make distribution assumptions; some of these are discussed in the guidance document entitled *NONPARAMETRIC METHODS*. If the samples are not independent (a common case in contaminated site studies) then the difference between the two means can be larger than the t-test assumes. If two groups of data fail the t-test under an assumption of independence — if their t-statistic is between -3 and +3 — then they would also fail a modified version of the test when correlation between the samples is taken into account.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Sinclair, A.J., "Selection of threshold values in geochemical data using probability graphs," *Journal of Geochemical Exploration*, v. 3, p. 129 – 149, 1974.

Understanding Robust and Exploratory Data Analysis, (Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds.), John Wiley & Sons, New York, 1983.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-8

OUTLIERS

A guide for data analysts and interpreters on how to evaluate unexpected high values

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

In contaminated site studies it is common to find that the data contain some surprisingly high values. Knowing that such high values are likely to have a profound effect on statistical analysis and interpretation, many of us are tempted to dismiss these unexpected (and possibly unwelcome) observations as "outliers" and to remove them from the data base. Discarding actual observations is not a good practice, however, since a thorough evaluation of the reasons for these unexpected values may lead to new insights into the data or to a reconsideration of underlying assumptions about the data and their distribution.

Whatever we decide to do with the outliers, this single decision will be one of the most critical in our study. If an erroneous high value is kept it may cause uncontaminated material to be misclassified as contaminated; such errors are costly because they lead to needless remediation. On the other hand, the decision to discard erratic high values may be even worse. If such values represent a previously unforeseen population, then arbitrarily discarding them will cause contaminated material to be left unremediated. With decisions on remediation often hinging on the proper evaluation and use of outlier values, it is necessary to have some consistency and objectivity in the treatment of outliers in contaminated site studies. This document aims to provide this much-needed consistency and objectivity by providing guidance on the identification and evaluation of outliers.

WHAT IS AN OUTLIER?

Barnett and Lewis (1984) give the following definition of an outlier: *An outlier in a set of data is an observation that appears to be inconsistent with the remainder of that set of data.* This definition identifies two aspects of the outlier problem: the prior decision to group data together and the apparent inconsistency that results. One possible solution to the outlier problem will be to rethink how we have grouped the data — maybe the outlier is providing clues to the existence of another previously unrecognized subpopulation. Another possible solution will be to revisit why it *appears* inconsistent — maybe we have faulty underlying assumptions about how the data should behave.

HOW TO IDENTIFY OUTLIERS

Figures 1 and 2 show two of the graphical displays that can assist with detection of outliers. Figure 1 is a probability plot of lead concentration measurements from the soil in the vicinity of a smelter. For most of the data values, their cumulative probabilities plot on a fairly straight line; at the high end, however, this trend breaks up. The highest few values fall to the right of the trend, meaning that these highest values are *even higher* than we might expect based on our observations of the rest of the data. Figure 2 shows a scatterplot of the same lead data plotted versus their distance from the smelter. The cloud of points shows a tendency for the lead concentration in the soil to decrease with distance from the smelter. There are some aberrant samples, however, that have abnormally high lead values when compared to other measurements taken at a similar distance from the smelter.

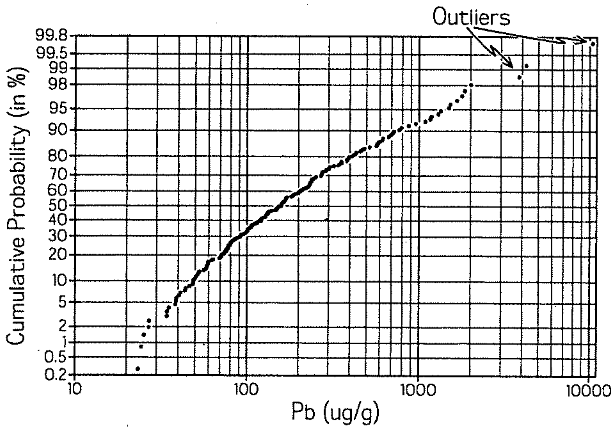


Figure 1 A cumulative probability plot for lead concentration measurements from the soil in the vicinity of a smelter.

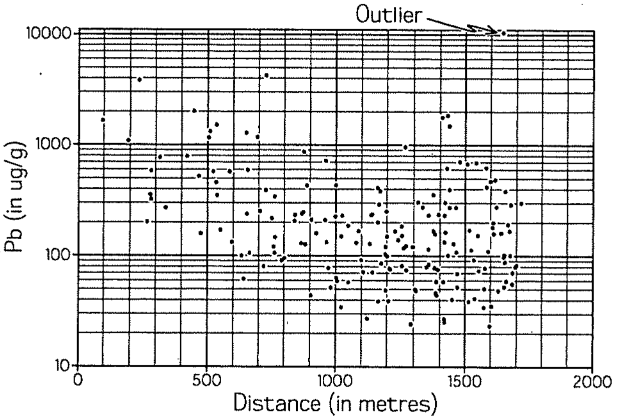


Figure 2 A scatterplot showing lead concentration data used in Figure 1 versus distance from the smelter.

The probability plot in Figure 1 and the scatterplot in Figure 2 complement each other since the sample(s) that appear as outliers on one plot do not necessarily appear as outliers on the other. A combination of statistical and spatial displays provides a more complete basis for identifying outliers than would any single plot.

In contaminated site studies, where there are several suspected contaminants, scatterplots of the concentration of one contaminant versus another may reveal that some samples have unusual ratios of the suspected contaminants. For example, if a plot of lead versus arsenic shows a narrow cloud in which the Pb/As ratio is quite consistent, then any samples that plot away from this main cloud could be classified as outliers.

Maps or cross-sections on which the data are colour-coded according to their order of magnitude can assist with identifying dubious samples that have moderate values but are inconsistent in their spatial context.

HOW TO EVALUATE OUTLIERS

Once an observation has been designated an outlier, we need to evaluate its significance. It should not be discarded as spurious until we have explored the possibility that our prior decision about the populations was reasonable and that our assumptions about how the distribution should behave are all appropriate. This examination of our prior decision and our assumptions must take into account not only the provenance of the data — where they came from and how they were collected and analyzed — but also the study objectives. Outlier data that might be appropriately dropped from one study may still be useful in another study with a different objective.

The lead data used in Figures 1 and 2 provide a good example of how data provenance and study objectives impinge on the treatment of outliers. In Figure 1, there are three values that might be treated as outliers due to their departure from the trend shown by the other values. When their distance from the suspected source of contamination, the lead smelter, is taken into account (Figure 2), only one of them remains suspicious — the value that approaches 10,000 ug/g far from the smelter.

In this lead study, the samples were located without consideration of the local site conditions. In addition to collecting the samples from their designated locations, the field staff also described the local conditions in the vicinity of each sampling site. The resulting set of field notes was an invaluable source of information for decisions regarding the treatment of outliers. These field notes record the fact that the value approaching 10,000 ug/g was collected from a junkyard that contained dozens of leaking car batteries. The knowledge that this sample is likely affected much more by a very localized source of contamination — leaking battery acid — than by the smelter allows an appropriate treatment of the outlier value.

Decisions about the handling of outliers are much easier to make if we maintain a clear audit trail that allows us to trace each and every data value back through the data base compilation, through the laboratory analysis, through the sample preparation procedure, and ultimately back to the specific time, location and conditions under which the sample was collected. Without a carefully maintained set of field notes and laboratory records, it may become impossible to make appropriate decisions about outliers.

Data errors

Some outliers are due to human error during sample collection, preparation and analysis; further errors can occur when analytical values are transcribed and compiled into a data base. Samples may be tagged and labelled incorrectly; they may be contaminated during handling in the field, during transportation to the laboratory or during the laboratory preparation process; the analytical procedure may not be implemented correctly. Even if a sample survives all of these possible humiliations, its analytical value may be transcribed incorrectly or it may be corrupted when it is electronically merged into a data base.

One of the few universal rules that we can make about handling outliers is this: we should not use data values that are clearly in error. At the same time, however, we should not be too quick to use the excuse of data errors to justify a decision to discard outliers. Data errors are a double-edged sword; while they can provide one of the few non-controversial reasons for discarding outliers, they also call into question the entire data base. If it becomes apparent that an outlier value is erroneous, then all data should be checked to see if any of the other data have been affected by the same problem. For example, if "suspected contamination" is the reason given for discarding an outlier, we must also question whether any of the other samples with more moderate data values might also be contaminated. Particularly in the case of cross-contamination between samples it is not appropriate to discard the high values and keep the low ones. Similarly, if "data transcription error" is the reason given for discarding an outlier, then we need to consider the possibility that some of the lower and more moderate numbers that remain in the data base are erroneously low for exactly the same reason.

Choice of population

If there is no reason to suspect that an outlier is due simply to human error, we should then consider the possibility that the value appears inconsistent with the rest of the samples because it does not belong in the same group — that we have made the mistake of mixing apples and oranges. This was the case with the very high lead concentration from the example shown in Figures 1 and 2; though the lead concentration in the rest of the samples might reflect the effect of the smelter, the lead concentration in the sample that came from the junkyard likely has very little to do with the smelter.

If qualitative information about the provenance of the data makes it clear that a particular value is not relevant according to the study objectives, then we should discard the irrelevant value from our study and document the reasons for doing so. In choosing to discard a particular observation for this reason, we must be clear about why it does not belong to the population under study. This requires that we have an unambiguous definition of what population *is* under study and that we know what other population the offending data value belongs to. In documenting the reasons for believing that an outlier belongs to a different population, we also need to reconsider the study objectives. If an outlier has revealed a previously unforeseen source of contamination, for example, should the study be broadened to address this new source, or should the objectives remain unchanged? In the example of the lead contamination in the vicinity of the smelter, field notes made it clear that the lead in one of the sample values was likely due to leaking battery acid. In addition to identifying this sample as part of a separate population, we should also consider whether the appearance of this unanticipated new population — leaking battery acid — affects the study objectives.

Distribution assumptions

If we cannot dismiss an outlier as simply erroneous and if we cannot dismiss it on the grounds that it belongs to a different population, we can then consider the possibility of rejecting it on statistical grounds. The decision to use a statistical argument for discarding an outlier is a desperate last resort, however, because virtually all of the statistical approaches to the problem rely entirely on our assumptions about the underlying distribution. Any statistical argument for the rejection of an outlier can be turned around into an argument that the underlying assumptions about the distribution are faulty. Before resorting to statistical arguments for rejecting outliers, we need to document why we continue to believe that our distribution assumptions are appropriate in spite of the outlier observations. One of the other documents in this series, entitled *CHOOSING A DISTRIBUTION*, discusses this issue in greater detail and provides recommendations for selecting a distribution and documenting the appropriateness of the choice.

Discarding outliers for statistical reasons

If, despite the observation of outliers, we are sure that the assumed distribution is appropriate, outlier values should be checked for consistency with the assumed distribution. By "assumed distribution", we mean the distribution that is assumed for all of the non-outlier values. For example, if we have decided that a normal distribution is an appropriate model for our data values, then the estimation of the mean and standard deviation of our assumed distribution should be based only on the non-outlier values and should not consider any outliers.

To justify the decision to discard an outlier, we should check two things. First, we should make sure that there is a very low probability that the outlier value belongs to the assumed distribution. Second, we should make sure that it is not part of a continuous tail of high values. Both of these checks require the ability to calculate percentiles of the assumed distribution.

To check that the outlier has a very low probability of belonging to the assumed distribution, we need to confirm that it falls in the upper 1% of the distribution. Tables and formulas in Johnson and Kotz (1970) allow us to calculate the percentiles for the distributions commonly used in contaminated site studies. For two of the more common choices, here are some rules of thumb:

Normal distribution. If the data are assumed to be normally distributed then any value more than three standard deviations above the mean will be in the upper 1%.

Exponential distribution. If the data are assumed to be exponentially distributed then any value more than five times the mean will be in the upper 1%.

Most of the distributions commonly used in contaminated site studies, including the two given above, have a maximum at infinity, so we can never completely reject the possibility that the outlier *might* belong to the assumed distribution. If it belongs to the upper 1%, however, it is an unlikely enough value that we can continue with the second check.

The upper 1% rule should not be the sole basis fa identifying outliers; in addition to confirming that the outlier value is in the upper 1%, we should therefore also make sure that it is aberrant even for a large value. One way of doing this is to check if there is an unexpectedly large gap between the value of the highest non-outlier and the outlier. Having made an assumption about the distribution, we can see what the assumed distribution would predict for the difference between the two largest values. If the actual difference is more than twice that predicted by the assumed distribution, then the outlier can be discarded.

To implement this gap check, we need to know what value the assumed distribution would predict fa the largest two values in a set of N observations. For the purposes of this gap check, we assume that the largest two values should correspond to the following percentiles of the assumed distribution:

$$\text{Assumed percentile for largest value} = \frac{N-1}{N} \times 100$$

$$\text{Assumed percentile for second largest value} = \frac{N-2}{N} \times 100$$

Once we know which two percentiles we are interested in, we then use the standard tables or formulas to find the two corresponding values from our assumed distribution. We are not particularly interested in how these two theoretical values compare to the two highest values that were actually observed; what we are interested in is their difference. If the difference actually observed between the outlier and the highest non-outlier value is more than twice the difference predicted by our assumed distribution, then the outlier can be discarded as inconsistent.

There are not many rules of thumb for what the difference described above should be; it will vary with the number of data and with the specific distribution model being assumed. For a quick approximation, however, one can assume that the tail of the distribution behaves like an exponential distribution, in which case the procedure described above depends only on the number of samples. Rather than explicitly checking that the absolute difference between the actual values is more than twice the absolute difference predicted by our assumed distribution, we can implement exactly the same gap check in terms of the relative difference:

$$\text{Predicted difference (in \%)} = \left[\frac{\log(N)}{\log(N) - \log(2)} - 1\right] \times 100$$

If the actual relative difference between the outlier and the highest non-outlier value is more than twice the predicted relative difference given above, then the outlier may be discarded.

Replacing outliers with new samples

Since there is a loss of information whenever sample values are discarded, we should always try to replace outlier samples with new samples. This is especially important if the sample values are being used for local mapping to support remediation planning. The new sample should be taken as close as possible to the discarded outlier, ideally within 1 m. If the new sample value is the same as the discarded outlier (within the tolerance predicted by QA/QC procedures on duplicate samples) then there is likely an unanticipated "hot spot" that needs to be better delineated. Even if the new sample value is quite different from the discarded outlier, we should still make an effort to understand why the original sample value was so unusual since this may lead to useful insights about the appropriate interpretation of the other data that we have decided to keep.

STATISTICAL METHODS FOR ERRATIC DATA

If outliers cannot be discarded for any of the reasons discussed above, then they must be used with care. Many common statistical tools are very sensitive to erratic high values. Any statistic that involves some type of averaging — such as the mean, the standard deviation and the correlation coefficient — will be strongly influenced by erratic high values. There is a large set of statistical tools and procedures that are said to be "robust" because they produce sensible results even in the presence of erratic high values. Huber (1981) is a standard reference for robust statistical methods; Hoaglin et al. (1983) provide an extensive discussion of robust methods for exploratory data analysis. Isaaks (1984) addresses the problem of mapping contaminant concentrations in the presence of erratic high values.

Before choosing more robust procedures, many of which are considerably more complicated than the less robust traditional alternatives, we can check to see if our remediation decisions are affected by the inclusion or exclusion of outliers By running every relevant calculation first with the outliers included and then with the outliers excluded, we can document the sensitivity of our final decision to the presence of the outliers. It is important in such sensitivity studies to keep in mind that it is not the actual statistics themselves that are of interest, but their effect on our remediation decision. If a statistic changes considerably when an outlier is included or excluded but the remediation decision remains the same, then the outlier has no real effect on the decision.

If sensitivity studies show that statistical tools and procedures being used in the study do not lead to different remediation decisions regardless of whether outliers are included or excluded, then there is no reason to explore the use of more robust alternatives. If such sensitivity studies do lead to different remediation decisions, then the outliers should remain in the data base and more robust statistical procedures should be used.

RECOMMENDED PRACTICE

1. Use probability plots, scatterplots and data postings to identify outliers.
2. Evaluate each outlier in its spatial context and consider whether the outlier requires any critical assumptions to be modified.
3. If an outlier is due to human error, then correct it if possible. If the correct value cannot be established, then discard the erroneous value and confirm that a similar error has not affected other data.
4. If an outlier is not due to human error, then consider the available qualitative information regarding the data provenance and the site history and discard the outlier only if there is documentation to support the belief that the outlier observation is not part of the population under study.

   In *all* such cases, describe the population that the outlier does belong to and justify why this population is not relevant according to the study objectives.
5. If an outlier is not due to human error and cannot be assigned to a different population based on the available qualitative information, then consider carefully the underlying assumptions about the distribution of the data values; if a re-examination of the available quantitative and qualitative data suggests that the assumed distribution is inappropriate then either choose a more appropriate distribution or adopt a non-parametric statistical approach.
6. If an outlier is not due to human error, and if the assumed distribution is believed to be correct despite the outlier, then two checks should be performed:
   (a) a check to see if the outlier value falls in the upper 1% of the assumed distribution; and
   (b) a gap check to see if the difference between the outlier value and the next highest non-outlier value is more than twice the value that the assumed distribution would predict.

   If an outlier is inconsistent with the assumed distribution for both of these tests, then discard it.
7. If an outlier cannot be discarded for any of the reasons given above, then use it in the statistical analysis and interpretation and, if necessary, choose robust statistical procedures that can produce sensible results even with distributions that have erratic high values.
8. In all cases where an outlier value is discarded, document the reason for this decision and give all relevant information about the sample value that was discarded.
9. In all cases where an outlier value is discarded, a new sample should be taken at a random location within 1 m of the discarded outlier sample.

REFERENCES AND FURTHER READING

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

Barnett, V., and Lewis, T., *Outliers in Statistical Data*, John Wiley & Sons, Chichester, 1984.

Huber, P.J., *Robust Statistics*, John Wiley & Sons, New York, 1981.

Isaaks, E.H., *Risk Qualified Mappings for Hazardous Waste Sites: A Case Study in Distribution Free Geostatistics.* M.Sc., Stanford University, Stanford, California, 1984.

Johnson, N.L. and Kotz, S., *Distributions in Statistics — Continuous Univariate Distributions, Volume 1*, Houghton Mifflin, Boston, 1970.

Sinclair, A.J., "Selection of threshold values in geochemical data using probability graphs," *Journal of Geochemical Exploration*, v. 3, p. 129 – 149, 1974.

Understanding Robust and Exploratory Data Analysis, (Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds.), John Wiley & Sons, New York, 1983.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-9

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

April 2001

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.

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:

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

$v_1, \ldots, v_n$ 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 ug/g, and that we have 25 randomly selected samples whose arithmetic average is 87 ug/g and whose standard deviation is 45 ug/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 $45 \div \sqrt{25} = 9$ ug/g. A pessimistically high estimate of the mean concentration of the stockpile would be $87 + 2 \times 9 = 105$ ug/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 ug/g even though the average of the available samples is only 87 ug/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.

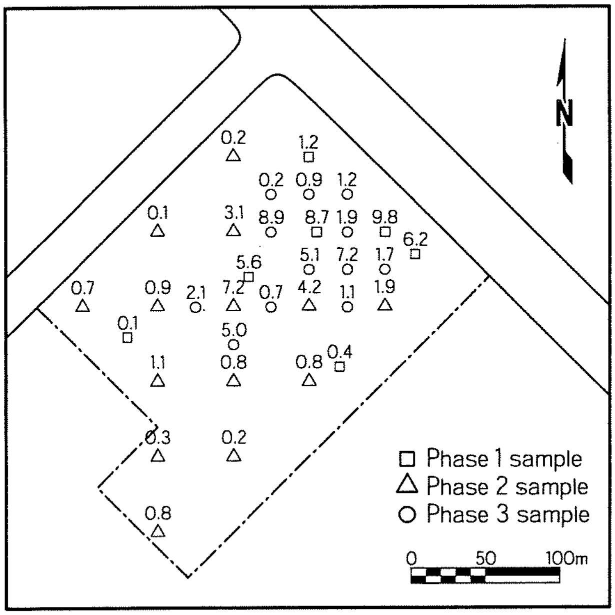


Figure 1 Mercury samples from three sampling campaigns.

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 ug/g. In the second group of 14 samples, the average dropped to 1.59 ug/g. In the third group of 12 samples, the average increased to 3.00 ug/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 ug/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 $v_1, \ldots, v_n$ are the n available sample values, and where $w_1, \ldots, w_n$ 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 50×50 m cells; but why didn't we choose 100×100 m, or 25×25 m, or even 100×50 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 50×50 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.

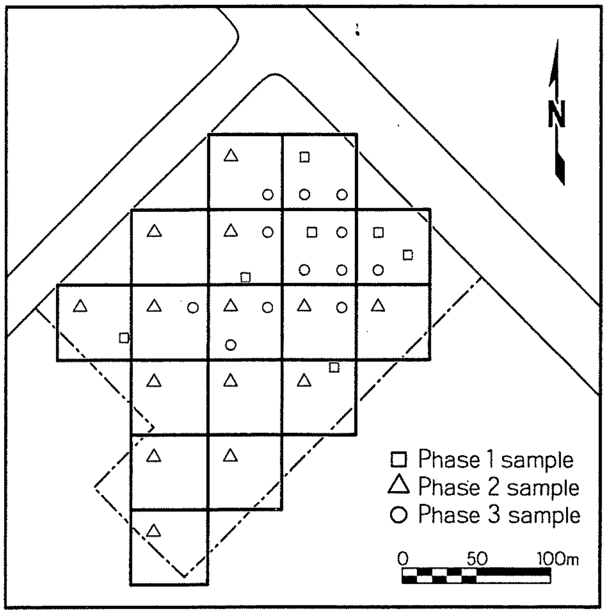


Figure 2 50×50 m cells over the site.

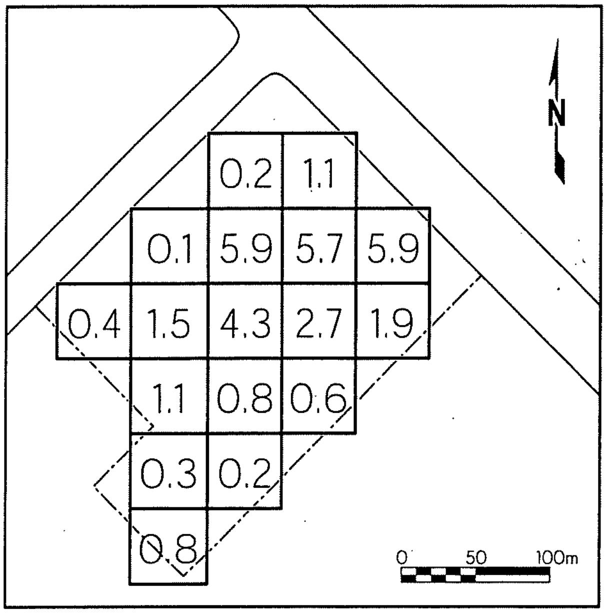


Figure 3 Average Hg concentration in each cell.

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 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.

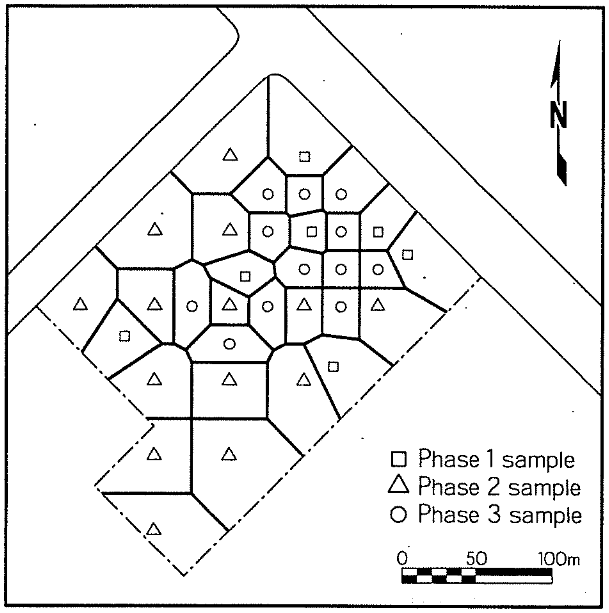


Figure 4 Polygons of influence.

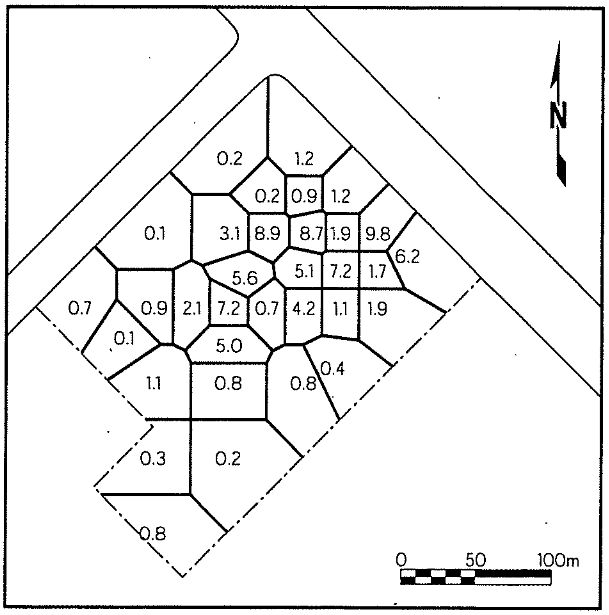


Figure 5 Sample values for polygons of influence.

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 ug/g. The remarkable agreement between this polygonal estimate of the global mean and the cell declustering estimate of 1.97 ug/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.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-10

COMPOSITE SAMPLES

A guide for regulators and project managers
on the use of composite samples

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

Samples are necessary in all phases of a contaminated site study. The money budgeted for analysis can be used more effectively if discrete samples from homogeneous areas are grouped together and combined into composite samples. For example, the separate analysis of five discrete samples is going to be about five times more costly than the analysis of a single composite sample created from the five discrete samples. If the samples are from an area that is thought to be uncontaminated, it is possible that the five separate analyses of the five discrete samples will be virtually identical, with each one confirming what we already suspected — that there is no contamination. The single analysis of a composite sample might be able to confirm the lack of contamination for a much lower cost.

The problem with such an approach is that it may complicate the task of classifying the material. For example, if we are trying to determine whether the arsenic concentration of discrete samples from a particular area ever exceeds 30 ug/g and if a composite composed of five discrete samples has an average concentration of 12 ug/g, then how do we know if all of the five discrete samples would also have been below 30 ug/g had they been analyzed individually? It is possible that all five discrete samples had concentrations of about 12 ug/g, as shown for the first composite in Table 1, and that there is no significant contamination in the area. It is also possible, however, that some of the five samples had arsenic concentrations above the 30 ug/g threshold while others had virtually no arsenic, as shown by the second composite sample in Table 1.

Table 1 Arsenic concentrations (in ug/g).

	Discrete Samples					Composite Average
Composite 1	11	9	12	11	17	12
Composite 2	1	3	35	16	5	12

This document discusses the practice of combining discrete samples into composites and provides some recommendations on when it is appropriate; it also provides two guidelines for assessing whether a composite analysis is compliant, i.e. whether all the discrete values likely would fall below a specified threshold. The first of these is a rather strict guideline that requires only that we know the number of discretes that went into the composite. The second is less strict but requires that we know in advance the variability between discretes within the composite. Though this second approach requires additional analytical work at the outset of the project, it may be more cost effective in the long run.

There is another common usage of the term "composite" that refers to an average of the analyses from contiguous samples, usually from the same well or borehole. With this type of compositing, the individual samples have already been separately analyzed and the goal of the compositing is either to reduce the number of data that need to be handled, to reduce their variability in order to facilitate statistical interpretation or to standardize samples of varying core length to a common length. This other type of compositing is less of a problem than the compositing of discrete samples prior to analysis since the separate analyses are, in fact, still available and, if necessary, can be used in statistical analysis. This guidance document does not address this other type of compositing but focuses instead on the compositing of discrete samples prior to analysis and on the interpretation of the analysis of such a composite sample.

Other guidance documents in this series provide additional information on related issues. In particular, the document entitled *SAMPLING PLANS* discusses the analysis of spatial variability and also discusses issues related to the number of discrete samples that will be needed to achieve a desired confidence in statistical predictions.

IS COMPOSITING APPROPRIATE?

The primary goal of compositing in contaminated site studies is to keep down the cost of analysis by analyzing fewer samples. Unfortunately, for those who are planning the remediation and for those responsible for ensuring that material has been classified correctly, fewer samples means less information. Planners and regulators could be more confident of the success of the remediation if analyses were available for every discrete sample.

The key to appropriate compositing is to ensure that the samples being combined together have similar concentrations, such as those in the first composite in Table 1. When discrete samples that go into a composite have concentrations that differ considerably, such as those in the second composite in Table 1, the analysis of the resulting composite sample is of little value to anyone. Neither remediation planners nor regulators can make much use of it since the individual discrete samples may represent entirely different levels of contamination that would be classified in different regulatory categories. In such an event, it is likely that the planners or regulators (or both) will eventually need to have separate analyses for the individual discrete samples, at which point the whole exercise of compositing has actually ended up costing more than the separate analyses of the discretes would have cost in the first place.

In situ characterization of the site is the best basis for delineating areas within which the material can be composited. The prediction of contaminant concentrations for unsampled material will be more accurate and reliable for *in situ* material than for material that has been disturbed or stockpiled. As long as the material remains *in situ*, models of the spatial distribution of a contaminant, such as contour maps, geostatistical simulation, or the results of flow simulation, will be able to benefit from historical, geological and hydrogeological information. If the material is carefully tracked as it is excavated and stockpiled, then *in situ* characterization will be useful in determining the sample-to-sample variability in the stockpiles. If stockpiled material cannot be traced to its *in situ* location, or if a careful *in situ* characterization was never performed, then the only way to assess the sample-to-sample variability in excavated and stockpiled material is through extensive (and costly) sampling.

The fundamental motivation for compositing is to reduce the money spent on analysis. We get the most for our sampling dollar when the samples we analyze are informative about all of the unsampled material that we could not analyze. We would be foolish to squander the opportunity to analyze and interpret *in situ* samples. An *in situ* sample is much more likely to be representative of its immediate surroundings than is a sample taken from excavated or stockpiled material. With *in situ* samples and *in situ* characterization, we can make more reliable predictions about the areas that will be sufficiently homogeneous to warrant compositing.

If we know that compositing is eventually going to be considered, and that *in situ* variability will become a key issue, we should attempt to document the spatial variation of the *in situ* material. In this series of guidance documents, there is one entitled *SAMPLING PLANS* that discusses methods for describing and documenting the spatial variability of *in situ* material.

Since homogeneity is the key to the technical and economic success of compositing, it is important to check on a regular basis the discrete samples within a composite to ensure that their values do not fluctuate too much. One in every ten composites should be chosen at random to have all of its discrete samples analyzed individually. As long as the information gathered from these regular checks of the within-composite variability continues to confirm that composites are homogeneous, then the compositing of samples can continue. If these regular checks demonstrate that there is much more within-composite variability than was originally assumed, then the compositing should stop and the discrete samples should be analyzed individually. Compositing should not resume until the reasons for the lack of homogeneity are well understood and documented.

HOW TO USE COMPOSITE ANALYSES

If composite homogeneity is not documented

If we are trying to decide whether any of the N individual discrete sample values might be above a specified threshold, T, and if no information exists on the variability of individual discrete sample values within a composite, then the only prudent approach is to compare the analytical value from the composite to T ÷ N. This T ÷ N rule is justified by the fact that if any single discrete value in the composite is larger then T, then the average of N equally-weighted discretes will be larger than T ÷ N. When we observe a composite average that is less than T ÷ N, we can be sure that none of the contributing discrete samples had a value greater than T (as long as each of the discrete samples contributed the same amount of material to the composite sample).

This particular approach is very strict in the sense that it frequently leads to false positive errors — cases in which we incorrectly reject a composite as non-compliant when all of its individual discretes were, in fact, compliant. The first composite shown in Table 1, for example, would have to be treated as non-compliant under this rule. With five discrete samples contributing to this composite, and with the threshold for the arsenic concentration in any individual sample being 30 ug/g, our composite would have to produce an analytical value of 6 ug/g or less before it would be considered as compliant according to the T ÷ N rule.

A further drawback of the T ÷ N rule is that it discourages compositing large numbers of discrete samples regardless of their homogeneity. Once the number of discrete samples in the composite reaches about ten, it becomes virtually impossible to satisfy the T ÷ N requirement. With many contaminants, the thresholds that define contaminated material are low enough that T ÷ N rapidly approaches the detection limit of the best available analytical procedures.

Documenting composite homogeneity

The strictness of the T ÷ N rule stems from the fact that it does not accommodate information about the variability (or lack of it) in the N discrete values that go into each composite. If we have gathered information on the actual variability of discrete sample values within a composite, then we can use this in a less strict rule.

To measure the variability of individual discrete sample values within a composite, we need to compare several discrete sample values to the analytical value of their corresponding composite. If we have N composite samples, each one consisting of M discrete samples, then we can calculate the variance of discrete sample values within the same composite and the corresponding standard deviation:

$$s^2_{\text{within composite i}} = \frac{1}{M}\sum_{j=1}^{M}\left[D_{i,j} - C_i\right]^2$$

$$s^2_{\text{within}} = \frac{1}{N}\sum_{i=1}^{N} s^2_{\text{within composite i}}$$

$$s_{\text{within}} = \sqrt{s^2_{\text{within}}}$$

where $C_1, \ldots, C_N$ are the N composite analyses and $D_{i,j}$ is the analytical value of the j-th discrete sample in the i-th composite.

Tables 2 and 3 show an example of this calculation from six composites, each of which contains five discrete samples. Table 2 shows the 30 discrete analyses and their corresponding composite analysis; note that the composite analysis may not be the same as the mean of the corresponding discrete sample values. Table 3 shows the values of $[D_{i,j} - C_i]^2$ for all 30 discrete samples along with the within-composite variance for each of the five composites. The average within-composite variance for these data is 7.33. The within-composite standard deviation based on these data is therefore 2.71 ug/g.

Table 2 Discrete and composite arsenic values used to calculate the within-composite standard deviation.

	Discrete Sample Analyses					Composite Analysis
Composite 1	12	7	10	12	16	11
Composite 2	10	13	15	12	15	13
Composite 3	22	16	15	16	18	18
Composite 4	7	10	2	5	10	7
Composite 5	17	9	15	12	11	12
Composite 6	8	12	7	13	6	8

Table 3 Values of squared differences for the 30 discrete analyses and their composite analysis in Table 2.

	Squared Difference from Composite Analysis					Within-composite variance
Composite 1	1	16	1	1	25	8.8
Composite 2	9	0	4	1	4	3.6
Composite 3	16	4	9	4	0	6.6
Composite 4	0	9	25	4	9	9.4
Composite 5	25	9	9	0	1	8.8
Composite 6	0	4	1	25	4	6.8
Average within-composite variance = 7.33						

This method requires at least 30 discrete samples in order to produce a good estimate of the within-composite standard deviation. As presented above, the calculation assumes that there are the same number of discrete samples in each composite. If the number of discrete samples varies from composite to composite, then the averaging of the N within-composite variances should be weighted by the number of discrete samples within each composite:

$$s^2_{\text{within}} = \frac{\sum_{i=1}^{N} n_i \cdot s^2_{\text{within composite i}}}{\sum_{i=1}^{N} n_i}$$

where n_i is the number of discrete samples in the i-th composite.

The within-composite standard deviation has several uses. Its first use is that it allows us to quantify the degree of homogeneity of the composites. A common yardstick for deciding that a population is reasonably homogeneous is to check to see if the coefficient of variation (CV) is bigger or smaller than 1. The CV is the ratio of the standard deviation to the mean; a CV less than 1 means that there are few erratic high values in the population. As long as the value of s_{within} remains less than the mean of the composited values, we have statistical support for our assumption that the material is sufficiently homogeneous to warrant compositing. For the data shown in Table 2, their within-composite standard deviation was calculated earlier as 2.71 ug/g; with the average composite value being larger than this, the coefficient of variation is certainly less than 1, and compositing of the discrete samples is acceptable.

If composite homogeneity is documented

The second, and more important, use of s_{within} is that it permits the development of a less strict rule regarding the interpretation of the composite's analytical value. Once we have an accurate estimate of s_{within}, when we are trying to decide if any of the discrete sample values in a composite might have a concentration above a threshold T, then we can compare the analytical value of the composite to the following quantity:

$$\text{Composite compliance threshold} = T - 3s_{\text{within}} \times \left[1 + \frac{1}{\sqrt{N}}\right]$$

The idea behind this rule is that we can be reasonably sure that no single discrete sample exceeds T if the mean of the discrete samples (which is assumed to be the same as the analytical value from the composite) is three standard deviations below T. There is some uncertainty on the mean, however, since we have only a few samples. The $1/\sqrt{N}$ term in the square brackets moves the composite compliance threshold a little bit lower so that even when the fluctuation on the mean is taken into account, we can still be reasonably sure that the threshold T is at least three standard deviations above the population mean.

Apart from the assumptions that the discrete samples contribute the same amount of material to the composite and that the discrete sample values are all uncorrelated with each other, there are no other assumptions hidden in this approach. If we are willing to be a little bolder, and assume that the discrete sample values follow a normal distribution (not a very defensible assumption since contaminant concentrations for discrete samples are usually quite skewed) then we can be more specific about the actual probability that a discrete sample value exceeds T when the composite's analytical value is below the compliance threshold provided by the formula. Under an assumption of normality, this probability is less than 1%. There is no particular need to assume normality, however; even with no assumption about the distribution of the discrete samples, this probability is never more than 10%. Further details on the calculation of these probability values can be found in the guidance documents entitled *DISTRIBUTION MODELS* and *NON-PARAMETRIC METHODS*.

As an example of the use of this formula, consider the situation from Table 1, where we are combining N=5 discrete samples in our composites and we have a threshold of T=30 ug/g for the arsenic concentration in a single sample. If we use the within-composite standard deviation that we calculated earlier, s_{within} = 2.71 ug/g, then the composite compliance threshold for this situation is:

$$\begin{aligned}\text{Composite compliance threshold} &= 30 - 3 \times 2.71 \times \left[1 + \frac{1}{\sqrt{5}}\right] \\ &= 18.2 \text{ ug/g}\end{aligned}$$

If a composite has an analytical value less than 18.2 ug/g it is very unlikely that any of its individual discrete sample values would exceed 30 ug/g. All six of the composites listed in Table 2 would count as compliant samples under this rule; under the T ÷ N rule, all of them would be viewed as non-compliant. From the actual discrete values listed in Table 2, we can see that none of the discrete samples is, in fact, above 30 ug/g. Although the rule based on s_{within} is less stringent than the T ÷ N rule, its false negative rate is still very low.

The key to this approach is the use of the actual within-composite standard deviation from composites whose discretes have also been analyzed as an estimate of the within-composite standard deviation for composites whose discretes have not been individually analyzed. This assumes that the composites from which the standard deviation is borrowed belong to the same population as the composites to which the standard deviation is being applied. We need to make sure that we are not mixing apples and oranges when we use a statistic calculated from one set of data as an estimate of a critical parameter for a different set of data.

We should analyze all of the discrete samples for one in every ten composites and use this information to monitor fluctuations in the statistics of the samples. If the composite mean or standard deviation changes unexpectedly, we should consider whether the within-composite standard deviation based on historical information remains an accurate estimate of the within-composite standard deviation of the composites we are currently creating. As new discrete samples and their corresponding composite analyses become available, we should also use this additional information to continuously update and improve our estimate of the within-composite standard deviation.

RECOMMENDED PRACTICE

1. If compositing is likely to be used on a project, use the available discrete samples to establish the degree of *in situ* spatial variability.
2. Use composite samples only after *in situ* characterization has established that all of the material within a particular area belongs to the same regulatory category. If an *in situ* characterization has not been done, then collect enough samples to document that the material has a coefficient of variation less than 1.
3. When compositing samples:
   (a) maintain a clear record of the samples that contribute to each composite;
   (b) homogenize each discrete sample before drawing the sub-sample that will contribute to the composite;
   (c) ensure that the individual discrete samples each contribute the same amount of material to the composite; and
   (d) archive a sufficient quantity of each sample to permit the discrete samples to be analyzed in the event that the composite is non-compliant.
4. From the first several composite samples, select a group that collectively contain at least 30 discretes, analyze the individual discrete samples as well as the composite sample and calculate the standard deviation of the discrete sample values about their respective composite values.
5. In every group of ten composites, randomly select one and, in addition to the analysis of the composite sample, also perform analyses on the individual discrete samples. Use this information to monitor the homogeneity of the composites and to improve the estimate of the within-composite standard deviation.
6. To decide if it is reasonable to suppose that the N discrete samples within a composite all have a concentration less than the threshold T, compare the composite's analytical value to the following quantity:

$$\text{Composite compliance threshold} = \mathrm{T} - 3 \cdot s_{\text{within}} \cdot \left[1 + \frac{1}{\sqrt{\mathrm{N}}}\right]$$

   where s_{within} is the standard deviation of the discrete samples about their corresponding composite analysis and is based on at least 30 actual analyses of discrete samples and the corresponding composites.
7. If there are less than 30 analyses of discrete samples and the corresponding composites, then compare the the composite's analytical value to the following quantity to decide if it is reasonable to suppose that the N discrete samples within a composite all have a concentration less than the threshold T:

$$\text{Composite compliance threshold} = \frac{\mathrm{T}}{\mathrm{N}}$$

REFERENCES AND FURTHER READING

In addition to the other guidance documents in this series, the following reference provides useful supplementary material on how the EPA views the issue of compositing:

Boomer, B.A., *Verification of PCB Spill Cleanup by Sampling and Analysis*, EPA-560/5-85-026, United States Environmental Protection Agency, 1985.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-11

STATISTICAL QA/QC

A guide for project managers, reviewers, data analysts and interpreters on statistical quality assurance and quality control

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

A statistical study is only as reliable as the data on which it is based; if the fundamental data are called into question, the entire study and its conclusions are also called into question. It is important, therefore, to be able to document how reliable the data are. Issues related to the reliability of data are often grouped under the general heading of "quality assurance and quality control" (QA/QC), a description that captures the idea that data quality can not only be documented but can also be controlled through appropriate practices and procedures.

Even with the most stringent and costly controls, data will never be perfect: errors are inevitable as samples are collected, prepared and analyzed. One goal of QA/QC is to quantify these errors so that subsequent statistical analysis and interpretation can take them into account. A second goal is to monitor the errors so that spurious or biased data can be recognized and, if possible, corrected. A third goal is to provide information that can be used to improve sampling practices and analytical procedures so that the impact of errors can be minimized.

This guidance document begins with a discussion of two concepts: accuracy and precision. It then presents statistical tools that can be used to study the accuracy and precision of existing data, and also presents ideas on practices and procedures that allow accuracy and precision to be monitored as the data are being collected. It closes with a brief discussion of some aspects of QA/QC that are often overlooked: the reliability of location information, the reliability of qualitative information and the reliability of computerized data bases.

ACCURACY AND PRECISION

Statistical QA/QC involves two separate but related concepts: accuracy and precision. Figure 1 captures the difference between these two concepts. A sample is accurate if repeated attempts are centered about the target value; it is precise if repeated attempts are all close to one another.

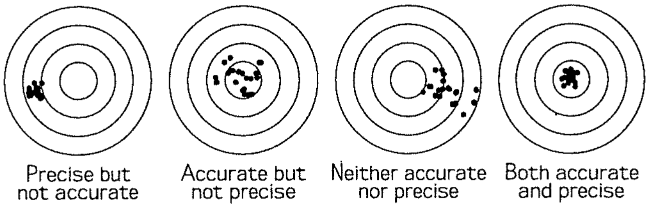
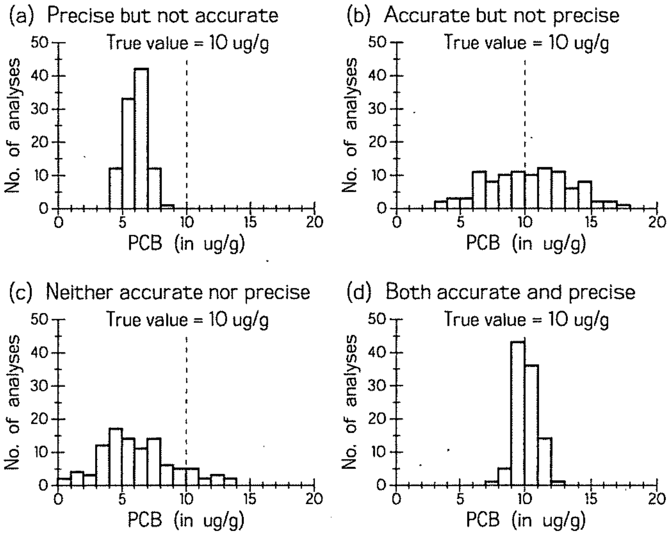


Figure 1 Examples of accuracy and precision.

For the specific case of analytical values, where repeated measurements of the same sample are possible (though somewhat expensive), we can imagine an experiment in which we reanalyze the same material 100 times. Figure 2 shows how accuracy and precision might manifest themselves on a histogram of the repeat analyses; in this example, the target we are aiming for is the true PCB concentration of 10 ug/g.


(a) Precise but not accurate

(b) Accurate but not precise

(c) Neither accurate nor precise

(d) Both accurate and precise

Figure 2 Histograms of 100 repeat analyses.

If the analytical procedure is inaccurate then the average of repeat analyses will be different from the true value; the difference between the average of repeat analyses and the true value is often called the bias. A histogram of repeat analyses from an inaccurate procedure will not be centered about the true value, as in Figures 2a) and 2c). If the analytical procedure is imprecise then repeat analyses will not be close to one another. As precision improves, the spread of the histogram of repeat analyses will decrease; Figures 2a) and 2c) both show repeat analyses that are inaccurate, but those in Figure 2a) are more precise because they show less scatter. Precise analyses are often referred to as "repeatable" because repeated analyses all come close to the same value. As Figure 2a) shows, precise or repeatable analyses are not necessarily accurate and may simply be coming close to the same wrong value.

Though the example in Figure 2 is built on repeat analyses of the same material, it should be noted that sampling errors are not solely due to the analytical procedure. The earlier steps of sample collection and sample preparation often contribute more to the total error than the analytical procedure used in the laboratory. Statistical QA/QC should attempt to document and control the accuracy and precision of each step in the sampling procedure, from the initial extraction of the material from the ground to the final analytical value produced by the lab.

Inaccuracy or bias in a sampling procedure is due to systematic errors that cause the sample values to be generally too high or too low. Imprecision, on the other hand, is a result of random errors that do not have any systematic bias but that cause the sample value to be different from the true value.

MISCLASSIFICATION

Though statistical QA/QC traditionally deals with accuracy, as reflected in the mean value of repeat analyses, and precision, as reflected in the variance of repeat analyses, these may not be the most critical statistical characteristics for remediation planning. For much of the data collected from contaminated sites, their purpose is to determine whether material is contaminated or not and our primary concern should be whether the sample values are above or below some critical threshold. An inaccurate or imprecise sampling procedure may have little consequence if it does not cause contaminated material to be misclassified as uncontaminated or vice versa.

Though improvements in precision and accuracy usually go hand-in-hand with improvements in classification, this is not necessarily the case. Statistical QA/QC for contaminated site studies should not focus exclusively on accuracy and precision but should also address the issue of misclassification.

MONITORING AND CHECKING DATA QUALITY

Control charts for reference standards

An ideal approach to studying the reliability of an analytical procedure is to reanalyze a prepared standard whose true value is known. Such reference material can be specifically prepared for a particular site; for many contaminants, reference material is also available commercially. The advantage of site-specific reference material is that its chemical and physical composition is representative of the particular site; if an analytical procedure is known to be sensitive to factors such as moisture or clay content, then site-specific reference material will provide the best opportunity for calibrating the analytical procedure. The advantage of commercial standards is that their true value has been well established; they have either gone through a battery of different analytical procedures by different laboratories or have been carefully prepared by spiking uncontaminated material with known concentrations of the contaminant.

At regular intervals during the course of a project, the reference material can be included for analysis along with other samples. The resulting repeated analyses of the reference material can then be plotted on a control chart that shows how the analytical values of the reference material fluctuate with time.

Figure 3 shows a control chart for reference material that was prepared for a site contaminated with mercury. With several hundred samples being collected and analyzed at an on-site laboratory over the space of a few weeks, it was decided that the reference material should be checked daily, so the control chart in Figure 3 shows one analytical value per day (except Sundays and some Saturdays). The dashed line on Figure 3 shows the accepted true value for the reference material, which was prepared commercially and was designed to have a true value of 20 ug/g.

In addition to showing the accepted reference value, a control chart should also show the range of acceptable fluctuations around this reference value. The limits for acceptable analytical values can either be established through an initial batch of repeat analyses or can be dictated by remedial design objectives. For the control chart shown in Figure 3, the minimum and maximum acceptable values are shown as dotted lines and are based on design objectives that require the analytical values to be within ±10% of the true value.

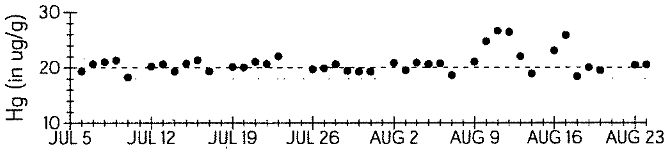


Figure 3 Example of a control chart.

The control chart in Figure 3 shows that the analytical procedure was generally acceptable in terms of its accuracy and precision. For a short period of time towards the end of the sampling exercise, the analyses became somewhat biased and more erratic. In this particular case, it took several days to determine that the cause of these unacceptable errors was operator error but, once identified, these errors were easily corrected. All samples that were initially analyzed during the troublesome period were reanalyzed to provide more reliable analytical values for remediation planning.

Though control charts are excellent for monitoring data quality, they generally focus on the analytical errors that accumulate after a sample has been collected and prepared. The errors that occur in the collection of the original sample material and the preparation of the subsample that is finally analyzed are often much greater than those that occur in the actual analysis of the prepared material. Even though control charts may show acceptable accuracy and precision, a thorough QA/QC program should also investigate errors that occur before the final prepared subsample is delivered to the analytical device.

Scatterplots and summary statistics for paired analyses

Another method for checking the quality of analytical data is to reanalyze several samples and to do a statistical study of the paired analyses using scatterplots and a few summary statistics. The statistical differences between the two sets of analyses will reflect the cumulative effect of all the differences in the way that the two sets of samples were collected, prepared and analyzed. A few examples of some different types of sample pairs will illustrate some of the different combinations of errors that such a study might address:

- The paired values can be one lab's reanalyses of the same prepared material, in which case the statistical comparison will reflect intra-laboratory analytical errors.
- The paired values can be reanalyses performed by different laboratories of the same prepared material, in which case the statistical comparison will reflect inter-laboratory analytical errors between labs.
- The paired values can be reanalyses performed by the same laboratory from different splits of the original sample material, in which case the statistical comparison will reflect

a combination of intra-laboratory analytical errors as well as the errors due to sample preparation.

- The paired values can be analyses performed by the same laboratory of two separate field samples that were very closely spaced, in which case the statistical comparison will reflect a combination of intra-laboratory analytical errors, sample preparation errors, sample collection errors and genuine short scale variations.

There are many ways that different samples, different splits of the same sample, different laboratories and different analytical techniques can be combined to provide pairs of experimental values. The interpretation of the paired values that result from such experiments is always easier if the QA/QC program is designed to isolate as much as possible the different factors that contribute to total sampling error.

	Lab A	Lab B	Linear correlation: 0.99
Mean	56 ug/g	41 ug/g	Rank correlation: 0.81
Standard Deviation	117 ug/g	118 ug/g	Misclassification: 27%

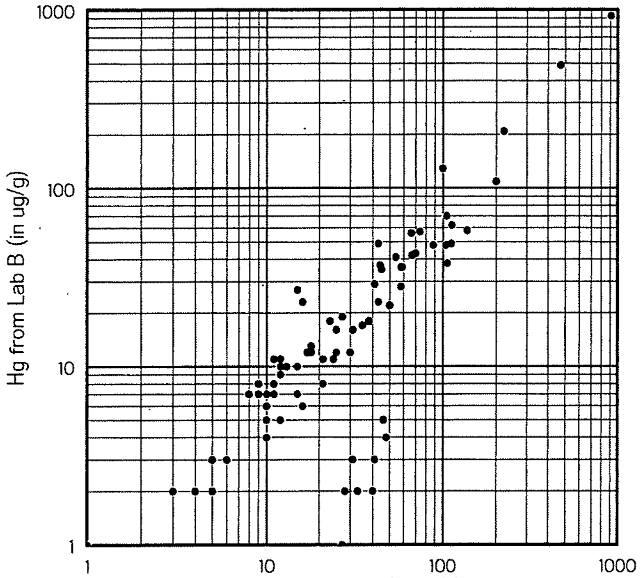


Figure 4 Comparison of analyses from different laboratories.

Figure 4 shows a scatterplot and some summary statistics for the mercury analyses that two laboratories produced for subsamples that were created by splitting the sample material in the field. This particular example reveals a systematic bias; the paired values tend to plot slightly off the main diagonal and the mean of the values reported by the two labs is noticeably different. The scatterplot also reveals that one of the labs may have a problem with inadvertent shifts of the decimal place; there is a handful of points along the bottom of the plot that would be more consistent if the Lab A value was lower by a factor of 10 or the Lab B value was higher by a factor of 10.

The example in Figure 4 also shows the advantage of reporting both the linear and the rank correlation coefficients. The strong skewness of the data makes the linear correlation coefficient quite sensitive to the extreme values; the rank correlation coefficient, a more stable statistic, shows that the high linear correlation coefficient in this example is due largely to the fact that the two labs were in very good agreement for the very highest few pairs of sample values.

Figure 4 reports the percentage of samples for which the labs disagreed on the classification for a remedial action threshold of 20 ug/g. For 21 of the 78 pairs, the two labs disagreed on whether the sample was contaminated; these 21 pairs plot in the shaded regions of the scatterplot. The 19 pairs that plot in the shaded region on the lower right were deemed contaminated by Lab A but not by Lab B; the 2 pairs in the upper left were deemed contaminated by Lab B but not by Lab A.

One of the shortcomings of a statistical analysis of paired observations is the ambiguity about which set of data is more reliable. Though the statistical summary of the mercury data shown in Figure 4 definitely reveals some problems, it is not clear if the problems lie with the analyses from Lab A or those from Lab B (or both). The best way to resolve such ambiguity is with control charts that directly address the accuracy and precision of each set of the paired sample values.

Another shortcoming of a statistical analysis of paired observations is that it may not reveal a systematic bias. Even if paired analyses show a strong agreement, this does not necessarily mean that both sets of values are accurate; it is possible that both sets of values share a common systematic bias.

Blank samples

QA/QC studies of data from contaminated sites need to pay particularly close attention to the possibilities of external contamination and cross-contamination between samples. With some of the contaminants being measured in trace quantities, external contamination can create considerable confusion in remediation planning if the materials used to collect, store and transport the samples are introducing measurable quantities of the contaminants of concern. Cross-contamination between samples can also create difficulties for remediation planning if material from an uncontaminated area becomes contaminated by material from elsewhere on the site.

Material that is known to be free of contamination can be inserted in the sampling procedure to provide experimental evidence of contamination. Such samples are usually called "blanks" and can be used to monitor contamination at various stages in the entire sampling procedure. The design of a program involving blank samples needs to consider all of the possible sources of contamination and all of the pathways for cross-contamination; without appropriate blank samples at each step, it may be difficult to interpret a finding of contamination and to develop an improved procedure that avoids the contamination. For example, trace amounts of chromium can be introduced by a variety of sources. Soil samples could be cross-contaminated by the chromium from refractory bricks if they are stored in the same area; chromium can also be introduced into samples by various metallic instruments. If blank samples prepared in the field reveal measurable increases in the chromium content, we may not know exactly where the trace amounts of chromium are coming from unless we have separate sets of blank samples that allow us to distinguish chromium cross-contamination during storage at the site from external chromium contamination introduced by metallic instruments in the lab.

OTHER ASPECTS OF QA/QC

QA/QC should not focus exclusively on the errors that affect the sample values. Statistical studies often depend on other quantitative information, such as sample location, and often also make use of qualitative information, such as descriptive logs of soil lithology. These other types of information often call for different QA/QC practices and procedures than those used for the sample values; the guiding principles, however, remain the same: we want to know the reliability of the information, we want to detect and correct spurious values and we want to minimize the impact of errors on the conclusions of our study.

Location information

Sample location errors can be minimized through careful surveying practices. Whenever pos sible, a standard reference coordinate system, such as UTM coordinates, should be used to record sample locations. If local coordinates are used, the procedure for converting these to a standard reference system should be documented. This can usually be accomplished by documenting the UTM coordinates of the origin of the local grid as well as any rotation between local north and UTM north.

If the only record of sample locations is a map, the sample locations will become increasingly unreliable as they are transcribed onto other maps. After several generations of copying and remeasuring, the original and correct sample locations are often so poorly known that the sample information is useless for location remediation planning. To prevent such problems, the coordinates of every sample location should be tabulated so that others can refer directly to the exact coordinates rather than trying to pick them off a copy of a map.

Qualitative information

Descriptive information, such as soil lithology, has a large component of subjectivity; the colour and texture that one person uses to describe a soil sample is often not the same as those that another person would use. As soon as it becomes apparent that descriptive information needs to be recorded for a particular site, we should standardize the collection of this information by preparing a form on which descriptive information can be recorded. When s everal people are collecting descriptive information, there will be more coherence between their descriptions if they are all given a standard set of reference materials, such as colour charts or grain size diagrams, that help them to calibrate their subjective visual judgement.

A complete photographic log of the samples is a very useful supplement for descriptive information and can be created using high quality film with relatively inexpensive photographic equipment. Variations in lighting conditions can be monitored and controlled by including a standard colour chart on every photograph. The existence of such a photographic record is often invaluable when old samples need to be relogged for descriptive information that has not yet been recorded because its importance was not initially recognized.

Mergi ng data bases from different sources

In large contaminated site studies, it is common to find that the available data were gathered in different sampling campaigns by different organizations. When data from different sources are merged into a single data base, it is important to maintain a record of the original source for each piece of information. Long after the data have been merged , a statistcal QA/QC study may detect that certain data are less reliable than others. For example, one of the organizations responsible for sampling may choose to use larger boreholes than those used by another organization that has also collected borehole samples; since the size of the sample may affect the reliability of subsequent analyses, it may eventually be important to be able to distinguish the information generated by one organization from that generated by another. Similar concerns arise with location information when different organizations use different surveying practic es, and with qualitative information when different organizations have different levels of expertise in recognizing and describing geological and geotechnical properties of the soil.

RECOMMENDED PRACTICE

1. A statistical study of contaminated site data should be accompanied by documentation of the reliability of any data that are critical to the study's conclusions.
2. The entire sampling procedure, including the collection, preparation and analysis of the sample, should not impart any systematic bias. For large studies in which more than 100 samples will be collected and analyzed, control charts should be used to monitor and control the accuracy and precision of the analyses. A t-test should be used to determine whether the average of repeat analyses is significantly different from the established reference value.
3. Sample precision should be monitored through control charts and through paired analyses of separate splits of the same sample material. For sample material that is split in the field, the paired analyses of the separate sample measurements should show a rank and linear correlation of 0.95 or greater for metallic and inorganic contaminants, and 0.90 or greater for organic contaminants.
4. When ever QA/QC reveals a significant systematic bias or an unacceptably high imprecision, specific corrective action should taken and the results documented.

REFERENCES AND FURTHER READING

The guidance document entitled *UNIVARIATE DESCRIPTION* provides information on the summary statistics used in this document and also presents a more detailed discussion of data base compilation and verification. *BIVARIATE DESCRIPTION* provides information on scatterplots and their summary statistics. In addition to the other guidance documents in this series, the following references provide useful supplementary material.

Cochran, W.G., *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York, 1977.

Merks, J., *Sampling and Weighing of Bulk Solids*, Series on Bulk Materials Handling, Volume 4, Trans Tech Publications, Clausthal, Germany, 1985.

Pitard, F., *Pierre Gy's Sampling Theory and Sampling Practice*, Volumes I and II, CRC Press, Boca Raton, Florida, 1989.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-12

SAMPLING PLANS

A guide for data analysts, project managers and reviewers on the design of sampling plans

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

The design of an appropriate sampling plan is a recurring concern throughout the course of a contaminated site project. The issue first arises when a site is suspected of being contaminated and preliminary reconnaissance is required. At this early stage, some thought needs to be given to samples that will provide useful information for more detailed studies that may follow. If the site is deemed to be contaminated, the sampling plan will frequently resurface as an important issue: when the site needs to be characterized so that a remediation strategy can be developed, when global estimates of contaminated volumes are needed for remediation planning, when "hot spots" are encountered during remediation and need to be delineated. At any of these stages, a poorly designed sampling plan can cause the remediation to be inefficient or, worse, ineffective. Sampling plan issues arise even after the remediation, when samples are needed to confirm whether or not the remediation was successful. A poorly designed sampling plan at this final stage can cause residual contamination to go undetected.

This document presents information and advice on designing an appropriate sampling plan. It begins with a discussion of what it means for a sample to be fair and then discusses the volume that a single sample can be assumed to represent. The remaining sections discuss sampling plans according to the goal of the study, and deal with statistical considerations that arise in designing sampling plans for initial reconnaissance, for global estimates, for local estimates and for detection of "hot spots".

All of the guidance given in this document, including the various formulas for calculating the number of samples and their spacing, assume that the the sample information is accurate and precise. In addition to defining the number, location and spacing of samples, a sample design should also identify QA/QC procedures that will ensure the reliability and repeatability of the sample information. These QA/QC procedures should not focus solely on the analytical values, but should also address the quality of the sample location information. Another document in this series, *STATISTICAL QA/QC*, discusses the documentation and monitoring of the reliability of sample information.

HOW FAIR ARE THE SAMPLES?

When we are designing a sampling plan, it often helps to think of the sampling plan in terms of an exercise in democracy in which we have to select samples that fairly represent some larger population. The samples themselves should not be the main focus of attention; they are interesting only insofar as they provide insight into the much larger remainder of the population that has not been directly sampled.

A chronic concern of statisticians is that sampling be fair in the sense that every member of the larger population has the same chance of being selected. If sampling is not fair then some members of the population have a greater chance of being selected than others; statistical studies based on such biased samples often reach erroneous conclusions. One of the best historical examples of an unfair sampling is the poll that was conducted by the newspaper that earned a lasting place in the photographic record of the 20th century with its "Dewey Defeats Truman" banner headline. This embarassing proclamation was based on a public opinion poll in which voters were randomly selected from the phone book. Though this method of sampling a population has now become conventional for many public opinion polls, it was not a fair method in the 1940's. At that time, owning a telephone was enough of a sign of affluence that wealthier voters had a slightly higher chance of being sampled than poorer voters. With the poll being inadvertently stacked with wealthier voters, the preference of the wealthy for Dewey, the Republican candidate, skewed the results enough that the pollsters reached a very wrong conclusion.

For many contaminated sites, the earliest samples are not fair since their selection is based on visual observation. A preliminary reconnaissance is more likely to sample material that looks "interesting" than material that is not visually distinctive. Though such samples may be very useful for establishing an understanding of the nature of the contamination, they usually impart awkward biases to statistical studies that aim at characterizing the entire site. For example, on a landfill site contaminated with glazing sludge from the manufacturing of ceramic products, visible layers of glazing sludge exposed in trenches may be preferentially sampled during a preliminary reconnaissance of the site since these layers are most likely to confirm the severity of lead, zinc and cadmium contamination. Later, when a remediation plan is being prepared for the entire site, which includes many landfill materials other than glazing sludge, the preponderance of highly contaminated samples from preliminary reconnaissance makes it difficult to develop an accurate 3D model of the contamination throughout the site. In this case, though the early samples may fairly represent the contamination within layers of glazing sludge, they do not likely fairly represent contamination in other layers.

Another example of unfair sampling is the siting of additional samples near anomalously high sample values from earlier sampling campaigns. As discussed later in this guidance document, this targetting of suspected "hot spots" does provide valuable information; nevertheless, it also compromises any statistical method that assumes the underlying population has been fairly sampled. When additional samples have been preferentially located in areas that are suspected of being highly contaminated, sample statistics should not be used as estimates of the corresponding parameters of the underlying population. The sample mean, for example, is usually a biased estimate of the mean of the underlying population if samples are preferentially clustered.

Dealing with preferential sampling

Samples are often collected from contaminated sites before anyone has started thinking about statistical issues, and those who are later responsible for statistical data analysis and interpretation typically have to cope with an unfair sampling that is preferentially biased towards certain regions. In such situations, statistical studies that address the remediation of the entire site should find some appropriate method for mitigating the effect of preferential sampling.

One solution for dealing with samples that preferentially target certain regions is to separate the site into subpopulations. The case described in the previous section, for example, might best be handled by building a model of the locations of the layers of glazing sludge. The many samples from these glazing sludge layers can be used to model the spatial distribution of contaminants within these layers, while the spatial distribution in other layers can be modelled using any samples from layers other than those that were visually recognizable as glazing sludge.

It is not always possible to split a statistical study into separate subpopulations, especially when the preferential sampling is the result of successive infill samples in anomalously high areas rather than the result of an intentional preference for visually distinct material. In such situations, statistical procedures may have to accommodate the effects of preferential sampling by assigning each sample a declustering weight. Samples from regions that have been densely sampled, are given low declustering weights to reduce their influence, while those from sparsely sampled regions are given higher declustering weights. The guidance document entitled *ESTIMATING A GLOBAL MEAN* discusses declustering and shows how declustering weights can be used to develop more accurate estimates of the mean of the underlying population from spatially clustered samples.

The least desirable way of dealing with preferential samples is to ignore or discard them; it is always better to try to limit their spatial influence, either by delineating the spatial extent of the population to which they belong, or by assigning them a declustering weight based on their proximity to neighbouring samples. Though discarding samples is not a good final solution to the problem of preferential sampling, it is often an expedient way to check the sensitivity of a statistical procedure to clustered sampling. If there is a concern that preferential sampling may be leading to erroneous conclusions, the procedure can be repeated with a regulary spaced subset of the available samples. If the conclusions based on all available samples are different from those based on a regularly spaced subset, then the effect of spatial clustering is severe enough to warrant specific attention. A subset that is more regulary spaced than the entire data set can be selected by overlaying a rectangular grid over the site and randomly choosing one sample from each rectangular cell; the document entitled *RANDOMIZATION* provides guidance on procedures for such a random subsampling.

WHAT VOLUME DOES A SAMPLE REPRESENT?

When samples are needed for local remediation planning, the sampling plan should consider the "range of influence" of a single sample. For spatially erratic contaminants, the range of influence is short and an individual sample represents only a small region in its immediately vicinity. For contaminants that are very spatially continuous, the range of influence is longer and an individual sample is representative of a larger region. If the range of influence is short, detailed local remediation planning will require more closely spaced samples than if the range of influence is long. The design of an appropriate sampling plan for detailed local remediation planning therefore requires a good understanding of spatial variation.

The intuitive notion of the "range of influence" can be expressed quantitatively by the "range of correlation", which is the distance at which pairs of sample values are no longer correlated. The range of correlation is usually determined by grouping pairs of samples according to their separation distance and plotting the correlation coefficient between the sample pairs as a function of the separation distance. Figure 1 shows such a plot, which is usually called a "correlogram", for a PCB contaminated site. In this example, the PCB concentrations have a range of correlation of 50m in the N–S direction and 30m in the E–W direction. Up to this distance, a single PCB analysis will have some correlation with the unsampled PCB concentrations in its immediate vicinity. Isaaks and Srivastava (1989) present a practical introduction to a variety of tools for analyzing and interpreting the pattern of spatial variation.

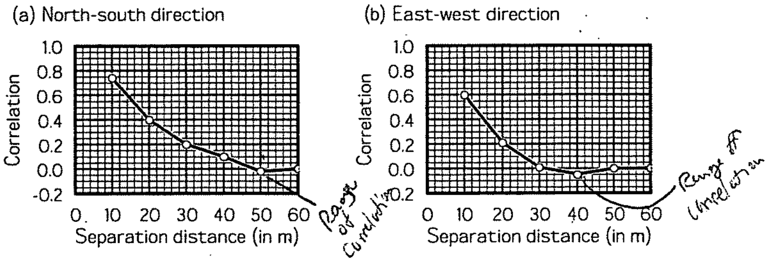


Figure 1 N–S and E–W correlograms of PCB concentration.

The example in Figure 1 shows one of the specific practical benefits of statistically analyzing the spatial variation: it may reveal that the sample spacing does not need to be the same in all directions. For many contaminated sites, the physical and chemical processes that influence the spatial distribution of the contamination are not isotropic — the contamination is likely to be more continuous in certain directions than in others. Without a study of spatial variation in different directions, we are not likely to recognize the opportunity for an efficient sample plan that takes advantage of the directional changes in the spatial continuity of the contamination.

Like all statistical characteristics, the correlogram may change locally; the range of correlation may not be the same throughout the site. On any site where erratic "hot spots" are suspected, we can check the possibility that high values have a shorter range of influence by calculating separate correlograms for high sample values and for low ones. Such information will be very useful for planning an efficient strategy to deal with any "hot spots" that may be encountered during remediation.

STATISTICAL CONSIDERATIONS IN SAMPLING

Preliminary reconnaissance samples

When a site is first suspected of being contaminated, reconnaissance is often performed to provide information on site history and usage. It is during this initial sampling that statistical rigour most frequently collides with other practical issues. Many statisticians disdain the haphazard approach to sampling that often characterizes reconnaissance, dismissing the samples as "grab samples" or mere "specimens" whose selection bias makes them of questionable value for statistical studies.

Unfortunately, preliminary site reconnaissance needs to address issues that are pre-statistical in the sense that statistical studies make little sense until these issues are resolved. For example, we should not initiate a statistical study until we know what we need to study; even without a statistically defensible sampling plan, preliminary reconnaissance can establish which contaminants exist in sufficiently high concentrations to warrant remedial action. Even when the contaminants of concern are already well known, a statistical study may be premature if the relevant populations are poorly understood. Reconnaissance can provide critical insights into the relevant statistical populations by gathering information on the physical and chemical factors the affect the spatial distribution of each contaminant.

Site reconnaissance need not be constrained by statistical concerns but should focus instead on establishing the contaminants of concern and their physical and chemical controls. For contaminants that occur naturally, such as most metals, the initial sampling of the site should also attend to the issue of establishing the statistical characteristics of background concentrations. By identifying and sampling material that should be uncontaminated, a site reconnaissance can document the distribution of background concentrations, information that can be used in subsequent statistical studies that try to split the data into "natural" and "affected" subpopulations.

Even though statistical issues need not be paramount during preliminary reconnaissance, subsequent statistical data analysis and interpretation will find the reconnaissance samples more useable if each sample is accompanied by documentation of the reasons for its collection. This may be a description of the visual appearance of the material, or may simply be a summary of the information that led to the belief that a particular location was contaminated. Reconnaissance samples will also be more useable in later studies if their locations are accurately recorded. If it is not possible at the time of the reconnaissance to survey the sample locations, they should be flagged or marked in the field so that they can be properly surveyed later.

Samples for global estimates

Once contamination on a site has been confirmed, the next statistical issues that arise usually pertain to global estimates, such as the average concentration of a contaminant over the site, or the overall proportion of the site that is contaminated. As discussed in *ESTIMATING A GLOBAL MEAN*, the uncertainty on the estimate of the mean of the underlying population is related to the number of samples used in the estimate. Uncertainty in an estimate is usually expressed in terms of its relative error, as a plus/minus interval around the estimate. The number of independent samples needed to ensure that the relative standard deviation on the estimate of the global mean is less than ±R% can be expressed in terms of the coefficient of variation of the individual samples:

$$\text{Number of samples} = [100 \times \text{CV} \div \text{R}]^2$$

As an example, if the available samples from a contaminated site show that the CV of individual lead samples is 1.5, and if we need a global estimate of the average lead concentration over the entire site that has a relative standard deviation of ±20% or less, then the number of independent samples we require is:

$$\text{Number of samples} = [100 \times 1.5 \div 20]^2 \approx 56$$

This calculation assumes the samples to be independent; this is rarely the case in contaminated site studies since there is usually some spatial correlation in the contaminant concentrations. In practice, the best we can do for a sampling grid that is intended for global estimates is to use a regular grid whose origin has been randomly selected.

Samples for local estimates

In addition to global estimates, statistical studies of contaminated sites also often call for local predictions. For example, remediation planning may require a map that shows the boundary between contaminated material that requires remediation and uncontaminated material that may be left *in situ*. Such a map is often constructed by contouring the available samples and then designating the remediation limit to be the contour line that corresponds to the remediation threshold.

Sampling plans for local estimation should take into account the range of correlation of the contaminant concentrations. For every location at which we need a local estimate, we should have at least one sample within the range of correlation. For any location where no samples fall within the range of correlation, local estimation will be fruitless since we do not have any information that directly correlates (even weakly) with the unsampled (and unknown) concentration we are trying to predict.

Once the range of correlation has been studied and documented through correlograms like the ones shown in Figure 1, this information can be used to design a sampling grid that will be appropriate for local estimation. In the same way that the formula given in the previous section allowed us to choose the number of samples that would limit the relative error in global estimates, the following formula provides insight into the sample spacing needed to limit the relative error in point estimates:

$$\text{Sample spacing} = \text{Range of correlation} \times [\text{R} \div (100 \times \text{CV})]^2$$

As with the corresponding equation given earlier, R is the relative error, expressed in percent and CV is the coefficient of variation of the individual sample values. The range of correlation is the distance at which the correlogram shows sample pairs to be uncorrelated; as noted earlier, this distance may change with direction. As an example of the use of this formula, suppose that we are dealing with lead contamination that has a CV of 1.2 and a range of correlation of 135 m; furthermore, we would like each of our point estimates to have a relative error of less than ±40%. The sample spacing should be

$$\text{Sample spacing} = 135 \times [40 \div (100 \times 1.2)]^2 \approx 12\text{m}$$

Though this formula provides a sample spacing that limits the relative error for point estimates, it is rare that we depend on point estimates in remediation planning. In practice, we do not segregate contaminated from uncontaminated material point by point; the equipment used to implement the remediation strategy limits the volume of material we can effectively segregate. We are not interested in whether the contaminant concentration at a specific point exceeds the remedial action threshold; instead, we usually need to know whether the *average* concentration over a small area is above the threshold. An estimate of the concentration at a single point is less certain than an estimate of the average concentration over some larger area. For this reason, the preceding formula gives a sample spacing that is usually smaller than we actually need in practice.

The following formula provides a sample spacing that should limit the relative error on estimates of the average concentration over a small square whose side is B:

$$\text{Sample spacing} = \text{Range of correlation} \times [\text{R} \div (100 \times \text{CV})]^2 + 0.75 \times \text{B}$$

As an example of the use of this formula, let us reconsider the previous problem, in which lead contamination had a CV of 1.2 and a range of correlation of 135 m. As before, we would like the relative error to be below ±40% but this relative error now refers to estimates of the average lead concentration over a 10m×10m square that represents the smallest amount of soil that can be practically segregated as clean or contaminated. For this situation, an appropriate sample spacing would be

$$\text{Sample spacing} = 135 \times [40 \div (100 \times 1.2)]^2 + 0.75 \times 10 \approx 20\text{m}$$

This formula will give an appropriate sample spacing for sample design problems in which the sample spacing is smaller than the range of correlation but larger than the size of the minimum volume of selective remediation. If the spacing calculated by this formula is larger than the range of correlation or smaller than B, then the result may not be appropriate; Isaaks and Srivastava (1989) present a more detailed discussion of estimation error and provide more general formulas that can be used to assist with the selection of an appropriate sample spacing.

Sampling for "hot spots"

Attempts to design a single sampling grid to delineate "hot spots" usually lead to a sample spacing that is so tight that the total number of samples becomes prohibitively costly and time consuming. The range of correlation is rarely constant throughout a contaminated site, but tends instead to be longer in areas with moderate and low contaminant concentrations and shorter in areas with high concentrations; similarly, the coefficient of variation is rarely constant throughout a contaminated site. As a result, the sample spacing needed to achieve a specified level of confidence in local estimates will vary throughout the site, with more samples usually being needed in anomalously high areas and fewer samples in moderate and low areas. Unfortunately, we usually can't take advantage of this fact in practice because we do not know where the anomalously high areas are located until we collect and analyze the samples. A practical and effective way around this problem is to design the sampling program so that the samples are collected in several stages rather than in a single campaign. With a multi-stage approach to sampling, the initial stage provides sample information on a relatively coarse grid, and each successive stage adds infill or step-out samples near the locations of the high sample values from earlier stages.

Another approach to identifying and delineating "hot spots" is to use a less costly and more rapid analytical procedure, such as X-ray fluorescence, to supplement the more costly and time consuming chemical analyses. Rapid and cost-effective analytical procedures are typically less reliable and can not be used without a careful QA/QC program that constantly monitors the reliability of the information they generate. If a rapid and cost-effective analytical procedure has been properly calibrated through statistical QA/QC, it can provide a dense sampling of the site that can then serve as the basis for selecting optimal locations for more reliable (and costly) chemical samples.

RECOMMENDED PRACTICE

1. Initial reconnaissance sampling may be based on visual inspection and should be designed to provide data on the contaminants that exist and their maximum concentrations. For any contaminant that requires remedial action, the initial samples should document its physical and chemical controls and, if the contaminant occurs naturally, should also serve to establish the distribution of background concentrations. Though the initial sampling need not be statistically based, a complete description of the location of each sample and the rationale for its collection should be compiled so that subsequent statistical analysis can make appropriate use of the sample information.
2. For sampling plans that aim to provide information for global estimates, the samples should fairly represent the underlying population; this can be accomplished with a regular grid whose origin has been randomized. The total number of samples should take into consideration the level of confidence that global estimates will need to have to meet the study objectives and this, in turn, requires that the coefficient of variation be taken into account.
3. For sampling plans that aim to provide information for local estimates, their design should take spatial variation into account by ensuring that the spacing between samples is smaller than the range of correlation. If closely spaced samples are not already available from earlier sampling campaigns, they should be added at this stage to provide data for quantifying spatial variation.
4. For sampling plans that are intended to detect "hot spots" or to check for residual contamination following a remediation exercise, a multi-stage sampling plan should used.

REFERENCES AND FURTHER READING

The guidance documents entitled *ESTIMATING A GLOBAL MEAN RANDOMIZATION* and *STATISTICAL QA/QC* provide more information on topics related to the sampling plan issues discussed in this document. In addition to these, the following references also provide useful supplementary material.

Cochran, W.G., *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York, 1977.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-13

CLASSIFICATION

A guide for data analysts, project managers and reviewers on the classification of contaminated soil

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

During the remediation of most contaminated sites, the affected material throughout the site eventually has to be classified into one of several contamination categories, usually according to whether the contaminant concentrations are above or below specified regulatory thresholds. In the system of classification used by the BC Ministry of the Environment, for example, the most contaminated material is assigned to the "special waste" and "waste" categories, and less contaminated material is classified from industrial quality to agricultural quality. The criteria for these categories are defined by the *BC Waste Management Act, and the Contaminated Sites* and *Special Waste Regulations*. When it comes time to assign contaminated material to one of these categories, we very rarely know the exact contaminant concentration of the material in question. Instead, we base our classification on samples that represent only a tiny fraction of the total amount of material that needs to be classified.

This guidance document discusses a variety of topics that all relate to the issue of classification. It begins with a brief presentation of some of the terminology that is commonly used when discussing misclassification, and moves on to a discussion of contouring, with a brief summary of the techniques that are commonly used to interpolate between the available sample data. It addresses the issue of quantifying the uncertainty on local estimates and presents a procedure for developing maps that directly show the probability of encountering contamination rather than showing estimates of the contaminant concentration. The volume–variance effect is then discussed, along with the related issue of selectivity and the concept of the volume of selective remediation.

This document focuses on the issue of classification of *in situ* material based on *in situ* samples. The document entitled *STOCKPILES* discusses classification of material that has already been excavated and is awaiting classification in stockpiles. For some small sites, the affected material does not need to be locally segregated into different categories. For such situations, where all the material on the site is assigned to a single category, the reader should consult the documents entitled *STOCKPILES* and *ESTIMATING A GLOBAL MEAN*.

MISCLASSIFICATION

Whenever we attempt to classify material as being above or below some specified threshold, there are two kinds of errors that may occur. The first, which is often called a "false negative error", occurs when we mistakenly classify material that is actually above the threshold as being below the threshold. The second, which is often called a "false positive error", occurs when we mistakenly classify material that is actually below threshold as being above the threshold.

For the remediation of contaminated sites, where classification is usually accomplished by comparing bcal estimates of the contaminant concentration to some regulatory threshold, a false positive error results in the remediation of material that did not, in fact, have to be remediated; a false negative error results in the failure to remediate material that should, in fact, have been remediated. These two types of errors have different impacts. False positive errors cost us the money that it requires to excavate and treat soil that could have been left unexcavated and untreated; residual contamination that results from false negative errors has the potential for damaging human health and the environment.

Though remediation should, ideally, minimize both types of misclassification, it is usually difficult to minimize both simultaneously. Decreases in the probability of a false negative error usually entail increases in the probability of a false positive error. Damage to human health and the environment is usually regarded as a much higher cost than the money spent on additional remediation. As a result, the focus of most statistical guidance on classification is to keep the false negative rate below some acceptable minimum.

CONTOURING

The most common approach to classification is to use the available sample data to contour or interpolate the contaminant concentrations into areas that have not been directly sampled. Figure 1 shows lead data from a site affected by airborne contamination from a lead smelter located roughly in the center of the map area. Using these data, a contour map, such as the one shown in Figure 2 can be constructed and used as the basis for classification. For this particular site, the target threshold for remediation was 500 ug/g, the first of the thicker contour lines shown in Figure 2.

Contouring is not a unique exercise: there are many different algorithms that can interpolate sample data into unsampled areas. The contour map shown in Figure 2 was created using a geostatistical procedure known as "kriging", which uses statistical information on the pattern of spatial continuity to calculate appropriate weights for the nearby samples in the vicinity of any point at which we need a local estimate. If there are enough samples on which to base an analysis of spatial continuity, as described in *SAMPLING PLANS*, then kriging will usually produce an excellent interpolation of the available data.

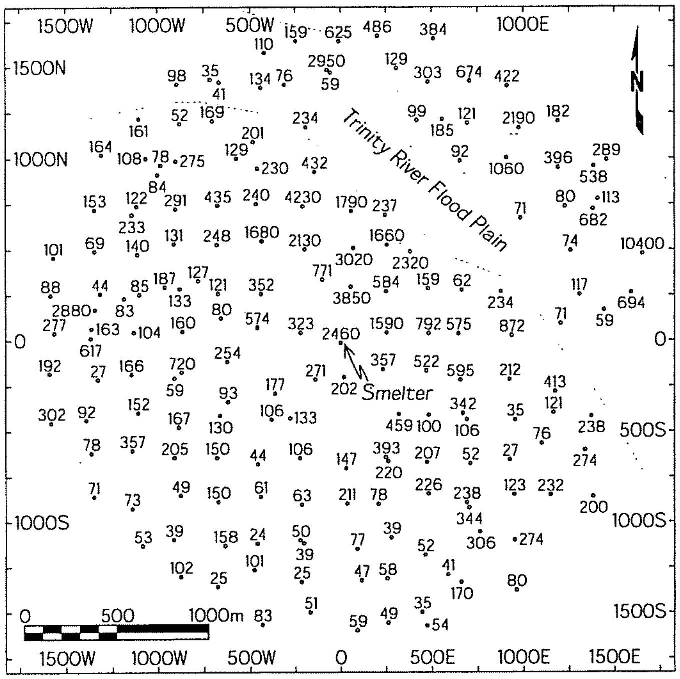


Figure 1 Lead sample data in the vicinity of a smelter.

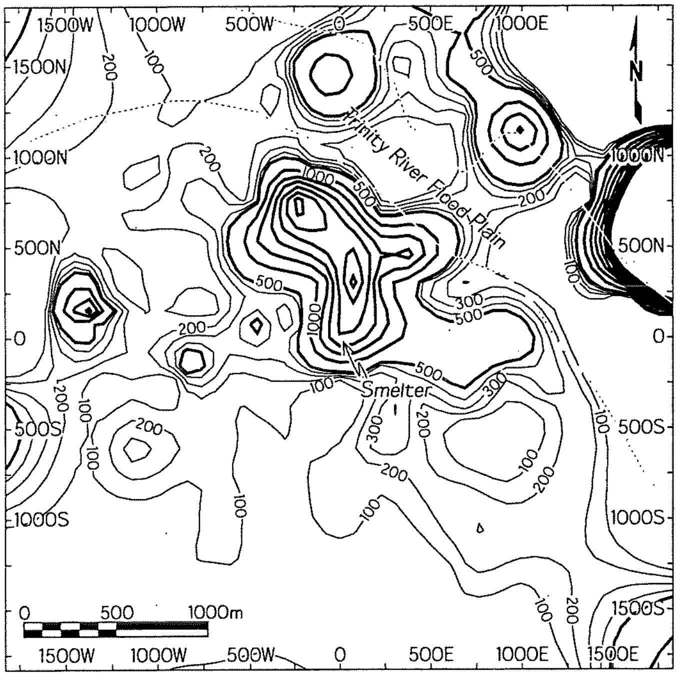


Figure 2 A kriged contour map of the lead data in Figure 1.

Other contour methods that are widely available in commercial software packages include the "inverse squared distance" method and "spline" or "minimum tension" interpolation. With an inverse distance procedure, each sample is given a weight that is inversely proportional to the squared distance from that sample location to the location of the point at which we need a local estimate. The "spline" or "minimum tension" approach to interpolation finds a surface that passes through all of the available sample data values and that has variations that are as smooth and gentle as possible.

Isaaks and Srivastava (1989) provide a practical introduction to kriging and to some of the more traditional and simpler alternatives, such as the inverse squared distance approach.

QUANTIFYING UNCERTAINTY

Whenever we make an estimate, regardless of the interpolation procedure we choose, one of the few things that we know about our estimate is that it is very likely wrong — we would be unimaginably lucky to hit the nail right on the head and predict the exact value of the unknown concentration. Since error is inevitable in any estimation procedure, our classification should not be based merely on estimated concentrations, but should also try to take into account the uncertainty on these estimates. In the same way that we consider a pessimistically high estimate of the global mean when we classify stockpiled material (see the document entitled *STOCKPILES*), so too should we consider the effect of uncertainty on any local estimates that we calculate.

Geostatistics provides one approach to incorporating estimation uncertainty into a classification procedure. The geostatistical interpolation known as kriging produces, along with an estimate of the concentration, a quantity called the "estimation variance" that is often used to build confidence intervals around local estimates. Unfortunately, this procedure assumes that the distribution of error is normal, an assumption that is rarely justified when the underlying population is skewed (as are the contaminant concentrations for most contaminated sites). Furthermore, the traditional calculation of the estimation variance does not take into account the sample data values themselves, but considers only their location. In many instances, the uncertainty in our predictions is linked as much to the actual data values as to their location, and the fact that the estimation variance does not depend on the data values makes it inappropriate as a measure of local uncertainty in such situations.

Probability maps

Geostatistics also offers an alternate method for quantifying uncertainty, a procedure known as "indicator kriging" that results in a map that directly displays the probability that the contaminant concentration exceeds a specified threshold.

Figure 3 shows the 50 ug/g "indicators" of the lead data shown earlier in Figure 1. These indicators are simply 0's and 1's that record whether each individual sample is above or below the 500 ug/g threshold. At any location where the lead concentration is known to be below 500, the corresponding indicator is 0, and at any location where the lead concentration is known to be above 500, the corresponding indicator is 1.

Figure 4 shows a kriged contour map of these indicators. By interpolating between a set of 0's and 1's, we end up with a map of intermediate values between 0 and 1; this contour map can be interpreted directly as a probability map. For example, in the region immediately north of the smelter, where the contour lines show values above 0.8, we can interpret this as a 80% probability that the contamination in this area is above the 500 ug/g threshold. Similarly, in the southern third of the map area, where the interpolated values drop below 0.1, there is less than a 10% chance of encountering lead contamination in excess of 500 ug/g.

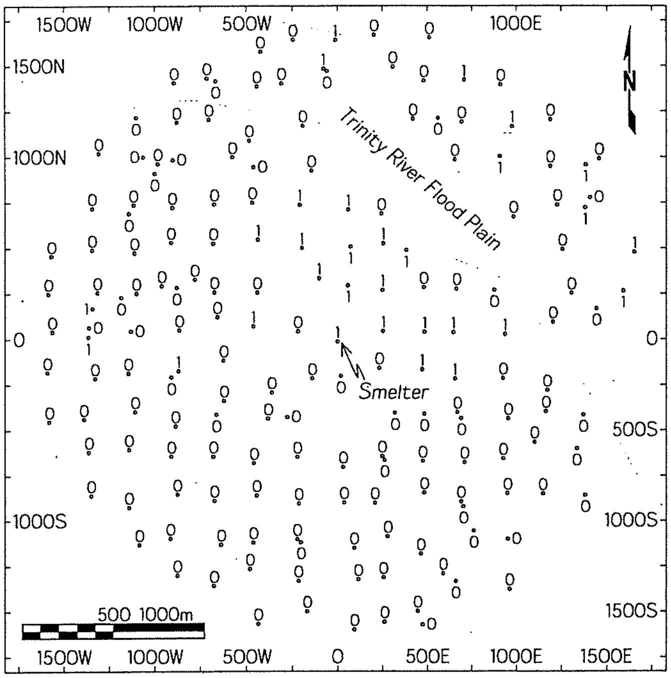


Figure 3 500 ug/g indicators for the sample data in Figure 1.

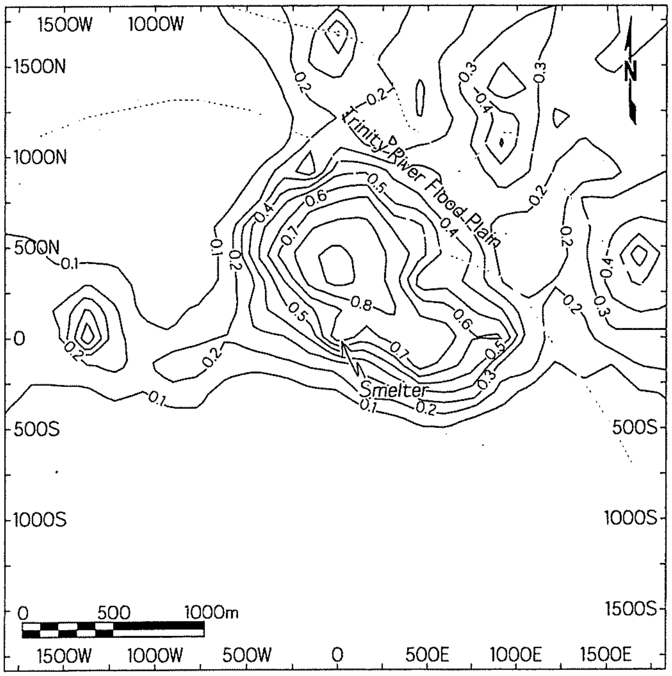


Figure 4 A probability map based on indicator data in Figure 3.

The reason that interpolated indicators can be viewed as probabilities is that individual indicators are themselves probabilities. In Figure 3, those locations where the indicator is 0 are locations where the chance of encountering lead in excess of 500 ug/g is 0. Similarly, those locations where the indicator is 1 are locations where the probability of encountering contamination in excess of 500 ug/g is 1, or certainty. With the indicator data being local probabilities, an interpolation of these data leads to an estimation of the probability of encountering contamination in excess of the threshold used to define the indicators.

Though the use of indicators was first presented in a geostatistical setting, with kriging as the interpolation procedure, the same approach can also be used with the more traditional and simpler interpolation procedures such as inverse squared distance or splines.

THE VOLUME–VARIANCE EFFECT

A histogram of contaminant concentrations for very small volumes, such as those typically collected by a split-spoon sampling device, will show a greater spread and more skewness than histograms of the same contaminant concentration measured on larger volumes, such as truckloads. The tendency for values based on larger volumes to be less variable is often called the "volume-variance" effect or the "support" effect ("support" being the word that statisticians often use for the size and geometry of a sample). This support effect is due to the fact that we are less likely to observe extreme high or low values in larger volumes of material since there is more opportunity for mixing high and low values together. The gradual disappearance of the extreme values causes the spread of the distribution to decrease and the symmetry of the distribution to increase.

The support effect has important implications for making predictions about remediable volumes. The problem in most contaminated site studies is that the remediable volume is quite different from the volume of the available samples. With the available samples coming from a population that is based on small volumes and is highly variable, and the remediation plan calling for estimates for a different population that is based on much larger volumes that should be much less variable, we have to be careful about how we use the sample information to make predictions. Parker (1979) and Isaaks and Srivastava (1989) provide an overview of the volume-variance problem, on its implications and on how to deal with it.

THE VOLUME OF SELECTIVE REMEDIATION

When we say that the goal of our classification is to correctly identify *"all material for which the lead contamination exceeds 500 ug/g"*, what exactly does this mean? By using a concentration-based criterion, such a statement carries some implicit assumption about the volume; without a volume of material, the concept of a concentration would have no meaning. What is the volume at which the remediation is intended to succeed? Do we clean up every spoonful of material whose lead concentration exceeds 500 ug/g? Or do we clean up every shovelful whose lead concentration exceeds 500 ug/g? Or does the statement pertain to entire truckloads? Or to even larger volumes? If we intend to go after every spoonful of material in excess of 500 ug/g then we'll have to do a lot of sampling or excavate virtually all of the soil as contaminated (or both). If we intend only that the remediation clean up truckload sized volumes, then fewer samples will be required and less soil will probably need to be excavated.

Without consensus on the volume to which a regulatory threshold applies, detailed remediation planning is usually premature. If the engineering plan is designed to correctly classify large truckloads of material while the regulatory agency intends that much smaller volumes be correctly classified, then the remediation, even if successful according to those who designed it, may not satisfy the regulatory agency. This is typically the case when remediation with large equipment is thought to be complete, only to discover that verification samples still encounter residual contamination. Even if the original remediation did, in fact, remove all contaminated material at the scale of large trucks and no entire truckload of contamined soil remains on the site, it is still possible that small surface samples will encounter contamination in excess of the regulatory threshold.

The "volume of selective remediation", or VSR, is the smallest volume of material that the remediation exercise intends to segregate into one category or another. When deciding what the VSR should be, the first consideration should be the basis for the regulatory thresholds. The adverse effects of a particular contaminant often depend on a variety of factors, including the exposure pathway, the chemical form of the contaminant and its physical form. The selection of a target threshold for a remediation should address the minimum volume for which the adverse effects can manifest themselves. For example, with lead contamination in the soil in a residential area, the primary focus of the remediation may be to minimize the risk of a child picking up a handful of dirt and eating it. For this situation, the VSR is usually considered to be very small, about a handful of dirt. As another example, the most adverse effect of mercury contamination in marine sediments may be the ingestion of mercury by bottom-feeding organisms that eventually become food themselves for larger aquatic life and, eventually, for humans. For this situation, the VSR may be the area over which a bottom-feeder is likely to browse in its lifetime.

In addition to considering the adverse effects of the various contaminants, the selection of an appropriate VSR will often also take into account the practical realities of the remediation exercise. Even though we would like to segregate each and every handful of lead contaminated soil, there is no equipment that can economically achieve such a fine level of selectivity. If the smallest loaders we can use have 3 m^3 buckets, the VSR might reflect this minimum loader size.

The VSR we achieve in actual practice may be much larger than the intended level of selectivity if we do not address the issue of selectivity in all of the different aspects of the remediation and verification design. For example, if the remediation plan is to excavate the upper one metre of soil wherever the available samples suggest contamination, and if the samples are spaced at 10 m, then the minimum volume that we actually end up segregating is roughly $10 \times 10 \times 1 = 100$ m^3 regardless of how small our equipment might be.

The volume of selective remediation and probability maps

When using a probability map, such as the one shown earlier in Figure 4, as the basis for classification, it should be recognized that the probabilities being shown on this type of map refer to the volume of the samples used to define the 0/1 indicators. For example, in areas where the probability map in Figure 4 shows a 50% probability of encountering contamination, single samples from these areas will have a 50% chance of being contaminated. This does not mean that there is a 50% chance that truckloads of material from the same areas will be contaminated.

The selection of an appropriate probability threshold for classification depends on the relative concentrations of the contaminated and uncontaminated material. For the lead example shown in Figure 4, the contaminated samples above 500 ug/g have an average lead concentration of 1850 ug/g, while the uncontaminated samples below 500 ug/g have an average lead concentration of 158 ug/g. With these relative contaminant concentrations, any area that has even a 20% chance of being contaminated will likely have VSRs whose average contaminant concentration exceeds 500 ug/g. The 0.2 contour line from the probability map in Figure 4 is therefore a reasonable basis on which to classify material as contaminated or not. Regions inside this 0.2 contour line have a good chance of containing VSRs whose average contaminant concentration is above 500 ug/g and which should, therefore, be classified as being above the threshold even if individual samples in these areas show lead concentrations less than the 500 ug/g threshold.

RECOMMENDED PRACTICE

1. Classification should be based on interpolation of *in situ* sample data and on the uncertainty in these estimates.

2. When affected material is being classified, the classification procedure should recognize that the regulatory threshold usually pertains to a volume of material that is different from the volume of material that is typically sampled. The "volume of selective remediation" should be explicitly stated during the design of an appropriate remediation strategy. The entire remediation strategy, including equipment selection and the sampling plan that will be used for local refinement of preliminary excavation limits, should be designed so that material can be effectively segregated at the intended level of selectivity.

3. The goal of a classification exercise should be to ensure that there is a less than 5% chance of making a false negative error on a volume of material equal in size to the intended volume of selective remediation. This goal can be met with the use of probability maps and the selection of an appropriate probability threshold based on the relative contaminant concentrations of material above and below the threshold of interest.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Jones, T., Hamilton, D. and Johnson, C., *Contouring of Geological Surfaces with the Computer*, Van Nostrand Reinhold, New York, 1986.

Parker, H.M., "The volume–variance relationship: a useful tool for mine planning," in *Geostatistics*, edited by P. Mousset-Jones, McGraw Hill, New York, 1980.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-14

STOCKPILING

A guide for data analysts, project managers and reviewers
on statistical issues related to stockpiled material

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

During the remediation of contaminated sites, the affected material is not always sampled and classified *in situ*; occasionally it is first stockpiled and then classified on the basis of samples taken directly from the stockpiles. Though this is generally not a good practice, it may be inevitable, unfortunately, when the logistics of the remediation require material to be excavated before chemical analyses are available. This guidance document addresses the issue of stockpiling and discusses why it is regarded as an inefficient and possibly ineffective approach to remediation. It also provides guidance on sampling stockpiles and on using these samples to classify the material.

Many of the issues raised in this guidance document are also discussed in other documents in this series. The reader should also read the documents entitled *ESTIMATING A GLOBAL MEAN*, *COMPOSITE SAMPLES*, *CLASSIFICATION*, and *DESIGNING A SAMPLING PLAN*, each of which contains information that is relevant to the issues discussed in this guidance document.

DRAWBACKS OF STOCKPILING

Loss of spatial context

The most effective and efficient approach to remediation is to use *in situ* samples, along with qualitative information about the site history and usage, to prepare maps or three dimensional models of the contamination as it exists *in situ*. Once material has been excavated and stockpiled, it usually has lost its spatial context and much of the information about site history and usage loses its value. For example, it may become apparent during the remediation of a site that the contamination has been dispersed along old and leaking waste-water lines that run underneath several buildings on the site. This insight could be used to refine models of *in situ* contamination so that the most contaminated material immediately adjacent to the old waste-water lines can be segregated from the less contaminated material on the site. If the affected soil has already been stockpiled, however, then we have likely mixed highly contaminated material with weakly contaminated material and have lost the opportunity to effectively segregate the affected soil into appropriate categories.

Even in studies where the relevant physical and chemical controls are understood, the remediation usually encounters unexpected "hot spots". When these are encountered *in situ*, it is possible through further sampling to delineate their lateral extent and to effectively segregate the "hot spot" in the appropriate contamination category. When a stockpile sample encounters an anomalously high analytical value, it is virtually impossible to tell where the rest of the "hot spot" has ended up; often, the only environmentally prudent approach is to condemn the entire stockpile as belonging to the contamination category of the single anomalous value.

Whether or not the stockpile samples encounter anomalously high values, the loss of spatial context reduces the confidence of any estimates we might need to make with the available sample data. Closely spaced *in situ* samples usually show more similarity in their contaminant concentrations than do closely spaced stockpile samples; the acts of excavating and stockpiling tend to destructure the *in situ* pattern of spatial continuity. The consequence of this for statistical studies of contaminated sites is that our predictions of the contaminant concentrations in material that we have not directly sampled will be more reliable when we use *in situ* samples to estimate nearby *in situ* concentrations than when we use samples from a stockpile to estimate contaminant concentrations elsewhere in the same stockpile.

Dilution

A second drawback of stockpiling is that it can result in highly contaminated soil being diluted with uncontaminated (or less contaminated) soil. If the relative proportion of the highly contaminated soil is low, then this dilution may cause material that should have been classified as contaminated to pass as uncontaminated. Since the adverse effects of many contaminants are directly related to the total mass or quantity of the contaminant, rather than to its concentration, such dilution of contaminated material is not an appropriate practice.

Difficulty of sampling

One of the other major shortcomings of stockpiling is that it is very difficult to obtain samples that fairly represent the entire stockpile. As discussed in *SAMPLING PLANS*, one of the important principles in sampling is that every member of the larger population has the same chance of being selected. Regrettably, stockpile samples are rarely fair; stockpile samples are commonly "grab" samples that are collected where the material is most accessible: usually from the surface of the pile, and often near the base. If the stockpile was homogeneous, then surface samples might be an acceptable basis for determining the average concentration of the entire pile. Stockpiles are rarely homogeneous, however, and there are a variety of reasons why surface samples might be badly biased.

When excavated material is dumped onto a pile, it inevitably segregates according to grain size as it cascades down the slope of the existing pile. If the contaminants are preferentially concentrated in the finer grain sizes — the silts and fine grained sands — then the material that accumulates along the toe of the pile will tend to show higher concentrations than the material that accumulates along the crest. Similarly, if the contamination is concentrated in the coarser fractions — the gravels and coarse sands — then the concentrations will tend to be lower at the toe and higher along the crest. Stockpiled material also segregates according to lithology; sticky lumps of clayey material, for example, will tend to end up in different parts of the stockpile than will loose dry soil.

Another reason that surface samples may be badly biased is that weathering often affects the contaminant concentrations on the exposed surface of a stockpile. With heavy metal contaminants, for example, rainwater may leach contaminants from the surface deeper into the pile. Similar problems arise when the contaminants are volatile organic compounds; the surface of the pile will often show considerably lower concentrations than the interior core since the surface material has had greater exposure to the air for a longer period of time.

STOCKPILE DESIGN

Given the various drawbacks of stockpiles as an intermediary step in remediation, the following principles should be used when designing stockpiles:

1. Wherever possible, stockpiling should always be preceded by *in situ* sampling and mapping of the contaminant concentrations. This *in situ* information should be used to identify areas that are sufficiently homogeneous that the mixing of material from different contamination categories is avoided.
2. To minimize the possibility of misclassification of contaminated material, the size of stockpiles must be kept relatively small, especially when the material is in the vicinity of any very highly contaminated *in situ* samples. Stockpiles should never exceed 50 m^3 when any of the stockpiled material is within 50 m of an *in situ* sample in which the contamination exceeds the concentration for BC Environment's "special waste" category. In no situation should stockpiles ever exceed 250 m^3.

STOCKPILE SAMPLING

Though proper stockpile sampling is difficult, it is not impossible. Three appropriate methods for sampling a stockpile are:

1. If the stockpile is small, create a random sub-sample by shovelling the pile into two separate piles, with one shovelful in every N shovels being randomly selected to go into the smaller pile that will form the sub-sample. With this approach, the selection of N depends on the size of sub-sample we can manage. If we need a small sub-sample, this random splitting of the entire pile may have to be repeated two or more times to obtain an appropriate sub-sample.
2. If the stockpile is too large for the previous procedure to be pragmatic, then collect samples at a regular spacing from vertical borings that completely penetrate the pile. These vertical borings should either be located randomly on the pile or should be located on a regular grid that covers the areal extent of the pile. In this approach, if the discrete samples from a single vertical boring are composited to produce a single analysis for each vertical boring, then the calculation of the average concentration in the entire stockpile should recognize that the borings have different lengths. *ESTIMATING A GLOBAL MEAN* describes how to accommodate different weights in the estimation of a global mean and in the quantification of the uncertainty on such an estimate.
3. As the stockpile is being created, whether by a shovel, by a loader, or by a series of dumped truckloads, the material can be randomly sampled as it accumulates. This approach to stockpile sampling ensures that some samples from the core of the ultimate stockpile will be available when it comes time to classify the material. As an example, if we are building a 200 m^3 stockpile by dumping 10 m^3 truckloads, then we could choose a random sample from each truckload immediately before it is dumped. By the time the entire stockpile has been created, we will have twenty samples that do a much better job of fairly representing the entire pile than any twenty samples we could collect from the surface of the ultimate pile.

Regardless of the method used to collect stockpile samples, the sampling program should be accompanied by a QA/QC program that monitors and documents the reliability and repeatibility of the sample analyses.

CLASSIFICATION OF STOCKPILED MATERIAL

When an entire stockpile is being classified, there are two questions that need to be considered:

1. Is the average contaminant concentration in the entire pile above or below the classification threshold?
2. Is the pile sufficiently homogeneous that classification on the estimated mean is appropriate?

Classification based on the global mean

The most straightforward check of whether the stockpile should be classified as being above or below any specific threshold is simply to check the mean of the available samples. If the sample mean is above the threshold, then the entire stockpile must be classified as being above the same threshold.

Even if the sample mean is lower than the target threshold, this does not ensure that the average concentration in the entire pile is below the threshold. The arithmetic average of the available samples is only an estimate of the true (but unknown) average concentration of the entire stockpile. As discussed in *ESTIMATING A GLOBAL MEAN*, the reliability of this estimate depends largely on two factors: the number of available samples and on the spread of the available sample values, usually measured in terms of the variance or standard deviation.

When the available samples fairly represent the entire stockpile, the uncertainty on the estimate of the global mean can be expressed through a quantity that is usually called the "standard error":

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

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.

Since there is always some uncertainty in our estimate of the true average concentration of an entire stockpile, the classification of stockpiled material should accommodate the fact that we don't really know exactly the true average concentration. For an entire stockpile to be classified as being below a specified threshold, we need to be at least 95% certain that the overall average concentration in the pile is below the target threshold. This can be accomplished by calculating a pessimistically high estimate of the average concentration that is often referred to as the "upper 95% confidence limit" of the global mean:

$$\text{Upper 95\% confidence limit} = \text{m} + 2 \cdot \sigma_{\text{global mean}}$$

where m is the arithmetic average of the available samples and $\sigma_{\text{global mean}}$ is the standard error described above. If this pessimistically high estimate of the global mean is below the target threshold, we have addressed the first of the questions given above, and can turn our attention to the issue of whether the pile is sufficiently homogeneous based on the global mean alone.

Classification for inhomogeneous piles

Whenever the stockpile samples suggest that even a pessimistically high estimate of the global mean is below the target threshold, it is important to address the issue of whether the material within the pile is sufficiently homogeneous to warrant classifying the entire pile as uncontaminated.

It is not appropriate to assume that a stockpile is homogeneous simply because the process of excavating and piling the material has mixed up the soil — stockpiling is not the same as blending. The blending that is accomplished in the stockpiles used in many industrial processes is not due to the casual mixing that occurs when the material is excavated and piled, but is the result of a carefully engineered stockpile. Blending piles typically are constructed with many thin layers and are reclaimed with specialized equipment that cuts across as many layers as possible to maximize the blending efficiency. The stockpiles used in contaminated site remediation exercises are not engineered as blending piles and homogeneity should not be assumed, but should be explicitly checked with the available sample data.

There are two recommended checks of homogeneity:

1. If any single analysis is more than twice the target threshold, then the entire pile should be classified as being above the threshold regardless of the estimated global mean.
2. If the coefficient of variation is larger than one, then classification of the entire pile should be based on the highest sample concentration regardless of the estimated global mean. Since the coefficient of variation is the ratio of the standard deviation of the samples to their mean, this second check entails that if the standard deviation of the samples is higher than their mean, then we should check to see if the highest sample value is above or below the target threshold.

Examples of use of classification criteria

Table 1 below shows examples of styrene analyses for discrete samples taken from stockpiled material that needs to be classified according to the BC Environment categories: "waste" if the styrene concentration exceeds 50 ug/g, and "industrial quality" if the styrene concentration does not exceed 50 ug/g. For each of the four cases shown in Table 1, the classification of the stockpile according to the criteria presented above is discussed below.

Table 1 Styrene measurements (in ug/g) from stockpiles.

Pile 1	Pile 2		Pile 3		Pile 4	
25	7	1	7	1	7	1
51	3	2	3	2	3	2
19	2	7	2	7	2	7
26	1	7	1	7	1	7
87	6	2	6	2	6	2
42	1	1	1	1	1	1
33	101	2	77	2	29	2
29	1	2	1	2	1	2
39	2	3	2	3	2	3
	4	2	4	2	4	2
	1	1	1	1	1	1
	2	7	2	7	2	7

Pile 1: With the nine available samples having a mean of 39 ug/g and a standard deviation of 18.9 ug/g, the standard error of the global mean is 18.9 $\div\sqrt{9}$ = 6.3 ug/g. The upper 95% confidence limit of the global mean is 39 + 2 × 6.3 = 51.6 ug/g. The chance that the true average concentration of the stockpile is above 50 ug/g is not negligible, and the entire stockpile would have to be classified as waste.

Pile 2: With the 24 available samples having a mean of 7 ug/g and a standard deviation of 19.5 ug/g the standard error of the global mean is 19.5 $\div\sqrt{24}$ = 4.0 ug/g. The upper 95% confidence limit of the global mean is 7 + 2 × 4 = 15 ug/g Even though the mean of the entire pile is almost certainly below 50 ug/g, there is a single sample that is more than twice the waste threshold of 50 ug/g. This indicates that some "hot spot" material has inadvertently been included in a pile that was only very weakly contaminated, and the entire stockpile would have to be classified as waste.

Pile 3: With the 24 available samples having a mean of 6 ug/g and a standard deviation of 14.7 ug/g the standard error of the global mean is 14.7 $\div\sqrt{24}$ = 3.0 ug/g. The upper 95% confidence limit of the global mean is 6 + 2 × 3 = 12 ug/g. As with Pile 2, the true average concentration of the entire pile is almost certainly below 50 ug/g. This example also has a single anomalous value, the 77 ug/g analysis, that causes the standard deviation, 14.7 ug/g, to be noticeably higher than the mean, 6 ug/g. In this case, the coefficient of variation is bigger than one and the classification should be based on the highest value and the entire pile would have to be classified as waste.

Pile 4: Like the previous two examples, the sample mean (4 ug/g) and standard deviation (4.9 ug/g) are both low enough that the upper 95% confidence limit of the global mean is well below the 50 ug/g target threshold. Like the previous example, the coefficient of variation is above one and the classification should be based on the highest value. With the highest value being only 29 ug/g, the entire pile would not have to be classified as waste, but in the next lower category (industrial waste).

In all of the three last examples, Piles 2 through 4, the fact that the coefficient of variation is above one should cause the stockpiling practice to be suspended until the reasons for the lack of sufficient homogeneity can be documented and corrective action can be taken.

RECOMMENDED PRACTICE

1. Stockpiles should be created only where *in situ* sampling has confirmed that the material being stockpiled is homogeneous, with a coefficient of variation of one or less.
2. Stockpile sampling programs should be designed to ensure a fair representation of the contaminant concentrations in the entire pile. Particular attention should be paid to the possibility that the concentrations in the core of the pile are different from those on the surface.
3. Classification of stockpiled material should be based on at least five separate analyses, some of which may be composite samples, and on the following statistical criteria:
   (a) If the "upper 95% confidence limit of the global mean", as described in *ESTIMATING A GLOBAL MEAN*, is above the classification threshold, then the entire stockpile must be classified as being above the threshold.
   (b) If any single analysis is more than twice the classification threshold, then the entire stockpile must be classified as being above the threshold.
   (c) If the standard deviation of the available analyses is larger than their mean, then the stockpiled material should be classified according to whether the highest analysis is above or below the classification threshold.

   If the stockpiled material is classified as being above the threshold for the reasons (b) or (c), and not for (a) alone, then the stockpiling practice is not accumulating homogeneous material; in this event, the stockpiling practice should not continue until the reasons for lack of homogeneity have been documented and corrective action has been taken.
4. Though the classification of stockpiled material may make use of composite samples, all of the discrete samples should be analyzed separately for at least one in every ten of the stockpiles. If the analyses of the discrete samples have a coefficient of variation greater than one, then the stockpiling practice should not be continued until further *in situ* sampling and data analysis allow more homogeneous regions to be identified.
5. As with any sampling program, stockpile samples should be accompanied by a QA/QC study that allows the quality of the analytical values to be monitored and documented.

REFERENCES AND FURTHER READING

In addition to the other guidance documents referenced on the first page of this document, the following references provide useful supplementary material.

Cochran, W.G., *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York, 1977.

Heuer, H., "Stockpiling and Blending of Bulk Materials", in *Stacking Blending Reclaiming of Bulk Materials*, edited by R.H. Wöhlbier, Series on Bulk Materials Handling, Volume 1, No. 5, Trans Tech Publications, Aedermannsdorf, Switzerland, 1977.

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-15

REPORTING

A guide for report writers, project managers and reviewers on reporting a contaminated site statistical application

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

Reports of various aspects of contaminated site studies may have several audiences, from regulators at the federal, provincial and local levels, to members of the affected community, to landowners and their consultants. In the particular case of the application of statistics to contaminated sites, reporting is often made more difficult by the fact that those who end up reading the report are not all likely to be familiar with the technical aspects of the statistical methods used in the study. It is often difficult to know what exactly to include and not to include in the reporting of a statistical study. In addition to discussing necessary elements of a good report of any kind, such as a clear statement of objectives and conclusions, this document also proposes the following general principle for statistical reporting: any reader of our report should be able to find out

- what data we used and why
- what assumptions we made and why
- what statistical tools and procedures we used and why

This document is not intended as a rigid prescription for reporting. There are many individuals and groups whose reporting practices are already excellent and whose reports already provide all of the information that most readers might want. Rather than trying to prescribe a common format for all reports, this document aims instead to provide ideas on what a good report should contain for those who are unfamiliar with reporting statistical studies or for those whose are looking for new ideas to improve their current reporting practice.

REPORT OUTLINE

Table 1 provides an outline of the major headings of a complete report of a statistical study. Not all of the sections listed in Table 1 may be necessary since our report may be an interim progress report or may be part of a larger report, other sections of which cover some of the background information. While we do not always need to generate a complete and comprehensive report, we should pay attention to the three what/why guidelines given earlier. If we know, for example, that the only people reading our report will already be familiar with the data we are using and why we are using it, then we might choose to leave out these details. If we are not sure who will be reading our report, however, then we should plan for the worst case: a reader who knows absolutely nothing about the project. While we certainly don't want all of our memos and progress reports ballooning into multi-volume sets of documents, we could still help a lot of the unprepared readers by having a brief introductory section that explains where they can find the information that we know to be relevant but that we choose not to include for the sake of brevity. If some of the missing information is not yet available, we should inform the reader of the preliminary or draft nature of our report and advise them on when and how they can get a more complete version.

Though it is often tedious to compile all of the ancillary information necessary for a complete report, it can also be an illuminating and beneficial exercise. We have to admit that our reports often lack critical information because of our ignorance about the data, the assumptions or the procedures we used, and not because we choose to omit this information. If we take the time to find out, we might be surprised at what we learn. In the process of trying to learn how the data base was verified, for example, we might discover that it was not. And in trying to find out why not, we might learn that at an early stage in the history of the site, two different studies had used conflicting sample numbers that complicated the checking of the laboratory's own report of its analytical results against the entries in our data base. And that might alert us to the possibility that sample values at certain locations have been transposed...

Such a story could go on and on. Similarly true stories could be told about embarassing last-minute discoveries of unstated critical assumptions, about missing data, and about the use of old software. Even though we may not actually produce a complete report, our statistical studies would benefit if we *planned* on writing a complete report, and gathered the necessary information. At the outset of the study, we should make a list of all the information that a complete report would ideally contain. During the course of the study, as time permits, we should find out where these various bits of information can be found. Even if we do not get the information itself into our various progress reports and memos, we will be able to direct interested readers to the appropriate sources and we may stumble across some information that has important implications for our statistical analysis and interpretation.

A report that reads well is written well, so as we put together our report, we should continually look at it from the point of view of our various readers. Is the presentation clear and informative? Are the graphical displays appropriately labelled and do they support the arguments in the text? Can regulators ensure that the contaminated site is being dealt with in an appropriate manner? Can concerned members of the community get a good appreciation for the rationale and justification for the remediation plan? Can the landowners and their consultants make decisions regarding their role in the remediation?

Table 1 Outline of a report of a statistical study of data from a contaminated site.

Section	Contents
SUMMARY	Salient facts and study results should be provided at the beginning of the report so that a busy reader can quickly get a good overall (but not overly detailed) feel for the objectives of the study, the conclusions and the recommendations.
OBJECTIVES	The goals of the study should be clearly stated so that the reader can judge the appropriateness of assumptions made throughout the study.
SITE DESCRIPTION AND HISTORY	The site's manmade, geographical, geological and hydrogeological features should be described, ideally with the support of maps and cross-sections; if there are off-site features that affect the study, these should also be described. The site history should include details of site usage, with maps showing current site usage along with the location of manmade features, such as roads and buildings and, if known, of any landfills or dumps. If the previous site usage is relevant to the study and significantly different from current usage, additional maps should be provided showing the historical evolution of the site usage.
DATA	The sample population should be described along with the sampling plan and the sampling protocol. Previous studies that contribute data to the study should be summarized; if previous studies contained data that may be perceived as useful and that were not used, the reasons for excluding these data should be discussed. If the study makes use of information that was not generated as part of the study — such as predominant wind direction, toxicity of a contaminant or mobility of a chemical compound — the us e of such auxiliary information should be justified and the source identified. The procedures used to confirm and verify the data base should be described.
EXPLORATORY DATA ANALYSIS	The relevant features of the data should be statistically summarized, ideally in a graphical format with the support of tables, so that for each important variable the reader has a good idea about its distribution, its relationship with other variables and its spatial distribution. Outliers should be identified and discussed individually.
STATISTICAL ANALYSIS AND INTERPRETATION	The statistical tools and procedures used to analyze and interpret the data should be described, along with their underlying assumptions. Every value that is not directly measured but is rather the result of some kind of estimation or prediction — such as the estimated volume of soil that requires remediation, the population standard deviation estimated for the purposes of a confidence interval calculation or the estimated average contaminant concentration in a stockpile — should be documented by explaining how it was calculated and what assumptions were involved in this calculation. All such estimated or predicted values should also be accompanied by a statement about their uncertainty.
CONCLUSIONS AND RECOMMENDATIONS	Each conclusion should be clearly stated with specific references to the statistical analysis and interpretation that support it. Each conclusion should also be accompanied by a discussion of how it is affected by any underlying assumptions, by the accuracy and precision of the available sample data and by the uncertainty in estimated or predicted values. Any recommendations for further work should be accompanied by a specific goal that sets up future objectives.
REFERENCES	All data sources, previous studies and other sources that contributed information to the study should be referenced, along with any technical literature that provides additional detail on statistical procedures used in the study.
APPENDICES	Analytical laboratory results should be provided, either in printed form or, if too voluminous, on a diskette. Laboratory QA/QC procedures, the sampling protocol and the results of check analyses should also be provided. Details of statistical computations omitted from the main body of the report should be included. The computer software used for the data base compilation and the statistical analysis should be documented by providing the name and version for commercial software, or by providing a brief description and a reference for any other non-commercial software used in the study.

DATA

Statistical studies depend on data and we owe it to our readers to be clear about what data we chose to use and why. There are three key steps in documenting the data. The first is how we chose sample locations, the second is how we got our sample values; the third is how we merged all of the information on sample locations and sample values, possibly from several different sources, into a data base.

Sample locations

Whenever we use sample information in a statistical study, it is very rare that we are interested in what the samples have to say about themselves; what we are really interested in is what the samples tell us about some much larger population. It is important, therefore, to explain to our readers the rationale for our choice of samples. Statistical inference about some larger population will be valid only if the available samples are representative of that larger population.

Few of us would put much faith in a public opinion poll that was based on someone going out and talking to a couple of friends. We expect a credible poll to be based on a systematic and unbiased sampling of the population; we also expect its conclusions to be appropriately qualified by the number of people that were actually surveyed. The readers of our report on the application of statistics to a contaminated site are going to be as demanding. They will want to know, for example, how the number of samples was chosen, how the locations were chosen and whether field conditions necessitated modifications to the original plan. Since all of these are questions that a thoughtful reader will ask, we should be sure to discuss the rationale behind the sampling plan and to supplement it with maps and cross-sections showing the sample locations.

Sample values

In describing the information needed to support statistical studies used as evidence in legal proceedings, Glasser (1988) writes

> *"... The meaning and proper interpretation of data cannot be divorced from the method of measurement that gave rise to the data. Different methods of measurement usually produce different statistical results. Hence it is essential to include a detailed description of the particular method or methods of data collection in a report of a statistical study. Such description should fully answer questions on how the data were collected and how they were recorded, and by whom..."*

Though these remarks were aimed at the type of medical data and social science data that are often used as evidence in legal proceedings, they apply equally well to data collected from contaminated sites.

Since errors are involved in every step of sample collection, preparation and analysis, we need to assure our readers that we know what biases are involved in these various steps. We should also show that we have made every effort to keep these biases as small as they can reasonably be and should document for the benefit of our readers the accuracy and precision of our sample values. If we don't document the reliability of the data that are the foundation of our statistical analysis and interpretation, then the reader is unlikely to have much confidence in our conclusions.

The report should discuss all samples that have been identified as outliers and explain, whenever possible, why these anomalous sample values were encountered. Since outlier values usually have a large influence on the analysis and interpretation of the data, the report should discuss the sensitivity of any conclusions to the outlier values. Discarding outlier values, rather than using them to better understand the nature of the p robkm, is generally a poor practice; if any outliers are discarded, the report must provide a rationale for this decision.

Data base compilation

With the data from contaminated site studies often having to be transcribed, keypunched or electronically merged from some other source, there are ample opportunities for human error. Our report needs to explain how the data base that we used in our study was created and how we verified that the data it contains is the same as the original data.

Verification of the data base is an area that is chronically overlooked in environmental studies. When it comes to the quality and integrity of the data, our attention is focused on laboratory quality assurance and quality control (QA/QC) issues. While we are right to demand that the analytical values from the laboratory are as precise and as accurate as they can be, we are wrong to believe that the lab is where most of our errors are occurring. Many errors occur before the lab ever gets the samples, and again after it has reported its analytical values. As much for our own benefit as for the assurance to the readers, we should make sure that we have verified the data base we are using and that we have documented how this verification was done.

ASSUMPTIONS

The underlying assumptions in our statistical analysis and interpretation are as important as the data that form the basis for our numerical calculations. Different assumptions about the distribution of the values, for example, can lead to quite different conclusions. We should not be shy or embarassed about having to state assumptions — all scienc e and engineering is based on assumptions and approximations. What we should be embarassed about is our failure to state them clearly. A clear statement of the underlying assumptions not only informs the reader that we have a good understanding of the tools we are working with, but it also allows others to improve on our work if future data suggest that a different assumption might be more appropriate. If we fail to state our assumptions, then the reader may believe that we don't really understand the limitations of the tools we are using, and others who have to work on the same site will be less able to use our work as a sensible point of departure.

STATISTICAL TOOLS AND PROCEDURES

The reason that statistical methods are commonly used on contaminated site studies is that they offer a variety of procedures for taking data, and, with a few assumptions, making inferences about the population from which the data were drawn. With the data clearly documented and our assumptions clearly stated, the one other statistical issue that we must be sure to address is what procedures we used to arrive at our results and why we chose those methods. While we could report, for example, that "we contoured the data and calculated volumes", it would be more informative to explain to the reader how we did that. Contouring is not a unique exercise; there are dozens of ways to contour a data set; some are based on statistical considerations, some are based on aesthetics. If we ran a program that did the contouring for us, it would help the reader to know what it was and what parameters we provided; if we contoured it manually then we should say so.

Whenever we use an estimated or predicted value, rather than one that was actually measured, we should document how the estimate or prediction was calculated and what assumptions were involved. Any such estimate or prediction has some uncertainty associated with it and our report should make some statement about the uncertainty. In some cases, there are statistical procedures that allow us to quantify the uncertainty; even if there are no such procedures readily available, it would help our reader to have some qualitative statement regarding the uncertainty of our various estimates and predictions. Based on our detailed knowledge of the data and the procedures that we used, do we believe our estimates to be very accurate? Or would we prefer the reader to think of them as ballpark figures?

Finally, through all of the uncertainties in our fundamental data, our various assumptions and our various choices of statistical procedures, what really matters is the bottom line: would the remediation decision change if we had more reliable data, or if we made different assumptions, or if we used a different statistical approach? In our report, we should try to summarize for the reader how sensitive are our conclusions and recommendations to the cumulative effect of all the uncertainties, assumptions and choices that are an inevitable part of any statistical study.

REFERENCES AND FURTHER READING

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

Brushaw, T.C., Alred, G.J. and Oliu, W.E., *Handbook of Technical Writing* St. Martin's Press, New York, 1987.

Glasser, G.G., "Recommended standards on disclosure of procedures used for statistical studies to collect data submitted in evidence in legal cases," Appendix II of Appendix F: "Recommendations on pretrial proceedings," in *The Evolving Role of Statistical Assessments as Evidence in the Courts*, S.E. Fienberg (ed.), Springer-Verlag, New York, 1989.

Kaltreider, R., et al., *Data Quality Objectives for Remedial Response Activities: (Volume 1) – Development Processes*, EPA-540-G-87-003, CDM Federal Programs Corporation, 1987.

CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-16

RANDOMIZATION

A guide for data analysts and field staff on how to select random samples and randomize sample locations

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

Randomization is a recurring aspect of sampling and data analysis in contaminated site studies. For example, during *in situ* characterization, if we identify a particular value as an outlier and need to replace it with another nearby sample, it is recommended that this replacement sample be located at random within a circle of radius 1 m from the original sample location. Other examples of situations that involve randomization are:

- For monitoring internal heterogeneity of composite samples, one in every ten composite samples should be chosen at random to have all of its discrete samples analyzed.
- In designing a sampling grid for a contaminated site, it may be necessary to randomize the origin of the grid.

With these and other similar situations that call for randomization, it is not appropriate to make the decision haphazardly; lack of forethought does not produce good random samples. Nor is it appropriate to leave the choice of a random sample to someone's guesswork; the "random" choices that people make usually turn out not to be suitable for statistical purposes. Proper randomization is a systematic and repeatable procedure that can be checked and verified.

This guidance document addresses the issue of randomization and presents procedures for making random choices. It begins with a discussion of uniform random numbers and then describes how a table of such numbers can be used for randomization. Other guidance documents in this series that make specific references to randomization are *COMPOSITE SAMPLES, SAMPLING PLANS, STOCKPILING* and *OUTLIERS.*

UNIFORM RANDOM NUMBERS

The cornerstone of randomization is a sequence of uniform random numbers between 0 and 1. Table 1 on the next page shows 500 uniform random numbers; information on how to generate such a table, or on other similar tables, can be found in the references listed at the end of this guidance document. The values shown in Table 1 are "uniform" in the sense that a histogram of the values, such as the one shown in Figure 1, will show that the numbers in the sequence are as likely to come from one particular class in the histogram as from any other class — all values are equally probable. The values are "random" in the sense that the next value in the sequence is always unpredictable. Whether we read across the rows or down the columns of Table 1, there is no pattern or clue that tells us what the 501st value might be; regardless of the past sequence, all values between 0 and 1 remain equally probable. One way of demonstrating the lack of predictability in the values is to plot consecutive pairs on a scatterplot, such as the one shown in Figure 2. In this figure, the first random number in each of the 250 possible pairs from Table 1 is plotted as the x value and the second in each pair is plotted as the y value. If the sequence is truly random then such a plot should show no correlation. Kennedy and Gentle (1980) and Bratley, Fox and Schrage (1980) discuss several other statistical criteria for sequences of random numbers; the two most important, however, are uniformity of the histogram (as shown in Figure 1) and lack of correlation on a scatterplot of consecutive pairs (as shown in Figure 2).

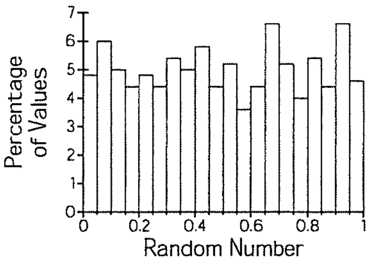


Figure 1 Histogram of the 500 values shown in Table 1.

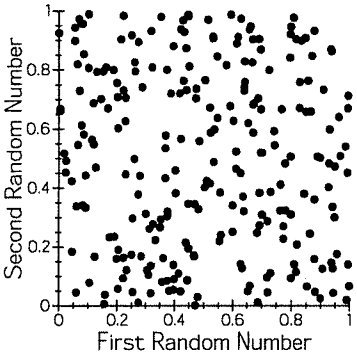


Figure 2 Scatterplot of the 250 consecutive pairs from Table 1.

USING UNIFORM RANDOM NUMBERS

A sequence of uniform random numbers, such as the one shown in Table 1, can be used as the basis for randomization procedures that are both systematic and verifiable. Two common randomization problems are discussed below; the first is the random selection of one of several samples, the second is the selection of a random location.

Selecting a random sample

The following procedure can be used to select one sample at random from a group of N samples:

1. Assign numbers 1 through N to the samples; which sample gets which number is unimportant, but each sample should have a unique index from 1 to N. The assignment of these unique indexes should be recorded for future reference in case another random sample is needed or in case the randomization needs to be checked.
2. Generate U, a uniform random number between 0 and 1. This can be done by taking the available next number from a table of random numbers, such as the one given in

Table 1 500 uniform random numbers between 0 and 1.

.534	.489	.800	.947	.641	.829	.719	.287	.432	.764
.222	.164	.640	.730	.600	.252	.936	.870	.086	.582
.305	.122	.188	.800	.744	.546	.405	.383	.820	.208
.701	.309	.556	.385	.466	.172	.503	.403	.377	.320
.829	.672	.230	.706	.397	.139	.559	.241	.200	.756
.233	.122	.310	.105	.434	.929	.653	.050	.531	.914
.404	.460	.612	.524	.870	.848	.086	.854	.480	.030
.093	.443	.878	.933	.104	.989	.128	.470	.993	.452
.485	.559	.471	.002	.124	.167	.263	.796	.469	.346
.274	.011	.636	.372	.848	.909	.439	.215	.121	.549
.146	.702	.718	.097	.357	.081	.660	.068	.995	.712
.274	.379	.064	.820	.354	.980	.800	.094	.888	.507
.536	.599	.423	.938	.687	.012	.261	.826	.730	.387
.696	.870	.654	.620	.214	.904	.274	.895	.999	.067
.673	.686	.578	.817	.397	.881	.670	.683	.334	.224
.521	.416	.175	.233	.474	.736	.780	.484	.879	.361
.318	.043	.628	.142	.374	.801	.521	.673	.692	.274
.810	.141	.799	.310	.370	.578	.163	.809	.160	.038
.774	.271	.362	.162	.546	.846	.473	.704	.692	.384
.938	.147	.451	.812	.547	.638	.447	.982	.070	.973
.854	.066	.591	.678	.701	.520	.510	.425	.230	.447
.388	.721	.986	.541	.114	.564	.060	.046	.369	.535
.945	.044	.002	.925	.671	.588	.072	.730	.205	.190
.083	.957	.280	.442	.153	.795	.990	.020	.720	.046
.941	.336	.762	.324	.254	.298	.045	.424	.131	.794
.318	.136	.949	.473	.103	.807	.106	.079	.930	.484
.900	.026	.684	.249	.006	.668	.369	.434	.066	.547
.928	.403	.169	.670	.447	.987	.448	.347	.662	.948
.415	.110	.796	.842	.376	.305	.422	.049	.725	.800
.026	.493	.301	.312	.229	.627	.284	.168	.351	.267
.079	.613	.006	.661	.482	.329	.899	.503	.222	.731
.603	.628	.857	.764	.202	.708	.808	.977	.382	.051
.532	.907	.056	.899	.690	.905	.191	.235	.157	.007
.642	.759	.448	.196	.419	.721	.083	.342	.251	.918
.742	.593	.056	.944	.619	.913	.554	.220	.612	.466
.994	.670	.667	.826	.907	.126	.815	.905	.250	.174
.120	.688	.696	.520	.326	.278	.222	.984	.774	.228
.844	.381	.078	.958	.376	.085	.199	.160	.851	.658
.658	.938	.514	.970	.204	.058	.639	.322	.879	.159
.019	.518	.204	.603	.260	.499	.829	.759	.288	.742
.024	.453	.397	.055	.061	.339	.697	.667	.633	.128
.963	.641	.523	.588	.781	.325	.442	.732	.323	.809
.817	.115	.764	.412	.863	.245	.440	.874	.505	.766
.910	.733	.353	.293	.968	.509	.948	.301	.614	.890
.400	.091	.669	.904	.430	.968	.628	.451	.957	.176
.254	.047	.881	.048	.093	.334	.221	.094	.629	.974
.468	.168	.594	.988	.798	.507	.600	.342	.768	.667
.319	.932	.730	.111	.995	.140	.932	.599	.445	.087
.423	.567	.313	.262	.865	.659	.046	.184	.938	.864
.943	.127	.640	.741	.834	.900	.687	.704	.797	.922

Table 1. As the tabulated random numbers are used, they should be crossed off so that it is immediately obvious which one to use when we next consult the table.

3. Turn U into an integer from 1 to N by multiplying it by N, adding 1 and dropping any digits after the decimal point:

$$\text{Random integer} = \text{Integer part of } [\text{U} \times \text{N} + 1]$$

4. Select as the random sample the one that was assigned this calculated random integer as its unique index.

As an example of the use of the procedure, let us go through the exercise of deciding which of the ten composite samples shown in Table 2 should be selected for separate analysis of all of its discrete samples. The first column in this table gives the sample number for the composite samples, the next four columns identify the discrete samples that make up each of the composite samples and the last column gives the unique index from 1 to 10 that we have assigned according to step 1 of the procedure given above. Table 3 shows our table of random numbers; in this example we are supposing that we have been using it for similar randomization exercises and have been crossing off the numbers as we use them; the next number on our list, 0.466, is the value of U. We take this number and turn it into a random index from 1 to 10:

$$\begin{aligned} \text{Random index} &= \text{Integer part of } [0.466 \times 10 + 1] \\ &= \text{Integer part of } 5.66 \\ &= 5 \end{aligned}$$

Our random sample from the group of ten would therefore be C130 since this is the one that was designated as number 5 when we assigned unique indexes from 1 to N.

Table 2 Ten composite samples and their discretes.

Composite Sample No.	Discrete Sample Numbers				Unique Index
C103	D562	D563	D564	D565	1
C104	D567	D568	D569	D570	2
C105	D572	D573	D574	D575	3
C129	D709	D710	D711	D712	4
C130	D714	D715	D716	D717	5
C131	D719	D720	D721	D722	6
C149	D831	D832	D833	D834	7
C150	D836	D837	D838	D839	8
C151	D841	D842	D843	D844	9
C172	D902	D903	D904	D905	10

Table 3 Uniform random numbers between 0 and 1.

~~.534~~	~~.489~~	~~.800~~	~~.947~~	~~.641~~	~~.829~~	~~.719~~	~~.287~~	~~.432~~	~~.764~~
~~.222~~	~~.164~~	~~.640~~	~~.730~~	~~.600~~	~~.252~~	~~.936~~	~~.870~~	~~.086~~	~~.582~~
~~.305~~	~~.122~~	~~.188~~	~~.800~~	~~.744~~	~~.546~~	~~.405~~	~~.383~~	~~.820~~	~~.208~~
~~.701~~	~~.309~~	~~.556~~	~~.385~~	.466	.172	.503	.403	.377	...

Selecting a random location

The following procedure can be used to select a random location within a rectangular area whose width is X_{width} and whose height is Y_{height} (see Figure 3):

1. Generate U_1 and U_2, two uniform random numbers between 0 and 1. This can be done by taking the next pair of numbers from a table of random numbers, such as the one given in Table 1. As the tabulated random numbers are used, they should be crossed off so that it is immediately obvious which one to use when we next consult the table.

2. Turn U_1 into an x-coordinate from 0 to X_{width} by multiplying it by X_{width} and turn U_2 into a y-coordinate from 0 to Y_{height} by multiplying it by Y_{height}:

$$X = U_1 \times X_{width} \qquad Y = U_2 \times Y_{height}$$

3. Using the corner of the rectangular area as the origin (0,0) select as the random location the point whose coordinates are (X,Y).

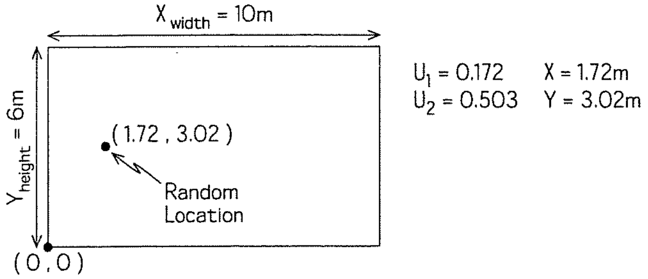


Figure 3 Random sampling from a rectangular area.

If the area within which we want a random sample is circular, rather than rectangular, a similar procedure can be used, with the first uniform random number being converted to a random azimuth from 0° to 360° and the second uniform random number being converted to a radius from 0 to R, the radius of the circular area (see Figure 4):

$$\text{Azimuth} = U_1 \times 360° \qquad \text{Radius} = U_2 \times R$$

As an example of the use of this procedure, let us go through the exercise described at the beginning of this document. At one of our existing sample locations there is an outlier value that we believe to be erroneous and we need to collect a replacement sample from a random location within 1 m of the existing sample. Table 4 shows the random numbers left after we have crossed off the one we used in the first example. The next two values are 0.172 and 0.503. Multiplying the first one by 360 gives us a random azimuth of 62° (N62°E). Since the radius of our circular area is 1 m, the second random number can serve directly as our radius. The replacement sample would therefore be taken at a distance of 0.503 metres from the location of the outlier sample in a direction of N62°E.

Table 4 Uniform random numbers between 0 and 1.

~~.534~~	~~.489~~	~~.800~~	~~.947~~	~~.641~~	~~.829~~	~~.719~~	~~.287~~	~~.432~~	~~.764~~
~~.222~~	~~.164~~	~~.640~~	~~.730~~	~~.600~~	~~.252~~	~~.936~~	~~.870~~	~~.086~~	~~.582~~
~~.305~~	~~.122~~	~~.188~~	~~.800~~	~~.744~~	~~.546~~	~~.405~~	~~.383~~	~~.820~~	~~.208~~
~~.701~~	~~.309~~	~~.556~~	~~.385~~	~~.466~~	.172	.503	.403	.377	. . .

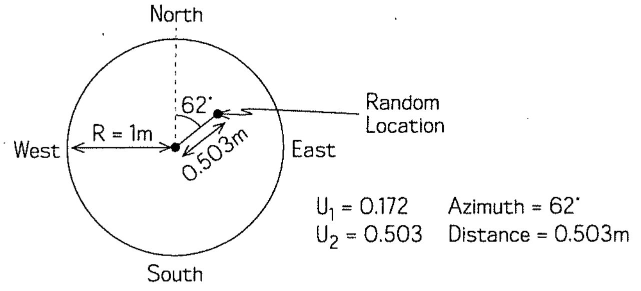


Figure 4 Random sampling from a circular area.

RECOMMENDED PRACTICE

1. When selecting a single random sample from a larger group or when selecting a random location within a specified area, use a systematic and verifiable procedure that is based on uniform random numbers.

2. Use a published table of random numbers or, if a computer or calculator is being used to create the random numbers, print an actual table of the random numbers it produces so that the procedure can be checked and verified.

REFERENCES AND FURTHER READING

The following references provide useful additional information on the specific problem of generating uniform random numbers and on the general problem of randomization.

Abramowitz, M. and Stegun, I.A., (eds.), *Handbook of Mathematical Functions*, Dover, New York, 1970.

Bratley, P., Fox, B.L. and Schrage, L.E., *A Guide to Simulation*, Springer-Verlag, New York, 1983.

Kennedy, W.J. and Gentle, J.E., *Statistical Computing*, Marcel Dekker, New York, 1980.