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		

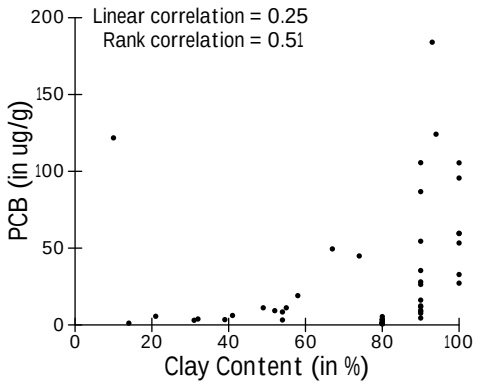

Figure 1 A scatterplot of the clay – PCB data in Table 1.

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 presentation 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.

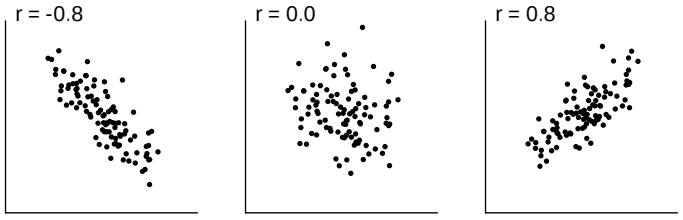

Figure 2 Examples of different correlation coefficients.

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.

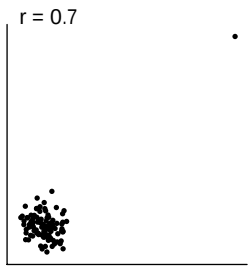

Figure 3 Example of aberrant sample enhancing an otherwise poor correlation.

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.

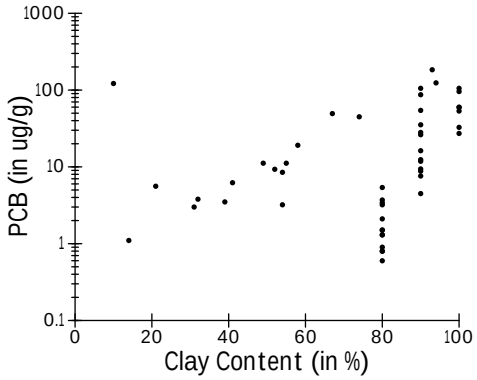

Figure 4 A scatterplot of the clay – PCB data in Table 1 with the y-axis logarithmically scaled.

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.