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

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; 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.

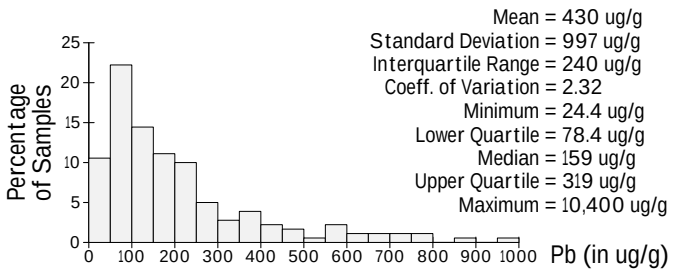

Figure 1 A histogram of 180 lead samples.

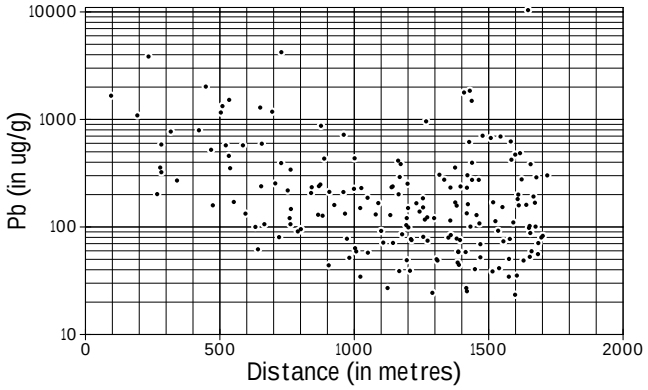

Figure 2 A scatterplot of lead versus distance from the smelter.

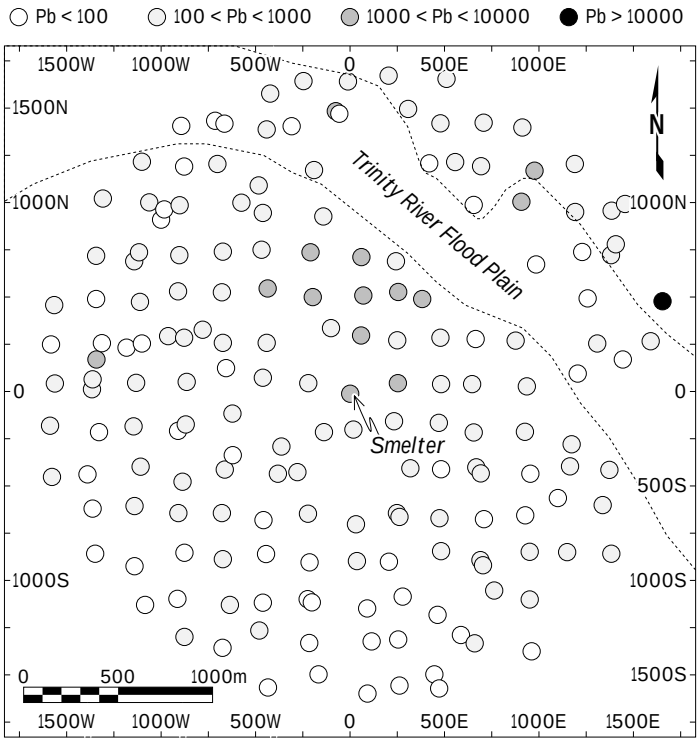

Figure 3 A greyscale map of the 180 lead samples.

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.

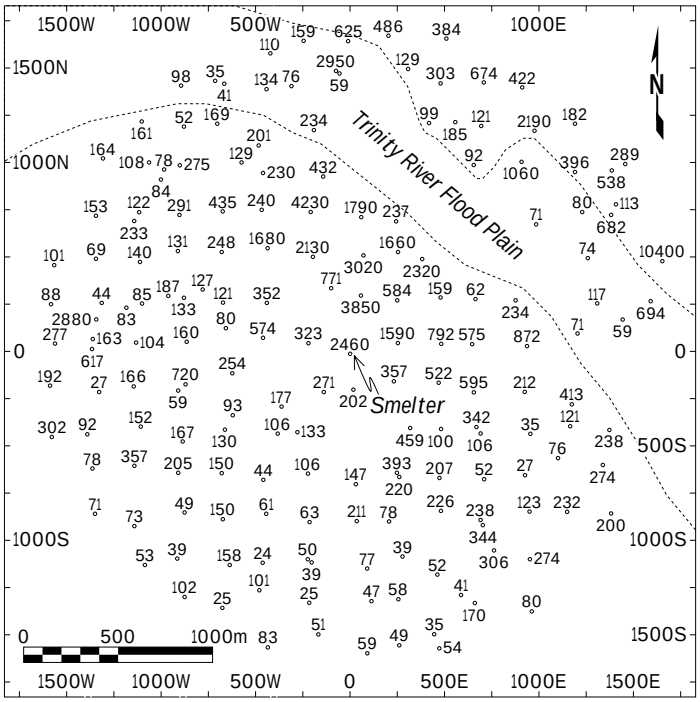

Figure 4 A data posting of the 180 lead samples.

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.

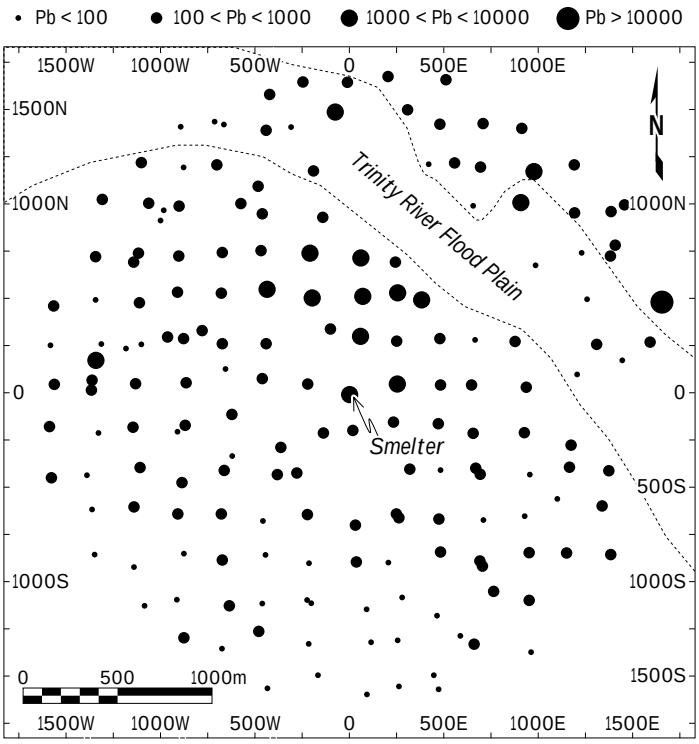

Figure 5 A map of the lead sample values on which the symbol size is proportional to magnitude of the lead concentration.

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.

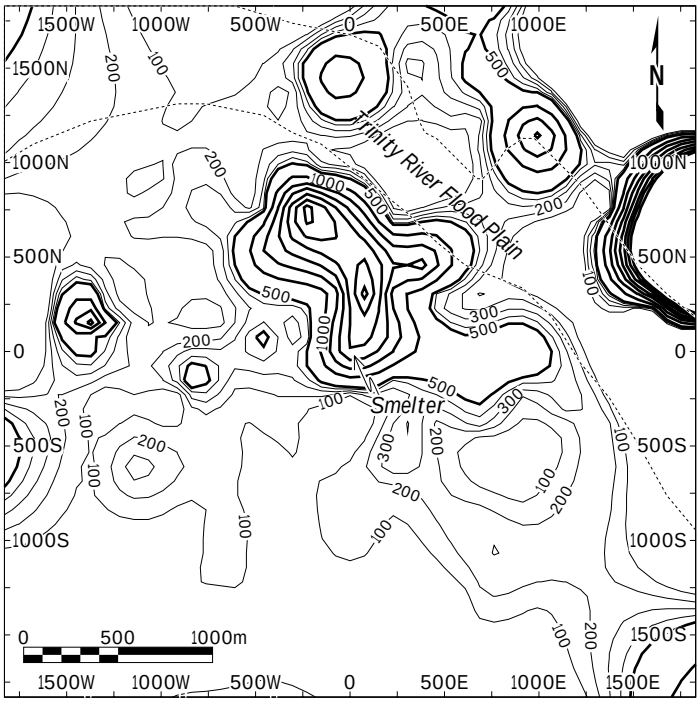

Figure 6 A contour map based on the lead data in Figure 4.

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 map 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 river's 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.