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.

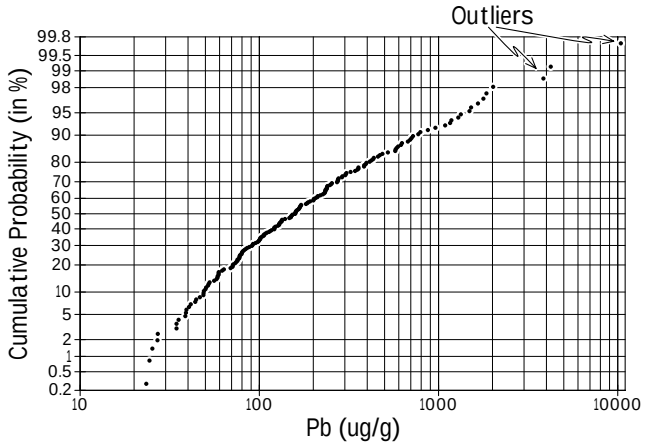

Figure 1 A cumulative probability plot for lead concentration measurements from the soil in the vicinity of a smelter.

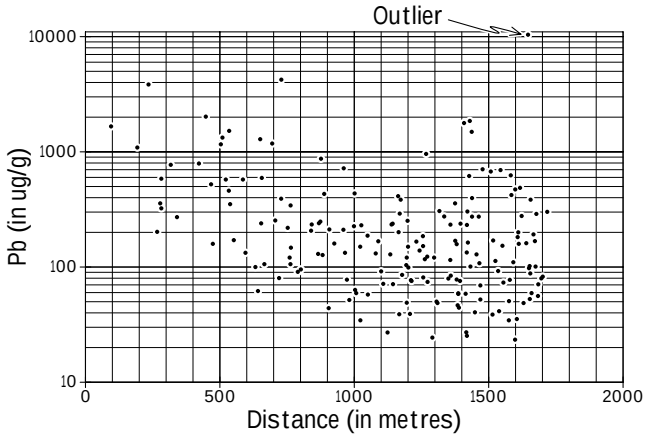

Figure 2 A scatterplot showing lead concentration data used in Figure 1 versus distance from the smelter.

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 for 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 for 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.