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 problem, 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 science 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.