STATISTICAL EVALUATION OF CLEANUP: HOW SHOULD IT BE DONE?*

Richard O. Gilbert
Statistical Design and Analysis Group
Analytic Sciences Department
Pacific Northwest Laboratory
Richland, Washington

ABSTRACT

This paper discusses statistical issues that must be addressed when conducting statistical tests for the purpose of evaluating if a site has been remediated to guideline values or standards. The importance of using the Data Quality Objectives (DQO) process to plan and design the sampling plan is emphasized. Other topics discussed are: 1) accounting for the uncertainty of cleanup standards when conducting statistical tests, 2) determining the number of samples and measurements needed to attain specified DQOs, 3) considering whether the appropriate testing philosophy in a given situation is "guilty until proven innocent" or "innocent until proven guilty" when selecting a statistical test for evaluating the attainment of standards, 4) conducting tests using data sets that contain measurements that have been reported by the laboratory as less than the minimum detectable activity, and 5) selecting statistical tests that are appropriate for risk-based or background-based standards. A recent draft report by Berger that provides guidance on sampling plans and data analyses for final status surveys at U.S. Nuclear Regulatory Commission licensed facilities serves as a focal point for discussion.

INTRODUCTION

Following the remediation of a site contaminated with radionuclides and/or hazardous chemicals, a final survey must be conducted to assure the site has been remediated to environmental concentration guidelines or standards. This paper discusses strategies and methods for sampling the site and statistically analyzing the resulting measurements to evaluate attainment of cleanup standards. We begin by discussing a draft report by Berger, which provides statistical design and testing procedures intended to provide a high degree of confidence that guidelines established by the U.S. Nuclear Regulatory Commission (NRC) for terminating the license of nuclear facilities have been attained. Berger's report, the focus of this "How Clean is Clean" workshop, provides a point of departure for discussing statistical design and testing issues associated with evaluating whether cleanup standards have been attained. Although this paper focuses on facilities contaminated with radionuclides, the discussion applies in general to sites and facilities contaminated with radionuclides and/or hazardous chemicals.

NUCLEAR REGULATORY COMMISSION DRAFT GUIDANCE

Berger (1) provides a draft of detailed procedures for planning, sampling, and data evaluation for the final status survey of an NRC licensed facility. The procedures include using a statistical test to evaluate if the radioactivity levels on building surfaces and radionulcide concentrations for soil and other bulk materials satisfy NRC guideline values for terminating operating licenses. The basic steps in the Berger report follow:

1. Estimate background radiation levels and determine NRC guideline values (risk-based standards) above background levels.
2. Divide the facility and open land areas into affected areas (those that have potential radioactive contamination) and unaffected areas (those that are not expected to be contaminated).
3. Establish a reference grid system (to facilitate selection of measuring/sampling locations and to provide a convenient means for determining average activity levels for $100m^2$ outdoor areas and $10m^2$ indoor areas).
4. Group $100m^2$ outdoor areas into survey units that have common history or are similar in some regard. Also, group $10m^2$ indoor areas into survey units.
5. Conduct 100% scanning surveys of all affected areas (structures and land) and at least 10% scanning surveys for unaffected areas.
6. Conduct cleanup of any local areas (hot spots of up to 100 cm^2) that have activity exceeding 3 times NRC guideline values.
7. Take measurements to compute the average concentration for each $100m^2$ outdoor area and each $10m^2$ indoor area. Conduct additional cleanup in any area where the average exceeds the NRC guideline value.
8. For soil, if an area has activity between 1 and 3 times the guideline value, the average concentration must be less than $(100/A)^{1/2}$ times the guideline value, where A is the area of the elevated activity in m^2.
9. After cleanup, for structures take at least 30 measurements on a grid system at spacing of 2m or less for each survey unit. For land areas take at least 30 surface (top 15 cm) soil samples in each survey unit.
10. Compute the upper one-sided 95% confidence limit on the mean for the survey unit using the method in EPA (2), which requires that the data be normally distributed.

* Prepared for the U.S. Department of Energy under contract DE-AC06-76RLO 1830.

11. If the upper confidence limit exceeds the guideline value, then 1) conduct more remediation, or 2) take one set of additional measurements [the number of which are determined using a statistically based formula in Berger (1)], combine them with the original measurements, and recalculate the upper one-sided 95% confidence limit on the mean. If the new limit is still greater than the guideline value, then conduct more remediation.

The above steps and the more detailed procedures in Berger (1) were developed in part using the Data Quality Objectives (DQO) process, which is discussed in the next section.

PLANNING USING THE DATA QUALITY OBJECTIVES PROCESS

Statistical design and data analysis issues should be addressed in the planning stages of the final status survey. Planning should be conducted using a thorough and structured approach, such as the DQO process, before any samples are collected. Rupp and Jones (3) define the DQO process as follows (page 29):

> Data Quality Objectives (DQOs, also called Data Performance Criteria) are the full set of constraints needed to design a study, including a specification of the level of uncertainty that a data user is willing to accept in the decision.

Rupp and Jones (3) provide an excellent detailed illustration of the DQO process for the problem of deciding whether drums of heterogeneous waste contain transuranic radionuclides. They define the process as consisting of the following steps:

1. State the problem to be resolved
2. Identify the decision or question
3. State the inputs (list of variables or characteristics to be measured and other information needed to make the decision)
4. Narrow the boundaries of the study (describe populations of interest and the spatial and temporal boundaries)
5. Develop a decision rule (set up hypotheses to be tested; develop a quantitative statement of how data will be used to make decisions)
6. Develop uncertainty constraints on the decision process (specify acceptable false positive and false negative decision error rates)
7. Optimize the design (use statistical methods to develop alternative designs that have the lowest cost and attain the uncertainty constraints in step 6.

Other examples of the DQOs process are provided by Neptune et al. (4) and Ryti and Neptune (5).

The DQO process is an important tool because it can be used to establish the technical basis for the statistical design, data analysis, and decision-making procedures. These design and data analysis procedures must have a technical basis *linked directly to an assessment* of human and ecological risk via environmental transport and dose (or risk) models. The inherent uncertainty in the predictions of such models can be quantified using computer-simulation uncertainty and sensitivity analyses [IAEA (6)]. Because these models are used in setting environmental concentration standards that must be attained by the remediation process, these standards are also uncertain. This uncertainty in standards should be quantified and used when statistical tests are used to determine if the standards have been achieved by the remedial action.

STATISTICAL ISSUES

In this section several statistical design and data-analysis issues are discussed.

Guidelines and Standards

The guideline value (standard) used by NRC [Berger (1)] is a risk (or dose) based standard above background. We shall refer to this standard as a background-plus-risk standard. An issue of interest is how to take into account the uncertainty in background concentrations and in the risk (or dose) portion of the standard when evaluating whether or not the site needs additional remediation. Typically, background is not a constant value throughout the background area. Also, as mentioned previously, a risk-based standard (a specified soil concentration that must not be exceeded) is typically (or should be) determined using environmental transport, dose, and risk models. The predictions of these models may be highly uncertain because of uncertainties in the model and model parameters. In practice, the uncertainty in background measurements tends to be ignored, i.e., the uncertainty in the mean background value is usually not considered.

One way of handling (in effect, avoiding the issue of) uncertainty in the risk component of the standard is to set the limiting soil concentration (standard) at a conservatively low value. This may be accomplished by specifying a conservatively low limiting dose and then solving the transport and dose model for the corresponding conservative limiting soil concentration. However, this approach may be acceptable only for preliminary screening studies. A more rigorous approach is to quantify the uncertainty of model predictions of environmental concentration limits using computer-simulation uncertainty and sensitivity analyses [IAEA (6)]. These simulations generate a probability density function (histogram) of potential risk-based limits. This uncertainty can then be combined with the variability of background measurements to arrive at a distribution of possible alternative applicable background-plus-risk standards (soil concentration limits). The data collected at the site following remediation is then compared to this distribution of potential standards, rather than to a single standard value. The details of how this testing procedure is done have yet to be developed.

When the objective is to compare concentrations with a background standard with no risk component, the variability of the background and site measurements can be taken into account by selecting appropriate statistical tests. The Wilcoxon Rank Sum test and *the Quantile* test discussed in a later section of this report are examples of appropriate testing procedures for this case. Gilbert and Simpson (7,8) and DOE (9) provide additional discussion.

Number of Samples and Measurements

Berger (1) indicates that after an "affected" land area is scanned and hot spots are removed, four soil samples are collected at locations equidistant from the center and each of the four grid block corners of each 10m-by-10m square of the entire land area. If the scanning detector is not sufficiently sensitive, soil samples are collected on a triangular grid

pattern for which the length of each side of the triangle is 5m. The 5m spacing on a triangular grid was derived on the basis of specifying a minimum size hot spot that must be detected with specified confidence (probability). [The method for determining grid spacing in this manner is given by Gilbert (10) Chapter 10.] This approach for determining the spacing of samples in a triangular pattern is a good example of using DQO to arrive at the number of samples to collect. In this case, the DQO are the hot spot size important to detect and the required confidence (probability) that a hot spot of that size will be detected. The important point here is that whenever samples will be collected, every effort should be made to established quantitative DQOs that provide the rationale for the number of samples to collect.

Testing Philosophy and Associated Hypotheses

There are two testing philosophies that may be used when evaluating whether a cleanup standard has been attained. The approach used in Berger (1) is to assume the site has not attained the risk-based standard (above background) until there is statistically significant evidence to the contrary. This is the "guilty until proven innocent" approach. This philosophy may be expressed by the following hypotheses:

H_o: Site Has Not Attained the Standard

H_a: Site Has Attained the Standard (Eq.1)

where we assume H_o is true unless the statistical test indicates otherwise. To illustrate, in Berger (1) it is not sufficient to reject H_o and accept H_a when the mean is less than the limit. It is also required that the upper one-sided 95% confidence limit on the mean be less than the standard. This confidence limit approach provides additional protection of public health and safety. This approach can be used for risk-based standards when background is absent or small relative to the risk-based soil concentration. But it is not appropriate for background-based standards, as discussed in the next paragraph.

The second testing philosophy is to adopt the "innocent until proven guilty" approach, which may be expressed by the following hypotheses:

H_o: Site Has Attained the Standard

H_a: Site Has Not Attained the Standard (Eq.2)

Note that these hypotheses are the reverse of those in Eq. (1), i.e. H_o and H_a are interchanged. Gilbert and Simpson (8) use the hypotheses in Eq. (2) when using statistical tests to determine if a remediated Superfund site has attained background standards (a risk standard is not used). Their rationale for using Eq.(2) instead of Eq.(1) is that if the hypotheses in Eq.(1) are used, then some or most site measurements would have to be less than background measurements before the test would indicate that the site has attained the standard and hence that no more remediation may be needed. Hence, using the hypotheses in Eq.(1) when testing for attainment of background standards could lead to remediating sites where concentrations are at background levels.

The next section discusses some criteria for selecting statistical tests.

Selecting Statistical Tests

The upper 95% confidence limit test used by Berger (1) requires the data to have a normal (Guassian) distribution. This assumption may not be appropriate, especially if remedial action has not been effective. Thus, consideration should be given to using nonparametric (distribution-free) tests, which do not require the data to be normally distributed. A simple, useful distribution-free test is the nonparametric upper one-sided 95% confidence limit on the median, as discussed by Gilbert (10, p. 173). The test is conducted by ordering from smallest to largest the measurements taken at the site. Then a simple table look-up procedure is used to determine which of the ordered measurements is the upper 95% confidence limit value. If this value is less than the standard, then the standard has been attained.

Another simple nonparametric test is the one-sided nonparametric tolerance limit test. This test is conducted by first collecting enough site measurements so the largest of those measurements is a one-sided upper nonparametric tolerance limit on a specified upper percentile of the distribution of site measurements. For example, if 59 representative measurements are taken at random locations over the site, we can be 95% confident that 95% of the distribution of possible site measurements is less than the maximum of those 59 measurements. The test consists of simply comparing the maximum site measurement with the background-plus-risk standard. If the maximum measurement is less than the standard, then the standard has been attained. This test is discussed by Conover (11) and applied to decommissioning and decontamination applications by Eger (12). It is important to note that this test is likely to give different results than the test used by Berger (1) or the nonparametric upper confidence limit on the median. The tolerance limit test is more likely to indicate additional cleanup is required than either of the latter tests. Different tests have different performance characteristics. Thus, tests must be selected with care.

The selection of statistical tests should be made on the basis of appropriate selection criteria. DOE (9) developed the following criteria for selecting tests that will be used to evaluate if an area of land is contaminated to levels greater than background levels: The test should

1. be applicable to testing the hypotheses in Eq.(2) above,
2. take into account uncertainty in the background standard,
3. have adequate (to DQO specifications) power to detect contamination problems that may not be easily detected by other tests in the suite of tests,
4. perform satisfactorily when applied to data sets for which some measurements are reported as below the minimum detectable activity (MDA), and
5. perform satisfactorily when the data are not normally distributed.

The use of two or more statistical tests on the same data set (called "tandem" testing) is used by Gilbert and Simpson (8) and DOE (9) for comparing site data with background standards. An important advantage of tandem testing is that the power (probability) that one or more of the tests will identify when the standard has not been attained will be greater than or equal to the power of any one test in the suite of tests. A disadvantage of tandem testing is that the suite of tests will tend to result in more false positive decision errors than if only one test method is used.

Measurements Less Than the Minimum Detectable Activity

When radionuclide concentrations are at very low levels, some measurements may be reported by the laboratory as less than the MDA. In this situation, it is common practice to report the MDA and use it or perhaps one half the MDA in statistical test calculations. Using MDAs in this way causes biased and possibly very misleading results. Berger (1) recommends reporting the actual measured value (even if it is negative) and using it in statistical calculations and tests. He also recommends the laboratory always (for all data) report the MDA value and the measurement uncertainty (95% confidence level) for the datum. The author totally agrees with these recommendations. Nevertheless, for data sets that contain < MDA values, it is important to use statistical tests that do not give misleading results.

One disadvantage of the testing procedure used by Berger (1) (one-sided upper 95% confidence interval on the mean) is that it may give a misleading test result if MDA values or some function of them are used in the calculations as if they were representative data. However, other tests can be used that do not suffer to the same extent from this problem. Examples are the nonparametric one-sided upper tolerance limit and the nonparametric upper one-sided 95% confidence limit on the median, tests which were discussed previously. The tolerance limit test uses only the maximum datum in the data set. Thus, the test can be conducted even when only one of the data are greater than the MDA. A similar situation applies to the confidence limit on the median. The tests discussed in following paragraphs for testing the attainment of background standards can also be used when data sets contain MDA values. See Gilbert and Simpson (8), Helsel (13), and Gilbert (10) for further information.

Nonparametric Statistical Tests for Background Standards

This section briefly describes three nonparametric statistical tests that can be used to test for attainment of background standards. Gilbert and Simpson (7) use these tests to test the hypotheses in Eq.(2) above. The tests are distribution-free and hence they can be used even when the data are not normally distributed. Also, these tests can be used when a moderate number of measurements are reported as MDA, as long as all MDA results (if accurate measurements could have been obtained for these samples) are really less than the smallest observed measurement in the data set. The power of these tests is discussed by Gilbert and Simpson (7,8).

The Wilcoxon Rank Sum (WRS) test [Gilbert and Simpson (8)] is performed by first listing the combined background and site measurements from smallest to largest and assigning the ranks 1, 2, ..., N to these ordered values, where N is the total number of background and site measurements. The ranks of the site measurements are then summed. If this sum is too large (determined by reference to a table of critical values), a potential contamination problem has been identified. A step-by-step procedure for conducting the test is provided in many statistics publications, including Gilbert (10), Gilbert and Simpson (8), and Conover (12).

The Quantile test is performed by first listing the combined background and site measurements from smallest to largest, as is done for the WRS test. Then, among the largest r measurements of the combined data sets, a count is made of the number of measurements, k, that are from the site. If k is sufficiently large, the background standard has not been attained. The Quantile test was originally developed by Johnson et al. (14). Gilbert and Simpson (8) present tables and a step-by-step procedure for determining the number of samples and the values of r and k that are used to achieve the specified false positive and false negative decision error rates.

The Slippage test is conducted by simply counting the number, K, of site measurements that exceed the maximum background measurement. If K exceeds the critical value obtained from the tables in Rosenbaum (15), a potential contamination problem has been identified. For example, suppose a false positive error rate of 0.05 (5 percent) is specified, and that 50 background measurements and 40 site measurements are obtained. Then, from Rosenbaum's tables, a critical value of four is indicated. That is, if four or more site measurements are larger than the largest background measurement, then H_o is rejected and H_a is accepted [Equation (2)]. The slippage test can be conducted even when a large proportion of the background measurements are less than the MDA.

CONCLUSION

The statistical sampling and data analysis aspects of evaluating compliance with cleanup standards should be planned using a structured approach such as the DQO process. The DQO process approach moves through the necessary steps of identifying the problem, determining the questions that must be answered, developing a decision rule (statistical hypotheses and test) based on specified acceptable levels of uncertainty, and interpreting the data and test results. This paper has briefly discussed some of the important issues that are addressed when selecting statistical tests. More details are provided in the references. Guidance on statistical aspects of environmental studies, such as that in Berger (1), are useful to the practitioner with little formal training in statistics. However, these documents will change as new knowledge is gained about which statistical procedures are optimum in various situations. A critical need is for statisticians and practitioners to work jointly in developing the needed statistical tools. The effective communication of problems and tools among all parties is the key to developing optimum statistical methods to meet real needs.

REFERENCES

1. BERGER, J.D., "Manual for Conducting Radiological Surveys in Support of License Termination," NUREG/CR-5849. U.S. Nuclear Regulatory Commission, Washington, DC (June 1992).
2. EPA. "Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media," EPA 230/02-89-042. U.S. Environmental Protection Agency, Washington, DC (February 1989).
3. RUPP, G.L. and R.R. JONES (editors). "Characterizing Heterogeneous Wastes: Methods and Recommendations," EPA/600/R-92/033. U.S. Environmental Protection Agency, Las Vegas, NV (February 1992).
4. NEPTUNE, D., E.P. BRANTLY, M.J. MESSNER, and D.I. MICHAEL. "Quantitative Decision Making in superfund: A Data Quality Objectives Case Study," Hazardous Materials Control, Vol. 3, No. 3, pp. 18-27 (May/June 1990).

5. RYTI, R.T. and D. NEPTUNE. "Planning Issues for Superfund Site Remediation," Hazardous Materials Control Vol. 4, No. 6, pp. 47-53 (November/December 1991).

6. IAEA. "Evaluating the Reliability of Predictions Made Using Environmental Transfer Models," Safety Series No. 100. International Atomic Energy Agency, Vienna (1989).

7. GILBERT, R.O. and J.C. SIMPSON. "Statistical Sampling and Analysis Issues and Needs for Testing Attainment of Background-Based Cleanup Standards at Superfund Sites," pp. 1-16. Proc. Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs. Pennsylvania State Univ., University Park, PA (November 1990).

8. GILBERT, R.O. and J.C. SIMPSON. "Statistical Methods for Evaluating the Attainment of cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media," PNL-7409 REV 1. Pacific Northwest Laboratory, Richland, WA (December 1992).

9. DOE. "Hanford Site Background: Part 1, Soil Background for Nonradioactive Analytes," DOE/RL-92-24, Revision 1, DRAFT. U.S. Department of Energy, Washington, DC (January 1993).

10. GILBERT, R.O. "Statistical Methods for Environmental Pollution Monitoring," Van Nostrand Reinhold, New York, NY (1987).

11. CONOVER, WJ. "Practical Nonparametric Statistics," 2nd edition. Wiley & Sons, New York, NY (1980).

12. EGER, K.J. "The Use of One-Sided Tolerance Tests for Surveys During Decontamination and Decommissioning," CONF-921029. Proceedings of the 8th Annual Oak Ridge Model Conference on Waste Management & Environmental Restoration. U.S. Department of Energy, Washington, DC (October 1992).

13. HELSEL, D.R. "Less Than Obvious: Statistical Treatment of Data Below the Detection Limit," Environmental Science and Technology 24: 1766-1774 (1990).

14. JOHNSON, R.A., S. VERRILL, and D.H. MOORE II. "Two Sample Rank Tests for Detecting Changes that Occur in a Small Proportion of the Treated Population," Biometrics 43:641-655 (1987).

15. ROSENBAUM, S. "Tables for a Nonparametric Test of Location," Annals of Mathematical Statistics 25:146-150 (1954).