Los Alamos

NATIONAL LABORATORY
Environmental Restoration Project
Los Alamos, New Mexico 87545
CLEAN UP LOS ALAMOS... faster, better, cheaper!

EPA REGION VI
WASTE
95 JAN 20 PM 4:23

Date: January 11, 1995
MS: M992
Phone: 505-665-4558
Refer to: EM/ER:95-J003

Ms. Barbara Driscoll
RCRA Permit Branch
US Environmental Protection Agency, Region 6
1445 Ross Ave., Suite 1200
Dallas, TX 75202-2733

Dear Barbara:

SUBJECT: GUIDANCE ON BACKGROUND COMPARISONS

Enclosed for your review is the Los Alamos National Laboratory (Laboratory) Environmental Restoration (ER) Assessment Council's position paper entitled, "Guidance on Making Comparisons to Natural Background Concentrations of Metals for the Los Alamos National Laboratory Environmental Restoration Project," dated January 9, 1995. Although you requested a copy of the references cited in this paper, the Longmire, et al. reference will not be ready until February 2. It will be sent to you then, under a separate letter. The Environmental Protection Agency references are also not included. One other reference cited in the paper, Gilbert 1987, is a book. Please let us know if you will need a copy of it. The Department of Energy and other Laboratory personnel are simultaneously reviewing this paper. Those comments are due back January 20, 1995. If at all possible, we would appreciate your comments by then as well. When all comments are resolved, we will distribute the paper to ER personnel for use in preparing Resource Conservation and Recovery Act Facility Investigation reports. The position paper will then be part of approved ER methodology and will be cited as such.

If you have any questions regarding this document, please call Alison Dorries at (505) 665-4791. Thank you.

Sincerely,

Tracy Glatzmaier
Environmental Restoration Project

TG:bp

Enc.: Position Paper

12569

Cy (w/o attach):
J. Jansen, EM/ER, MS M992
D. McInroy, EM/ER, MS M992
A. Dorries, TSA-11, MS K557
T. Taylor, LAAO, MS A316
C. Fesmire, LAAO, MS A316
RPF, MS M707

Guidance on Making Comparisons to Natural Background Concentrations of Metals for the Los Alamos National Laboratory Environmental Restoration Project

External Review Draft

January 9, 1995

1.0 Introduction

Comparisons of site data to background concentrations are needed as part of the RCRA corrective action process. This document summarizes the primary regulatory requirements relevant to comparing site data to background data. In addition, the approaches taken at other DOE facilities to characterize and statistically compare site data to background data will be briefly discussed. Sufficient detail is provided for data analysts to conduct statistical comparisons between site data and background data, and the key background comparison issues for the ER Project Management Staff and other interested parties are discussed.

The main purpose of this document is to provide guidance on the recommended statistical comparisons between Potential Release Site (PRS) data and naturally occurring concentrations of metals in soils. This guidance may also be applicable to background comparisons of certain anthropogenic compounds (e.g., radionuclides distributed from nuclear fallout or organic chemicals associated with urban activities).

The strategy at LANL has been to collect soil and tuff samples that are representative of LANL-wide background metal concentrations. These data form the default background data base for all PRSs. One requirement in assembling the LANL background data is to ensure comparability of the physical soil conditions where background constituents are detected relative to the assumptions inherent to the conceptual exposure model. For example, ecological risk may be related to constituent concentrations in the plant root zone. Ideally, PRS sample data should be compared to the appropriate subset of LANL background data. Typically, detailed information on the soil physical characteristics is not available, which means that most PRS soil data will be compared to the aggregate LANL soil background data. Another requirement is the background samples have to be analyzed by the same methods as the PRS samples. A third requirement is that the proposed background data should represent sites that are not impacted by LANL operations. This last requirement makes collection of PRS-specific background data more difficult, since the most PRS-specific soil conditions occur within the PRS. All of these issues are discussed in more detail in the LANL background data report (Longmire et al. 1994, 1142).

If comparisons to LANL background data yield inconclusive results relative to the No Further Action (NFA) criteria, the data analysis team may recommend that "site-specific" background data are collected. It is beyond the scope of this document to provide detailed guidance on the number of samples required for site-specific background comparisons. If site-specific background data seem warranted, data analysts are encouraged to arrange a meeting with the background

specialists that have been designated by the LANL ER Project Assessments and Earth Sciences Council Chairs.

2.0 Regulatory Literature Review

EPA guidance documents that support the CERCLA and RCRA programs provide specific information on how to design background studies and how to statistically compare site data to background data.

The CERCLA guidance document, "Guidance on Data Useability in Risk Assessment (Part A)" (EPA 1992, ????), recommends collecting background data prior to collecting site data. If the comparison of background data and site-derived data are not statistically different for a given constituent, that constituent is eliminated from further analyses. The CERCLA guidance suggests that the number of background samples collected from a site should be based on the "minimum detectable difference" procedure EPA (1989, ????). Data analysts unfamiliar with this approach should contact the statistical specialists that have been designated by the LANL ER Project Assessments Council Chair for more information.

The RCRA guidance document, "The RFI Guidance" (EPA 1989, 0088), is nearly mute on the subject of background comparisons except for a discussion of background comparisons for groundwater monitoring data. The RCRA groundwater statistical analysis guidance document (EPA 1989, 1141), referenced in the RFI guidance, presents statistical methods for comparing data derived from upgradient wells to downgradient well data. These statistical methods are codified in 40 CFR Part 264, "Statistical Methods for Evaluating Ground-Water Monitoring from Hazardous Waste Facilities: Final Rule", Federal Register Tues. Oct. 11, 1988.

The RCRA groundwater monitoring guidance document is applicable to making background comparisons for data derived from other media. The preface of EPA (1989, 1141) states:

> "This scenario **can be applied to other non-RCRA situations involving the same spatial relationships and the same null hypothesis**. The **explicit null hypothesis** for testing contrasts between means, or where appropriate between medians, is that the means between groups (here monitoring wells) are equal (i.e., **no release has been detected**), or that the group means are below a prespecified action level (e.g., the ground-water protection standard). Statistical methods that can be used to evaluate these conditions are described in Section 5.2 (**Analysis of Variance**), 5.3 (**Tolerance Intervals**), and 5.4 (Prediction Intervals)." **Bold added.**

The RCRA groundwater monitoring guidance suggests that the specific approach proposed by the owner/operator should be submitted to EPA for approval, especially where methods other than those presented in the guidance are used. The statistical methods presented below can be found in the analysis of variance and tolerance interval sections of the RCRA groundwater monitoring guidance (EPA 1989, 1141).

3.0 Background Comparison Approaches at Other DOE Facilities[1]

Most DOE facilities have funded a facility-wide background analysis of all potentially impacted media (soil, sediment, surface water, groundwater). Most DOE facilities have information on the soil horizon or other supporting data that would help define a more site-specific comparison to background. Lastly, there is no single statistical test used to compare site data to background data at DOE facilities.

Background comparison approaches at specific facilities include:

1) Hanford Site (Richland, WA): a background conceptual model has been developed. This model includes metals (including certain radionuclides) and considers transport between soil, sediment, surface water and groundwater.
2) Oak Ridge Reservation (Oak Ridge, TN): a statistically significant difference between site data and background data is required before including the constituent in a risk assessment.
3) Sandia National Laboratories (Albuquerque, NM): upper tolerance limits (UTL) of metals and radionuclides are calculated based on historical data collected within and outside of PRSs. These UTL calculations are made if there are an adequate number of "detects" reported by the laboratory, and either a parametric or nonparametric tolerance limit is calculated depending on the distributional properties of the analyte.

4.0 Proposed methods

Since background comparisons are used from site screening to corrective measure implementation under RCRA, data analysts must have statistical tools that can be applied to this broad range of decisions. This guidance document defines two classes of statistical methods for

[1] Based on information presented at the 1994 Technical Information Exchange (TIE) Workshop, which is sponsored by DOE-EM.

background comparisons that meet the needs of the RCRA decisions for the LANL ER Project. The first comparison is a "hot measurement" test, where site concentration data are compared to a statistic that represents the upper end of the background concentrations. The second comparison is a "distributional shift test," which determines if the mean of the site data are statistically greater than the mean of the background data. Used together or separately, these tests help demonstrate if there was a release at a PRS, and help define what risk consequence a release may have. Figure 1 illustrates the kinds of differences between site data and background data that are detected by these two methods.

The kind of decision that the background comparison supports will determine which test is more appropriate. The hot measurement test is more appropriate in the initial RFI screening when a single high value should trigger further analysis. The distribution shift test is more appropriate when extensive data are collected to support a risk assessment and a change in the average concentration should lead to further action at the site.

4.1 Common Data Analysis Issues

As certain methods depend on the statistical distribution of site data and background data, analysts should prepare graphical data displays. Box plots (see Figure 1), where the background data and the site data can be compared side-by-side, are most useful. Analysts should also consider using histograms and probability plots to graphically display results. These graphs provide the data analyst or the report reader with tangible evidence of the similarity or differences between site data and background data.

4.2 Hot measurement test

The hot measurement test defines the threshold value that represents the upper range of the background concentrations. This threshold can be exceeded by a background sample, because no matter what value is chosen for the threshold, there is a probability that a background measurement will exceed the hot measurement threshold. The frequency of false positive results is minimized by using a threshold that is statistically related to the higher background concentrations. The tolerance limit is such a value, and is one of the background comparison methods recommended in EPA (1989, 1141). A tolerance limit is a confidence limit on a percentile of the distribution. Given that the tolerance limit is a value that should be rarely exceeded, the 99th percentile is a reasonable value of the distribution to estimate. EPA recommends calculating an upper 95% confidence limit for the target percentile (EPA 1989,

1141). The Upper Tolerance Limit (UTL) for the 99th percentile at 95% confidence is calculated by using Eq. (1).

$$UTL_{0.99,0.95} = \text{mean} + \text{standard deviation} * k_{0.99,0.95} \quad (1)$$

The "k-factors" depend on the number of background samples, and complete tables of k-factors are published in EPA (1989, 1141) and Gilbert (1987, 0312). Table 1 presents some selected k-factors that represent the range of values used to compute the UTLs for LANL background soil samples. The background data must be either normally distributed or transformed to normality (e.g., by using log-transformation) to be able to apply Eq. (1). There are nonparametric methods available for calculating tolerance limits (e.g., as described in Gilbert [1987, 0312]), if data deviates sufficiently from normality to warrant use of nonparametric tolerance limits. In lieu of nonparametric tolerance limits, the analyst may trim outliers, when appropriate, from the distribution and calculate the UTL based on the trimmed mean and standard deviation.

The LANL soil background data were used to calculate the UTL values that are presented in Table 2. The soil concentration distributions that were skewed (median not roughly equal to the mean) were log-transformed. The UTLs calculated for the lognormal distribution were backtransformed into the original units to simplify comparisons with site data. No values were trimmed from the distributions to calculate the mean and standard deviation. Values below the laboratory detection limit, if four or fewer non-detects were reported for the analyte, were replaced by one-half of the detection limit (EPA 1989, 1141). The UTL was not calculated for any analyte with more than four non-detect values (>10% non-detects).

An alternative to using the UTL as the hot measurement threshold is the observed maximum concentration in the background data. The maximum is not recommended, because it will tend to underestimate the upper range of background when few background samples are available. If more data are added to the LANL background data, the maximum value will have to increase, but the UTL will, on average, not change. The maximum is <u>not</u> a recommended hot measurement threshold, given its extreme sensitivity to the background sample size. Analytes that are rarely detected are an exception to this general recommendation. For this limited subset of analytes, the maximum detected background concentration can be used as the hot measurement threshold.

Exceeding the UTL does not prove that a release occurred at the PRS. There is nominally a 1% chance that the UTL will be exceeded by each sample that is collected from the PRS (assuming

the PRS is at background). In addition, a typical metals suite will require comparison of 23 analytes to background. Given that we have not adjusted the probability values for these multiple comparisons[2], the results of the hot measurement test must be carefully evaluated. The concentrations that exceed the UTL must be considered in relation to human health or ecological screening levels. Also, the possibility of an exceedance due to an unusual, but naturally occurring soil matrix must also be considered. These factors will help determine what kind of further background analysis is warranted.

The hot measurement test is recommended if only a single statistical comparison to background is performed. The UTL is the simplest comparison and is functionally most similar to comparisons of site data to target risk levels or screening action levels (SALs).

4.3 Distributional shift test

The second background comparison test determines if the site data are systematically greater than background data. There are a number of statistical tests that can be used to determine if site data are significantly greater than background data. The Student's t-test is a parametric statistical two-sample test that tests whether the mean concentration of the site data is statistically greater than the mean concentration of the background data. The nonparametric equivalent to the t-test is the Wilcoxon's rank-sum test. The Wilcoxon's test pools site and background data into one aggregate set and then determines if the average rank of the site data are greater than the background data. The Wilcoxon's test is recommended when the site data consists of few samples or non-detects are frequent.

Whether a distribution shift should be tested is based on the number of samples available for this comparison. In general, there should be 10 site sample concentrations to compare to the background data to provide adequate confidence in detecting a shift. Thus, most data collected in RFI Phase I may not have adequate numbers of samples to warrant a distributional shift comparison.

A 95% confidence level is recommended to infer a significant result for the distributional shift test. Given that multiple comparisons will also be performed with the distributional shift test, the same statistical interpretation issues are also present. In addition, the human health and ecological consequences of above background PRS concentration data must be considered along with

[2] If the concentrations of the 23 metal analytes vary independently, the nominal 1% chance increases to a 21% chance that one of the 23 UTLs will be exceeded.

differences in metal concentrations between soil horizons. If site data were collected exclusively from a soil horizon that is naturally enriched in a metal, then these data would appear to be greater than the LANL background data.

5.0 References

EPA (US Environmental Protection Agency) 1989. "Ecological Assessment of Hazardous Waste Sites: A Field and Laboratory Reference," EPA/600 3-89/013, Environmental Research Laboratory, Corvallis, OR. (EPA 1989, ????)

EPA (US Environmental Protection Agency) 1989. "Interim Final RCRA Facility Investigation (RFI) Guidance, Volume I of IV, Development of an RFI Work Plan and General Considerations for RCRA Facility Investigations," EPA/530-SW-89-031, OSWER General Directive 9502.00-6D, Office of Solid Waste, Washington, DC. (EPA 1989, 0088)

EPA (US Environmental Protection Agency) 1989. "Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities. Interim Final Guidance," Office of Solid Waste, Waste Management Division, US Environmental Protection Agency, Washington DC. April 1989. (EPA 1989, 1141)

EPA (US Environmental Protection Agency) 1992. "Guidance for Data Useability in Risk Assessment (Part A)," Office of Emergency Remedial Response, US Environmental Protection Agency, Washington DC. April 1992. (EPA 1992, ????)

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Von Nostrand Reinhold Company Inc. New York. 320 pp. (Gilbert 1987, 0312)

Longmire, P., S. Reneau, P. Watt, J. Gardner, and C. Duffy. 1994. Geomorphology, Pedogenesis, and Geochemistry of Background Bandelier Tuff and Selected Soil Profiles, Los Alamos, New Mexico, 1994. (Longmire et al. 1994, 1142)

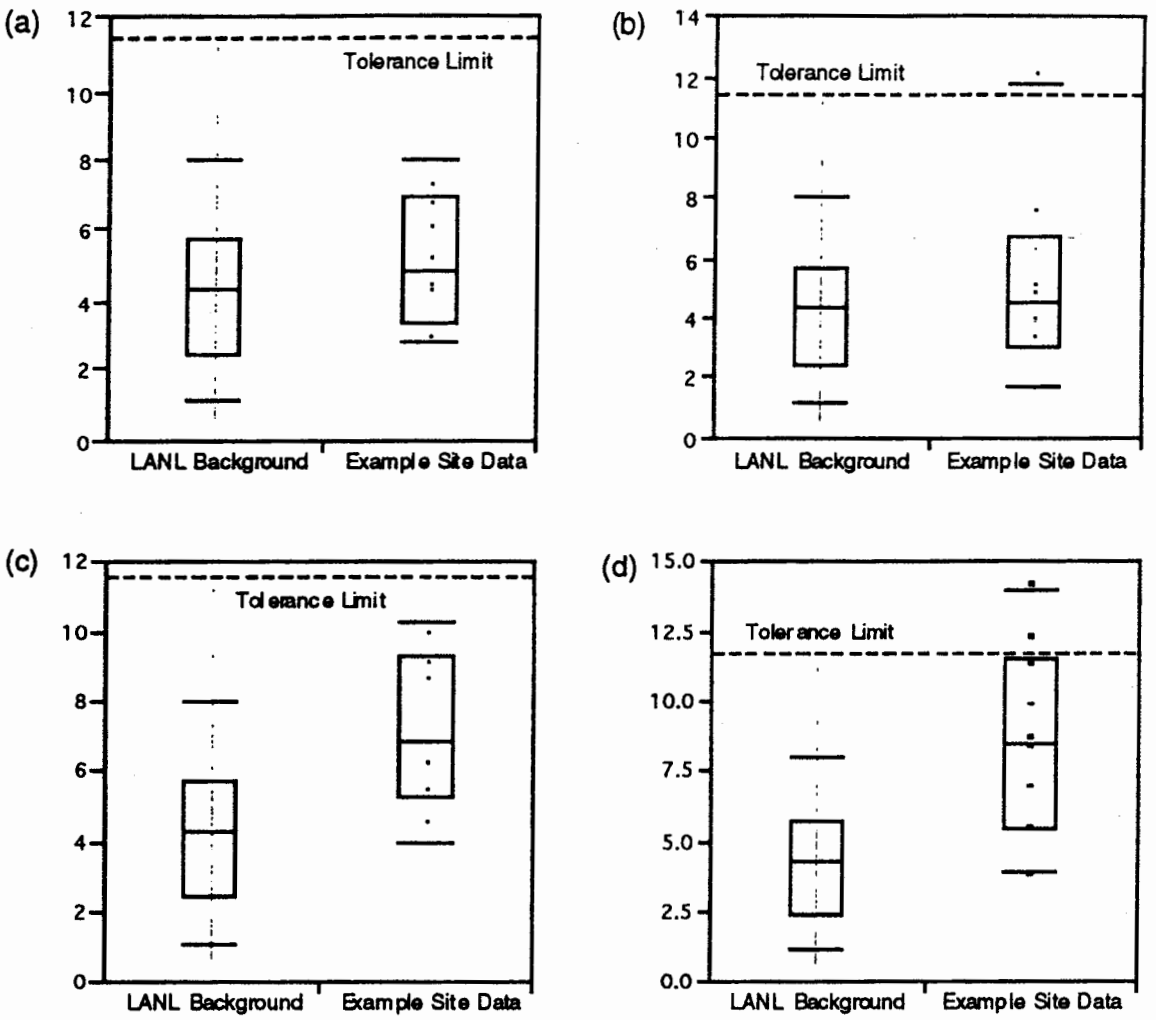

Figure 1. Box plot comparisons of example site data to LANL background data.
(a) Site data is within the range of background: no distribution shift or hot measurements (i.e., values greater than the tolerance limit).
(b) Site data fail hot measurement test: 1 of 11 site arsenic concentration values exceeds the tolerance limit of 11.6 mg/kg.
(c) Site data show a distribution shift: the Wilcoxon rank sum test show that site data tend to be greater than the background data.
(d) Site data show a distribution shift and fail the hot measurement test: 2 of 10 site arsenic concentrations exceed the tolerance limit of 11.6 mg/kg and the site data tend to be greater than the background data.

Table 1 - Selected "k-factors" that are used to calculate Upper Tolerance Limits (* reprinted from Gilbert 1987, 0312).

Number of background samples	$k_{0.99,0.95}$
45	2.897 *
46	2.8902
47	2.8834
48	2.8766
49	2.8698
50	2.863 *

Table 2 - List of Upper Tolerance Limits (UTL) for LANL soil background data.

Analyte	SAL (mg/kg)	Mean *	Standard deviation	UTL $_{99\%,0.95}$	N	N > DL
Aluminum		19000	13800	58900	47	45
Antimony	32	2.45	0.36	2.5 (MAX)	46	0
Arsenic		4.4	2.5	11.6	46	45
Barium (LT)	5,600	161	129	1140	47	45
Beryllium		1.15	0.75	3.31	47	45
Cadmium	80	0.39	0.54	2.7 (MAX)	47	3
Calcium (LT)		5790	12500	54400	47	45
Chromium (Total) †		11.7	7.8	34.2	47	45
Cobalt (LT)		15.2	7.6	51.1	47	45
Copper	3,000	5.3	3.6	15.7	47	43
Iron		14500	7320	35600	47	45
Lead	400	15.0	8.3	39.0	47	42
Magnesium (LT)		2920		16100	47	45
Manganese	11,000	343	238	1030	47	45
Mercury	24	0.05	0.01	0.1 (MAX)	48	1
Molybdenum	400	NA	NA	NA	NA	NA
Nickel	1,600	9.7		26.7	47	43
Potassium		2420	1304	6180	50	50
Selenium	400	0.43	0.41	1.7 (MAX)	46	22
Silver	400	NA	NA	NA	NA	NA
Sodium		577	453	1880	47	45
Strontium	48,000	NA	NA	NA	NA	NA
Thallium	6.4	0.27	0.24	0.9 (MAX)	45	20
Vanadium	560	25	14	66	47	45
Zinc	24,000	41	21	101	47	45
^{40}Potassium (1)		21.6	5.07	36.1	50	50
^{90}Strontium (2)	5.9	0.34	0.27	1 (MAX)	29	29
^{137}Cesium (2)	4	0.42	0.31	1.4 (MAX)	79	79
^{232}Thorium (1)	5	1.71	0.34	2.68	50	50
^{234}Uranium (1)	86	1.21	0.29	2.03	50	50
^{235}Uranium (1)	18	0.052	0.012	0.088	50	50
^{238}Uranium (1)	59	1.14	0.27	1.90	50	50
^{238}Plutonium (2)	20	0.0013	0.0024	0.014 (MAX)	76	62
$^{239/240}$Plutonium (2)	18	0.0083	0.0079	0.052 (MAX)	88	85

* Concentration values <DL (detection limit) were replaced by 1/2 of the DL
† - SAL for Chromium-III is 80,000 mg/kg and for Chromium-VI is 400 mg/kg
LT - UTL is based on log transformed data
NA - data not available for LANL background
MAX - Maximum value is reported, rather than the UTL
(1) - Data are converted from elemental concentrations reported in the LANL background report. Units are in pCi/g.
(2) - Data are from the Environmental Surveillance Reports (1974 - 1990). Units are in pCi/g.

DRAFT

Statistical Comparisons to Background

Part I

Los Alamos National Laboratory

Environmental Restoration Project
Assessments Council

March 28, 1995

DRAFT

INTRODUCTION

The main purpose of this policy is to provide guidance to the Los Alamos National Laboratory (the Laboratory) Environmental Restoration (ER) personnel on recommended statistical comparisons between Potential Release Site (PRS) data and naturally occurring concentrations of metals (including radionuclides) in soils. Any background comparison approach consists of two components. The first is to assemble a defensible set of background data. This document describes comparisons to the Laboratory wide set of background data collected by Longmire (Longmire at al. 1994, 1142). The second is the selection of the statistical method(s) used to compare site data to background data. Two statistical methods are presented. In the first method, site concentration data are compared with a statistic representing the upper percentile of background concentrations, the upper tolerance limit (UTL). The UTL is used as a screening value, or extreme (hot) value, to determine if a significant release has occurred at the site. The second is a group of methods designed to detect a distributional shift between site data and background data. Although guidelines for the application of these methods are presented in this document, each data analysis report should briefly describe the statistical analysis method chosen and justify its application to the data in question. In PRS data, particular attention should be paid to background comparisons of beryllium and arsenic, because background concentrations of these elements exceed risk-based screening action values. Each data analysis report should also justify the use of Laboratory-wide background concentration data, or present the rationale for using site-specific background concentration data.

Comparisons of PRS data with background concentrations are needed as part of the Resource Conservation and Recovery Act (RCRA) corrective action process. The guidance on statistical comparisons between PRS data and naturally occurring concentrations of metals (including radionuclides) in soils provided by this document may also be applicable to background comparisons of certain anthropogenic compounds (e.g., radionuclides distributed from nuclear fallout or organic chemicals associated with urban activities). It should be emphasized that the approach presented in this document should be applied to naturally-occurring metals only.

REGULATORY LITERATURE REVIEW

The Environmental Protection Agency (EPA) guidance documents supporting the Comprehensive Environmental Response Compensation and Liability Act (CERCLA) and RCRA

programs provide specific information on how to design background studies and how to statistically compare site data with background data.

The CERCLA document, *Guidance on Data Useability in Risk Assessment (Part A)* (EPA 1992, 1166), recommends collecting background data prior to collecting site data. If the comparison of background data with site-derived data for a given constituent does not show a difference statistically, that constituent is eliminated from further analyses. The CERCLA guidance also suggests basing the number of background samples collected from a site on the "minimum detectable difference" procedure (EPA 1989, 0303). Data analysts unfamiliar with this approach should contact the statistical specialists designated by the Assessments Council Chair.

Background comparisons for groundwater monitoring data are addressed in the RCRA document, *The RFI Guidance* (EPA 1989, 0088). Methods for comparing data derived from upgradient wells with data from downgradient wells is presented in the RCRA groundwater statistical analysis document (EPA 1989, 1141), referenced in the EPA RCRA facility investigation (RFI) guidance (EPA 1989, 0088). These statistical methods are codified in 40 CFR Part 264, *Statistical Methods for Evaluating Ground-Water Monitoring from Hazardous Waste Facilities: Final Rule* Federal Register Tues. Oct. 11, 1988.

Statistical methods used for background comparisons of groundwater can be applied to background comparisons for data from other media as stated in the preface of EPA (1989, 1141):

> "This scenario **can be applied to other non-RCRA situations involving the same spatial relationships and the same null hypothesis.** The **explicit null hypothesis** for testing contrasts between means, or where appropriate between medians, is that the means between groups (here monitoring wells) are equal (i.e., **no release has been detected**), or that the group means are below a prespecified action level (e.g., the ground-water protection standard). Statistical methods that can be used to evaluate these conditions are described in Section 5.2 (**Analysis of Variance**), 5.3 (**Tolerance Intervals**), and 5.4 (Prediction Intervals)." **Bold added for emphasis.**

The RCRA groundwater monitoring guidance suggests that the specific approach proposed by the owner/operator should be submitted to EPA for approval, especially where methods other than those presented in the guidance are used. Statistical methods presented below are consistent with those found in the analysis of variance and tolerance interval sections of the RCRA groundwater monitoring document (EPA 1989, 1141).

DRAFT

BACKGROUND COMPARISON APPROACHES AT OTHER DOE FACILITIES

Based on information presented at the 1994 Technical Information Exchange (TIE) Workshop, most DOE facilities have funded a facility-wide background analysis of all potentially impacted media (soil, sediment, surface water, groundwater). Most have information on the soil horizon or other data that support a site-specific comparison with background. No single statistical test for comparing site data with background data is used throughout the DOE facilities studied.

Background comparison approaches at specific facilities include:

1) Hanford Site (Richland, WA): a background conceptual model has been developed. This model includes metals (including certain radionuclides) and considers transport between soil, sediment, surface water and groundwater.

2) Oak Ridge Reservation (Oak Ridge, TN): a statistically significant difference between site data and background data is required before including the constituent in a risk assessment.

3) Sandia National Laboratories (Albuquerque, NM): upper tolerance limits (UTL) of metals and radionuclides are calculated based on historical data collected within and outside of PRSs. UTL calculations are made if there is an adequate number of "detects" reported by the analytical laboratory. The UTL is calculated using a method dependent on the distributional properties of the analyte.

LABORATORY BACKGROUND DATA

The strategy at the Laboratory has been to collect samples that are representative of Laboratory-wide background metal concentrations in soil and tuff (see Longmire at al. 1994, 1142) for comparison with PRS data. (Readers interested in more detail on the statistical distribution of naturally-occurring metals are referred to the Appendix of this paper.) Longmire's data, representing Laboratory-wide variation in soil and volcanic tuff, are used as the default background data for making comparisons in the initial RFI screening assessment process. The Laboratory-wide background data were collected from sites representing the range of soil conditions observed at the Laboratory. At present, Longmire has 47 soil samples (A, B or C soil horizon) analyzed by EPA SW 846 methods, and 50 soil (A, B or C soil horizon) and 38 tuff samples analyzed by Instrumental Neutron Activation Analysis (INAA). The INAA data represent

DRAFT

total elemental concentrations and are useful for background comparisons only at selected PRSs with data using a comparable method of analysis. Additional soil, sediment and volcanic tuff background analyses will be added to the Laboratory-wide background data base during 1995. The new data will be made available to ER Project personnel and other interested parties as soon as validated data are received.

Variation in certain elements makes comparisons with Laboratory-wide background less valid than comparisons with site-specific background concentrations. For example, due to natural variability in element concentrations of Bandelier Tuff, background soils from Technical Areas (TAs) at the east and west ends of the Laboratory are likely to be enriched, or depleted, in certain elements relative to Laboratory-wide background values. Bandelier Tuff was derived from a zoned magma chamber in which some elements were concentrated at the top and others were concentrated at the bottom. During the eruption resulting in deposition of the Bandelier Tuff, magmas at the top of the chamber were erupted first. Consequently, those elements that were concentrated in the upper magmas are in higher abundance at TAs to the east of the Laboratory because the Bandelier Tuff subunit exposed in that region represents the earlier-erupted magmas. Thus, soils derived from tuff located to the east of the Laboratory are likely to have higher abundance of certain elements than Laboratory-wide background soils. For example, at TA-33 uranium in background soils is at higher levels than Laboratory-wide background because it is more abundant in the stratigraphically lowest units of the Tschirege Member of the Bandelier Tuff. Thus, it is important to develop site-specific background for some sites.

All users of background data must follow some simple guidelines to ensure that site data is being compared with Laboratory background data in a scientifically valid manner. A primary requirement is that PRS data and background data be collected from soil or tuff having the same physical properties. If PRS data are collected from sediments, background data should also come from sediments. For example, if site data were collected exclusively from a soil horizon naturally enriched in a metal, these data would appear to be greater than Laboratory background data. A second requirement is that PRS samples be assayed by the same analytical methods as used for background samples.

Each data analysis report should demonstrate that the above guidelines were considered before using Laboratory-wide background distributions. Failure to meet these guidelines may be an indication that site-specific background be collected or that a subset of Laboratory-wide background data be compared with PRS data. Lack of data for a particular analyte could be justification for the collection of site-specific background data. Before recommending the

collection of site-specific background for an analyte, data analysts should confirm whether the analyte in question could be present by looking at historical information. On the other hand, lack of defensible historical information might justify the collection of site-specific background data.

Should site-specific background data seem warranted, data analysts are encouraged to seek technical advice from the background specialists designated by the Laboratory ER Project Assessments and Earth Sciences Council Chairs. These specialists will be able to provide detailed guidance on the location and number of samples required for site-specific background comparisons.

PROPOSED STATISTICAL METHODS

Because background comparisons are used to make decisions throughout the RCRA process from site screening to corrective measures implementation, data analysts must have statistical methods that can be applied over a broad range of decisions. This guidance defines two statistical methods for background comparisons. Both methods meet the requirements for RCRA decision-making. In the first method, the "hot measurement" test, site concentration data are compared with a statistic representing the upper percentile of background concentrations. In the second method, the "distributional shift test," the mean of site data is compared with the mean of background data to determine if the former is statistically greater than the latter. Used together, or separately, these tests help demonstrate if a release at a PRS occurred, and help define what risk consequence the release may have. Figure 1 illustrates the differences between site data and background data detected by the two methods.

The decision to be supported by the background comparison determines which test is more appropriate. In an initial RFI screening, when a single high value should trigger further analysis, the hot measurement test is typically more appropriate. When extensive data are collected to support a risk assessment and a change in the average concentration should lead to further action at the site, the distributional shift test may be more appropriate. A data analysis report should clearly indicate the rationale for selecting a statistical method that differs from those presented in this guidance document.

It is emphasized that the level of effort spent in evaluating potential differences between PRS data and background data should be related to the site-specific information available. For example, if historical information indicates that beryllium was released at a firing site, the potential

DRAFT

differences between the beryllium concentration data at the firing site and Laboratory or site-specific background data should be carefully evaluated and presented in the data analysis report.

COMMON DATA ANALYSIS PROCEDURES

Because certain methods depend on the statistical distributions of site and background data, analysts are encouraged to prepare graphical data displays to facilitate the communication of the results of data comparisons. Box plots (see Figure 1), in which background data and site data can be compared side-by-side, are most useful. Analysts should also consider using histograms and probability plots. These graphs provide tangible evidence of the similarity or differences between site data and background data.

HOT MEASUREMENT TEST

The hot measurement test defines a threshold value that represents high background concentrations. No matter what parameters are chosen to define the threshold, there exists a probability that a background measurement will exceed the hot measurement threshold. The frequency of false positive results is minimized by using a threshold statistically related to higher background concentrations. The confidence limit on a percentile of the distribution, termed the tolerance limit, is such a value, and is one of the background comparison methods recommended by EPA (1989, 1141). The Laboratory has selected the 99th percentile for calculating the upper tolerance limit, based on the general guidance in the RCRA groundwater document. If the underlying distributional model is correct, the upper tolerance limit based on the 99th percentile is rarely exceeded. EPA recommends calculating an upper 95% confidence limit for the target percentile (EPA 1989, 1141). The Upper Tolerance Limit (UTL) for the 99th percentile at 95% confidence can be calculated using Equation (1).

$$UTL_{0.99,0.95} = \text{mean} + \text{standard deviation} * k_{0.99,0.95} \qquad (1)$$

The "k-factor" depends on the number of background samples, and complete tables of k-factors are published in EPA (1989, 1141) and Gilbert (1987, 0312). Table 1 presents k-factors selected to represent the range of values used to compute UTLs for Laboratory background soil samples. To apply Equation (1), the background data must be normally distributed or transformed to normality (e.g., by using log-transformation). If data deviates sufficiently from normality, nonparametric methods for calculating tolerance limits should be considered (e.g., as described in

DRAFT

Gilbert [1987, 0312]). Alternatively, when appropriate, the data analyst may trim outliers from the distribution and calculate the UTL based on the trimmed mean and standard deviation.

The Laboratory soil background data were used to calculate the UTL values presented in Table 2. As discussed in section 4.0, use of the default is emphasized, or Laboratory-wide background data should be justified for the specific background comparison being performed. Table 2 UTL values should not be used without considering the guidelines discussed in section 4.0. A minimalist approach to data preparation was used for the initial UTL calculations. Because some soil concentration data were sufficiently skewed, a log-transformation was applied to improve the fit of these data to a normal distribution. (Readers interested in the details of these data distributions are referred to the Appendix of this policy paper.) The UTLs calculated for the lognormal distribution were backtransformed into original units to simplify comparisons with site data. No values were trimmed from the distributions used to calculate the mean and standard deviation. If four or fewer non-detects were reported for an analyte, values below the laboratory detection limit were replaced by one-half of the detection limit (EPA 1989, 1141). The UTL was not calculated for any analyte having more than four non-detect values (>10% non-detects).

The observed maximum concentration in the background data is an alternative to using the UTL as the hot measurement threshold. However, when few background samples are available, using the maximum will result in an underestimation of the upper background percentile. In general, the sample maximum (for "n" samples) is an estimate of the $\left[\frac{n-0.5}{n}\right]$th percentile. Thus, if 10 samples are collected, the sample maximum is an estimate of the 95% percentile. As more Laboratory-wide background data become available, the maximum value will increase, but the UTL will typically not change. Because the maximum is extremely sensitive to background sample size, it is <u>not</u> recommended for use as a hot measurement threshold. Rarely detected analytes, which include: antimony, cadmium, mercury, selenium and thallium, are an exception to this general recommendation. For this limited subset of rarely detected analytes, the maximum detected background concentration can be used as the hot measurement threshold.

Exceeding the UTL does not prove that a release occurred at a PRS. Assuming the PRS is at background and the statistical model is correct, there is a 1% probability that the 99th percentile will be exceeded by each sample collected from the PRS. Furthermore, a typical metals suite requires comparison of 23 analytes to background. If the concentrations of the 23 metal analytes vary independently, the 1% probability that each PRS sample exceeds the 99th percentile increases to a 21% probability that at least one of the 23-99th percentiles will be exceeded in a

single sample. Additionally, given that the probability values for these multiple comparisons have not been adjusted, the overall level of confidence for 23 analytes will be substantially less than 95%. Consequently, the results of a hot measurement test must be carefully evaluated. The possibility of exceeding a UTL due to an unusual, but naturally occurring, soil matrix is a further consideration.

The results of the UTL comparison should also be evaluated relative to potential human health or ecological screening levels. Some analytes, arsenic and beryllium in particular, represent a special case. Because background levels at the Laboratory exceed risk-based screening levels, no screening action levels (SALs) for beryllium or arsenic have been calculated for the Laboratory ER Program. Seven metals (antimony, barium, chromium, lead, manganese, thallium, and vanadium) and one radionuclide (thorium-232) have background UTLs that appear close to their SAL values. The UTLs of this group represent a significant fraction of the SAL (8 to 50%). If, in a comparison similar to a multiple constituent test, each of these metals had a concentration equal to the SAL, the total of the metal UTLs divided by the corresponding SAL (in effect normalizing the UTL to SAL ratio) would equal 89%. Thus, the concentrations of most naturally-occurring metals are significantly lower than their respective SALs.

Both the multiple constituent evaluation and the UTL-to-SAL comparison will help determine what level of effort should be expended to evaluate deviations from background. For most naturally-occurring metals, when only a single statistical comparison to background is performed, the UTL will be adequate because probability levels are not compromised. Under this circumstance, the UTL is the simplest comparison and is functionally most similar to comparisons of site data to target risk levels or SALs.

DISTRIBUTIONAL SHIFT TEST

The distributional shift test is used to determine if site data are systematically greater than background data. Several types of distributional shift tests are available. The Student t-test is a parametric, statistical, two-sample test that determines whether the mean concentration of site data is statistically greater than the mean concentration of background data. The Wilcoxon rank-sum test is the nonparametric equivalent to the t-test (Gilbert 1987, 0312; Gilbert and Simpson 1992, 0974). The Wilcoxon test pools site and background data into one aggregate set and determines if the average rank of site data is greater than that of the background data. The Wilcoxon test is recommended when site data consists of few samples or when non-detects are frequent. Another useful distributional shift test is the Quantile test (Gilbert and Simpson 1992,

DRAFT

0974). This test, which compares the upper quantile (e.g., 25%) of background data with that of PRS data, is more capable of detecting a difference when only a small number of PRS concentrations are elevated. The Quantile test is the most useful distributional shift test for PRSs at which samples from a release represent a small fraction of the overall data collected at the PRS because it does not artificially reduce statistical significance. For example, to detect contamination from historical spills at unknown locations, an RFI work plan may call for samples to be collected from a grid. Most sample results show no contamination, but those in or near spill locations show elevated concentrations.

Use of the distributional shift test is dependent on the number of samples available for comparison. In general, at least 10 sample concentrations for comparison with background data are needed to provide adequate confidence for detecting a shift. Frequently, in RFI Phase I, inadequate numbers of samples are collected to warrant a distributional shift comparison.

To infer a significant result in a distributional shift test, a 95% confidence level is recommended. Given that multiple comparisons will be performed with the distributional shift test, the same statistical interpretation issues cited above for the hot measurement test are also present. In addition, the human health and ecological consequences of a PRS concentration data above background must be considered along with differences in metal concentrations between soil horizons. In particular, multiple comparison tests with SALs must be performed and ecological SAL comparisons must be made.