Title: Statistical Methods for Background Comparisons 1998

Author(s): Dewart, Jean M.

Intended for:
NMED
Report
Environmental monitoring and surveillance
Remediation
Reading Room
Consent

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy.  Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Los Alamos Environmental Restoration Records Processing Facility

ER Record I.D.# 59596

# LOS ALAMOS NATIONAL LABORATORY ENVIRONMENTAL RESTORATION

## Records Processing Facility
## ER Records Index Form

*ER ID NO.* 59596 *Date Received:* 11/13/98 *Processor:* YCA *Page Count:* 25

*Privileged:. (Y/N)* N *Record Category:* P *Record Package No:* 0

*FileFolder:* N/A

*Correction: (Y/N)* N *Corrected No.* 0 *Corrected By Number:* 0

*Administrative Record: (Y/N)* Y

*Refilmed: (Y/N)* N *Old ER ID Number:* 0 *New ER ID Number:* 0

*Miscellaneous Comments:*

N/A

*THIS FORM IS SUBJECT TO CHANGE. CONTACT THE RPF FOR LATEST VERSION. (JUNE 1997)*

25

59596

# STATISTICAL METHODS FOR BACKGROUND COMPARISONS

DRAFT

ER ID # 59596

Los Alamos National Laboratory
Environmental Restoration Analysis and Assessment Focus Area

September 30, 1998

## INTRODUCTION

The purpose of this technical paper is to provide guidance to Los Alamos National Laboratory (LANL/the Laboratory) Environmental Restoration (ER) Project personnel on the ER Project's approach to conducting background comparisons.

The background comparison approach consists of two steps. The first step is the assembly of a defensible set of background data. The Laboratory-wide set of background analytical data from samples of soils, sediment and tuff is summarized in "Inorganic and Radionuclide Background Data for Soils, Canyon Sediments and Bandelier Tuff at Los Alamos National Laboratory." (Ryti et.al. 1998, 58093). This document presents a simple decision logic to select geologically defensible subsets of these data. All ER Project reports that evaluate differences from background will justify the use of Laboratory-wide background concentration data or present the rationale for using site-specific background concentration data. The second step is the selection of the statistical method(s) used to compare site data with background data. Two statistical methods are presented. The first compares the site concentration data with a background value (BV), a statistic (or detection limit) representing the largest concentration representative of natural background concentrations. 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 ER Project report that includes background comparisons will briefly describe the statistical analysis method chosen and justify its application to the data in question. Background comparisons should support revisions to the conceptual site model. In particular, background comparisons provide the basis for understanding the nature of inorganic contamination. Other analyses are also required to develop an understanding of contamination extent.

The organization of this paper includes the following general sections: (1) the summary of regulations and guidance governing statistical comparisons to background, (2) the selection of background data for (a) inorganics, and (b) radionuclides, (3) the recommendation of methods for background comparisons.

## 1. SUMMARY OF REGULATIONS AND GUIDANCE GOVERNING STATISTICAL COMPARISONS TO BACKGROUND

The EPA guidance documents supporting the Comprehensive Environmental Response Compensation and Liability Act (CERCLA) and Resource Conservation and Recovery Act (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 for Data Usability in Risk Assessment (Part A)* (EPA 1992, ER ID 54947), 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 evaluation. The CERCLA guidance also suggests that the number of background samples collected from a site be based on the "minimum detectable difference" procedure (EPA 1989, ER ID 54947). Data analysts unfamiliar with this approach should contact the statistical specialists within the ER Project's Data Analysis and Assessment Team.

Background comparisons for groundwater monitoring data are addressed in the RCRA document, *The RFI Guidance* (EPA 1989, ER ID 08794). Methods for comparing data derived from upgradient wells with data from downgradient wells are presented in the RCRA groundwater statistical analysis document (EPA 1989, ER ID 54946). 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 the RCRA groundwater statistical analysis document (EPA 1989, ER ID 54946):

> "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)."

The RCRA groundwater monitoring guidance states 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 statistical analysis document (EPA 1989, ER ID 54946).

## 2. Selection of Background Data Sets

### (a) Selection of Background Data for Inorganics

Selection of the appropriate Laboratory background data set(s) for making statistical background comparisons is essential for potential release site (PRS) decision-making. PRS decisions are ultimately based on samples collected from a number of primary geomorphic units, including mesa top, hill or canyon slope, and canyon bottoms. In addition there are subdivisions within the primary geomorphic units. For example, the geomorphic unit designated as canyon slopes is a mixture of mesa top soils and Bandelier Tuff. Although not inclusive of all Laboratory geomorphic units, existing Laboratory-wide background data include samples of mesa top soils, Bandelier Tuff, and canyon sediments. The purpose of this section is to guide the selection of appropriate subsets of those background data.

To support RCRA facility investigation (RFI) data review or other data analyses; background data are often selected after site characterization samples have been collected. However, background data selection should also be considered in planning for sampling. The planning team should consider what quality of background data is needed to meet their specific sampling objectives. Members of the Data Analysis and Assessment (chemistry, statistics) and Physical Modeling and Characterization (pedology, geology, geochemistry, geomorphology, and stratigraphy) Teams should be consulted to provide guidance on the selection and uses of background data.

Caveat: Comparability of Analytical Methods

The sample preparation and analytical methods used for all LANL background sample analyses are listed in Inorganic and Radionuclide Background Data for Soils, Canyon Sediments, and Bandelier Tuff at LANL (Ryti et.al. 1998, ER ID 58093). Selection of comparable methods for site samples should be made before sample collection and analysis. Consultation with a trained chemist is recommended for determination of appropriate analytical methods. In the event that different methods were chosen for the site samples, determine if the methods are comparable to those used for LANL background samples. The conclusions regarding comparability are documented in RCRA facility investigation (RFI) data review and other ER Project reports.

Be aware that there are two distinct data sets for some analytes (potassium, thorium, and uranium) from LANL background. Samples that underwent complete sample dissolution or analysis are identified as "total". For example, there are separate soil background data sets for the analytes identified as "Uranium" and "Total-Uranium" in the LANL Background document (Ryti et.al. 1998, ER ID 58093). These background data sets can be identified by the BKGD_DATA_SET_REF codes of 'U_SOIL' or 'U_TOTAL_SOIL' in the FIMAD table

BKGD_DATA_SET_INFO. The typical ER Project sample preparation by SW846 Method 3050A was used on the samples identified as "Uranium". Sample preparation by hydrofluoric acid digestion to digest the silicates in the soil media was used on the samples identified as "Total-Uranium." Choice of the appropriate background data set depends on the preparation method used for the site data. For example, there are separate tuff background data sets for "Potassium" and "Potassium-TOTAL". The analytical method of ICPES was used on samples identified as "Potassium". "Potassium-TOTAL" concentrations were determined by instrumental neutron activation analysis (INAA). Choice of the appropriate background data set depends on the analytical method used for the site data.

Caveat: Comparability of Analytical Detection Limits

In addition to analytical methods, it is important to request analytical detection limits for site analyses that will produce results which are comparable to those obtained for the LANL background data set(s), especially for analytes which are rarely detected (e.g.: antimony, thallium, mercury). Nondetected chemicals that are reported at detection limits above background values are problematic for this reason. All site results (concentrations of detected chemicals and detection limits of nondetected chemicals) are compared to background values. Additionally, evaluation of differences between PRS and background concentrations is more straightforward when detection limits are comparable.

Caveat: All Background Tuff Samples were collected from Unweathered Tuff

If a tuff sample from a PRS is identified as weathered, a geologist or geochemist should be consulted to verify that the weathered tuff sample from the PRS is comparable to the unweathered tuff from LANL background. In some cases, it may be more appropriate to compare the PRS samples from weathered tuff to soil or canyon sediment background data depending on the sample locations and characteristics.

**Decision Process:** The process for selecting the most appropriate Laboratory background data set is summarized in Figure 1. The selection of background data sets is based on the sampling media groups used in the LANL Background document (Ryti et.al. 1998, 58093). In addition to the decision points shown in Figure 1 and discussed below, it is essential that comparable sample preparation and analytical methods be used for background and PRS samples and that detection limits are adequate, as discussed above.

The LANL background data sets described below are available to data analysts in FIMAD table BKGD_DATA_SET_INFO.

### Decision 1. Were Pajarito Plateau soils[1] and/or (geological) fill material sampled at the potential release site (PRS)?

**"No" Decision.** If Pajarito Plateau soils and/or fill material were not sampled, move on to decision 2, which pertains to Bandelier Tuff.

**"Yes" Decision.** Any soil samples, including identified soil horizons (A, B, C) and/or (geological) fill material, are compared to the combined set of Laboratory-wide background soil samples from all soil horizons, referred to as the soil background data.

The term "soil" refers to material overlying intact bedrock that has been subject to soil-forming processes such as the addition of organic matter, the vertical translocation of clay-sized particles, or the development of ferric oxyhydroxides. Thus, soils are the typical surficial material on mesa tops and hillslopes, and are widespread in canyon bottoms. Soils across the Laboratory are highly variable spatially and in complexity due to variation in ages of soil parent material. Most PRSs on mesa tops consist of a mixture of native soils and (geological) fill material. The amount of fill material can vary (0 to 100%). Geological fill material refers to fill derived from geological media.

[1] Because most Laboratory PRSs are located on the Pajarito Plateau, Pajarito soil samples form the bulk of the soil samples included in the Laboratory-wide background soil database. One exception is Fenton Hill, which is located in the Jemez Mts.

Fill typically consists of disturbed soils with crushed Bandelier Tuff, but other rock types also may be present. Soil consists of layers or horizons of mineral and/or organic matter of variable thickness that parallel the land surface and differ from their parent material in morphological, physical, chemical, and mineralogical properties and in biological characteristics. Soil horizons are identified by a master horizon designation for the soil samples in the LANL background data. These horizons are indicated for data analytes in the field MEDIA_GROUP of the FIMAD table BKGD_DATA_SET_REF.

**Decision 2. Was Bandelier Tuff sampled?**

**"No" Decision.** If the Bandelier Tuff was not sampled, move on to decision 4 that addresses Laboratory background canyon sediment data.

**"Yes" Decision.** For the purpose of statistical background comparisons, the stratigraphic units have been combined into more general groups as listed in Decision 3.

Bandelier Tuff (Tshirege Member) rock units and additional stratigraphic units can be identified in the field by mapping and/or by evaluating core samples. Data from the individual stratigraphic units are summarized in the ER Project background report *Natural Background Geochemistry and Statistical Analysis of Selected Soil Profiles, Sediments, and Bandelier Tuff, Los Alamos, New Mexico* (Longmire et al. 1995, ER ID 52227). The stratigraphic units for the tuff background samples are indicated for data analytes in the field MEDIA_GROUP of the FIMAD table BKGD_DATA_SET_REF. Be aware that all tuff background samples were collected from unweathered stratigraphic sections.

Continue to Decision 3 for determination of the appropriate data set for the specific rock units.

Decision 3.

**Was Qbt 2, Qbt 3, and/or Qbt 4 sampled?**

**"Yes" Decision.** Units Qbt 2, Qbt 3, and Qbt 4 are the upper Bandelier Tuff cooling units or glassy tuffs which underlie all mesa top PRSs. If site tuff samples were taken from Qbt 2, Qbt 3 and/or Qbt 4, compare site samples to combined set of all Laboratory-wide background samples from units Qbt 2, Qbt 3, and Qbt 4.

**Was Qbt 1v sampled?**

**"Yes" Decision.** Cooling Unit 1v of Bandelier Tuff should only be encountered in deep drilling investigations. If site tuff samples were taken from unit Qbt 1v, compare to the full set of unit Qbt 1v samples from LANL background.

**Was Qbt 1g, Qct, and/or Qbo sampled?**

**"Yes" Decision.** Qbt 1g (Bandelier Tuff cooling unit 1g), Qct (Cerro Toledo), and Qbo (Otowi member of Bandelier tuff) are the lower Bandelier Tuff cooling units or non-glassy tuffs. These units should only be encountered in deep drilling investigations. If site tuff samples were taken from Qbt 1g, Qct, and/or Qbo, compare site samples to combined set of all Laboratory-wide background samples from units Qbt 1g, Qct, and Qbo.

**Decision 4. Were canyon sediments sampled and can Laboratory sediment data be used?**

**"Yes" Decision.** If canyon sediments were sampled at the PRS and the Laboratory sediment data can be used, compare the PRS data to the set of canyon sediment data from LANL background.

At present, the canyon sediment data and BVs are being reviewed by the NMED Surface Water Quality Bureau. Contact the ER Project Office to check approval status prior to use of data.

The LANL background data set includes samples of canyon sediments from Ancho Canyon, Indio Canyon (Longmire et al. 1995, ER ID 52227), Los Alamos Canyon, Guaje Canyon and Pueblo Canyon (Ryti et.al. 1998, ER ID 58093).

**"No" Decision.** A "no" decision indicates that none of the existing subsets of Laboratory-wide background data (soil, Bandelier Tuff, and canyon sediments) are obviously applicable. Other background data options should be considered, including evaluation of data through interelement correlations, or generating site-specific (local) background.

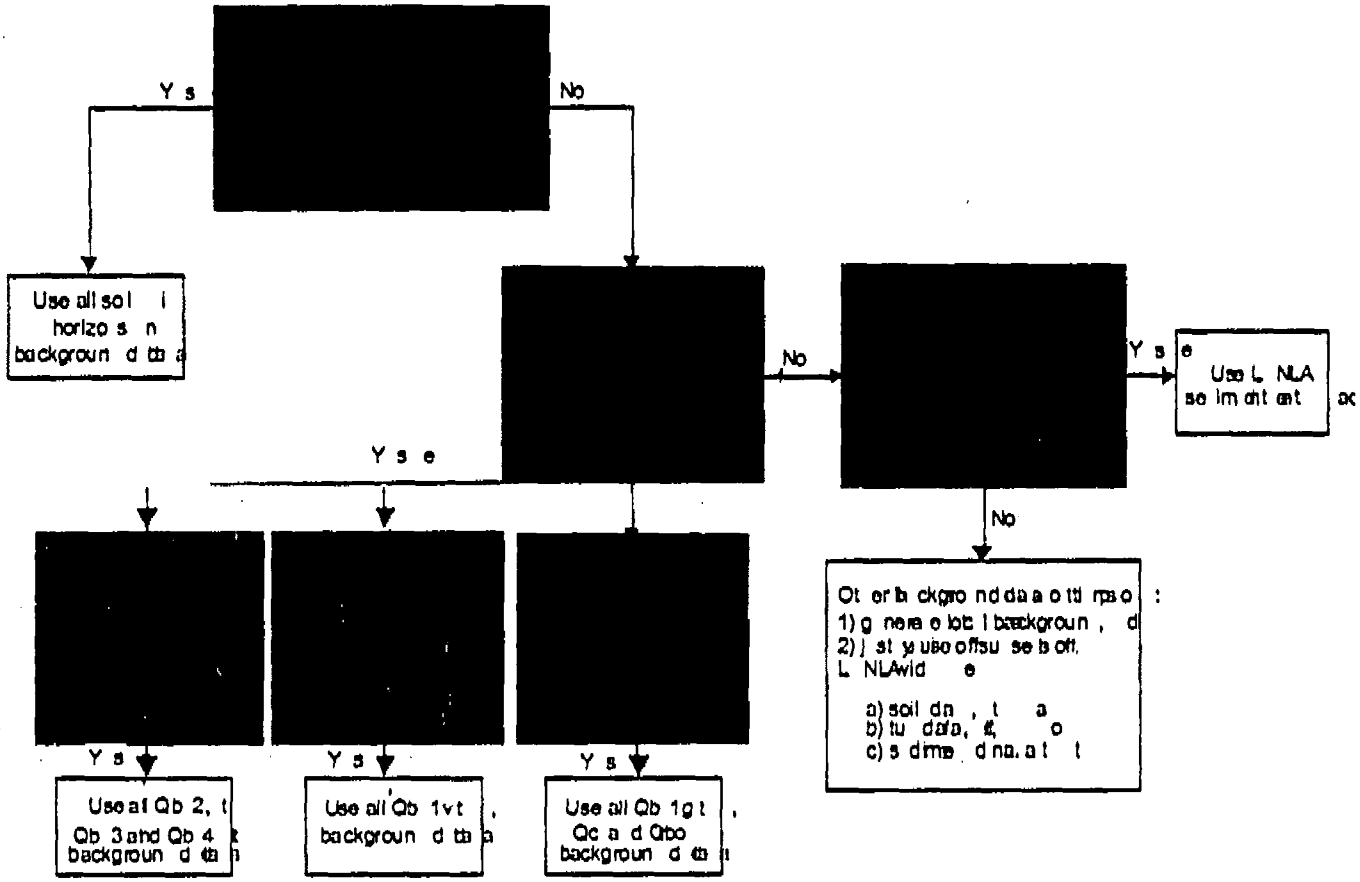

Y s
No
Use all so l i horizo s n backgroun d da a
No
Y s e
Us L NLA so im ent ant ac
Y s e
No
Ot er b ckgro nd da a o ti ns o :
1) g nera e lot l background , d
2) j st y use of su se b ot.
L NLA wid e
a) soil da , t a
b) tu data, t o
c) s dime d na. a t t
Y s
Use al Qb 2, t Qb 3 and Qb 4 k backgroun d da n
Y s
Use all Qb 1v t , backgroun d da a
Y s
Use all Qb 1g t , Qc a d Qbo backgroun d da i

**Figure 1. Flow chart for selecting inorganic background data sets.**

Interelement correlations

One way to justify the use of Laboratory-wide background data is to evaluate the data through interelement correlations. Typically, there are significant correlations between major elements (aluminum, iron, and potassium) and trace elements (arsenic, beryllium, copper, nickel, vanadium, and zinc). The correlations are presented and the geochemical basis is detailed in *Natural Background Geochemistry and Statistical Analysis of Selected Soil Profiles, Sediments, and Bandelier Tuff, Los Alamos, New Mexico* (Longmire et al 1995, ER ID 52227). For most inorganic chemicals, these strong correlations result in a consistent ratio of trace to major elements. A significantly elevated ratio of a given trace to a major element can be used to document a release of that trace element. Bivariate plots of trace elements to major elements are one way to visually display the ratios for background and PRS data. Any points distant from the cluster of points exhibiting strong correlation should be examined as possible indicators of contamination. An example data display is presented in Figure 2. This plot shows the bivariate

relationship between beryllium and iron for Technical Area 10 surface samples and Laboratory-wide soil background data.

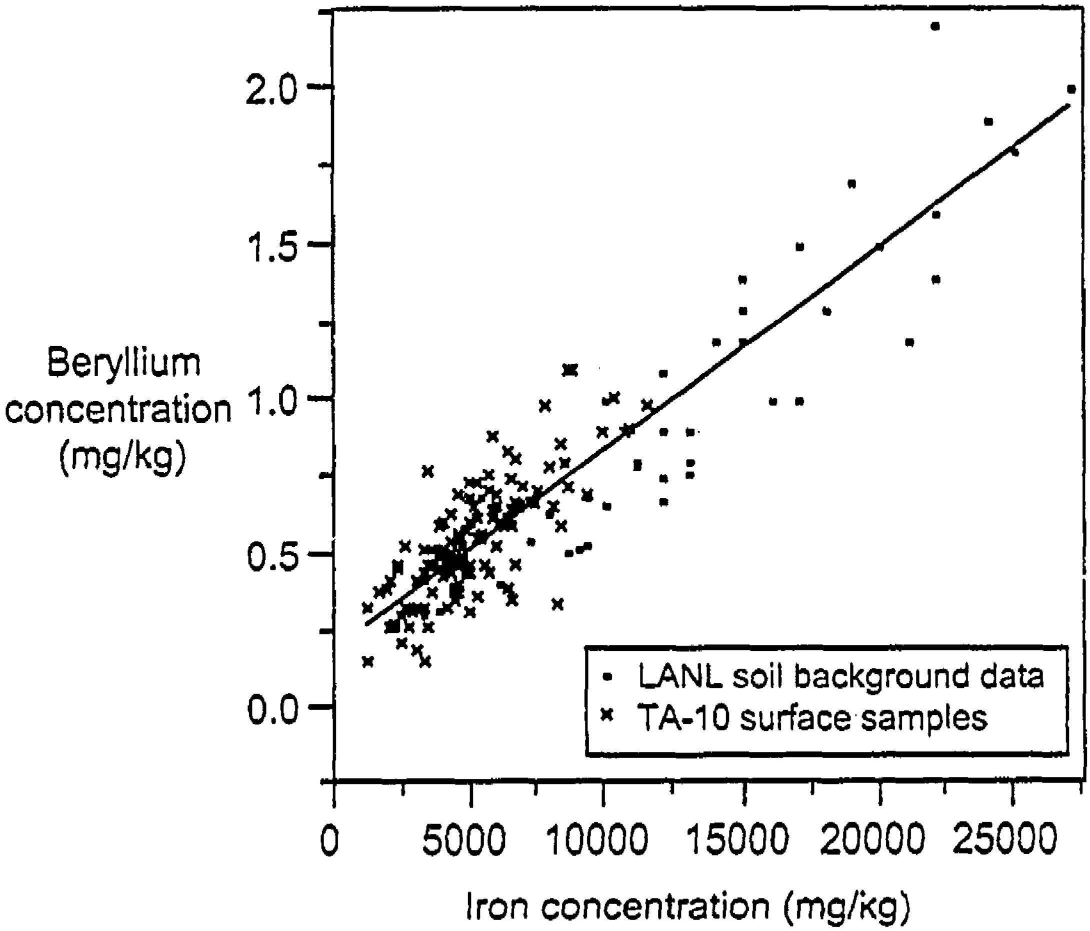


Figure 2. BIVARIATE PLOT OF BERYLLIUM AND IRON (NITRIC ACID FRACTION) FROM THE LABORATORY SOIL BACKGROUND DATA AND TA-10 SURFACE SAMPLES. CORRELATION COEFFICIENT IS 0.916 FOR 174 BACKGROUND SAMPLES.

Another example is the strong correlation between concentrations of thorium and uranium in the Bandelier Tuff, presented in Figure 3. The bivariate plot shows that each rock unit has similar ratios of thorium to uranium (the uranium concentration is roughly 30% of the thorium concentration).

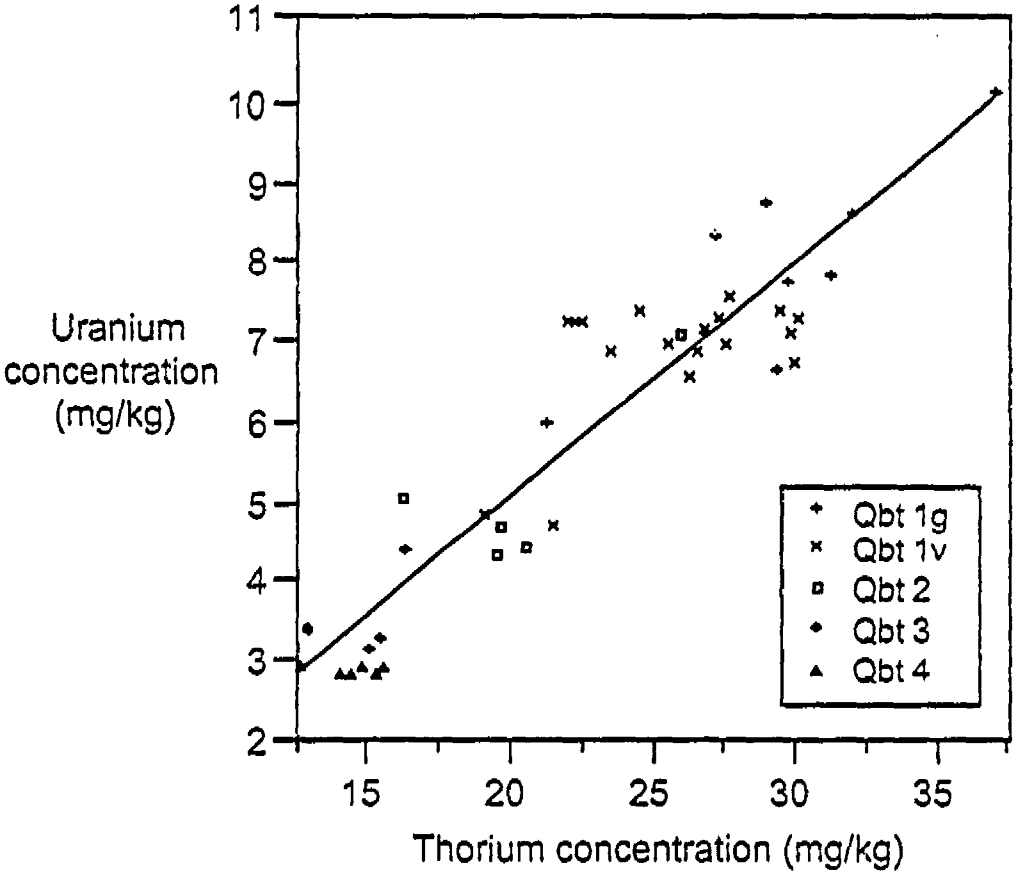


Figure 3. BIVARIATE PLOT OF URANIUM AND THORIUM (WHOLE ROCK ANALYSIS) FROM BANDELIER TUFF SAMPLES IDENTIFIED BY BANDELIER TUFF UNIT. CORRELATION COEFFICIENT IS 0.933 FOR 44 SAMPLES.

Generating appropriate subsets of background data can be performed cost-effectively by using interelement correlations to statistically subsample Laboratory-wide data to create a conditional set of site-specific background data. At a minimum, this statistical subsampling requires that the concentration of one or more of the major inorganic elements (aluminum, iron, or potassium) can be shown, through archival information, to have never been released at a PRS, and that other inorganics are highly correlated to at least one major element. The concentration range and statistical distribution of the major element results at a PRS are used to subsample the expected concentration of a trace element in the Laboratory-wide background data. For example, if a PRS had uniform concentration of iron between 5000 and 10000 mg/kg, the expected range of beryllium concentrations would be predicted to be between 0.3 and 1.1 mg/kg. PRS beryllium concentrations greater than 1.1 mg/kg would be outside the range of a statistically-based subsample of the Laboratory-wide data. This approach utilizing conditioning on covariates is more completely discussed in Campbell (1994, ER ID 54949). Data analysts unfamiliar with this statistical subsampling approach should contact the Data Analysis and Assessment Team for more information.

Site-specific background data

If site-specific background data are needed, statistical guidance can be used to help determine an appropriate number of background samples. One such approach, the minimum detectable difference procedure (EPA 1989, ER ID 54945), is mentioned in the Summary of Regulations and Guidance Governing Statistical Comparisons to Background section of this paper. This procedure requires three types of input: 1) the difference between the mean concentration of site data and background data that is desired to be detected (e.g., 50% of the background mean); 2) the desired probability of detecting that difference (e.g., 20%); and 3) the expected variability in the concentration data (usually expressed as the relative variability or coefficient of variation, e.g., 100% is typical). Given these inputs, 20 samples per background media are typically considered adequate for making background comparisons. The New Mexico Environment Department (NMED) position paper on the use of tolerance intervals to estimate background concentrations requires collection of at least eight background samples per media. In addition, NMED approval of the background data set, calculation methods and the background (or baseline) values must be obtained prior to their use in (site-specific) background comparisons (NMED 1998, ER ID 59376). In light of the time and effort required to obtain NMED approval, collection and use of site-specific background data is generally discouraged. Before collecting site-specific background data, the potential use of the existing Laboratory data should be fully explored. If baseline samples are collected, it is critical that consistent sample digestion and analytical methods are used for the baseline data and the PRS data.

Analysis of Laboratory soil background data indicates that naturally occurring levels of inorganic chemicals will vary as a function of certain soil properties (e.g., clay and iron content, see Longmire et al., 1995 ER ID 52227). PRS sample information on particle size, organic carbon and pH may support the understanding of fate and transport. Recording appropriate key soil information during sample collection is easily achievable by following a simple checklist that your Physical Modeling and Characterization Team representative can provide.

The Data Analysis and Assessment and Physical Modeling and Characterization Teams can provide further guidance and technical support for sampling and analysis plan development and to support sampling teams in the field.

**(b) Selecting Background Data for Radionuclides**

The following are three ways in which the consideration of radionuclide background is different from that of inorganics. First, the spatial distribution and concentration of radionuclides is derived from natural background sources (primordial and cosmogenic), globally elevated concentrations from atomic weapons test fallout, regionally elevated radioactivity from past Laboratory operations, and PRS-specific releases of radionuclides. Both natural radionuclide sources and fallout-related activity are considered to represent background radioactivity. Locally elevated values from Laboratory operations represent baseline radioactivity. Comparing PRS radionuclide concentrations to background or baseline concentrations would identify a PRS-specific release. Second, the standard practice for reporting radiological analyses is to report all results in the analysis library regardless of the analytical detection limit. Third, the Department of Energy (DOE) regulatory guidance (DOE 1990, ER ID 54216.5; DOE 1993, ER ID 22361) for establishing cleanup levels for radionuclides is based on a dose above background.

The following are potential uses for radionuclide background data in three different decision-making situations. The first and most common use of a background data set is to determine whether a release from a specific PRS has occurred. This is the process followed in data review to support risk management decisions in the ER Project. The statistical methods used in the decision are described in the following section. A second use of background data is to determine the portion of dose attributable to background. This evaluation may involve a correction (such as subtracting the mean radioactivity of the background data from the mean radioactivity of the site

data) or a comparison (such as comparing the dose resulting from site radioactivity with dose resulting from background radioactivity). Each of these methods can lead to different recommendations for future actions at a site. Regulatory guidance for determining radionuclide dose limits is provided in DOE documents (DOE 1993, ER ID 22361). A third use of radionuclide background data is for making certain types of waste classification decisions. In particular, low-level radioactive waste is defined as material containing added radioactivity from DOE operations. Added radioactivity would take into consideration natural background levels of radionuclides. Readers interested in more information on radioactive waste classification decisions are referred to Radioactive Waste Management Procedure for ER Project Field Operations SOP 1.11 (LANL, ER ID 55939.23).

This technical paper provides the decision logic for selecting background data sets for radionuclides. Making comparisons to radionuclide background provides the ER Project with a mechanism for determining if a release has occurred, for deriving cleanup levels, and for evaluating the attainment of cleanup goals. Existing radionuclide background data include the Laboratory's annual Environmental Surveillance Reports (Campbell 1998, ER ID 57585), isotopic activities estimated from total element abundance measured in rock samples (Longmire et al. 1995, ER ID 52227), and background samples of canyon sediments (McDonald et.al. 1997, ER ID 55532.1). All sets of radionuclide background data are summarized in the LANL Background document (Ryti et.al. 1998, ER ID 58093).

General considerations

Certain radionuclides in the PRS-specific data set are extraneous and should be eliminated before an appropriate background data set can be selected for comparison. These radionuclides include analytical laboratory quality control indicators and radionuclides for which all reported activities are less than minimum detectable activity.

Radiological data packages returned by analytical laboratories typically contain indicators designated to ensure the quality control and quality assurance of the data package. These indicators do not represent suspected site contaminants, and thus warrant no further assessment relative to background.

The measured activity of the radionuclide must be greater than the minimum detectable activity for the sample. Radionuclides for which all reported activities are less than the minimum detectable activity do not warrant further evaluation as potential site contaminants. Frequently, the counting time and the presence of other radionuclides can cause an estimated result for a radionuclide to be less than the sample and radionuclide-specific minimum detectable activity. Thus, the counting time should be reviewed to make sure that the detectable quantity was not arbitrarily inflated due to inadequate counting time. In addition, the spectral quantification windows should be reviewed for possible radionuclide interference.

Selection of an appropriate background data set is based on the type of media from which a sample was collected and whether the sample was collected from the surface or the subsurface. Naturally occurring radionuclides are distributed at different concentrations in various soil and rock matrices. Thus, as for inorganic chemicals, the type of media from which a sample is collected (tuff versus all other solid media) is an important factor in determining an appropriate background data set. Background concentrations of radionuclides resulting from fallout are typically associated with surface or near-surface depths. Thus, soil background data for radionuclides resulting from fallout apply to samples collected from the surface (or the near surface) only.

Caveat: Comparability of Analytical Methods

[illegible]

[illegible]

[illegible]

The sample preparation and analytical methods used for all LANL background data sets are listed in Inorganic and Radionuclide Background Data for Soils, Canyon Sediments, and Bandelier Tuff at LANL (Ryti et.al. 1998, ER ID 58093). It is important to identify appropriate analytical methods before sample collection and analysis. This issue is an even greater pitfall for radionuclide analyses than for inorganic analyses, because radiochemical analyses are not as standardized as inorganic analyses. For example, americium-241 may be analyzed by either gamma spectroscopy or alpha spectrometry with multiple options for sample preparation, counting geometries, and counting times possible for each analytical method. Consultation with a radiochemist is recommended for determination of appropriate analytical methods for comparability with methods used in background sample analyses. In the event that different methods were chosen for the site data, determine if the methods are directly comparable to those used for LANL background data. These conclusions regarding comparability are documented in ER Project reports. Analytical methods for radium, thorium and uranium isotopes also warrant careful review to ensure comparability with the relevant set of background data. For example, the background activity of uranium isotopes in tuff samples was estimated from total elemental concentrations of uranium, using the conversion formulas listed in the LANL background document (Ryti et.al. 1998, ER ID 58093). The resultant estimated background radioactivity of uranium-235 is less than the typical minimum detectable activity for this isotope when it has been analyzed by gamma spectroscopy.

Caveat: Use of Radioactivity Estimated from Total Abundances

There are several data preparation issues relating to estimation of naturally occurring radionuclide activity from total abundances. These issues include: 1) constants used in the calculations (isotopic abundance, specific activity), 2) secular equilibrium - or how far down the decay chain can radioactivity be estimated (disruption of secular equilibrium at radon daughters or the special case of depleted uranium), 3) guidance for assembling radionuclide data sets (converting totals to isotopic, mixed data sets with some totals and some isotopic).

Recommendation: The details regarding conversion of totals to isotopic as applied to LANL tuff background data are given in the LANL Background document (Ryti et.al.1998, ER ID 58093). It is recommended that the same procedure be followed for site data to maintain comparability.

**Decision process.** The process for eliminating extraneous PRS data and for selecting the most appropriate set of radionuclide background data is summarized in Figure 4. The selection of background data sets reflects the sampling media categories used in the LANL Background document (Ryti et.al. 1998, 58093). Before using the decision process shown in Figure 4 (and discussed below), it is essential that the sample collection, preparation and analytical methods used for PRS samples are comparable to those used for the background samples.

**Decision 1. Is the radionuclide a radiological indicator?**

**"No" Decision.** The data for the remaining radionuclides warrant further assessment. Proceed to decision 2.

**"Yes" Decision.** Remove these radionuclides from consideration.

Radiological indicators are used for quality control and quality assurance evaluation of analytical laboratory data packages. Thus it is not appropriate to compare radiological indicators to background, dose- or risk-based health protection standards. Radiological indicators used for QC evaluations by the ER Project include: annihilation radiation, cadmium-109, cerium-139, mercury-203, potassium-40, protactinium-231, protactinium-234m, tin-113, strontium-85, and yttrium-85. It should be noted that if potassium-40 is not included as a radionuclide indicator (i.e., was identified for investigation at the PRS), it should be treated in the same manner as naturally occurring radionuclides (i.e., it must be compared to background values). If site data was analyzed using gamma spectroscopy, professional judgement should be used for evaluation of some of the radionuclides included in the standard reported analyses, including radionuclides considered as

"not reliably measured" by gamma spectroscopy and some radionuclides with half-lives less than 365 days. Consult a radiochemist for information about the radionuclides reported in the gamma spectroscopy suite.

**Decision 2. Is the activity of the radionuclide greater than the minimum detectable activity?**

**"Yes" Decision.** The radionuclide requires further evaluation based on the activity reported by the analytical laboratory. Proceed to decision 3.

**"No" Decision.** None of the results reported for the radionuclide are greater than the minimum detectable activity value. In most cases, this indicates that an insignificant quantity of the radionuclide is present in the PRS-specific data and no further assessment of the radionuclide is necessary. However, before eliminating the radionuclide, the analytical data report should be reviewed to ensure that inadequate counting time and interferences did not lead to erroneous elimination of the radionuclide.

In cases for which the analytical laboratory does not report a minimum detectable activity for the sample, but does report an analytical uncertainty estimate, a value of three-times the analytical uncertainty may be used as an estimate of the sample-specific minimum detectable activity. Using the minimum detectable activity as a data screening tool is valid only when the data package from the analytical laboratory meets the general guidelines available in the Statement of Work for radiochemical analyses (LANL 1995, ER ID 49738). Using different criteria for determining whether a radionuclide has detection status may compromise the comparability of the data sets. Note that data analysts should not assume that the uncertainty value reported in FIMAD matches the analytical uncertainty as reported in the analytical data package. The uncertainty from the analytical data package is the value needed for the calculation.

Input from a radiochemist is recommended for determination of detection status.

**Decision 3. Is the radionuclide associated with fallout?**

**"Yes" Decision.** Radionuclides resulting from fallout are expected to be associated with surface or near surface samples only. At the Laboratory, radionuclides resulting from fallout include tritium, cesium-137, americium-241, plutonium-238, plutonium-239/240, and strontium-90. Proceed to decision 4.

**"No" Decision.** Proceed to decision 5 for evaluation of other classes of radionuclides.

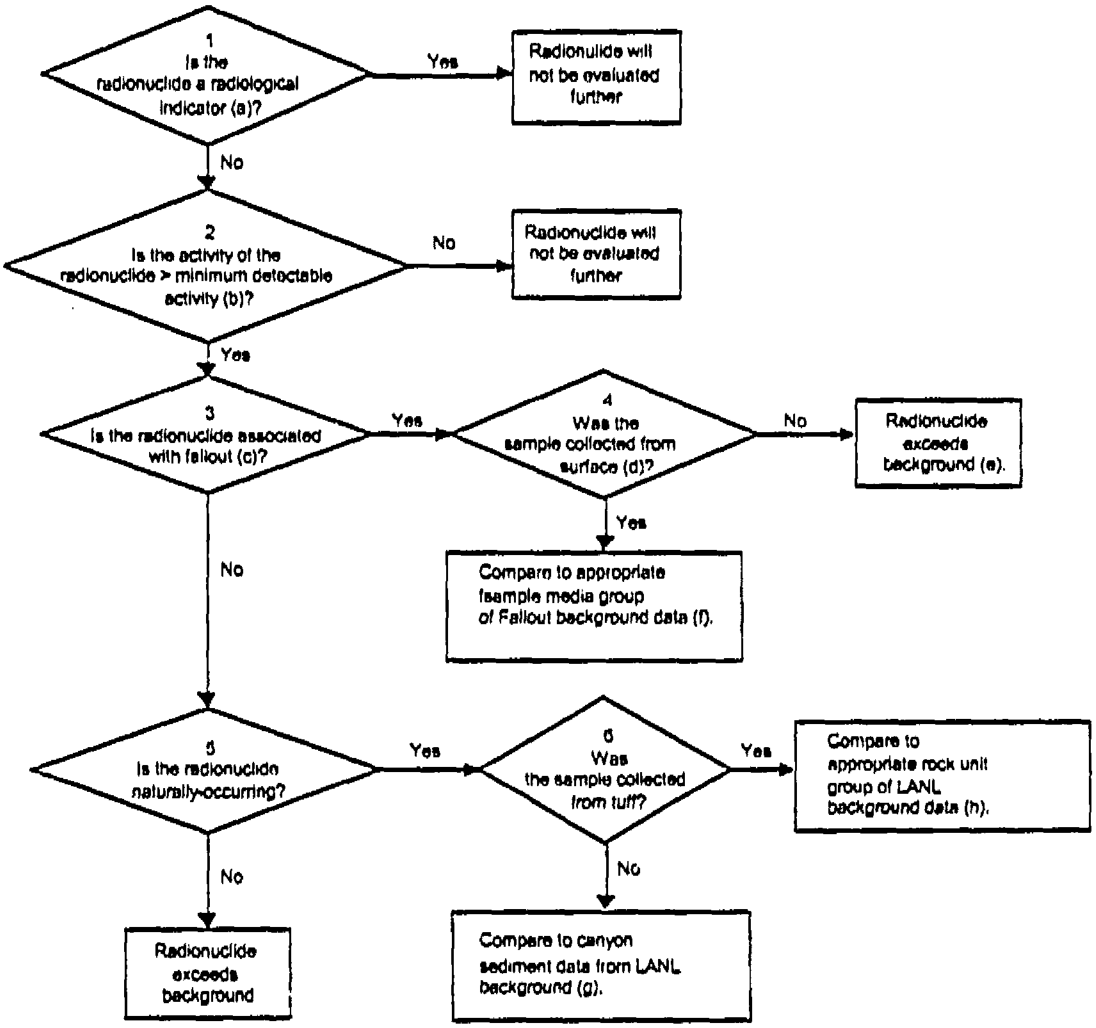
**Decision 4. Was the sample collected from the surface?**

Because radionuclides resulting from fallout are associated with atmospheric deposition, the background activity of this class of radionuclide is limited to surface samples at relatively undisturbed sites. In this context, surface samples are defined as intervals starting at a 0 inch depth and extending no deeper than 6 inches.

**"Yes" Decision** Compare surface soil PRS data to the background data associated with Laboratory operations and global fallout. These data are summarized in Campbell (1998, 57585) and LANL (Ryti et.al. 1998, ER ID 58093). This fallout background data set is appropriate for all surface samples taken from soil media or geological fill material. If canyon sediments (any depth) were sampled at the PRS, the canyon sediments background data is the appropriate choice. If fallout radionuclides were detected in surface (or any depth) tuff samples, identify them as COPCs and carry them forward to a dose- or risk-based assessment. Fallout radionuclides were not analyzed in tuff samples. The BVs listed in the LANL background document (Ryti et.al. 1998, ER ID 58093) are nominal detection limits and should not be used to determine if observed activities exceed background.

At present, the NMED Surface Water Quality Bureau is reviewing the canyon sediment data. Contact the ER Project Office to check approval status prior to use of canyon sediment data.

**"No" Decision.** Once this point is reached, it is likely that the radionuclide exceeds background and the radionuclide should be carried forward to a dose- or risk-based assessment. Note: that the depth restriction does not strictly apply to canyon sediment data.


1
Is the radionuclide a radiological indicator (a)?
Yes
Radionulide will not be evaluated further
No
2
Is the activity of the radionuclide > minimum detectable activity (b)?
No
Radionuclide will not be evaluated further
Yes
3
Is the radionuclide associated with fallout (c)?
Yes
4
Was the sample collected from surface (d)?
No
Radionuclide exceeds background (e).
Yes
Compare to appropriate fsample media group of Fallout background data (f).
No
5
Is the radionuclide naturally-occurring?
Yes
6
Was the sample collected from tuff?
Yes
Compare to appropriate rock unit group of LANL background data (h).
No
Compare to canyon sediment data from LANL background (g).
No
Radionuclide exceeds background

(a) Radiological indicators include annihilation radiation, potassium-40, cadmium-109, cerium-139, mercury-203, tin-113 and strontium-85. Others from gamma spec suite?

(b) The minimum detectable activity value should be reviewed to ensure that adequate counting time and interferences did not cause inappropriate elimination of radionuclides.

(c) Fallout radionuclides include tritium, cesium-137, americium-241, plutonium-238, plutonium-239/240, and strontium-90.

(d) Surface samples are defined as intervals starting at 0 inch depth and extending no deeper than 6 inches.

(e) Conclusion does not apply to PRS sediment samples.

(f) Surface soil samples are compared to Fallout background data from Environmental Surveillance Program. Sediment samples are compared to canyon sediment data for these analytes. Tuff samples should be evaluated on the basis of detection status alone for Fallout radionuclides. There are no background samples for tuff. BVs listed for tuff samples represent typical minimum detectable activity levels for these radionuclides and should not be used to determine if observed activities exceed background.

(g) The sediment background data set for naturally occurring radionuclides is the background data set for use with sediment data and with all soil data (see discussion of decision point 6).

(h) For the purpose of background comparisons, the stratigraphic units have been combined into more general categories. Compare site tuff samples to background samples in the category that includes the site sample's stratigraphic unit. The categories are (1) Qbt 2, Qbt 3, and Qbt 4, (2) Qbt 1v, (3) Qbt 1g, Qct, Qbo. Use the combined set of all background samples from the given category (for more information, see discussion of decision point 6, "Yes" decision).

**Figure 4. Flow chart for selecting radionuclide background data sets.**

**Decision 5. *Is the radionuclide naturally occurring?***

**"Yes" Decision.** In the context of selecting radionuclide background sets, naturally occurring radionuclides are limited to uranium, uranium decay-chain daughters, thorium, and thorium decay-chain daughters. Proceed to decision 6 for further evaluation of naturally occurring radionuclides.

**"No" Decision.** Once this point is reached, it is likely that the radionuclide exceeds background, and the radionuclide should be carried forward to a dose- or risk-based assessment.

**Decision 6. Was the sample collected from tuff?**

**"Yes" Decision.** Compare PRS data to radionuclide background data associated with the geologic unit. For the purpose of background comparisons, the stratigraphic units have been combined into more general categories. The categories are (1) Qbt 2, Qbt 3, and Qbt 4, (2) Qbt 1v, (3) Qbt 1g, Qct, Qbo. If site tuff samples were taken from Qbt 2, Qbt 3 and/or Qbt 4, compare site samples to combined set of all Laboratory-wide background samples from units Qbt 2, Qbt 3, and Qbt 4. If site tuff samples were taken from unit Qbt 1v, compare to the full set of unit Qbt 1v samples from LANL background. . If site tuff samples were taken from Qbt 1g, Qct, and/or Qbo, compare site samples to combined set of all Laboratory-wide background samples from units Qbt 1g, Qct, and Qbo.

The source of these tuff background data is total abundance measured in rock samples (Longmire et al. 1995, 1266). The total abundance is converted to isotopic activity through isotopic abundance of uranium isotopes and the specific activity of these isotopes. Thus, the geologic unit radionuclide background data represent a surrogate data set for radioactivity of naturally occurring isotopes.

When PRS data are compared to background data by geologic unit, special attention should be paid to the analytical methods used for both the PRS data and the estimated background data. For example, the estimated background radioactivity of uranium-235 is less than the typical minimum detectable activity for this isotope when it has been analyzed by gamma spectroscopy.

Although there are potential problems in applying these estimated background values to naturally occurring radionuclides, the geologic unit background data are based on the natural variation of uranium and thorium present in various tuff cooling units. Thus, these data will assist in correctly interpreting results from boreholes that intersect multiple tuff cooling units.

**"No" Decision.** For solid media other than tuff (soil, geological fill materials and sediments), compare PRS data to background data associated with canyon sediment. Because abundances of naturally-occurring uranium, thorium and their daughters are expected to be similar in sediment, soil, and fill material, the background data collected for naturally-occurring radionuclides in canyon sediment (McDonald et al. 1997, ER ID 55532) are viewed to represent a conservative background data set for comparing all these media. The use of the canyon sediment data as a surrogate for soil and fill is considered preferable to using estimated isotopic activities from total uranium and thorium analyses for mesa top soil background sampling locations.

If canyon sediment was sampled and the Laboratory sediment data can be used, the sediment data are the appropriate choice. At this time, NMED Surface Water Quality Bureau is reviewing the canyon sediment background data. Contact the ER Project Office to check approval status prior to use of BVs or Fallout Values for evaluating sediment samples from a PRS.

Other Background Data

If none of the existing subsets of Laboratory-wide background data (soil, Bandelier Tuff, and canyon sediments) are obviously applicable, other background data options may be considered, including: evaluation of data through interelement correlations, or generating site-specific (local) background. The recommendations and requirements for these options are discussed in the inorganics section above. In addition, to determine the adequacy of site-specific background for decision-making, the analytical suite used should be examined to insure that all potential radionuclide contaminants have been included. The number of samples taken should be evaluated to insure that site-specific background conditions have been adequately represented (EPA 1989, ER ID 54947; NMED 1998 (draft), ER ID 59376).

## 3. RECOMMENDED STATISTICAL METHODS FOR BACKGROUND COMPARISONS

Because background comparisons are used to make decisions throughout the RCRA process, from site screening to corrective measures implementation, data analysts must use statistical methods that can be applied to a broad range of decisions. This guidance defines two methods for background comparisons, which meet the requirements for RCRA decision making.[2] In the first method, the Hot Measurement comparison, site concentration data are compared with a statistic that is an estimate of the largest concentration that could be considered representative of the set of background concentrations. In the second method, the distributional shift test, the mean (mean rank, quantile) of site data is compared with the mean (mean rank, quantile) of background data to determine whether the former is statistically greater than the latter. These tests help demonstrate whether a release has occurred at a PRS and help define what risk consequence the release may have. Figure 5 illustrates the differences between site data and background data detected by the two methods.

[2] The methods are among those discussed in the RCRA groundwater monitoring guidance document.

This EPS image does not contain a screen preview.
It will print correctly to a PostScript printer.
File Name : fig1qp.ps
Title : (S-PLUS Graphics)
Creator : S-PLUS
CreationDate : Wed Sep 23 19:24:18 1998
Pages : 1

(a) Site data are within range of background: no distributional shift or hot measurements (i.e., no value is greater than the UTL).
(b) Site data fail hot measurement comparison: one of eleven arsenic concentrations exceeds the UTL of 8.17 mg/kg. The site data does not fail the distribution shift test.
(c) Site data show a distributional shift: the Wilcoxon rank sum test shows site data tend to be greater than background data. This difference was not detected by the hot measurement test.
(d) Site data show both a distributional shift and a failure of the hot measurement comparison: ten of thirty-two arsenic concentrations exceed the UTL of 8.17 mg/kg and the site data tend to be greater than the background data.

Figure 5. BOX PLOT COMPARISONS OF EXAMPLE SITE DATA WITH LABORATORY BACKGROUND DATA.

The decision to be supported by the background comparison determines which test is appropriate. The hot measurement test, or comparison to background values (BVs), is required for all analytes evaluated in a data review that supports risk management decisions in the Laboratory's ER Project. Additional statistical tests (distritional shift tests) are recommended for use in conjunction with the Hot Measurement test. Use of the Hot Measurement test alone may

lead the data analyst to different conclusions about which analytes to retain as contaminants of potential concern (COPCs). In the examples presented in Figure 5(b) and 5(c), the two tests would lead to different conclusions. In Figure 5(c), the hot measurement test does not detect the shift in site concentrations above background that may be indicative of a release. In Figure 5(b), there is one hot measurement (one site concentration larger than the BV) but the distribution shift test indicates that the site concentrations are not statistically (shifted) larger than background. After noting that all site measurements are within the range of background, it would seem appropriate to conclude that site concentrations are not elevated. Before dismissing an analyte with a hot measurement, the magnitude and location of the large concentration would be considered in terms of the site operational history. When extensive data are collected to support a risk assessment or corrective action and a shift in the distribution could lead to further action at the site, the distributional shift test is more appropriate. The rationale for selecting a statistical method that differs from those presented in this guidance will be clearly indicated in the ER Project report that summarizes the background comparison.

Because the selection of a particular statistical method depends on the statistical distributions of site and background data, data analysts are encouraged to prepare graphical data displays to communicate the results of data comparisons. Box plots, in which background and site data can be compared side-by-side, are most useful. The box plots in Figure 5 show actual values (as filled square dots) for each data group (Laboratory background and example PRS data). The ends of each box represent the "inter-quartile" range, which is specified by the 25th and 75th percentiles of the data distribution. The line within the box represents the median (50th percentile) of the data distribution. Thus the box indicates concentration values for the central half of the data. Concentration shifts can be assessed by comparing the relative positions of the boxes. If the boxes do not overlap each other's median positions, the distribution shift test will most likely detect a statistical difference. If the majority of the data are represented by a single concentration value (usually the detection limit), the box is reduced to a single line. In addition to box plots, data analysts should also consider using histograms and probability plots to provide tangible evidence of similarities or differences between site and background data.

The level of effort spent to evaluate potential differences between PRS 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 differences between beryllium concentration data from firing site activities and Laboratory-wide or site-specific background data should be carefully evaluated to determine the levels of anthropogenic beryllium added to the environment. In all cases, data comparisons will be documented in the appropriate ER Project report.

### Hot Measurement Comparison

The Hot Measurement comparison uses a threshold value that represents high natural background concentrations. This threshold value is known as the background value (BV), and there exists a probability that a natural background measurement will exceed the hot measurement threshold. Using a threshold statistically related to higher background concentrations reduces the frequency of false positive results. 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, ER ID 54946). The ER Project has selected the 95th percentile for calculating the UTL (upper tolerance limit) based on the general guidance in the RCRA groundwater document. EPA recommends calculating an upper 95% confidence limit for the target percentile (EPA 1989, ER ID 54946). The details regarding calculations of the BVs for LANL background data are given in the LANL Background document (Ryti et.al. 1998, 58093). For the analytes that were rarely detected in background samples, the BV is the detection limit specified in the analytical services statement of work for the analysis method used on the LANL background data.

The hot measurement comparison is made between the maximum detected site sample (or detection limit of a nondetected chemical, if that is the maximum result) and the background value

(UTL or detection limit). Exceeding the UTL as a background value is not definitive evidence that a release has occurred at a PRS. Assuming the PRS is at background and the statistical model is correct, there is a 5% probability that the 95th percentile will be exceeded by each sample collected from the PRS. Furthermore, a typical inorganic chemical suite requires comparison of 23 analytes with background. If the concentrations of the 23 inorganic analytes vary independently, the 5% probability that each PRS sample exceeds the 95th percentile increases to a 69% probability that at least one of the 23 ninety-fifth percentiles will be exceeded in a single sample. Additionally, given that the probability values for these multiple comparisons have not been adjusted, the overall confidence level for 23 analytes will be substantially less than 95%. In addition to the strictly probability-based discussion presented above, the possibility of exceeding a UTL due to an unusual, but naturally occurring, soil matrix is a further consideration. Consequently, the results of a hot measurement comparison must be carefully evaluated.

### Distributional Shift Tests

Instead of comparing to a single value representing background (a BV), distributional shift tests compare the site data to the distribution of background concentrations. A distributional shift test is used to determine whether site data are systematically greater than background data. Several types of distributional shift tests are available, and these tests are presented in two groups below. The preferred statistical method in each group is indicated where there are multiple options.

For detecting an overall distribution shift between all PRS data and the appropriate subset of Laboratory-wide background data, the following statistical tests can be used:

- The Student t-test is a parametric, two-sample test that determines whether the mean concentration of site data is statistically greater than the mean concentration of background data (Gilbert 1987, ER ID 55619). Data analysts should be aware that the t-test performs well for some deviations from normality but increased power may be obtainable through nonparametric methods. A nice discussion regarding the robustness limitations of the t-test are found in Miller (1986, ER ID 59375, p. 40–44). Normality can be visually assessed using normal qq-plots or probit plots. Formal tests for normality may be performed first, such as the Shapiro-Wilk W test or the Kolmogorov-Smirnov test (ref Gilbert 1987, 55619, p. 158). In general, the t-test is not recommended because it assumes that the data being compared are normally distributed and environmental data are rarely fit by a normal distribution.
- The Wilcoxon rank sum test (same as the Mann-Whitney U-test) is the nonparametric equivalent to the t-test (Gilbert 1987, ER ID 55619; Gilbert and Simpson 1992, ER ID 54952). The Wilcoxon test pools site and background data into one aggregate set and determines whether the average rank of site data is greater than that of the background data. The Wilcoxon test is recommended when nondetects are relatively infrequent (<10%) and all have the same detection limit. The nondetects are treated as tied at a value less than the smallest detected concentration.
- The Gehan test uses a modified ranking of the sample results to accommodate nondetected chemicals and then applies the Wilcoxon rank sum test. It is recommended when non-detects are relatively frequent (>10% and <50%). It handles multiple detection limits in a statistically robust manner (Gehan 1965, ER ID 54950; Millard and Deverel 1988, ER ID 54953). Further explanation including an example comparing the Wilcoxon and Gehan ranking procedures is provided in Appendix A. The test is not recommended if there are more than 50% non-detects in either of the two data sets. It is identical to the Wilcoxon rank-sum test when applied to results containing no non-detects. The Gehan test is the preferred test because of its applicability to a majority of environmental data sets.
- There are other variations of the rank sum test adapted to two sample problems with multiple nondetect limits. Among those studied by Millard and Deverel (1988,

ER ID 54953), they recommended two methods on the basis of performance in Monte Carlo simulations for various sample sizes and censoring mechanisms under an assumed lognormal distribution. The normal scores (Van der Waerden) test is the locally most powerful rank test for data that are normally or lognormally distributed. The Peto-Prentice test is another variation of the Wilcoxon rank-sum test that performed as well as the normal scores test using an asymptotic variance estimate. It is identical to the Wilcoxon rank-sum test when applied to results containing no non-detects.

For detecting distribution shifts between the upper range of PRS data and the appropriate subset of Laboratory-wide background data, the following statistical tests can be used:

- The Quantile test (Gilbert and Simpson 1992, ER ID 54952), which compares a selected upper quantile of background data with that of PRS data, is capable of detecting a statistical 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. 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. The Quantile test is applied at a prespecified quantile or threshold, usually the 80th percentile. It can be used when the frequency of non-detects is approximately the same as the quantile being tested. For example, in a case having 75% non-detects in the combined background and PRS data set, application of a quantile test comparing 80th percentiles would be appropriate. If the relative proportion of the two populations being tested is different in the top 20% of the data than in the remainder of the data, then there is reason to believe that the distributions are partially shifted due to a subset of the site. However, this implies that this test cannot be performed if more than 80% (or the threshold percentage) of the combined data are nondetected values. The threshold percentage can be adjusted to accommodate the detection rate of the analyte, or to look for differences further into the tails of the distributions. It is more powerful than the Wilcoxon (or Gehan) test for detecting a difference when only a small percentage of the PRS concentrations are elevated.
- The slippage test is based on the maximum observed concentration in the background data set and the number ("n") of site concentrations that exceed the maximum concentration in the background set (Gilbert and Simpson 1990, ER ID 55612, pages 5-8). The result (p-value) of the slippage test is the probability that "n" site samples exceed the maximum background concentration by chance alone. The test takes into account the numbers of samples in each data set (the number of samples from the site and the number of samples from background) and determines the probability of "n" exceedences if the two data sets came from identical distributions. This test is similar to the hot measurement test in that it is evaluating the largest measurements. It is more useful than the BV comparison because it is based on a statistical hypothesis test and not simply a statistic of a distribution. However, it is not applied if there are no site results larger than the maximum in the background distribution.

The ability to use the distributional shift tests 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, during Phase I of an RFI, inadequate numbers of samples are collected to warrant a distributional shift comparison. When planning in advance of sample collection, a better estimate of specific sample size requirement follows from specifying data quality objectives (DQOs) and calculating samples sizes based on the DQOs.

For those analytes which are rarely detected in LANL background (e.g.: mercury, antimony, and thallium in soil samples), an increase in the detection rate at the site may be evidence of a release. The following test is recommended.

- The Chi-square Goodness of Fit Test can be used to check for differences between proportions from data sets that fall into given categories. When proportions from data sets are categorized on the basis of two attributes, this is also referred to as a test for independence of attributes. For example, it can be used to test whether the attribute of detection [proportion of detected chemicals out of analyses performed] is the same from the site data set and the background data set [attribute of belonging to "site" or "background" data sets]. If these proportions are not statistically significantly different, then the detection rate attribute is "independent" of the categorization into "background vs. site" sets. (Box, Hunter and Hunter, ER ID 56653, pages 149-150) The test on detection rates as stated is most appropriate when the two data sets were analyzed using similar methods and had similar detection limits.

To infer a significant result in a single 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 relevant.

In addition, when more than one test is performed on the same set of data there is an increased possibility of observing a p-value of less than 0.05 by random chance alone. If a p-value is much less than 0.05 there is some reason to suspect that there is a difference between the distributions. If the p-value is much greater than 0.05, no difference is indicated. If the p-value is close to 0.05, then further evaluation is usually indicated. In particular, the nominal significance level for multiple or simultaneous tests can be adjusted using a method attributed to Bonferroni (Keppel 1982, ER ID 56652, pages 145-150). The procedure is to conclude that there is a difference between the data set distributions when applying *n* tests if any of the *n* tests results in a p-value less than $p=0.05/n$. Assuming independence between the outcomes (p-values) for the set of tests being applied to the data, this is a conservative procedure that preserves the overall or simultaneous error rate at the desired nominal level of 0.05. Judgement should be applied as the test results may be correlated, with the degree of correlation depending on the data set distributions and the tests. For example, there is greater correlation between the quantile test and slippage test than between the quantile test and Gehan test [reference Kathy Campbell's simulations published in the Rocky Flats document]. In general, division by two would not be too much of an adjustment for the set of common distribution shift tests (Gehan and quantile, with or without the slippage test). It is always appropriate to simply plot the data distributions and use the test results to back up what is observed in the plots. The corresponding adjustment for application of tests to multiple analytes for each sample is more complicated. It would involve correlation analyses of the analytical results and/or multivariate methods. No adjustment is recommended for use in ER Project reports, but the information is pertinent for interpretation of results.

In addition to test results described above, the data should be plotted spatially and evaluated relative to the conceptual site model. Specific aspects of the conceptual model that warrant a statistical assessment include the collocation or correlation of concentrations of contaminants. Another important step in revising the conceptual site model is evaluating geochemical or geologic patterns in the data. For example, evaluate concentrations as a function of distance down a borehole using information regarding documented fractures or position relative to a suspected source term, such as the depth at which an angled borehole extends below a materials disposal unit.

## REFERENCES

Box, G. E. P., W. G. Hunter and J. S. Hunter, 1978. Statistics for Experimenters. John Wiley and Sons, New York. 653 pp. (Box, Hunter and Hunter 1978, ER ID 56653)

Broxton, D.E., R.T. Ryti, D. Carlson, R.G. Warren, E. Kluk, and S. Chipera. March 20, 1996. "Natural Background Geochemistry of the Bandelier Tuff at MDA P, Los Alamos National Laboratory," Los Alamos National Laboratory Report LA-UR-96-1151, Los Alamos, New Mexico. (Broxton et al. 1996, ER ID 54948)

Campbell, K., 1994. "Estimation of Local Background Concentrations for Identification of Environmental Releases," Los Alamos National Laboratory Report LA-UR-94-2274, Los Alamos, New Mexico. (Campbell 1994, ER ID 54949)

Campbell, K. 1997. "Baseline Data for Fallout Radionuclides at LANL," Los Alamos National Laboratory Report LA-UR-98-958, Los Alamos, New Mexico. (Campbell 1998, ER ID 57585)

DOE (US Department of Energy), June 5, 1990. "Radiation Protection of the Public and the Environment," DOE Order 5400.5 (Change 1), Washington, DC. (DOE 1990, ER ID 54216.5)

DOE (US Department of Energy), March 25, 1993. "Radiation Protection of the Public and the Environment, Proposed Rule," Title 10, Part 834, Federal Register, Vol. 58, No. 56. (DOE 1993, ER ID 22361)

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, ER ID 54945)

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, ER ID 08794)

EPA (US Environmental Protection Agency), April 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. (EPA 1989, ER ID 54946)

EPA (US Environmental Protection Agency), April 1992. "Guidance for Data Usability in Risk Assessment (Part A)," Office of Emergency Remedial Response, US Environmental Protection Agency, Washington DC. (EPA 1992, ER ID 54947)

Gehan, E. A., 1965. "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, Vol. 52, Nos. 1 and 2, pp. 203-223. (Gehan 1965, ER ID 54950)

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. Von Nostrand Reinhold Company Inc. New York. 320 pp. (Gilbert 1987, ER ID 55619)

Gilbert, R. O., and J. C. Simpson 1990. "Statistical Sampling and Analysis Issues and Needs for Testing Attainment of Background-Based Cleanup Standards at Superfund Sites," in *Proceedings of The Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs*, Environmental Protection Agency, Arlington, Virginia. (Gilbert and Simpson 1990, ER ID 55612)

Gilbert, R. O., and J. C. Simpson, 1992. "Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media," Pacific Northwest Laboratory, Richland, Washington 99352. (Gilbert and Simpson 1992, ER ID 54952)

Keppel, G., 1982. DESIGN AND ANALYSIS A Researcher's Handbook, Second Edition. Prentice-Hall, Inc., Englewood Cliffs, New Jersey. 669 pp. (Keppel 1982, ER ID 56652)

LANL (Los Alamos National Laboratory), July 1995. "Statement of Work–Analytical Support," Revision 2, RFP No. 9-XS1-Q4257, Los Alamos, New Mexico. (LANL 1995, ER ID 49738)

LANL (Los Alamos National Laboratory), May 1997. "Los Alamos National Laboratory Environmental Restoration Program Standard Operating Procedures, SOP 1.11," Los Alamos National Laboratory report, Los Alamos, New Mexico. (LANL 1997, ER ID 55939.23)

Longmire, P. A., D. E. Broxton, and S. L Reneau (Eds.), October 1995. "Natural Background Geochemistry and Statistical Analysis of Selected Soil Profiles, Sediments, and Bandelier Tuff, Los Alamos, New Mexico, Los·Alamos National Laboratory Report LA-UR-95-3486, Los Alamos, New Mexico. (Longmire et al. 1995, ER ID 52227)

Millard, W. P., and S. J. Deverel, 1988. "Nonparametric Statistical Methods for Comparing Two Sites Based on Data with Multiple Nondetect Limits," *Water Resources Research*, Vol. 24, No. 12, pp. 2087-2098. (Millard and Deverel 1988, ER ID 54953)

Miller, R.J.,1986. Beyond ANOVA, Basics of Applied Statistics. John Wiley and Sons New York. 317 pp. (Miller 1986, ER ID 59375)

Palachek A.D., D.R. Weier, T.R. Gatliffe, D.M. Splett, D.K. Sullivan, "Statistical Methodology for Determining Contaminants of Concern by Comparison of Background and Site Data with Applications to Operable Unit 2", April 1993, SA-93-010, EG&G Rocky Flats report. (Palachek et.al. 1993, ER ID XXXXX)

Ryti, R, E. McDonald, and D. Carlson, June 1997. "Natural Background Geochemistry of Sediments, Los Alamos National Laboratory," Los Alamos National Laboratory Report, Los Alamos, New Mexico. (McDonald et.al. 1997, 55532.1)

Ryti, R.T., P.A. Longmire, D.E. Broxton, S.L. Reneau, and E.V. McDonald, March 1998. "Inorganic and Radionuclide Background Data for Soils, Canyon Sediments and Bandelier Tuff at Los Alamos National Laboratory." (Ryti et.al. 1998, ER ID 58093)

## APPENDIX A

The following is taken from Palachek et.al. 1993, ER ID XXXXX

Explanation of Scores Methodology:

In a standard Wilcoxon application, two samples which are to be compared are combined into a single sample and the observations are then ranked as a single sample. The ranks resulting from one of the two samples are then summed to see if they generally were larger or smaller than would be expected if the samples were taken from the same distribution. If so, the null hypothesis of no difference in the two underlying distributions would be rejected in favor of an alternative hypothesis of one distribution being shifted with respect to the other.

As a simple example consider the following where one-sided test of whether the sample 2 values come from a distribution of larger values is of interest:

| | | | | | | |
|---|---|---|---|---|---|---|
| Sample 1: | 1 | 4 | 5 | 7 | 12 | 15 |
| Sample 2: | 4 | 8 | 17 | 18 | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| The combined sample is then: | 1 | 4 | 4 | 5 | 7 | 8 | 12 | 15 | 17 | 18 |
| with respective ranks: | 1 | 2.5 | 2.5 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

The sum of the ranks for the second sample is therefore 2.5 + 6 +9 +10 = 27.5 ( note that tied values receive average ranks ). This rank sum of 27.5 is compared to values expected under the null hypothesis of equal distributions to determine if the sum is sufficiently large to be deemed statistically significant.

The Mann-Whitney/Wilcoxon approach can be applied to censored data only if the censoring values are smaller than all detects. In this case all nondetects would be treated as tied.

The situation gets more complicated when multiple detection limits ( censoring values ) are present in the two samples. Not all values can then be ranked with respect to each other. For example, it is unknown whether a nondetect with a detection limit of 10.0 is greater or less than a detect at 5.0, so their relative ranks cannot be determined. Similarly, the ranking of two nondetects with different detection limits cannot be determined.

One simple approach for determining a ranking of such data that has been suggested in statistical literature is to treat every measurement that is less than the largest nondetect as a tied value whether it is a detect or a nondetect. This clearly has the shortcoming of not using all the information that is available. For example, with nondetects at 5.0 and 10.0 and detect at 7.5, it is known that the 7.5 valued detect is clearly greater than the nondetect at 5.0. This information would be ignored in this approach.

An improvement is given in Millard and Deverol (1988, ER ID 54953). The scores approach proposed in this report is developed in that paper. While several variations are discussed, they generally behave comparably. The "Gehan" variation is proposed for use in this report largely since its derivation is the simplest to understand.

To see how the scores approach works, consider another example. The notation "<12" represents a nondetect at the detection limit of 12 and therefore a value less than 12.0.

| | | | | | | |
|---|---|---|---|---|---|---|
| Sample 1: | 1 | <4 | 5 | 7 | <12 | 15 |
| Sample 2: | 2 | <4 | 8 | 17 | 24 | |

| This gives the combined sample as | 1 | 2 | <4 | <4 | 5 | 7 | 8 | <12 | 15 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| (initial rank: disregarding detection status) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Although the scores approach is not specifically defined in terms of ranking, it can most easily be explained in terms of the ranks it equivalently ends up generating. Again, it is analogous to the Mann-Whitney/Wilcoxon approach and equivalent to it in the presence of no nondetects. The ranking for the scores method occur as follows. They are most easily assigned from largest to smallest:

The values 17 and 15 get ranks 10 and 9 respectively as they are know to be the two largest values, even with the presence of nondetects.

The <12 value is taken to be tied with all seven values below it and thus receives as its rank the average of the ranks 1 to 8 which is 4.5.

The value 8 is clearly greater than the six values below it and clearly less than the values 15 and 17. It is treated as a tie with the value <12 and therefore receives the average of the ranks 7 and 8 which is 7.5.

The value 7 is clearly greater than the five values below it and clearly less than the values 8, 15, and 17. It is treated as a tie with the value <12 and therefore receives the average of the ranks 6 and 7 which is 6.5 (i.e.: the rank of the value 7 can't be less than 6, and it can't be greater than 7 if it is tied with one other value, a nondetect with a detection limit above it)

The value 5 is clearly greater than the four values below it and clearly less than the values 7, 8, 15, and 17. It is treated as a tie with the value <12 and therefore receives the average of the ranks 5 and 6 which is 5.5.

The two values <4 are clearly less that the values 5, 7, 8, 15, and 17 and are treated as tied with each other as well as with the values 1, 2, and <12. They therefore receive the average of the ranks 1, 2, 3, 4, and 5 which is 3.5.

The value 2 is clearly greater than the value 1 and clearly less than the values 5, 7, 8, 15, and 17. It is treated as a tie with the values <4, <4, and <12 and therefore receives the average of the ranks 2, 3, 4, and 5 which is 3.5.

The value 1 is treated as tied with the values <4, <4, and <12 and therefore receives the average of the ranks of the ranks 1, 2, 3, and 4 which is 2.5.

In summary the following ranking results:

| Sample values | 1 | 2 | <4 | <4 | 5 | 7 | 8 | <12 | 15 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranks | 2.5 | 3.5 | 3 | 3 | 5.5 | 6.5 | 7.5 | 4.5 | 9 | 10 |

Note that with no ties or nondetects in this example, the sum of the resulting ranks 1, 2, . . ., 10 would be 55. The sum of the "scores" ranks in the example is also 55. This is always a property of this scores ranking scheme.

The test statistic, as in the Wilcoxon/ Mann-Whitney approach, can be considered as the sum of the ranks of the sample values from one of the samples. If the sample used is the site sample, then large values of this statistic would indicate that the site is generating samples that are large relative to the background samples, and the associated analyte would be classified as a COC.

Distributional properties for the statistic can be obtained through the usual approach used for rank methods. This considers all permutations of the resulting rankings since all such permutations are equally likely under the null hypothesis of no difference in the underlying site and background populations. If the statistic takes on a value in the upper five percent of the resulting values, it would be taken as statistical evidence that the analyte is elevated in the site relative to background and is therefore considered a COC. Note that this would provide the standard 0.05 Type I error probability.