

**NAVY RESPONSE TO EPA COMMENTS DATED MAY 22, 2006 ON THE
DRAFT SUMMARY REPORT FOR ENVIRONMENTAL BACKGROUND CONCENTRATIONS
OF INORGANIC COMPOUNDS APRIL 2006
NAVAL ACTIVITY PUERTO RICO CEIBA, PUERTO RICO**

EPA REGION II COMMENT

The Background Report was reviewed to determine if it complies with EPA's Guidance for Comparing Background and Chemical Concentrations I Soil for CERCLA Sites (EPA 540-R-01-003) (CERCLA Guidance). The CERCLA Guidance extensively references the Guidance for Data Quality Assessment - Practical Methods for Data Analysis (EPA QA/G-9) (DQO Guidance). Thus, the Background Report was also reviewed to verify compliance with DQO Guidance, where relevant. The data sets and background statistics presented in the Background Report were reviewed to identify any potential concerns.

Based on Booz Allen's and our own reviews, EP A has identified a number of concerns regarding compliance with the CERCLA and DQO Guidance. Concerns regarding several of the data sets and associated statistic were also identified. These concerns are discussed in the enclosed Technical Review.

Response: Comment noted. Please see the responses to BAH comments below.

BAH TECHNICAL REVIEW

General Comment

1. A technical review has been performed on the April 2006 Draft Summary Report for environmental Background Concentration of Inorganic Compounds (Background Report) at Naval Activity Puerto Rico (NAPR) in Ceiba, Puerto Rico. The Background Report was reviewed to determine if it complies with EPA's Guidance for Comparing Background and Chemical Concentrations I Soil for CERCLA Sites (EPA 540-R-01-003) (CERCLA Guidance). The CERCLA Guidance extensively references the Guidance for Data Quality Assessment - Practical Methods for Data Analysis (EPA QA/G-9) (DQO Guidance). Thus, the Background Report was also reviewed to verify compliance with DQO Guidance, where relevant. The data sets and background statistics presented in the Background Report were reviewed to identify any potential concerns.

The technical review identified a number of concerns regarding compliance with the CERCLA and DQO Guidance. Concerns regarding several of the data sets and associated statistic were also identified. These concerns are presented in the following Specific Comments.

Response: Noted. The specific comments are addressed below. When necessary, the Background Summary Report was also modified to reflect the comment responses.

Specific Comments

1. As indicated in the CERCLA Guidance (page 3-12), it is generally difficult to judge the adequacy of a background data set without first making certain basic decisions regarding the statistical comparison of background and site data. Of particular importance are decisions regarding the desired power and confidence levels of the statistical analysis. These inputs, particularly the desired power of the statistical tests, are closely related to the number of background samples required to achieve the required statistical performance. The adequacy of the number of background samples can only be judged in the context of the specific comparisons being made and the decisions made regarding the desired statistical performance. However, the number of background samples included in some of the data sets raise concerns over the adequacy of these data sets. For example, weathered bedrock soil background set only contains samples and the surface water and sediment background data sets only contain seven to ten samples. These relatively small data sets may limit significantly the statistical performance of any statistical analysis used to compare site data to background. After a more complete review of the implication of these limited background data sets, the Navy may want to consider the collection of additional background data.

Response: Additional background data was pulled into to the background data sets for the surface water and sediment data sets, both estuarine and open water, as requested in the conference call on June 2, 2006.

The weathered bedrock data set was not expanded. This data set was only included for completeness and very little data is available for weathered bedrock in the first place. However, it is highly unlikely that a background comparison would be necessary for this media, since much of this formation is located in areas where no risk is posed to human health or the environment. Therefore, it is likely that this background data set will never be used.

Discussion related to the above response was included in the text of the report (see Sections 1.5 and 3.4.3). Surface water and sediment background data sets were modified to include additional samples (see Section 5 tables).

2. When discussing the treatment of censored data (non-detects), the Background Report (page 1-6) indicates that for data sets with a frequency of detection (FOD) greater than 50 percent, descriptive statistics were developed using surrogate values for the censored data. This does not appear consistent with the CERCLA Guidance (page 4-7), which indicates that if less than 15 percent of the background samples are non-detects, the distributions of the background sample may be determined using surrogate values. However, if more than 15 percent but less than 50 percent of the measurements in the background sample set are non-detects, the CERCLA Guidance recommends the use of specialized methods for analyzing non-detects and refers the reader to Section 4.7 of the DQO Guidance. The approach that was used in the Background Report to treat background data sets with between 15 and 50 percent non-detects does not appear to conform with the those recommended in the DQO Guidance. NAPR should ensure that the approach used to handle background data sets with between 15 and 50 percent non-detects is consistent with the CERCLA Guidance.

In addition, the Background Report (page 1-6) indicates that for data sets with a FOD of 50 percent or less, "the data set is truncated such that non-detect and blank results are not considered in the calculation of descriptive statistics." The Background Report further indicates that "although this will reduce the power of the calculated statistics, the use of non-

detect or blank results could yield an unacceptably large bias of any calculated statistics." This approach does not appear consistent with CERCLA Guidance. The CERCLA Guidance (page 4-7) indicates that for data sets with more than 50 percent nondetects, "it may not be possible to compare the means of two distributions," and indicates that "an alternative approach is to compare the upper percentiles of two distributions by comparing the proportion of the two populations that is above a fixed level." The DQO Guidance (page 4-50) suggests the use of the Test of Proportions to perform such a comparison. NAPR should ensure that the approach used to handle background data sets with greater than 50 percent non-detects is consistent with CERCLA guidance.

Response: CERCLA guidance indicates that background data sets should be used in parametric or non-parametric statistical tests to determine if site data sets are significantly different (above some predetermined level of confidence) from background data sets. The treatment of non-detects used in the Background Summary Report was only done in order to establish screening values for INITIAL screening of site data against background data. No statistical tests of comparison with site data were done with truncated data sets in the Background Summary Report. Those comparison tests, in accordance with CERCLA guidance, should be done during site investigation reporting if questions arise as to whether site data exceed background data.

Discussion was added to the text in Section 1.4 in order to address USEPA's concern above.

3. When discussing the use of background data sets, the Background Report (page 1-9) indicates that "the use of the upper limit of the means is warranted as an initial step in screening the analytical results for inorganics, consistent with the previous use of background data sets." It is not clear that this approach is consistent with the CERCLA Guidance. The use of this approach to initially screen site data relative to background should be justified based on the CERCLA Guidance.

Response: The use of setting an initial screening value is warranted based on the intended purpose of establishing inorganic background values, and falls under the "identify the decision" step. The purpose is to determine if one or more site concentrations are in excess of the background concentrations, or outside the distribution of the background data set. There are some recommended ways to do this. One is to use two times the average value, or mean, of the background concentrations. This is recommended in RAGS (USEPA. 2000. Supplemental Guidance to RAGS: Region 4 Bulletins, Human Health Risk Assessment Bulletins. EPA Region 4, originally published November 1995, Website version last updated May 2000: <http://www.epa.gov/region4/waste/oftecser/healthbul.htm>).

Another way is to use some percentile value of the data, such as the 90th or 95th percentile, as given by the distribution of the background data set. If a site concentration is less than or equal to this percentile value, there is no doubt that it represents background concentrations. This is the approach used in this document. A typical normal distribution is encompassed by 6 standard deviations, three on each side of the mean. In fact, an EPA Engineering Forum Issue (EPA/540/S-96/500, December 1995) states that:

"In some cases, it may be of interest to establish an upper limit of background for the site. This would be useful if the investigator wanted to compare single values for a soil type from the hazardous waste site with the background population for a similar soil. The mean background concentration plus 3 standard deviations comprises a reasonable maximum allowable or upper limit."

Since we are only interested in the higher concentrations, and since the data is not always normally

distributed, a conservative approach is to use the mean plus 2 standard deviations as an initial screening level. This method actually takes into account the scatter in the data set, as opposed to simply multiplying the mean by two. In addition, it is more conservative than using the mean plus three standard deviations.

In both methods, a site concentration in exceedance of the background screening level does not by itself warrant its determination to be non-background. Statistical tests, whether parametric or nonparametric, should be used to make that determination.

Discussion reflecting the above response was added to Section 1.5 of the Report

4. When discussing the analysis of outliers, the Background Report (page 1-7) states that "the discordance test is one of four recommended outlier tests," while referencing Navy guidance. Although the discordance test is referenced, the text does not clearly state how outliers were identified. It should be noted, however, that the DQO guidance (page 4-29) indicates that the discordance test is only suitable for identifying outliers for normally distributed data. The Background Report should clearly identify how each data set was analyzed to identify outliers. The Background Report should also verify that the outlier tests that were used are suitable for the distributions of the data sets tested.

Response: The DQO guidance gives four examples of outlier tests. Three of them require that the data set is normally distributed. The other one (Walsh's test) can be used for data that is not normally distributed, but requires data sets in excess of 60 values for a significance level of 0.10. None of the data sets in this document have more than 60 values. Therefore, all the data sets were assumed to be normally distributed for the purpose of conducting outlier tests.

Text reflecting the above response was added to Section 1.4.3.

5. The Background Report (page 1-7) indicates that an outlier test was conducted on data sets with a FOD of more than 50 percent. The text further indicates that "in general outliers should not be removed from the data set unless clear evidence shows that they are not based on elements of the population being studied and should not have been included in the data set." As indicated in the tables presenting the results of the background analysis for the individual media, outliers have been identified in a number of the data sets. However, none were removed from the data set because "no errors were found in the sample results." Although these data were not removed from the data set because no errors were found, the outlier tests indicate that these data likely do not belong to the statistical population being studied

When discussing outliers, the CERCLA guidance (page 4-6) indicates that:

The use of nonparametric hypothesis tests for background comparisons greatly reduces the sensitivity of test results to the presence of outliers. Parametric tests based on the lognormal distribution may yield results that are extremely sensitive to the presence of one or more outliers.

The CERCLA Guidance (page 5-6) further indicates that:

If the data sets contain outliers or non-detect values, an additional level of uncertainty is faced when conducting parametric tests. Since most environmental data sets do contain outliers and non-detect values, it is unlikely that the current widespread use of parametric tests is justified,

given that these tests may be adversely affected by outliers and by assumptions made for handling non-detect values.

Thus, the retention of the outliers in the background data sets will likely require that nonparametric tests be used when comparing these sets with site data, although distributional tests may identify the populations as normal or lognormal.

Response: An additional outlier test was performed (Dixon test) in order to determine if outliers were indeed present in the data sets, as requested in the June 2, 2006 conference call. However, for data sets with less than 20 values, the outliers were only removed if both outlier tests were positive for outliers. For data sets with greater than 20 values, the outliers were removed if either the Discordance test or the Dixon test were positive for outliers. Modified statistics were calculated on data sets with the outliers removed and are provided in the tables in the Report.

6. The Background Report (page 1-7) indicates that the Shapiro-Wilk's W-test was performed on all data sets with frequencies of detection over 50 percent. The text further indicates that "the W test is a 'goodness-of-fit' test considered to be effective for determining whether a data set can be described as 'normally' or lognormally distributed for sample sets with 50 or fewer samples." This statement is in agreement with the test of normality presented in the CERCLA Guidance (page 4-2). However, the CERCLA Guidance (page 5-3) adds further qualifications to the use of the W-test for determining normality by also indicating that:

Tests for the distribution of the data (such as the Shapiro- Wilk test for normality) often fail if there are insufficient data, if the data contain multiple populations, or if there is a high proportion of non-detects in the sample. Test for normality lack statistical power for small sample sizes. In this context, "small" may be defined roughly as less than 20 samples, either on site or in background areas.

Therefore, for small sample sizes or when the distribution cannot be determined, non parametric tests should be used to avoid incorrectly assuming the data are normally distributed when there is not enough information to test this assumption.

Many of the background data sets presented in the Background Report have less than 20 samples. Thus, it does not appear appropriate to use the results of the W -test to identify normally or lognormally distributed populations for purposes of later recommending the use of parametric over nonparametric tests. For those sample populations with less than 20 samples that are found to be normally or lognormally distributed using the W -test, the Background Report should either remove their designations as normally or lognormally distributed or clearly identify these designations as qualified based on sample size.

Response: Text was added to the report in order to clearly identify the normal or lognormal distributions as qualified based on sample size.

7. The data sets used to establish background for groundwater do not appear to include multiple measurements from the same background well over the period of a year or more. Consequently, these data sets may not adequately include any temporal variability inherent in background groundwater quality, such as that introduced by seasonal effects. NAPR should demonstrate that data sets used to establish groundwater background adequately represent seasonal and other temporal effects.

Response: A check of the sampling dates for groundwater revealed that the samples were taken in five different months, and most of the wells were only sampled once. Therefore, no temporal variability is able to be determined from the background groundwater data set. It would seem fairly intuitive that temporal variability would not be as important in a setting such as Puerto Rico which has a fairly uniform, year round climate. However, data from NAPR Landfill (SWMU 3) semi-annual sampling was analyzed for temporal variability. Typically this data is collected in March and September of each year. Data from March was compared to data from September using the Krustal-Wallis test to determine if the two data sets are statistically different. The H statistic was found to be 0.915 compared to the H statistic from the chi-square distribution for 1 degree of freedom of 2.705 (0.10 significance level). Since the H statistic was less than the chi-square statistic, the null hypothesis is accepted that they are both from the same distribution. No discussion is provided in the text.

8. Background for groundwater has been established without any apparent regard for the geologic strata from which the groundwater samples were derived. Frequently, groundwater quality is influenced by geochemical differences between the various geologic materials through which groundwater passes. NAPR should demonstrate that it is not necessary to establish separate groundwater backgrounds for each of the various strata present at the former Roosevelt Roads site. Otherwise, a separate groundwater background should be established for each geologic strata in which groundwater is present and in which contamination is present at the facility.

Response: Upon inspection of the boring logs in Appendix A, the majority of the groundwater samples for inclusion in the background data set were from formations primarily composed of clay. Since most of NAPR shallow groundwater geology is clay in nature, it is expected that the inorganics in the clay will represent an adequate background data set for comparison to site data for groundwater.

9. Table 5.3 indicates that the mean copper concentration in the data set used to establish background for estuarine wetland surface water is 12.2 micrograms per liter ($\mu\text{g/L}$). This is nearly three times the chronic marine ambient water quality criteria for copper (3.1 $\mu\text{g/L}$). NAPR should provide further discussion and/or analysis that demonstrate the suitability of the data set proposed for establishing background estuarine wetland surface water.

Response: The background database for NAPR estuarine surface water was expanded as indicated in Specific Comment No. 1. The mean copper concentration for estuarine wetland surface water was recalculated and found to be 8.89 $\mu\text{g/L}$ (based on 23 samples), slightly lower than the original value of 12.2 $\mu\text{g/L}$, but still higher than 3.1 $\mu\text{g/L}$. The outlier tests run on the data set indicated no outliers present. Therefore, it is concluded that copper is present naturally in higher concentrations at NAPR than elsewhere on the island.

10. Based on a comparison to EPA's National Coastal Assessment (NCA) data, concentrations of cadmium and selenium in NAPR's estuarine background sediment samples appear to be somewhat greater than typical background levels observed in Puerto Rico. For example, the mean cadmium concentration reported in Table 5-9 is 0.527 milligrams per kilogram (mg/kg), while only two of 43 samples in the NCA data set had detected cadmium concentrations of 0.5 mg/kg or greater. (NCA data were obtained in June 2005 from John Macauley of EPA's Environmental Effects Research Laboratory in Gulf Breeze, Florida). NAPR should discuss possible reasons for elevated cadmium and selenium concentrations in the background estuarine sediment samples and provide adequate justification for continued use of the data set proposed for background for cadmium and selenium in estuarine

sediments.

Response: The estuarine sediment background databases were expanded as indicated in response to Specific Comment No.1. The mean cadmium concentration in estuarine sediment was lowered slightly to 0.45 mg/kg and the mean selenium concentration was recalculated and found to be 0.67 mg/kg. It is concluded that these chemicals are present naturally in higher concentrations at NAPR than elsewhere on the island. In addition, a comparison to the Threshold Effects Level (TEL) reveals that these mean concentrations are below the TEL for those two compounds (see Baker, 2006, [SWMU 45 Screening Level Ecological Risk Assessment and Step 3A of the Baseline Ecological Risk Assessment](#)). Since the TEL is higher than the mean background, it would likely be used as a screening level, as opposed to the background level.

**EPA ADDITIONAL COMMENTS TO JULY 28, 2006 WORKING DRAFT
DATED AUGUST 15, 2006**

I GENERAL COMMENT

Our review of NAPR's July 28 Draft Response to Comments and Draft Revised Background Report indicates that the NAPR has adequately addressed many of our early comments. However, a few issues remain. These issues center around the proposed use of the background data to calculate a screening value that can be used to identify which downgradient data sets require further statistical analysis to determine if statistically significant increases over background have actually been observed. While acknowledging that Superfund Guidance does not clearly provide for the prescreening of data, NAPR has cited two documents in support of this approach. The first document cited was the Supplemental Guidance to RAGS:Region 4 Bulletins, Human Health Risk Assessment Bulletins. NAPR indicates that this guidance recommends that two times the background levels can be used for screening downgradient data. However, the Region 4 Web Site <http://www.epa.gov/region4/waste/ots/healthbul.htm> indicates that this suggested approach was for use when eliminating potential contaminants as COPCs during the early phases of risk assessment. Furthermore, the Web Site indicates that

Although RAGS allows the use of statistics in data evaluation, the use of statistics may not be sufficiently conservative at this stage of the BRA. In most cases, a sufficient number of samples will not be available for conducting a statistical analysis with appropriate power. Therefore, the OTS recommends the use of the twice background criterion. OTS should be consulted before using any type of statistical approach for comparison to background.

It's not clear if this situation directly applies to those in which the proposed background data are intended for use. Moreover, the data sets presented in the Background Report are ultimately intended for statistical analysis and presumably should be suitable for use in statistical analysis. Regardless, it is clear that the Region 4 guidance is intended to provide a conservative approach to identifying COPCs.

The second document quoted by NAPR is an Engineering Forum Issue (EPA/540/S-96/500). This document is indicated to suggest the use of some percentile value of the background data, such as the 90th or 95th percentile, of the background data set be used for screening.

This document could not be located for this review. Consequently, there is no way of evaluating the context of this guidance.

Regardless of the lack of precedent or guidance for the screening approach proposed by NAPR, it may provide a reasonable approach for limiting the number of complex statistical analyses that must be undertaken. The approach suggested establishes a screening level by adding two times the standard deviation to the mean of the background data set. This is equivalent to using tolerance intervals for evaluating downgradient data. The use of tolerance intervals is an established statistical procedure presented in a number of EPA guidance documents. The tolerance factor of two suggested for multiplying the standard deviation by is a reasonably conservative value and may well result in a number of positives that may subsequently be eliminated or confirmed by more sophisticated statistical analysis.

However, tolerance intervals are based on the assumption that the data are normally or lognormally distributed. Testing of the data sets presented in the Background Report generally indicates that the background data sets consisting of more than 50% detects are normally or nearly normally (or lognormally) distributed. Thus, the proposed screening approach appears acceptable for these background data sets.

The applicability of the proposed screening approach to the background data sets with less than 50% detects is less clear. These data sets have not been tested for normality. Moreover, when calculating the screening value for these data sets, NAPR has truncated the data, using only the detect values. This approach clearly removes a large portion of the sampled population (i.e., that portion below detection limits). Depending on the detection limits involved with the censored data, this approach likely biases the mean significantly, resulting in a higher mean than the characteristic of the sampled population. This approach is clearly not conservative. Since the initial screening of downgradient data should be conservative, the non detects should be included in the background data set used to calculate the screening values. The normal practice in such calculations is to substitute half the reporting limit for the undetected value. In addition, with the high number of non detect values, these populations are most likely to be best represented by a lognormal, rather than normal, distribution. Consequently, population statistics for these background data sets (mean and standard deviation) should be computed using the logs of the constituent values.

Based on the above considerations, the Navy may use the screening approach proposed in the Background Report. However, for all data sets with less than 50% detects, the entire data set must be used when computing population statistics. One half of the reporting values should be substituted for nondetect values, and all population statistics should be computed using the logs of the data.

Response: No changes were made to data sets with a frequency of detection greater than 50 percent (FOD category "D"). The Navy revised the data sets with the frequency of detection category of "C" (more than one detection but less than 50 percent detections) to include surrogate values for the non-detects. The surrogate value was one-half the reporting limit. No presumption of normality or lognormality was made for these revised data sets. Instead, the W-test for normality was performed on the data set. If the data set was found to be represented by a normal distribution, the mean and standard deviation of the revised data sets were calculated in a straightforward manner. If the data set was represented by a lognormal distribution, the data was transformed to their natural logs, and the mean and standard deviation of the transformed data set were calculated. A screening value of the mean plus two standard deviations was used for these compounds as indicated above, and the tables in the report were revised to reflect the new screening values for the FOD Category "C" compounds.

Slight textual modifications in Section 1.0 were necessary to reflect changes made by implementing the above process.

A few other concerns were noted in our review. In response to Specific Comment No. 1, NAPR has stated that “in general, the higher the number of samples in a data set, the stronger the statistics. However, for comparison purposes, if the number of site samples (e.g. from an investigation of a contaminated site) is approximately the same as the number of background samples, valid statistical comparisons can be made with confidence.” The basis for this statement is not evident; and the statement does not appear to be consistent with traditional statistical analysis which generally indicates that the power and confidence associated with statistical comparisons depend on the number of samples available to characterize the statistical populations undergoing comparison. The above cited statement should be removed from the final response to comments.

RESPONSE: *The above statement was removed from the final response to comments.*

In response to Specific Comment No. 2, NAPR has stated that CERCLA guidance indicates that “background data sets should be used in parametric or non-parametric statistical tests to determine if site data sets are significantly different (above some predetermined level of confidence) from background data sets. Further guidance is not given.” However, as the quotes included in the original comments indicate, the CERCLA guidance and the referenced DQO guidance provides a great deal of guidance regarding the use of parametric and nonparametric methods when comparing background and downgradient data. The above cited statement should be removed from the final response to comments.

RESPONSE: *The above statement was removed from the final response to comments.*