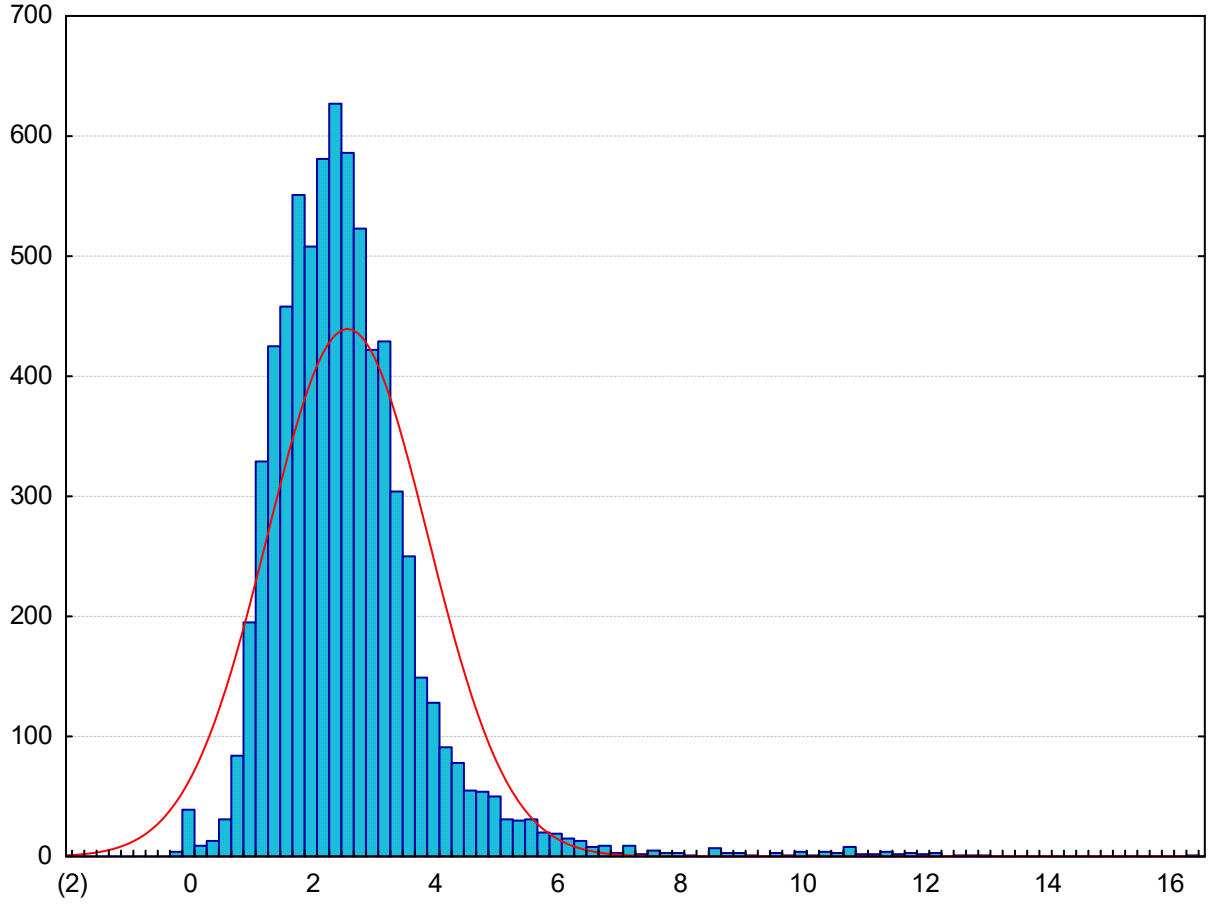
Statistical Methods Used in the Idaho National Laboratory Annual Site Environmental Report



A Supplement to the
INL Site Environmental Report for 2020

Statistical Methods Used in the Idaho National Laboratory Annual Site Environmental Report

Relatively simple statistical procedures are used to analyze data collected by the Idaho National Laboratory (INL) Environmental Monitoring Program. This supplement presents the methods used to evaluate sample results for this annual report.

Guidelines for Reporting Results

The results reported in the quarterly and annual reports are assessed in terms of data quality and statistical significance with respect to laboratory analytical uncertainties, sample locations, reported INL releases, meteorological data, and worldwide events that might conceivably have an effect on the INL environment.

Data Validation

First, field collection and laboratory information are reviewed to determine identifiable errors that would invalidate or limit use of the data. Examples of field observations that could invalidate the data include insufficient sample volume, torn filter or mechanical malfunction of sampling equipment.

Laboratory Qualification of Results

The analytical laboratory also qualifies the results and may reject them for reasons, such as:

- Uncertainty is too high to be accepted by the analyst
- Radionuclide has no supporting photopeaks to make a judgment
- Photopeak width is unacceptable by the analyst
- Result is below the decision critical level
- Other radionuclides display gamma-ray interferences
- A graphical display of analyzed photopeaks showed unacceptable fitting results
- There is no parent activity; therefore, the state of equilibrium is unknown and the radionuclide could not be quantified
- Radionuclide is a naturally occurring one with expected activity.

Evidence of laboratory cross-contamination or quality control issues could also disqualify a result (see Chapter 12).

Data that pass initial screening are further evaluated prior to reporting.

Guidelines for Interpreting Results of Radiochemical Analyses

The goal of the Environmental Surveillance, Education, and Research Program is to minimize the error of reporting a constituent is absent in a sample population when it is actually present. The approach used by the Environmental Surveillance, Education and Research Program to interpret individual analytical results is based on guidelines outlined by the U.S. Geological Survey in Bartholomay et al. (2019), which are based on an extension of the methodology proposed by Currie (1984). Most of the following discussion is taken from Bartholomay et al. (2019).

For radiological data, individual analytical results are usually presented in this report with plus or minus one sample standard deviation (± 1*s*). The sample standard deviation is obtained by propagating sources of analytical uncertainty in laboratory measurements. The uncertainty term, "*s*," is an estimate of the population standard deviation "σ," assuming a Guassian or normal distribution.

The laboratory measures a target sample and a laboratory-prepared blank. Instrument signals for the sample and blank vary randomly about the true signals. Therefore, it is essential to distinguish between two key aspects of the problem of detection: 1) the instrument signal for the sample must be greater than the signal observed for the blank before a decision can be made that the radionuclide was detected, and (2) an estimation must be made of the minimum radionuclide concentration that will yield a sufficiently large observed signal before a correct decision can be made for detection or nondetection of the radionuclide. The first aspect of the problem is a qualitative decision based on an observed signal and a definite criterion for detection. The second aspect of the problem is an estimation of the detection capabilities of a given measurement process.

In the laboratory, instruments must exceed a critical level (L_C) before the qualitative decision can be made as to whether the radionuclide was detected. Using algorithms in Currie (1984) that are appropriate for our data, the L_C is 1.6*s*. At 1.6*s*, there is about a 95-percent probability that the correct conclusion—not detected—will be made. Given a large number of samples, as many as 5 percent of the samples with measured concentrations greater than or equal to 1.6*s*, concluded as detected, might not contain the radionuclide. These measurements are referred to as false positives and are errors of the first kind in hypothesis testing.

Once the critical level has been defined, the minimum detectable concentration, or detection level (L_D), may be determined. Using the equations in Currie (1984), concentrations that equal 3s represent a measurement at the minimum detectable concentration. For true concentrations of 3s or larger, there is 95 percent or larger probability that the radionuclide was detected in a sample. In a large number of samples, the conclusion, not detected, will be made in 5 percent of the samples that contain true concentrations at the minimum detectable concentration of 3s. These are referred to as false negatives or errors of the second kind in hypothesis testing.

Actual radionuclide concentrations between 1.6s and 3s have larger errors of the second kind. That is, there is a larger-than-five-percent probability of false negative results for samples with true concentrations between 1.6s and 3s. Although the radionuclide might have been detected, such detection may not be considered reliable; at 2s, the probability of a false negative is about 50 percent.

In this report, radionuclide concentrations less than 3s are considered to be below a reporting level. The critical level, minimum detectable concentration, and reporting level aid the reader to interpret analytical results and do not represent absolute concentrations of radioactivity, which may or may not have been detected. In this report concentrations equal to or above 3s are reported as “detected”. Results between 2s and 3s are considered to be “questionable” detections. Results below 2s are considered to be “undetected.”

Each result is reported with the associated 1s uncertainty value for consistency with other INL reports. To determine if an analytical result is statistically detected (i.e., at or above the reporting level), the result must equal or exceed three times the uncertainty. For example, a radionuclide concentration of 10 ± 2 picocuries per liter (pCi/L) would be considered to be detected because 10 > 3*2.

Statistical Tests Used to Assess Data

An example data set is presented here to illustrate the statistical tests used to assess data collected by the Environmental Surveillance, Education, and Research contractor. The data set is the gross beta environmental surveillance data collected from January 8, 1997, through December 26, 2001. The data were collected weekly from several air monitoring stations located around the perimeter of the INL Site and air monitoring stations throughout the Snake River Plain. The perimeter locations are termed “boundary,” and the Snake River Plain locations are termed “distant.” There are seven boundary locations: Arco, Atomic City, Birch Creek, Federal Aviation Administration (FAA) Tower, Howe, Monteview, and Mud Lake; and five distant locations: Blackfoot, Blackfoot Community Monitoring Station (CMS), Craters of the Moon, Idaho Falls, and

Rexburg CMS. The gross beta data are of the magnitude 10^{-15}. To simplify the calculations and interpretation, these have been coded by multiplying each measurement by 10^{15}.

Only portions of the complete gross beta data set will be used. The purpose of this task is to evaluate and illustrate the various statistical procedures, and not a complete analysis of the data.

Test of Normality

The first step in any analysis of data is to test for normality. Many standard statistical tests of significance require that the data be normally distributed. The most widely used test of normality is the Shapiro-Wilk W-Test (Shapiro and Wilk 1965). The Shapiro-Wilk W-Test is the preferred test of normality because of its good power properties as compared to a wide range of alternative tests (Shapiro et al. 1968). If the W statistic is significant ($p<0.00001$), then the hypothesis that the respective distribution is normal should be rejected.

Graphical depictions of the data should be a part of any evaluation of normality. The following histogram (Figure 1) presents such a graphical depiction along with the results of the Shapiro-Wilk W-Test. The data used for the illustration are the five years of weekly gross beta measurements for the Arco boundary location. The W statistic is highly significant ($p<0.0001$), indicating that the data are not normally distributed. The histogram shows that the data are asymmetrical with right skewness. This skew suggests that the data may be lognormally distributed. The Shapiro-Wilk W-Test can be used to test this distribution by taking the natural logarithms of each measurement and calculating the W statistic. Figure 2 presents this test of lognormality. The W statistic is not significant ($p=0.80235$), indicating that the data are lognormal.

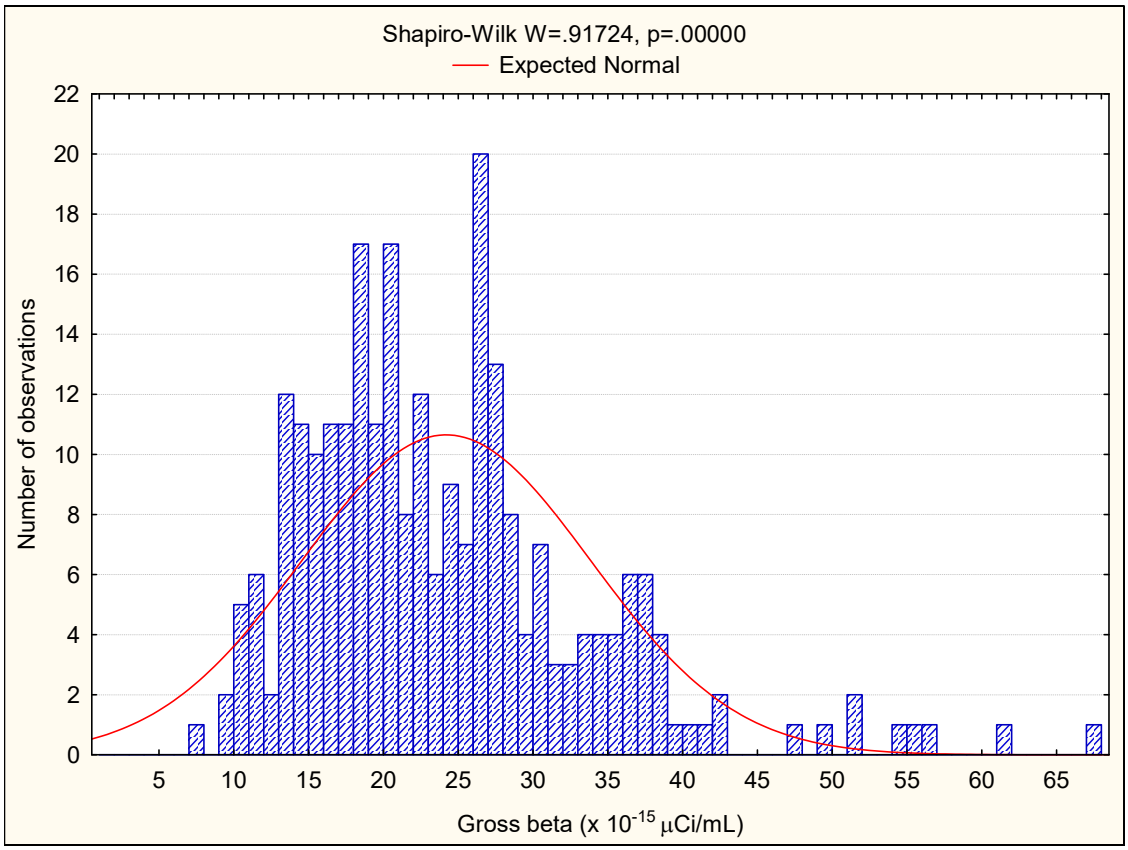


Figure 1. Test of Normality for Arco Gross Beta Data.

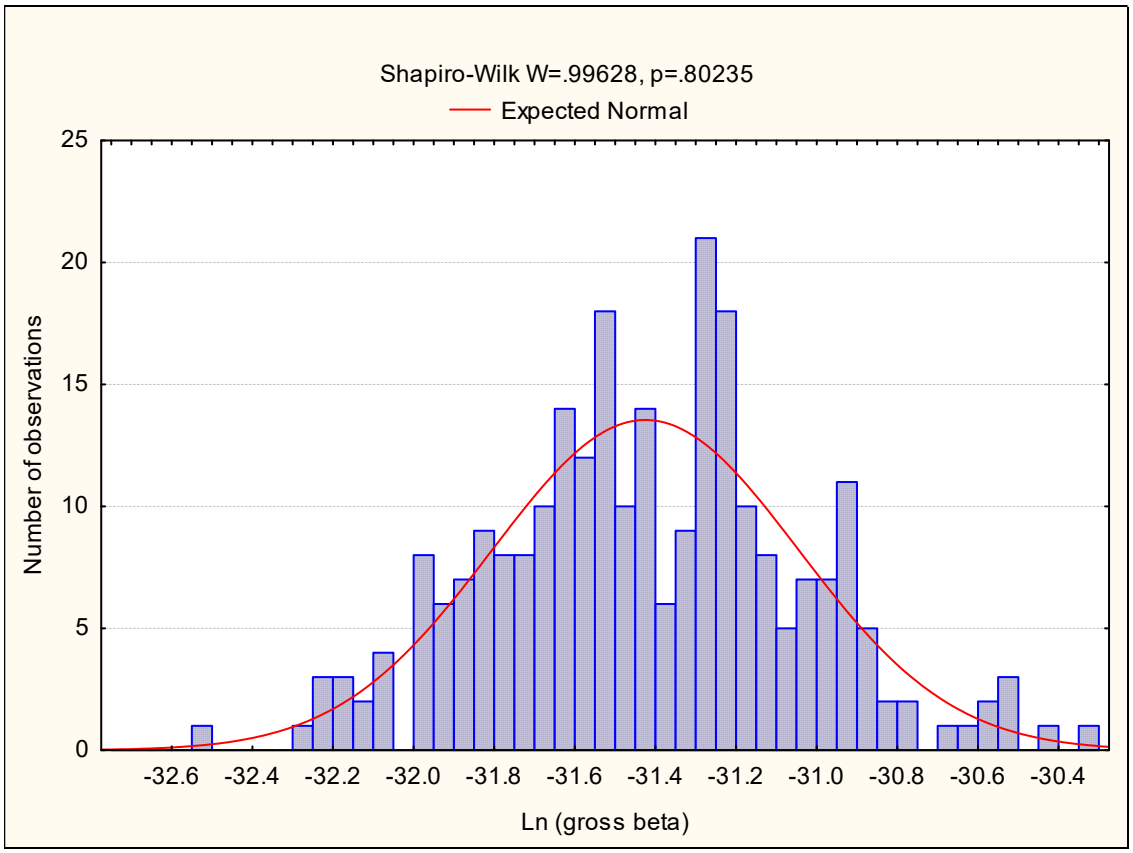


Figure 2. Test of Lognormality for Arco Gross Beta.

To perform parametric tests of significance such as Student's T-Test or One-Way Analysis of Variance (ANOVA), all data are required to be normally (or lognormally) distributed. Therefore, if one desires to compare gross beta results of each boundary location, tests of normality must be performed before making such comparisons. Table 1 presents the results of the Shapiro-Wilk W-Test for each of the seven boundary locations.

From Table 1, none of the locations consist of data that are normally distributed, and only some of the data sets are lognormally distributed. This is a typical result and a common problem when one desires to use a parametric test of significance. When many comparisons are to be made, attractive alternatives are nonparametric tests of significance.

Table 1. Tests of Normality for Boundary Locations.

	Normal		Lognormal	
Location	**W Statistic**	**p-Value**	**W Statistic**	**p-Value**
Arco	0.9172	<0.0001	0.9963	0.8024
Atomic City	0.9174	<0.0001	0.9411	<0.0001
Birch Creek	0.8086	<0.0001	0.9882	0.0530
FAA Tower	0.9119	<0.0001	0.9915	0.1397
Howe	0.8702	<0.0001	0.9842	0.0056
Monteview	0.9118	<0.0001	0.9142	<0.0001
Mud Lake	0.6130	<0.0001	0.9704	<0.0001

Comparison of Two Groups

For comparison of two groups, the Mann-Whitney U-Test (Hollander and Wolfe 1973) is a powerful nonparametric alternative to the Student's T-Test. In fact, the U-Test is the most powerful (or sensitive) nonparametric alternative to the T-Test for independent samples; in some instances, it may offer even greater power to reject the null hypothesis than the T-Test. The interpretation of the Mann-Whitney U-Test is essentially identical to the interpretation of the Student's T-Test for independent samples, except that the U-Test is computed based on rank sums rather than means. Because of this fact, outliers do not present the serious problem that they do when using parametric tests.

Suppose one wants to compare all boundary locations to all distant locations. Figure 3 presents the box plots for the two groups. The median is the measure of central tendency most commonly used when there is no assumed distribution. It is the middle value when the data are ranked from smallest to largest. The 25th and 75th percentiles are the values such that 75 percent of the measurements in the data set are greater than

the 25th percentile, and 75 percent of the measurements are less than the 75th percentile. The large distance between the medians and the maximums shown in Figure 3 indicates the presence of outliers. It is apparent that the medians are of the same magnitude, indicating graphically that there is probably not a significant difference between the two groups.

The Mann-Whitney U-Test compares the rank sums between the two groups. In other words, for both groups combined, it ranks the observations from smallest to largest. Then, it calculates the sum of the ranks for each group and compares these rank sums. A significant p-value ($p<0.05$) indicates a significant difference between the two groups. The p-value for the comparison of boundary and distant locations is not significant ($p=0.0599$). Therefore, the conclusion is that there is not strong enough evidence to say that a significant difference exists between boundary and distant locations.

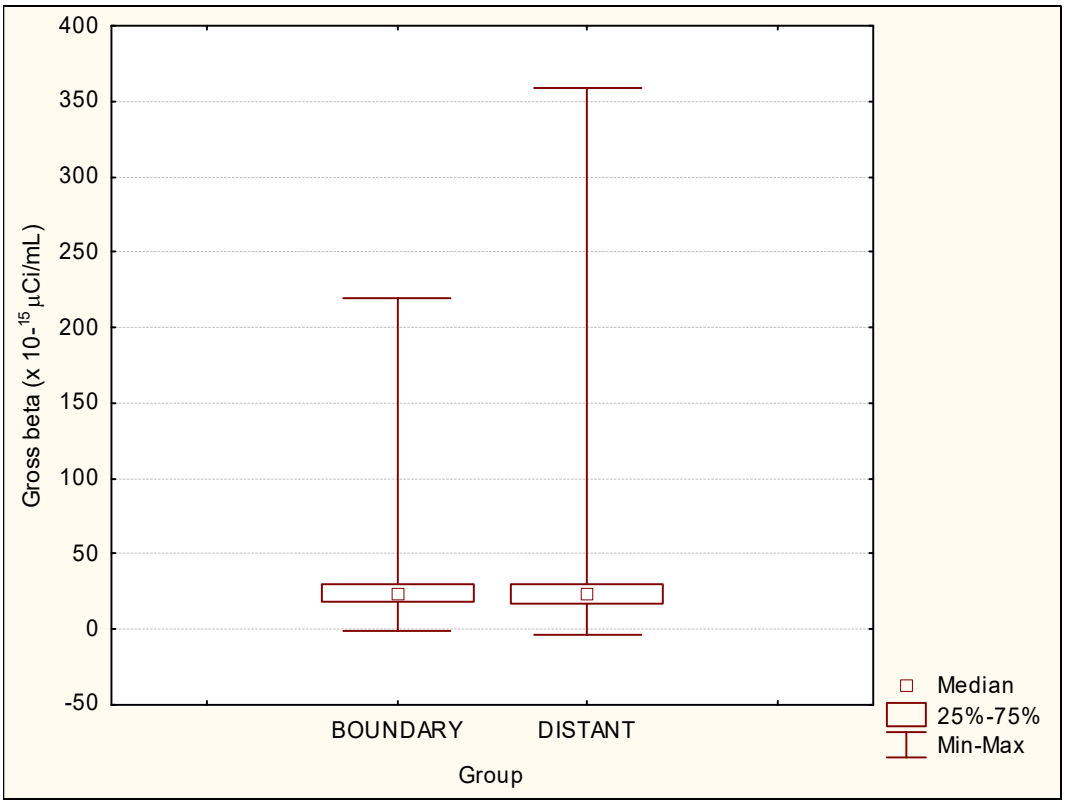


Figure 3. Box Plot of Gross Beta Data from Boundary and Distant Locations.

Comparison of Many Groups

Now suppose one wants to compare the boundary locations among themselves. In the parametric realm, this is done with a One-Way Analysis of Variance (ANOVA). A nonparametric alternative to the One-Way ANOVA is the Kruskal-Wallis ANOVA (Siegel and Castellan 1988). The test assesses the hypothesis that the different samples in the comparison were drawn from the same distribution or from distributions with the same median. Thus, the interpretation of the Kruskal-Wallis ANOVA is basically identical to that of the parametric One-Way ANOVA, except that it is based on ranks rather than means. That is, each of the results is replaced by a rank. The smallest value is replaced by 1, the next smallest by 2, and the largest by rank N, where N is the total number of independent observations. The average rank is then computed for each location group. If the samples are from the same populations, the average ranks should be about the same. If the populations are from different populations, the average ranks should differ.

Figure 4 presents the box plot for the boundary locations. Table 2 gives the number of samples, medians, minimums, and maximums for each boundary location. The Kruskal-Wallis ANOVA test statistic (*KW*) is highly significant ($p<0.0001$), indicating a significant difference among the seven boundary locations. When the obtained value of *KW* is significant, it indicates that at least one of the groups is different from at least one of the others. It does not identify which ones are different.

A post-hoc comparison of mean ranks of all pairs of groups can be conducted to determine which groups are different (see Siegel and Castellan, 1988). The differences between mean ranks are first calculated for all pairs of groups. When there are k groups, k(k-1)/2 comparisons are possible. To find which of the comparisons is significant, the critical value of z for multiple comparison is used. The differences between each pair of groups is significant if the estimated value of *z* exceeds the critical *z* value for multiple comparisons. Table 3 presents the results of a multiple comparison of ten years (2009-2018) of gross beta activity at current air sampling locations. Location by location comparisons show that less than 12% of the comparisons show a significant difference. The greatest number of statistical differences with other locations are associated with Craters of the Moon and Mud Lake. The lowest concentrations are associated with Craters of the Moon and the highest with Mud Lake.

Table 2. Summary Statistics for Boundary Locations.

Location	Number of Samples	Median (10^{-15} μCi/mL)	Minimum (10^{-15} μCi/mL)	Maximum (10^{-15} μCi/mL)
Arco	258	22.49	7.53	67.66
Atomic City	260	23.61	1.13	72.20
Birch Creek	234	23.15	-0.52	117.00
FAA Tower	260	21.90	3.59	72.78
Howe	260	24.55	3.95	90.10
Monteview	260	25.30	1.03	80.10
Mud Lake	260	24.85	4.30	219.19

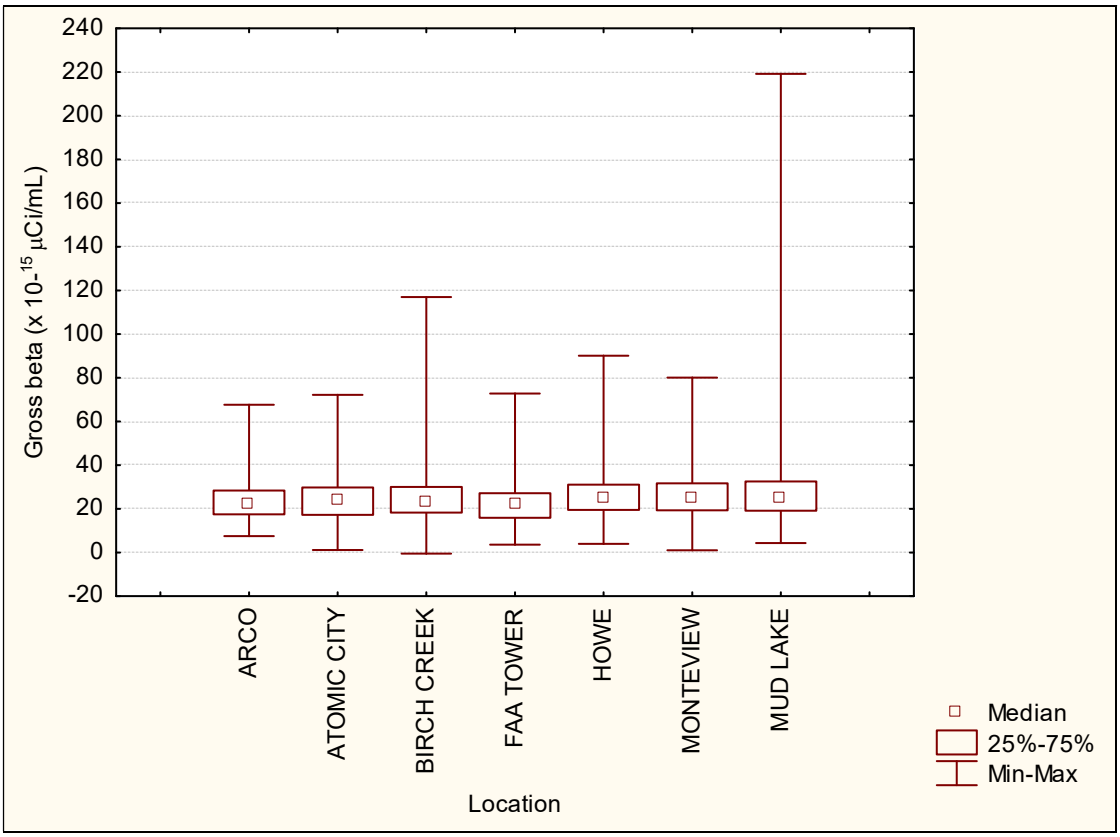


Figure 4. Box Plot of Gross Beta Data for Each Boundary Location.

Table 3. Multiple Comparisons of Ten Years (2009-2018) of Gross Beta Results by Location. A p-value (probability value) greater than 0.05 signifies no statistical difference between data groups. Any values below 0.05 are indicated in red.

Multiple Comparisons p values (2-tailed); Transformed result (x E-14) (TenYearData.sta in TenYearDataBeta)
Independent (grouping) variable: Location
Kruskal-Wallis test: H (15, N= 8184) =76.65357 p =.0000

Depend.: Transformed result (x E-14)	MUD LAKE R:4526.7	MONTEVIEW R:4124.4	EFS R:4220.2	FAA TOWER R:3859.0	MAIN GATE R:4288.0	VAN BUREN GATE R:4289.0	ARCO R:4163.0	ATOMIC CITY R:4237.1	IDAHO FALLS R:3937.0	HOWE R:4247.6	BLACKFOOT R:4023.3	DUBOIS R:3880.9	BLUE DOME R:3762.9	Craters of the Moon R:3646.1	REXBURG/ SUGAR CITY R:3978.5	**JACKSON WY R:4326.1**
MUD LAKE		0.745797	1.000000	0.000670	1.000000	1.000000	1.000000	1.000000	0.007028	1.000000	0.075449	0.001305	0.000025	0.000000	0.022744	1.000000
MONTEVIEW	0.745797		1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.139201	1.000000	1.000000
EFS	1.000000	1.000000		1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.225572	0.011413	1.000000	1.000000
FAA TOWER	0.000670	1.000000	1.000000		0.448686	0.411559	1.000000	1.000000	1.000000	0.993657	1.000000	1.000000	1.000000	1.000000	1.000000	0.259277
MAIN GATE	1.000000	1.000000	1.000000	0.448686		1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.703824	0.046499	0.001726	1.000000	1.000000
VAN BUREN GATE	1.000000	1.000000	1.000000	0.411559	1.000000		1.000000	1.000000	1.000000	1.000000	1.000000	0.649761	0.041248	0.001458	1.000000	1.000000
ARCO	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	1.000000	1.000000	1.000000	0.787212	0.053340	1.000000	1.000000
ATOMIC CITY	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	1.000000	1.000000	0.153707	0.007176	1.000000	1.000000
IDAHO FALLS	0.007028	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
HOWE	1.000000	1.000000	1.000000	0.993657	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	0.118942	0.005249	1.000000	1.000000
BLACKFOOT	0.075449	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	1.000000	1.000000	1.000000
DUBOIS	0.001305	1.000000	1.000000	1.000000	0.703824	0.649761	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000	1.000000	1.000000	0.410341
BLUE DOME	0.000025	1.000000	0.225572	1.000000	0.046499	0.041248	0.787212	0.153707	1.000000	0.118942	1.000000	1.000000		1.000000	1.000000	0.026087
Craters of the Moon	0.000000	0.139201	0.011413	1.000000	0.001726	0.001458	0.053340	0.007176	1.000000	0.005249	1.000000	1.000000	1.000000		1.000000	0.000961
REXBURG/SUGAR CITY	0.022744	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000		1.000000
JACKSON WY	1.000000	1.000000	1.000000	0.259277	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.410341	0.026087	0.000961	1.000000	

If desired, one can identify pairs of locations of interest and test those for significant differences using the Mann-Whitney U-Test. It is cautioned that all possible pairs should not be tested, only those of interest. As the number of pairs increases, the probability of a false conclusion also increases.

Suppose a comparison between Arco and Atomic City is of special interest due to their close proximity to each other. A test of significance using the Mann-Whitney U-Test results in a p-value of 0.7288, indicating no significant difference exists between gross beta results at Arco and Atomic City. Other pairs similarly can be tested, but with the caution given above.

Tests for Trends over Time

Regression analysis is used to test whether or not there is a significant positive or negative trend in gross beta concentrations over time. To illustrate the technique, the regression analysis is performed for the boundary locations as one group and the distant locations as another group. The tests of normality performed earlier indicated that the data were closer to lognormal than normal. For that reason, the natural logarithms of the original data are used in the regression analysis. Regression analysis assumes that the probability distributions of the dependent variable (gross beta) have the same variance regardless of the level of the independent variable (collection date). The natural logarithmic transformation helps in satisfying this assumption.

Figure 5 presents a scatter plot of the boundary data with the fitted regression line superimposed. Figure 6 presents the same for the distant data. Table 3 gives the regression equation and associated statistics. There appears to be slightly increasing trends in gross beta over time for both the boundary and distant locations. A look at the regression equations and correlation coefficients in Table 4 confirms this. Notice that the slope parameter of the regression equation and the correlation coefficient are equal. This is true for any linear regression fit. So, a test of significant correlation is also a test of significant trend. The p-value associated with testing whether or not the correlation coefficient is different from zero is the same as for testing if the slope of the regression line is different from zero. For both the boundary and distant locations, the slope is significantly different from zero and positive, indicating an increasing trend in gross beta over time.

Also of importance in Figures 5 and 6 is the obvious cyclical trend in gross beta. It appears as if the gross beta measurements are highest in the summer months and lowest in the winter months. Because the regression analysis performed above is over several years, a positive trend over time can still be detected even though it is

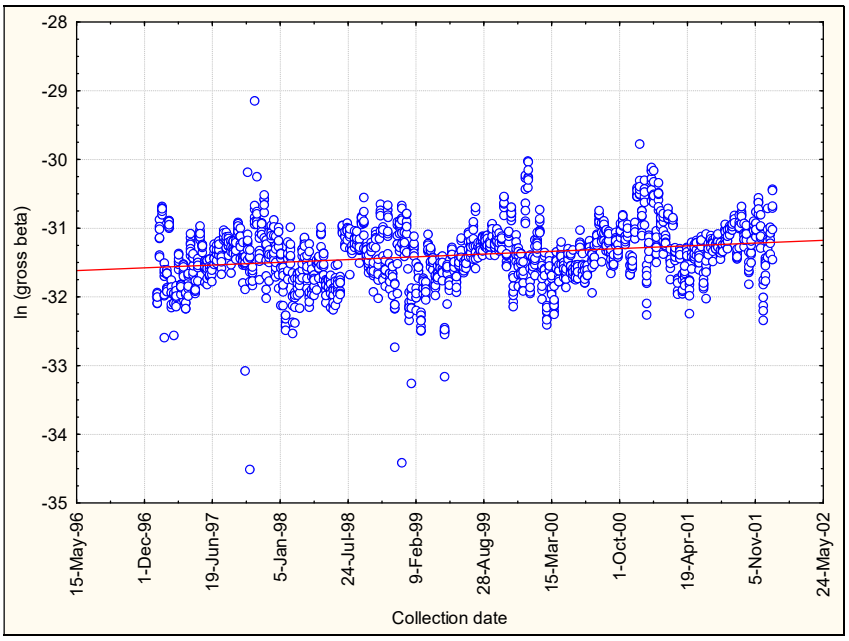


Figure 5. Scatter Plot and Regression Line for ln (Gross Beta) from Boundary Locations.

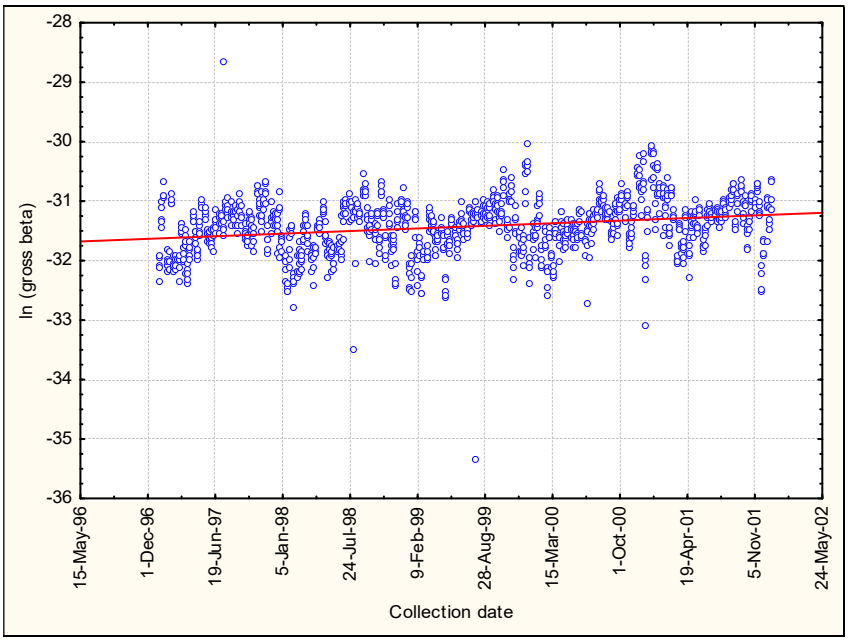


Figure 6. Scatter Plot and Regression Line for ln (Gross Beta) from Distant Locations.

Table 4. Regression Equations and Associated Statistics for Boundary and Distant Locations.

Sample Group	Regression Equation	Correlation Coefficient	p-value
Boundary	ln (gross beta) = -38.7 + 0.245 × (date)	0.245	<0.0001
Distant	ln (gross beta) = -39.4 + 0.253 × (date)	0.253	<0.0001

confounded somewhat by the existence of a cyclical trend. This ability to detect a positive trend is important because a linear regression analysis performed over a shorter period may erroneously conclude a significant positive or negative trend, when in fact, it is a portion of the cyclical trend.

Comparison of Slopes

A comparison of slopes between the regression lines for the boundary locations and distant locations indicates if the rate of change in gross beta over time differs with location. The comparison of slopes can be performed by constructing 95 percent confidence intervals about the slope parameter (Neter and Wasserman 1974). If these intervals overlap, it can be concluded that there is no evidence to suggest a difference in slopes for the two groups of locations.

A confidence interval for the slope is constructed as shown in Equation (1):

$$b - t_{0.025,n-2}s_b \leq \beta \leq b + t_{0.025,n-2}s_b \quad (1)$$

Where

b = point estimate of the slope

$t_{0.025,n-2}$ = the Student's t-value associated with two-sided 95 percent confidence and n-2 degrees of freedom

s_b = the standard deviation of the slope estimate, b

β = the true slope, which is unknown.

Table 5 gives the values used in constructing the confidence intervals and the resulting confidence intervals. As seen in the fifth column of Table 4, the confidence intervals for the slope overlap, and it can be concluded that there is no difference in the rate of change in gross beta measurements for the two location groupings, boundary and distant.

Table 5. Ninety-five Percent Confidence Intervals on the True Slope.

Sample Group	b	z[a]	s_b	95% CI[b]
Boundary	0.245	1.96	0.0229	[0.200, 0.290]
Distant	0.253	1.96	0.0269	[0.200, 0.306]

a. For large sample sizes, the standard normal z-value is used instead of the Student's t-value.
b. CI = confidence interval.

Calculating Upper Statistical Limits

It is valuable to compare current measurements with historical data to help decide if any results exceed expected values and thus require further evaluation. To establish background levels and determine outliers in data sets, ESER has adopted the use of the upper tolerance level (UTL). The 99%/95% UTL is a value such that 99% of the population (all possible air measurements) is less than the UTL with 95% confidence. With a 99%/95% UTL it is expected that approximately 1% of the measurements will exceed the UTL if the result is within the normal range. This means that if a concentration exceeds the UTL it does not necessarily indicate that the result is outside of the normal range. Rather, it indicates that the measurement should be closely examined to determine if it is unusually high.

The ProUCL statistical software package (https://www.epa.gov/land-research/proucl-software), initially developed by the Environmental Protection Agency, was used to compute the UTLs used by the ESER Program as decision limits. For example, Table 5 shows the UTLs for gross alpha and gross beta activity in air, calculated using ten years (2009-2018) of historical data.

Table 5. Decision limits for gross alpha and gross beta concentrations in air, based on ten years (2009-2018) of air monitoring data.

Constituent	UTL (µCi/mL)
Gross Alpha	3.98E-15
Gross Beta	6.38E-14

Figure 7 shows the UTL for gross beta concentrations compared to measurements made during the January 2019.

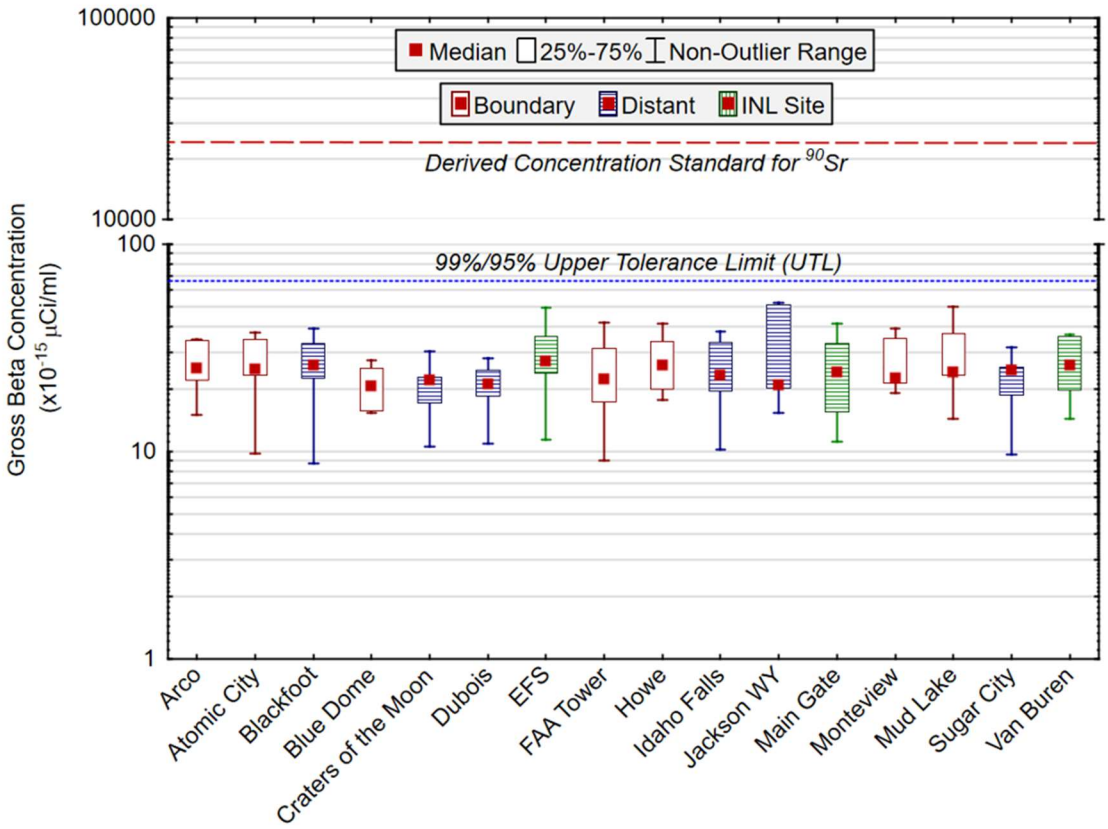


Figure 7. January 2019 gross beta concentrations in air at ESER INL Site, Boundary, and Distant locations. Number of samples (N) = 5 at each location.

References

Bartholomay, R. C., Neil V. Maimer, Gordon W. Rattray, and Jason C. Fisher, 2019, An Update of Hydrologic Conditions and Distribution of Selected Constituents in Water, Eastern Snake River Plain Aquifer and Perched Groundwater Zones, Idaho National Laboratory, Idaho, Emphasis 2016–18, Scientific Investigations Report 2019-5149 (DOE/ID-22251), U.S. Geological Survey.

Currie, L. A., 1984, Lower Limit of Detection-Definition and Elaboration of a Proposed Position for Radiological Effluent and Environmental Measurements, NUREG/CR-4007, U.S. Nuclear Regulatory Commission.

Neter, J. and W. Wasserman, 1974, Applied Linear Statistical Models, Homewood, Illinois: Richard D. Irwin, Inc.

Siegel, S. and N. J. Castellan, Jr., 1988, Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Company.

Shapiro, S. S. and M. B. Wilk, 1965, “An Analysis of Variance Test for Normality (complete samples),” Biometrika, Vol. 52, pp. 591 – 611.

Shapiro, S. S., M. B. Wilk, and H. J. Chen, 1968, “A Comparative Study of Various Tests of Normality,” Journal of the American Statistical Association, Vol. 63, pp. 1343 – 1372.