A Statistical Summary and Visualization Tool for a 30-year Background Soil and Sediment Metals Data from North Carolina Superfund Sites

Matthew C. Ogwu[1*], Frances M. Nilsen[2*], William F. Hunneke[2], Landon Norris[2], Robert J. Kelley[2], Paul P. Goodwin[3], Matthew A. Nichols[4], Alexis R. VanVenrooy[5], James T. Bateson[2]

[1]Goodnight, Family Department of Sustainable Development, Appalachian State University, 212 Living Learning Center, 305 Bodenheimer Drive, Boone, NC 28608, USA; ogwumc@appstate.edu
[2]North Carolina Department of Environmental Quality, Division of Waste Management, Superfund Section, 217 W Jones St, Raleigh, NC 27603; frannie.nilsen@deq.nc.gov
[3]North Carolina State University, Graduate Program, 2101 Hillsborough Street, Raleigh, NC 27695
[4]Florida Institute of Technology, Dept of Marine Biology, 150 W. University Blvd. Melbourne, FL 32901
[5]Rice University, Chemistry Dept, 6100 S Main, Houston, TX 77005
*Corresponding author

Supplementary Information I: NC Soil Metals Dashboard Features and User Information

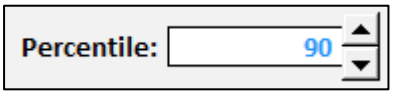
1. Accessibility

The dashboard's layout, color schemes, and overall design aesthetics promote user-friendliness while the placement of key elements such as menus, filters, and interactive components ensures intuitive navigation. The features of the dashboard comply with Web Content Accessibility Guidelines (WCAG) for users with disabilities. However, there is no availability of language options or translations for a diverse user base.
The NC Soil Metals Dashboard can be found in the Supplementary Information III in the form of an Excel spreadsheet (Microsoft Corporation, Bellevue, Washington, USA) and on the NC DEQ website at ……

2. Dashboard Overview

The dashboard (*Figure 1*) provides several statistical (tabular and graphical) highlights for each of the "Background Metals in Soil at North Carolina CERCLIS Sites." The dashboard is interactive and provides options to customize the presented material for the 18 different background metals (units for all metals and statistics are in mg/kg). The following sections will cover the dashboard's usage along with a summary of any statistical methods used to describe the data.
It should be noted that some versions of Microsoft Excel may not support the heat map. To view this dashboard, it is recommended to update to the most recent version of Office 365. In the case where the heat map is not viewable, there is a check box ("Show Map" on the chart), which can be unchecked to change the heat map to the NCDEQ logo.

3. Dashboard Components



Menu Bar

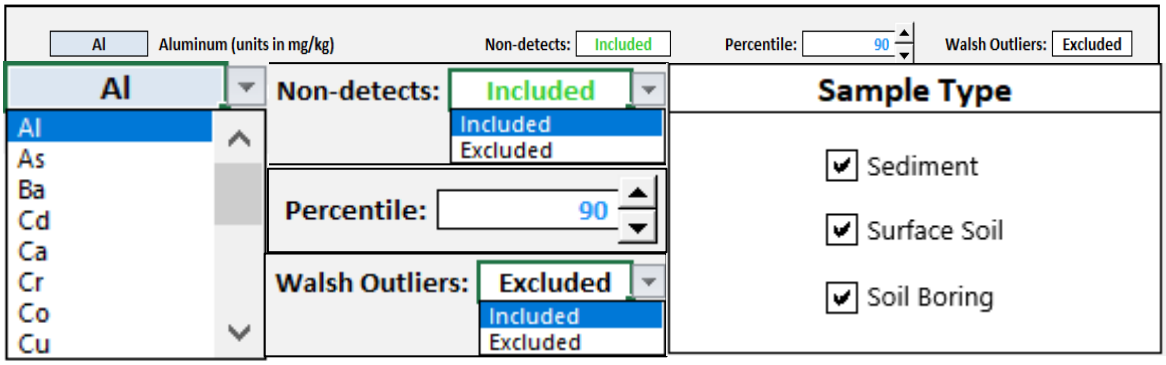


Figure A1. Menu bar options of the NC Soil Metals Dashboard

- The first menu option (*Figure A1; top and left*) is in the top left corner of the gray-filled box and allows you to change the metal presented by the dashboard. Clicking on this option will show a drop-down list of the available metals.

- The second menu option (*Figure A1; center top)* available is the inclusion or exclusion of non-detects. This can be found in the top middle (to the right of the first option). The box to include or exclude non-detects is directly to the right of "Non-detects:" with green text. The default is to have non-detects included, however, if non-detects are chosen to be excluded the text will turn red. Changing the non-detects option impacts the following:
    - The "Nonparametrics" section changes between performing calculations on "All Data" (for including non-detects) and "Detects Only" (for excluding non-detects). This option will only change calculations for "Median" and "90th %" (or whichever percentile is currently selected).
    - The county "median" heat map will change between displaying values from "All Data" (for including non-detects) and "Detects Only" (for excluding non-detects).

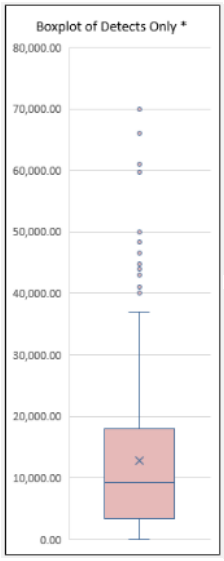
- The third menu option (*Figure A1; center middle*) allows the user to choose a specific percentile (with the default being 90%). This option, located to the right of the second option, computes the specified percentile for a selected metal. This number can be set to any value between 1-99 (anything above or below will throw an error). You can set this number by either manually typing it in, or by clicking on the up/down arrows to increase/decrease the chosen percentile in increments of 1. The value for the percentile will be shown under the "Nonparametrics" box in the dashboard.

- The fourth menu option (*Figure A1; center bottom*) is the inclusion or exclusion of Walsh outliers, located in the top right corner of the gray-filled box. Details for the determination of these outliers are explained in the "Walsh Outliers" section. The default for this option is to exclude the Walsh outliers. If you choose to include Walsh outliers, a yellow fill will appear to highlight this decision, this fill will also appear in the bottom left corner of the dashboard. Including the Walsh outliers retains certain extreme values in certain calculations.

- All outputs that this option impacts have a "*" next to their names and are listed below:
    - Boxplot of Detects Only [Chart]
    - Boxplot of Detects [Table]
    - Non-parametrics
    - Summary of Detects
    - Kaplan-Meier
    - County Heat Map

- The fifth (last) menu option (*Figure A1, right*) is the selection of sample type(s) to subset/filter the analysis. It is located directly under the listed Walsh Outliers [Table]. The data is based on the following sample types: sediment, surface soil, and soil boring. The results use all the data (all three sample types) by default. However, the user can uncheck unwanted sample types by clicking the sample type(s) that they want to exclude. Excluding a sample type will change all aspects of the dashboard.

Tables and Charts – Boxplots



Boxplot of Detects *	
Statistic	Value
Q1 (25th %)	3,400.00
Median	9,200.00
Q3 (75th %)	18,000.00
Bot. Whisker	33.00
Top Whisker	37,000.00

Boxplot Definitions	
Box	IQR (Q3 - Q1)
Box line	Median
Bot. Whisker	Q1 - 1.5*IQR
Top Whisker	Q3 + 1.5*IQR
X	Mean
o	Outlier

Figure A3. Boxplot and associated data tables in the NC Soil Metals Dashboard.

Boxplot of Detects Only [Chart]
The "Boxplot of Detects Only" shows a boxplot of the detected values for a selected metal. This box plot is automatically generated in Microsoft Excel and appears as presented in *Figure A3 (left)*. The boxplot is impacted based on the inclusion or exclusion of Walsh Outliers. Information about the symbols used in the boxplot is described in the "Boxplot Definitions" section.

Boxplot of Detects [Table]
The "Boxplot of Detects" provides the statistical values for various components of the boxplot (*Figure A3; right top*).

- The first statistic is Q1 (25th %). This represents the first quartile, or 25th percentile, which can be visualized on the boxplot as the bottom line of the box (where the bottom whisker ends).
- The median is the middle number in the data and is the middle line in the box.

- Q3 (75th %) represents the third quartile or the 75th percentile. This is seen as the top line of the box (where the top whisker begins).
- The bottom whisker starts at the smallest value at or above Q1-1.5*IQR (spread of the middle half of the distribution) and goes until it reaches Q1, where IQR (interquartile range) = Q3 – Q1. The top whisker starts at Q3 and goes until it reaches the greatest value below or at Q3+1.5*IQR.

Boxplot Definitions [Table]

"Boxplot Definitions" provides the definitions for different statistics of the boxplot (*Figure A3, right, bottom*).

- The boxed rectangle of the boxplot shows the area between the first quartile and the third quartile. This area is the interquartile range (or IQR for short).
- The line that appears within the box represents the median.
- The bottom and top whiskers are the same as described in the "Boxplot of Detects [Table]" section.
- The 'X' symbol represents the mean (or average).
- The 'o' symbol represents outliers (computed by Microsoft Excel as any observation that goes above or below the top or bottom whiskers, respectively).
  Note: Microsoft Excel-generated outliers are different than Walsh outliers.

Tables and Charts – Summary Data Tables

Summary of All Data		Summary of Detects *	
Statistic	Value	Mean	12,774.24
Total (N)	852	Std Dev	12,248.80
Detects	408	Max	70,000.00
Non-detects	442	Min	33.00
Missing	2		

Table A4. "Summary of All Data and All Detects output tables in the NC Soil Metals Dashboard.

Summary of All Data [Table]

"Summary of All Data" provides information on the observations for a selected metal (*Figure A4; left*).

- Total (N) represents the total number of observations including missing values and non-detects.
- Detects is the number of observations that have a detected value (does not include missing values or non-detects).
- Non-detects is the number of observations that were non-detected values, this was represented by "#" in the original data but is represented by "-999" in the "Raw Data."
- Missing is the total number of blank observations (did not contain a value) in the original data (represented in "Raw Data" as "-888").
  Note: Summary of All Data will always include non-detects and Walsh outliers.

Summary of Detects [Table]

"Summary of Detects" provides summary statistics on the detected values of a selected metal (*Figure A4; right*). For the selected metal, each of the following is calculated: arithmetic mean, standard deviation,

minimum and maximum values. These numbers will change if Walsh outliers are included or excluded (except for the minimum value).
Note: Non-detects are always excluded from "Summary of Detects."

Tables and Charts – Statistical Tables

Nonparametrics [All Data] *		Kaplan-Meier All Data *		Walsh Outliers [>=86,000.00]	
Statistic	Value	Mean	6,125.33	Value	Count
Median	Non-detect	Std Dev	10,586.30	107,900.00	1
90th %	21,580.00			91,300.00	1
				86,000.00	1

Figure A5. Statistical output tables in the NC Soil Metals Dashboard.

Nonparametrics [Table]
"Nonparametrics" calculates the median and the specified percentile for the chosen metal (*Figure A5; left*). There are options to include or exclude non-detects and/or Walsh outliers at the top of the dashboard. There is also an option, highlighted in blue, to set the percentile to any number between 1-99. Details on this option are described in the "Menu Bar" section. If you choose to include non-detects, the median and/or chosen percentile may be calculated as "non-detect."

Kaplan-Meier [Table]
Within the NC Soil Metals Dashboard "Kaplan-Meier All Data" calculates the mean and standard deviation for the selected metal using the Kaplan-Meier method (*Figure A5; center*). Kaplan-Meier is a nonparametric method that is used to handle censored data that has multiple detection limits. Since Kaplan-Meier is used for censored data, it will always include non-detects. However, it is impacted by the inclusion or exclusion of Walsh outliers.

Walsh Outliers [Table]
"Walsh Outliers" are identified using Walsh's Outlier test in the NC Soil Metals Dashboard (*Figure A5; right*). Walsh's test determines if a group of observations (above a certain value) are outliers. The determination of the outliers was calculated, using external statistical software, to test for the most extreme group of outliers. This group was determined by inspection of the data. Some Walsh Outlier values appeared more than once in the dataset. Therefore, the "Count" column is included to show the number of times each specific Walsh Outlier value appears in the dataset.

County Median Heat Map
The county heat map shows the median concentration of the selected metal by North Carolina County (*Figure A6*). If a county is grey, that means there was no data collected for the selected metal in that county. The heat map can include or exclude non-detects. Therefore, if non-detects are included, the median for a particular county may be a non-detect. If a county's median is a non-detect, it will show up as the lowest detected value for that metal. Hence, not every county showing the lowest value may be a non-detect (since the actual median for a county could be the lowest reported value). The heat map will be impacted by the inclusion or exclusion of Walsh outliers. If a user wishes to exclude the heat map, the box in the top left corner that says "Show Map" can be unchecked to exclude the map. The NCDEQ logo is displayed by default in the heat map area, unless the "Show Map" box is checked.

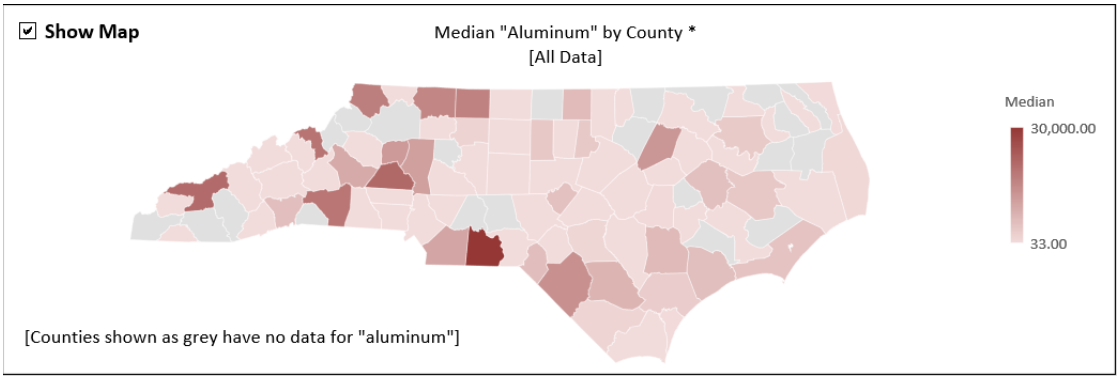


Figure A6. County heat map showing median concentration of Aluminum across North Carolina counties in the NC Soil Metals Dashboard.

Supplementary Information II: Methods

1. Procedure used for qualified data

When using qualified laboratory data, the guidelines provided by the EPA were followed (1996); (2020). In summary,

"U" flagged values – The analyte was analyzed for but was not detected at or above the Contract Required Detection Limit (CRDL) or Method Detection Limit (MDL). In the data report source documents used, these values were marked as < X, where X is the CRDL or MDL.

- Due to the lack of CRDL/MDLs in early laboratory reports and the variability of CRDL/MDLs over time, the U-flagged value was replaced with the minimum detected value listed in the dataset for that analyte.

"UJ" flagged values: The analyte was analyzed for but was not detected at or above the CRDL or MDL. The reported CRDL or MDL is an estimate.

- The reported value is treated as a "U" flagged value and replaced with the minimum detected value listed in the dataset for that analyte.

"B" flagged samples: The analyte was detected, but the associated blank for this sample reported a detection of the same analyte. To avoid any potential environmental interference, B-flagged samples were processed as follows:

- If the analyte concentration in the sample was greater than 10x the concentration of the same analyte in the associated blank, then the value was reported as a detection.
- If the analyte concentration in the sample failed the 10x test, the data point was discarded.

"J" flagged, values: The analyte was detected above the CRDL/MDL but below the laboratory contract required quantitation limit (CRQL) or method reporting limit (MRL), therefore the analyte is present, but the concentration is an estimate. Because the concentration is an estimate, bias may be present, and data were further analyzed for quality based on the following:

- If no qualifier report or explicit description of the J flag was given, the value was adjusted for an unknown bias by multiplying the value by a correction factor (see 1996 EPA guidelines).
- If a qualifier report was given, the information was used to determine if the data was usable or required adjustment for bias. This was done by either multiplying the value by a correction factor or leaving the value as is. Matrix spike recovery and serial dilution issues were common causes for high, low, and unknown bias. Further details on how data was evaluated for quality and adjusted are included in Table 1 below.

2. Treatment of non-detects

Many analytical results in the dataset were identified as non-detects, which practitioners typically address using a variety of techniques (Harter, 2006). Some of those techniques rely on the use of reported detection limits to estimate a substituted value. However, the laboratory reports for this study did not all include detection limits, particularly for reports of earlier CERCLA site investigations. The data were produced by different laboratories for 30 years, and detection limits likely changed through that period. USEPA (1989) recommended the non-omission of non-detects. To address this issue the minimum detected value was substituted for non-detects since exclusion of non-detects can cause an overestimation of the mean, standard deviation, and median.

Ogwu, Nilsen, et al. (*In Prep*)

In preparing the data for the NC Soil Metals Dashboard, when calculating non-parametric summary statistics (the median and other percentiles, and numbers for the heat map), values listed as non-detect are replaced with the minimum detected value listed in the dataset for that metal. In cases where any of these types of summary statistics are calculated to be equal to the minimum detected value for the metal, the dashboard will present the statistic as "non-detect". The dashboard also includes the option to include or exclude non-detects for these calculations.

When creating the county heat map, a county's median can be a non-detect (if the dashboard user includes non-detects). In this case, that county's median value will appear as the minimum detected value for the specified metal so many counties may share the same heat level (median). Thus, the replacement of non-detects with the minimum detected value will not affect heat map computations but will impact the heat map graphically.

Other (parametric) computations represented in the dashboard were performed using only the detected values for the metals. These computations include the mean, standard deviation, minimum, maximum, and all the calculations for the boxplot.

The simple expedients described above, while not commonly practiced, reflect shortcomings of the available data. These shortcomings may be less concerning given the intended purpose of this study, which is to provide general information to site investigators and risk assessors who will be able to rely on non-parametric descriptive statistics of the upper (right) end of the datasets, such as simple percentiles.