RMP
REGIONAL MONITORING
PROGRAM FOR WATER QUALITY
IN SAN FRANCISCO BAY

sfei.org/rmp

Re-evaluation of the Floating Percentile Method for Deriving Dredged Sediment Screening Guidelines

Prepared by:

Don Yee, Adam Wong

San Francisco Estuary Institute

Suggested Citation
Yee, D.; Wong, A.; 2023. Re-evaluation of the Floating Percentile Method for Deriving Dredged Sediment Screening Guidelines. SFEI Contribution #1143. San Francisco Estuary Institute, Richmond, CA.

CONTRIBUTION NO. 1143 / June 2023

Table of Contents

Introduction

This document summarizes a study conducted for the Regional Monitoring Program for Water Quality in San Francisco Bay (RMP) to re-evaluate the use of the Floating Percentile Method (FPM) to derive sediment screening guidelines for dredged material reuse in the San Francisco Bay Region. The Long Term Management Strategy (LTMS) has a goal to use at least 40% of the sediment dredged from San Francisco Bay for beneficial reuse (USACE, 1998). The suitability of dredged sediment for beneficial reuse is in part determined by concentrations of toxic pollutants.The San Francisco Regional Water Quality Control Board (SFB-RWQCB) issued draft screening criteria in 2000 to categorize the suitability of sediment for reuse as either "surface" sediment, that may be placed near the surface for re-use in wetlands, or "foundation" sediment, that is buried under sediment that meets surface criteria. Contaminant concentration guidelines for surface sediment are lower than foundation sediment, based on the assumption that biota are more likely to be exposed to surface sediment than deeper foundation sediment.

A RMP workshop on beneficial reuse guidelines in September 2019 resulted in the recommendation that the Floating Percentile Method (FPM) be used to re-evaluate and potentially update the sediment screening guidelines (Foley et al., 2020). One concern was that the draft sediment guidelines for beneficial reuse (SFBRWQCB, 2000) may be restricting the amount of sediment that is determined as suitable for beneficial reuse purposes, which is urgently needed to restore marshes around the Bay to make the area more resilient to sea level rise.

The Floating Percentile Method (FPM) was developed by Science Applications International Corporation (SAIC) and Avocet Consulting (SAIC/Avocet, 2002) and used paired sediment toxicity and chemical concentration data to develop sediment quality guidelines for the Washington State Department of Ecology. The FPM was also used by the Oregon Department of Environmental Quality Regional Sediment Evaluation Team for its Freshwater Sediment Workgroup (Michelsen & Anderson, 2011). A previous study using Bay data (Germano & Associates, 2004) examined the FPM and Receiver Operating Characteristic approaches as alternatives to the updated SFB-RWQCB sediment screening guidelines (SFB-RWQCB, 2000).

The basic concept behind the FPM is to use locally derived data for paired toxicity and chemistry tests to determine suitable criteria, rather than relying on nationally derived criteria. The methodology of the FPM uses the local test data to set criteria for surface sediment with low percentages of false negatives (e.g., < 5% of actual toxic sediments below a presumed non-toxic concentration). In doing so, any sediment designated for "surface" reuse would likely have a low (< 5%) probability of toxicity due to its chemical composition. In contrast, for foundation sediment, the desire is to allow a higher percentage (e.g., 75%) of toxic results to be

included below the given concentration, to maximize the amount of sediment that can be beneficially reused, while still excluding the most highly contaminated sediments.

The percentiles selected for determining the criteria are largely a management choice, informed but not dictated by fixed percentiles of the available data. The criteria chosen should ideally provide a balance between the probability of toxic effects due to the chemistry of the sediment and the percentage of dredged sediment that would be categorized as unsuitable for reuse. For the purposes of this study, concentrations for surface and foundation sediment criteria were calculated for the 5th and 75th percentiles as was used in the prior application of the FPM, but different values can be applied if desired (for example, if the resulting criteria would categorize very little of the dredged sediment as suitable for reuse).

Between 2002 and 2012, the RMP conducted annual benthic sediment sampling at randomized sites distributed throughout the Bay, with concurrent chemical analysis and sediment toxicity. Since 2012, the RMP has not regularly conducted toxicity testing of surface sediment samples, so the 2002 to 2012 data were used to refine and update the prior FPM analysis of SF Bay data. The data from RMP studies primarily include surface sediment (0-5 cm depth) composited and subsampled for chemical analysis and toxicity testing, so there is generally a 1:1 correspondence between samples with both types of analyses.

Another source of recent and potentially ongoing chemical and toxicity paired data is available from dredging projects that are required to assess contaminant risks and evaluate appropriate sediment placement locations. The Dredged Material Management Office (DMMO) requires data to be uploaded to a queryable database. The DMMO database was made available to the public in 2019, and nearly all of the data present in its initial release were collected from projects occurring between 1998 and 2016, with new projects continuing to be added. Thus it would be expected that a majority of data in the DMMO database were obtained after the 2004 application of FPM to SF Bay, and similarly, much of the RMP monitoring data was also collected subsequent to that report.

Approach

We obtained a copy of the FPM software (FPMData.xls - December 28, 2008 Revision, FPMDataGroups.xls - December 28, 2008 Revision, FPMAnova.xls - July 29, 2008 Revision, FPMCalc.xls - December 6, 2008 Revision) from the Washington State Department of Ecology. The package consists of four Excel workbooks, with embedded Visual Basic macros to perform calculations. The first Excel workbook takes raw input data from paired toxicity and chemistry tests and processes (sorts and filters) the data to yield a final formatted data table. The next workbook takes the formatted input table, and calculates chemical concentrations for various percentiles of the input data, with the starting point and intervals of the percentile calculations set by the user. The software package as currently implemented is capable of handling up to 10 chemical analytes simultaneously, with calculations for up to 10 percentile values (e.g., percentiles between 5% and 50% can be calculated at 5% intervals).

Input data from toxicity and chemistry tests from sites in the San Francisco Bay used for the FPM calculation were downloaded from the California Environmental Data Exchange Network. A majority of the available data were from the RMP, with all available data on surface sediment since the beginning of the program included. Toxicity and chemistry for ambient monitoring samples generally had a 1:1 correspondence, as in the RMP and other monitoring programs, samples sent for toxicity testing are generally subsamples of the same material sent for chemical analysis.

Data on dredged sediment tests available from the DMMO database were also added, with chemistry and toxicity data originating from the same project, site, and sampling date presumed to be related. However, because the objectives of dredged material testing are slightly different from those for generating monitoring data, a 1:1 correspondence between samples analyzed for chemistry and toxicity is less certain; in some cases discrete or core samples may be analyzed for chemistry, while toxicity tests from the same site might be performed on composited samples. When it was unclear whether chemistry and toxicity tests were performed on subsamples of the same material, we excluded the samples since the FPM requires paired data. We also excluded Z-layer (post dredging) and dilution series test results in the DMMO database in this study because they would not possess characteristics typical of ambient Bay sediments in the mid- to long-term.

Results from either source reported as non-detects—primarily organic contaminants—were substituted using a very small value above zero. This is contrary to the documentation for the tool, which recommends excluding the data. However, the FPM is a non-parametric calculation, so the exact values of substituted results are immaterial unless the substitution occurs at or above a calculated percentile, and the exclusion of all non-detects (typically with expected lower concentrations of an analyte in question) would effectively bias the distribution of results to appear to have a higher central tendency.

The first task in application of the FPM was to define a "toxic" response. In this effort, we continued to define a "toxic" result using the "reference envelope" approach previously described (Hunt et al., 1998), and employed in the prior application of the FPM (Germano & Assoc., 2004). The steps in the categorization of the response were as follows:

1. "Normalizing" reference site responses versus test controls (e.g., because controls can have < 100% survival)
2. Calculating percentiles and their confidence intervals (CI) for responses of "reference" site samples
3. Selecting appropriate response thresholds balancing the percentage of reference data that would be deemed toxic against the uncertainty in percentiles.

In the prior application of FPM, for the amphipod *Ampelisca abdita*, the lower 95% CI of the 10th percentile (lowest "reference" survival) was used as the threshold for a toxic result. For the amphipod *Eohaustorius estuarius*, the 20th percentile was used instead, due to the lower 95% CI of the 10th percentile yielding a very low reference survival rate. If the 10th percentile were used, it would result in an overly liberal assignment of very few responses as showing toxic effects worse than reference conditions. For this effort, we used the same percentiles of the reference site data for these respective species (as in the 2004 study) for designation of toxic responses, but did include additional new reference site data (in tests collected since 2004) in deriving those response percentiles. The toxic or non-toxic designation for each of the test amphipod species was then combined, where a toxic result for either species yielded a categorization of the sample as toxic for a single pooled amphipod toxicity indicator. The data used in this study also included results from mollusk toxicity tests, for which the reference 10th percentile lower 95%CI was used to denote a toxic response. Because the modes of toxicity may differ for mollusks, we elected to analyze the mollusk data in a separate FPM calculation, rather than combine all species into a single pooled generic toxicity indicator.

Once all toxicity test results had been assigned as either toxic or non-toxic, the FPM tool derived surface criteria by finding the maximum concentration of a chemical constituent that did not increase the occurrence of toxic responses above a base rate of "false negatives" (i.e., tests with toxic responses despite a chemical concentration being below the given value), while ideally reducing the occurrence of "false positives" (i.e., tests without toxic responses, despite concentrations above a given value). Using the approach, fewer non-toxic results would erroneously be categorized as toxic as the selected concentration was increased, up to the maximum concentration of the analyte found in any non-toxic sample. In practice, this was done in the FPM by:

1. Selecting a target concentration of the chemical that provided a given low (e.g., 5%) false negative rate
2. Adjusting individual chemical values upward until false positive rates were optimized (decreased to their lowest possible level) while retaining the same level of false negatives.

If the false negative rate was examined in a concentration range where the toxic response was fairly insensitive to a given analyte (i.e., no additional toxic responses occurred despite increasing concentrations), the resulting chemical limit could be notably higher than the lowest concentration with the same false negative rate. In contrast, if the incidence of toxicity rose continually with increasing concentration, the lower and upper concentrations with a given false negative rate would be effectively nearly identical. Values for foundation sediment were derived using the same method as used for surface values, but using a higher percentile (e.g., a concentration including 75% of toxic results as a target value).

Results

Table 1 lists results from the current application of the FPM tool, with columns for the pooled amphipod toxicity tests and mollusks reported separately. Additional columns provided for comparison list the surface and foundation results from the prior application of the FPM (Germano & Assoc. et al., 2004), as well as the 2000 SFB-RWQCB criteria.

Table 1. Comparison of sediment surface guidelines. Surface (5th) and foundation (75th percentile) FPM results for amphipod and mollusk tests in this study are presented in separate columns, along with Germano et al. (2004) FPM values and existing SFB-RWQCB (2000) draft surface and foundation guidelines.

Analyte	Amphipod FPM surface	Amphipod FPM foundation	Mollusk FPM surface	Mollusk FPM foundation	2004 FPM Surface	2004 FPM foundation	2000 SF Bay surface criteria	2000 SF Bay foundation criteria
PAH µg/kg	193	2688	38	2175	6.3 µmol/kg	32 µmol/kg	3390	44792
PCB µg/kg	1.61	16.9	0.369	15.4	600	600	22.7	180
DDT µg/kg	0.59	4.97	0.052	5.77	250	250	7.0	46.1
Chlordanes µg/kg	0.0672	0.306	0.001	0.884	69.2	69.2	2.3	4.8
BHCs µg/kg	0.0050	0.0244	0.00585	0.0110	2.0	2.0	—	—
Arsenic mg/kg	4.78	11.0	2.80	11.0	40	40	15.3	70.0
Cadmium mg/kg	0.065	0.349	0.106	0.497	0.25	0.62	0.33	9.6
Chromium mg/kg	58.1	118	38.2	97.3	119	320	112	370
Copper mg/kg	18.9	48.7	17.5	55.6	50	150	68.1	270
Lead mg/kg	9.90	26.6	5.87	28.4	200	200	43.2	218
Mercury mg/kg	0.0784	0.313	0.040	0.327	1.18	1.18	0.43	0.70
Nickel mg/kg	48.5	90.3	37.9	94	230	230	112	120
Silver mg/kg	0.015	0.339	0.0270	0.353	0.28	2.00	0.58	3.7
Zinc mg/kg	64.8	124	52.6	141	1200	1200	158	410

In the prior FPM effort, values were derived using the software in a multi-analyte analysis mode. For each set of linked results (the pooled results combining toxicity tests from either amphipod species, and chemical concentrations for as many analytes as reported in a given sample), up to 10 analytes (the limit in the software) were considered together at once. Thus for any given observed toxic result, the presence of one or more chemical analytes above their respective estimated surface or foundation percentile would be considered to represent a correctly identified "true positive". In principle, this multi-analyte application of FPM could reduce the apparent incidence of "false" positives for a given analyte by accounting for "true" positives caused by other analytes. Consider the simple case of a two chemical mixture, with only one of the chemicals reported or analyzed at a time. In applying the FPM one analyte at a time, the examined analyte may be at a relatively low concentration in some samples, yet the apparent toxicity was actually caused by the (unexamined) second chemical at a high concentration. Without knowledge of the other chemical concentration, the examined single analyte is considered to have possibly caused the toxicity, and as a result, the calculated concentration for the reported analyte at which a given percentile of false negatives occurs appears lower than that which would be calculated if samples with toxic concentrations of the second analyte were already factored in.

However, in practice, the use of the FPM software in a multi-analyte mode as currently implemented proved to be unstable. In a test of the software using a subset of data with completely reported values for five metal analytes (with no unreported values, thus reducing the likelihood of cases where undetected or unreported analytes within the group could have caused the observed toxicity), the results obtained were dependent on the sequence of analyte names used in the header of the input data table. When analyte names in the table header were swapped without moving any of the data (e.g., a column of Arsenic data relabeled as Zinc and vice versa), the concentrations reported for the percentiles of each column of data changed, despite the analyte label being the only information changed. The cause of this artifact, which appears to be a software bug, was not easily identified in a cursory inspection of the code.

Another characteristic of the software routines was observed in its handling of covarying data; results appeared to be sensitive to the order of evaluation (internal to the algorithm). This was illustrated in a test dataset constructed with chemistry results for a single analyte, adding replicated data with dummy variable names (e.g., DDT, DDT1, DDT2). The concentrations corresponding to the calculated percentiles for the primary analyte were always the same, and the surface and foundation percentiles for the remaining (dummy) analytes were all reported as their maximum values for non-toxic samples (the same maximum concentration for all of the variables, both original and dummy, since they were exact duplicates). This behavior is likely as expected or intended, because once the false negative percentiles (proportion of toxic hits below a concentration value) were calculated for a first analyte, the additional (identical) dummy analytes had no influence on the assignment of additional false negatives or false positives. However, even in this simple test case, the software application was unstable; depending on the number of additional dummy variables, and the order of the dummy analyte

names, the analyte handled as the primary driver for assigning percentiles was not predictable (i.e., not always the first column, nor the alphabetic first analyte).

Although these bugs and instabilities in the FPM program could potentially be fixed through careful examination of the source code and redesign of the calculation logic and code, such an effort was beyond the originally planned scope of the project. We thus elected to run the FPM routines for each analyte one at a time, which appeared to yield accurate percentile results based on a visual cross check of the distributions of individual analytes plotted graphically.

For the chemicals with low proportions of non-detects (~10% or less, which was the case for nearly all the trace elements, PAHs, and PCBs), the 5% false negative rate occurred in a range with some detectable concentrations. For some pollutants the incidence of additional toxic hits in lower percentiles initially increased only very slowly with increasing concentrations (e.g., in the plot of false negatives in Figure 1 for PAHs), while for others, even small additional increases in concentration immediately resulted in more toxic results (e.g., Figure 2 for lead). Plots for all the analytes, for both amphipod and mollusk toxicity tests, are provided in the Appendix.

The definitions for the reported metrics in Figures 1 and 2 (and similar plots in the Appendix) are provided below. As in the prior study, sample results could be assigned to one of four groups for being above or below an evaluated concentration (predicting non-toxic below that concentration, and toxic above):

A = toxic samples predicted as toxic (true positives)
B = non-toxic samples predicted as toxic (false positives)
C = toxic samples predicted as non-toxic (false negatives)
D = non-toxic samples predicted as non-toxic (true negatives)

If the concentration evaluated as a candidate toxicity threshold is below the lowest concentration reported in any samples, all samples are at higher concentrations, and thus predicted to be toxic. All results will therefore be either A (true positives) or B (false positives). If the evaluated concentration is above the maximum concentration, all samples are predicted to be non-toxic, and thus either C (false negatives) or D (true negatives).

For each of the analytes, the plotted metrics included the following percentages:

Percentile: all sample results below a given concentration (C+D)/(A+B+C+D)
False positive: non-toxic samples above a given concentration B/(B+D)
False negative: toxic samples below a given concentration C/(A+C)
Predicted hit reliability: samples above a concentration that were actually toxic A/(A+B)
Predicted no hit reliability: percentage of samples below a concentration actually non-toxic D/(C+D)
Overall reliability: samples correctly predicted (as toxic or non-toxic) out of the total number of samples) (A+D)/(A+B+C+D)

It should be noted that the overall reliability had an initial value of the percentage of toxic results, and a final value of the percentage of non-toxic results. With a guideline set at the minimum concentration, all samples would be predicted to be toxic (C and D = 0), while if set at the maximum concentration, all would be predicted non-toxic (A and B = 0).

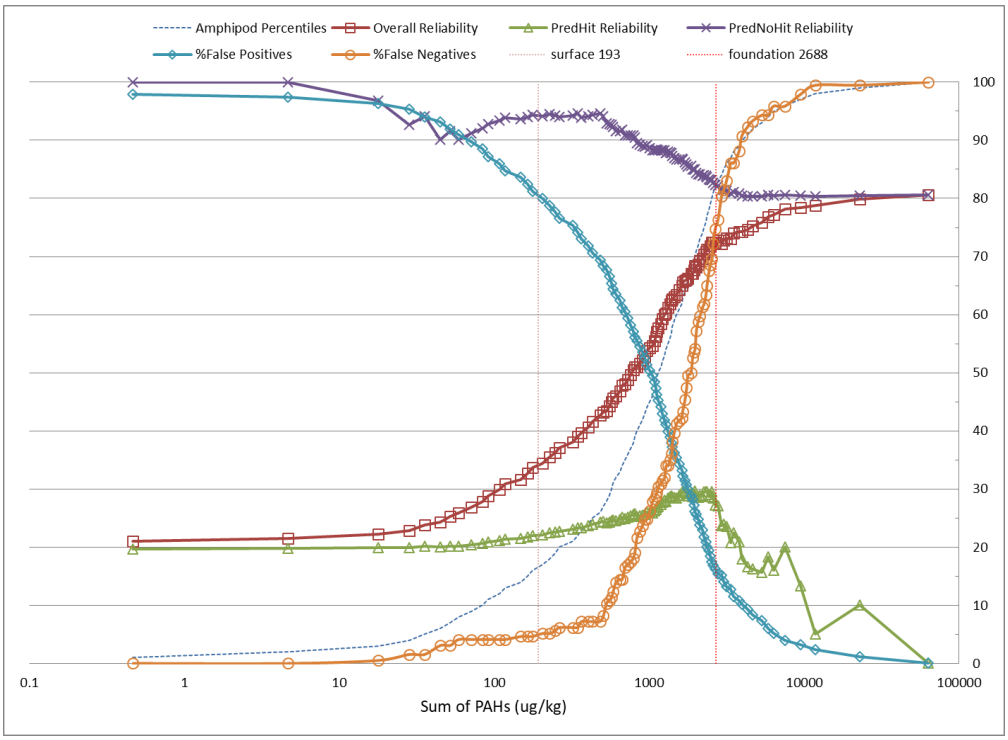


Figure 1. Plot of percentile metrics for PAHs in amphipod toxicity tests.
The y-axis is in units of percent (%) for each metric.

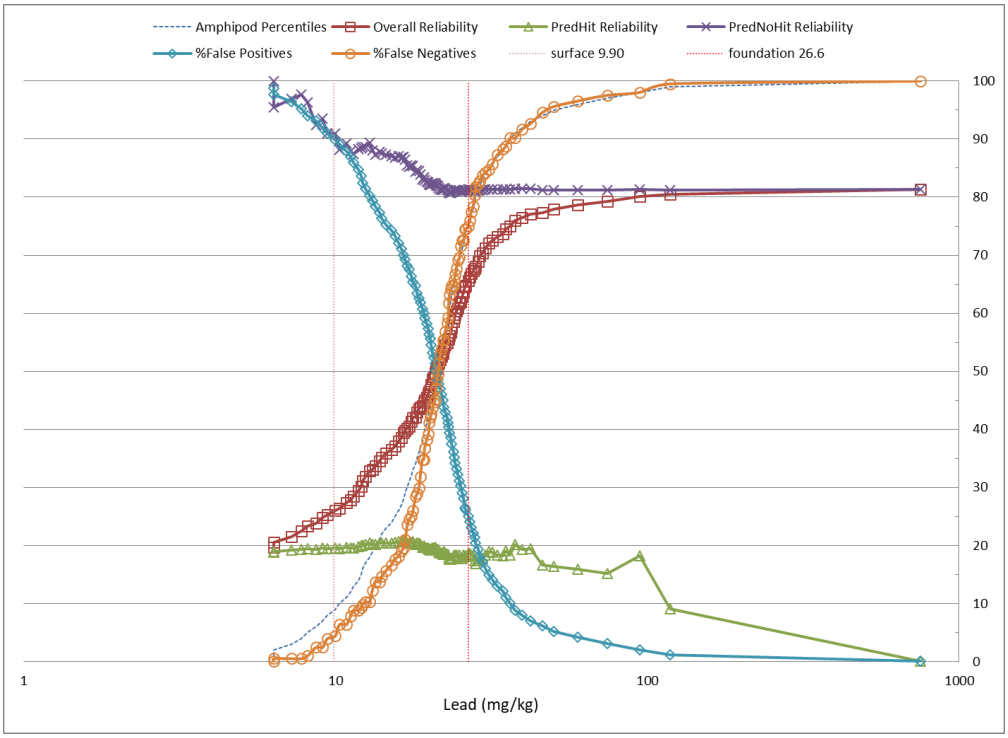


Figure 2. Plot of percentile metrics for lead in amphipod toxicity tests.
The y-axis is in units of percent (%) for each metric.

At higher percentiles of the distribution, such as the concentrations for each analyte where 75% of the toxic results were found (marked as a potential "foundation" sediment values, e.g., in Figures 1 & 2), even small increases in chemical concentrations generally resulted in continually and rapidly increasing occurrences of toxicity. Thus for determining foundation values, step 5 described previously (seeking a higher concentration with the same false negative rate) was usually effectively moot. Once the lowest concentration for a given analyte containing the 75th percentile of toxic results was determined, only very small increases in the concentration were possible before the false negative percentile increased (i.e., additional toxic results accrue below the evaluated concentration).

For many of the analytes, concentration distributions in samples with toxic and non-toxic responses differed very little. For example, for PAHs (Figure 1), the median concentration of non-toxic samples (the 50th percentile of false positives) was only slightly lower than the median of toxic samples (50th percentile of false negatives), and around 40% of toxic samples had concentrations below the median non-toxic sample. Similarly, the medians of lead in toxic and non-toxic samples were about the same (Figure 2). This was also the case for most other analytes (Figures in the Appendix), with 40% to 50% of the toxic samples having concentrations lower than the median of non-toxic samples.

Many of the pollutant concentrations were highly significantly correlated with each other, so in many samples, identification of individual analytes primarily driving toxicity may be difficult.

A scatterplot matrix of all the analytes compared pairwise is shown in Figure 3. Scatterplots for analyte pairs are shown in the lower panel, and correlation coefficients, with significant correlations ($p < 0.05$) marked with an asterisk (*) shown in the upper panel of the matrix.

The majority of these pollutants were highly correlated to each other within the Bay sediment samples. Nearly all the trace elements were significantly correlated to each other ($p < 0.001$ in most cases) and to one or more of the organic pollutant classes. Similarly, aside from the BHCs (which may be influenced by having over 70% NDs), the organic pollutants were generally significantly correlated with each other as well as with many of the trace element pollutants.

As a result of the tendency for many of these analytes to covary, there are likely relatively few samples where any one pollutant would be elevated without concurrent higher concentrations of one or more of the other analytes. This could confound the interpretation of results from the FPM algorithm when conducted in multi-analyte mode, as was illustrated in the test case with replicated data, where the primary cause of toxicity cannot truly be distinguished, and a lower surface value appears to have been initially assigned to one of the analytes, with surface values (as well as foundation values) near the maximum concentration assigned to the rest of the analytes.

To examine the degree to which the correlation among pollutants might affect the determination of the surface values using FPM analytes one at a time, we estimated the average expected concentrations of all other analytes that would co-occur with copper concentrations set at its surface value concentration (5th percentile of false negatives). Copper was used because it was among the most frequently measured chemicals, and it showed a very significant ($p < 0.001$) positive correlation with all other analytes aside from BHCs, for which the correlation was not significant ($p \cong 0.2$) and had a negative slope.

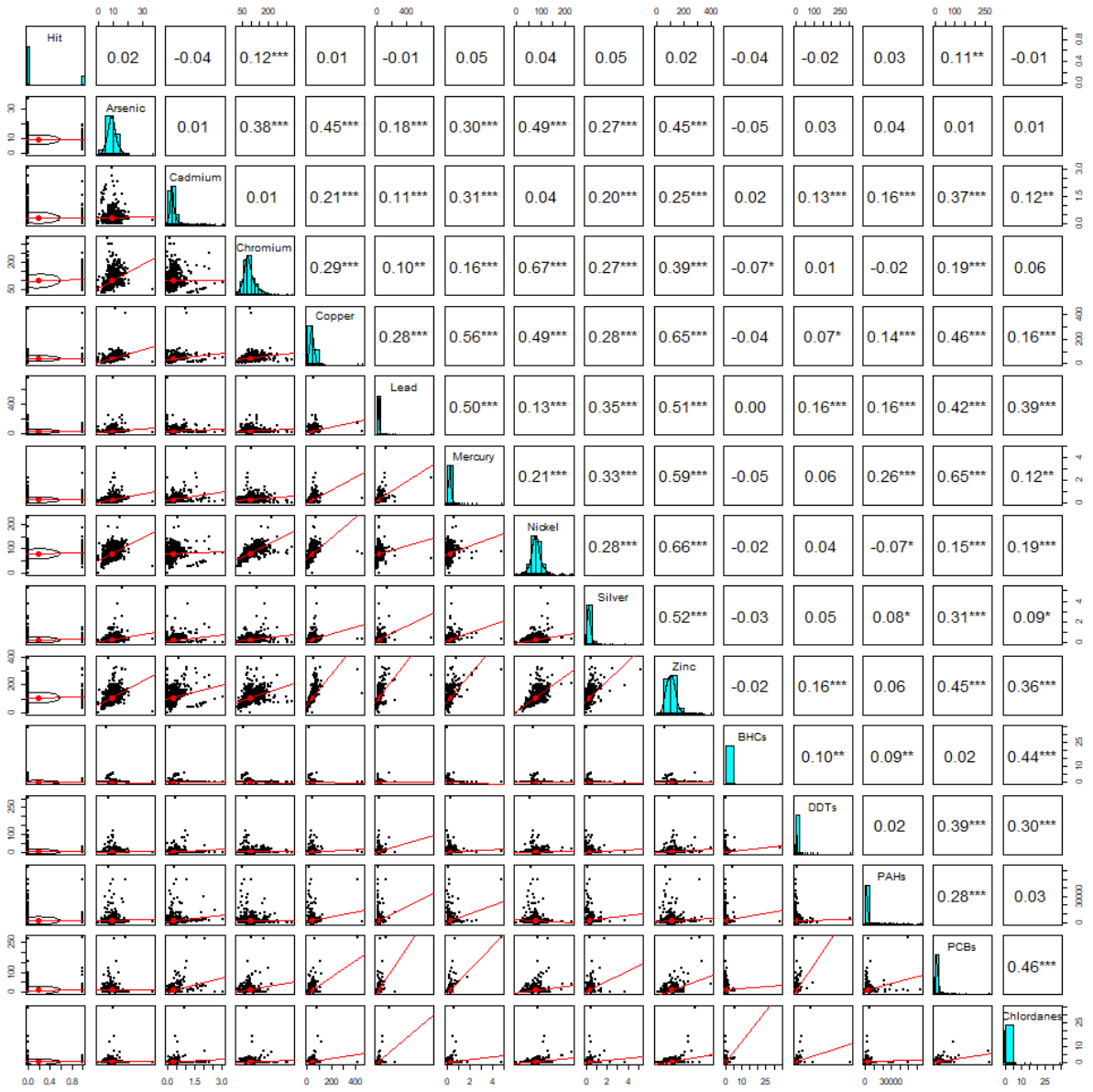


Figure 3. Matrix of correlations among analyzed chemicals. Lower triangle contains scatter plots between chemicals in the data used for this study (amphipods, all toxicity test outcomes). The red line represents a linear regression between the parameters. Upper triangle shows correlation coefficients, with significance indicated by number of asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

The copper surface value was substituted into the linear regression equation with each of the other analytes individually. The result was an expected average concentration for the analyte (that would co-occur in samples containing copper around its FPM surface value), which could then be compared to the prospective surface value of the analyte directly derived by the FPM. If the FPM surface value for an analyte was lower, it suggests the analyte is more toxic than copper; the surface value for that analyte (its 5th percentile lowest concentration in toxic

samples) would on average likely be exceeded before copper reached its respective surface (5th percentile false negative) value. Conversely, if the analyte's FPM surface value was higher than the estimated value from its correlation with copper, it is likely on average less toxic.

Table 2 lists the regression estimated average concentrations of all the other analytes likely co-occurring with samples containing copper at its FPM derived surface and foundation values. For the most part, these estimated average surface concentrations (co-occurring with copper at its FPM surface value) for other analytes are higher than their respective surface and foundation results derived directly by the FPM. Thus most other analytes would contribute to earlier (lower concentration) observances of toxicity more frequently than for copper, although some of the differences are small (less than a factor of two). For the foundation concentrations, the FPM calculated values for individual analytes were lower than those co-occuring with the copper foundation values. The majority of the differences were less than 15%, however, and the foundation values (75th percentile of toxic results) were often similar to the 75th percentile of all samples, as relatively few non-toxic results (less than 20% for a majority of analytes) occur above the FPM derived foundation values.

Table 2. Estimates of other contaminants co-occurring with copper - estimates made at copper surface (5% of toxic samples) and foundation (75% toxic) values, compared to the respective surface and foundation values for the other analytes. Total BHC values were not estimated because its correlation to copper was not significant.

	Intercept	Slope	Slope p	FPM surface	Est. @Cu surface	Estimate %ile	FPM found.	Est. @Cu found.	Estimate %ile
Arsenic	6.6777	0.0596	6.2E-56	4.78	7.80	30	11	9.58	54
Cadmium	0.196495	0.00266	5.1E-12	0.065	0.247	59	0.349	0.326	74
Chromium	76.30657	0.434555	1.3E-18	58.1	84.5	25	118	97.5	51
Lead	10.25918	0.350929	4.9E-21	9.9	16.9	22	26.6	27.3	78
Mercury	0.026369	0.005763	9.9E-92	0.0784	0.135	14	0.313	0.307	74
Nickel	61.58735	0.42753	1.2E-66	48.5	69.7	23	90.3	82.4	51
Silver	0.123737	0.003411	2.2E-20	0.015	0.188	29	0.339	0.290	61
Zinc	62.44841	1.087485	3.0E-130	64.8	83.0	18	124	115.4	62
DDTs	2.628663	0.039206	0.019	0.59	3.37	58	4.97	4.54	71
PAHs	1050.175	26.36558	8.1E-06	193	1548.5	42	2688	2334.2	64
PCBs	-6.24915	0.420365	5.2E-36	1.61	1.70	5	16.9	14.22	67
Chlordanes	-0.12157	0.011898	4.1E-05	0.0672	0.103	43	0.306	0.458	81

Discussion

One of the primary motivations for embarking on the exercise of repeating the FPM calculation using an updated and larger dataset was to verify and refine the previous findings (Germano et al., 2004), with a goal of improving the confidence in the suitability of the resultant FPM values for use in dredged sediment placement decisions. In addition to a greater number of reported results, the expanded dataset included a smaller proportion of non-detect results, so it was expected that the reanalysis would provide a more representative and quantitative sample of the population of interest, the sediment from subtidal areas of San Francisco Bay.

A likely significantly impactful departure from the prior effort was the handling of non-detects. The FPM tool documentation recommends discarding non-detect data, but doing so effectively shrinks the population of low concentration samples for any given analyte with some sample concentrations near their detection limit. Aside from silver, each of the trace elements reported had less than 10% non-detects, so the substitutions for those analytes likely had little or no impact. In contrast, for organic analytes in samples with amphipod tests, results for > 10% of sums for PCBs, > 20% for DDTs, > 50% for chlordanes, and > 70% for BHCs were non-detects for all contributing compounds. The distributions of the truncated datasets for these analytes excluding non-detects would appear dramatically different. For example, if non-detects were substituted at the MDL value or another nominal value < MDL, medians for chlordanes and BHCs for the un-truncated sets would have been below the lowest reported results for the truncated datasets. As a result, any FPM calculated percentiles from the truncated datasets would have been shifted to higher concentrations, since the counts of all false negatives (toxicity, despite unquantified or undetected concentrations) below the lowest quantified value would be ignored.

Although the choice for inclusion and substitution for non-detect results could have severely impacted calculated FPM surface values for some analytes (if non-detects were well over 5% of results, mostly the organic analyte groups), even those analytes with few non-detects had surface values well below those derived from the prior FPM application. This may in part be the result of evaluating the analytes one at a time; some “false negative” toxic hits for each analyte could have been “true positive” toxic hits for another analyte in the sample, as illustrated in the hypothetical two analyte example described earlier. For analytes in the 2004 work that did not have identical surface and foundation values (PAHs, cadmium, chromium, copper, and silver), the new surface values were generally less than 10-fold lower than past surface values. In contrast, the remaining analytes (that previously had identical surface and foundation FPM values) often had new surface values more than 10-fold lower than prior surface values, despite also being set using a 5% false negative rate.

Thus contrary to initial expectations based on results from the prior application of FPM, these new FPM derived surface and foundation values are all well below the respective SFB-RWQCB 2000 surface and foundation criteria. In many cases, the newly derived foundation values (75th percentile concentrations of samples with toxic hits) for many analytes were lower than the

previous FPM calculated surface (5th percentile) values, and similar to or lower than even the existing SFB-RWQCB 2000 surface guidelines. This latter finding is perhaps not surprising, since a large portion of the RMP data used in this analysis was also used in the characterization of ambient conditions for the derivation of the 2000 sediment guidelines, and the distributions of most legacy contaminants have not shown significant evidence of change in Bay sediment (SFEI, 2015).

Newer pesticides such as pyrethroids and neonicotinoids were seldom included in the reported data until recently in the RMP, and almost never reported for dredging data, so were not useful for assessing this historical dataset. Sample toxicity due to the constantly evolving landscape of current use pesticides or other more recent emerging pollutants could be incorrectly attributed with older pollutants that might co-occur or correlate with these new toxicants. Even if hypothetically these correlations to legacy contaminants were significant in a large dataset, the variance around the central tendency of the correlation could be very large (e.g., in the Figure 3 scatterplots among various legacy contaminants, often < 50% of the variation was related to the concentration of another), making the prediction of one analyte based on the presence or quantity of another analyte highly uncertain.

The approach of the FPM itself is fraught with challenges for application to real world data. The case for which it is best constructed is a set of data where 1) the concentrations of nearly all the constituents are minimally contributing to toxicity in many of the samples, but 2) a select analyte varies largely independently of the others and occurs over a wide enough concentration range that increases the propensity for toxic outcomes for numerous samples. The varying analyte becomes the final factor pushing samples into observably higher rates of toxic outcomes. The scenario described (one analyte varying, with the rest constant or very low) fits that of a gradient from a point source of a single contaminant, with the rest of the contaminants at low or constant ambient concentrations. The likelihood of both conditions occurring together in the analyzed dataset are fairly low, as there are relatively few loading pathways or source locations delivering isolated individual pollutants.

Even if the FPM routines could be successfully modified to eliminate the existing bugs, the modified program would likely not yield surface guidelines higher than the existing SFB-RWQCB surface guidelines, which had occurred for several analytes in prior usage of the FPM in 2004. Nearly all the existing SFB-RWQCB surface guidelines are higher than even the new derived FPM foundation guidelines set at 75% false negatives (toxic hits occurring below that concentration), which are often similar to the overall 75th percentile of results in the current compiled data. The lone exception is cadmium, but even in that case, the existing sediment surface guideline is only ~5% lower than the newly derived FPM foundation (75th percentile of toxic hits) value.

The prior results with the FPM producing high concentration surface guidelines thus appear to be due in large part to the smaller total number of samples available at that time (studies from 2001 and prior), combined with the exclusion of sample results with non-detects. The

implementation of the algorithm used likely also contributed, with all the previous cases where the calculated FPM surface values exceeded the existing SFB-RWQCB surface guidelines being analytes for which the FPM derived surface and foundation values were identical.

Given the LTMS goal of increased reuse of dredged sediment (USACE et al., 2001) for wetland restoration projects and mitigating the loss of existing intertidal habitats due to sea level rise, deviation from the lower range of "reference" conditions may be overly stringent as the determinant of toxicity (both in this study and in the 2004 application of FPM). A majority of samples showing toxicity had concentrations (median and often 75th percentile) below the existing surface guidelines for all the target analytes. Thus short of choosing a high rate of false negatives (occurrence of toxicity despite concentrations below guidelines) in excess of 50-75% for surface values, application of the FPM will not yield higher surface guidelines.

The 75th percentile concentrations of these legacy contaminants in toxic samples are often only slightly higher than the 75th percentile for all (both non-toxic and toxic) samples, and the difference shrinks with increasing concentration percentiles. Thus FPM derived foundation values effectively become similar to setting targets just based on overall percentiles without regard to toxicity. For cases where multiple chemicals depart from their usual relative abundance, a mean hazard quotient approach such as the sediment quality guideline quotient, or SQG-Q1 (Fairey et al., 2001) may be useful, at least for samples primarily impacted by known legacy contaminants. SQG-Q1 or other similar hazard index or dose addition methods (USEPA 1986) provide a simple means to address potential multiple stressor impacts by combining estimates of their individual contributions to toxicity. This allows prioritization of samples in which more stressors are near or over their respective concern thresholds, although non-additive interactions are still not accounted for. Such methods may be useful for further refinements in dredged material management decision making.

Ultimately, guidelines developed for decisions on dredged sediment reuse need to find a balance between the relative importance of possible toxicity from the sediment and the resultant quantity of sediment that can be designated for reuse. Even if the method of evaluation or ranking is quantitative (FPM, mean risk quotients, or more complex models), the determination of what level of risk is acceptable remains subjective. Rather than evaluating just the risk associated with placed sediment, it may be useful to evaluate the problem from the opposite direction: in the absence of placed sediment, what is the current sediment quality in a given area, and what are the likely current effects to resident species? Even if reused sediment is projected to be potentially toxic, so long as its quality is equal to or better than the sediment currently in the area (or redistributed from the nearby Bay in the absence of reused sediment), the increased sediment quantity alone could be reason enough to justify a greater portion designated for reuse, at least for that specific area. In addition, if dredged sediment is not beneficially reused, will the effects to mudflat and marsh species be greater due to habitat loss from sea level rise rather than contaminant toxicity?

References

Fairey, R., E.R. Long, C.A. Roberts, B.S. Anderson, B.M. Phillips, J.W. Hunt, H.R. Puckett, and C.J. Wilson. 2001. An evaluation of methods for calculating mean sediment quality guideline quotients as indicators of contamination and acute toxicity to Amphipods by chemical mixtures. Environmental Toxicology and Chemistry, 20: 2276-2286.

Germano & Assoc., TerraStat, and Avocet Consulting. 2004. An Evaluation of Existing Sediment Screening Guidelines for Wetland Creation/Beneficial Reuse of Dredged Material in the San Francisco Bay Area Along with a Proposed Approach for Alternative Guideline Development. Prepared for the San Francisco Regional Water Quality Control Board and Port of Oakland.

Hunt, J. W., B. S. Anderson, B. M. Phillips, J. Newman, R. Tjeerdema, M. Stephenson, M. Puckett, R. Fairey, R. W. Smith, and K. Taberski. 1998. Evaluation and use of sediment reference sites and toxicity tests in San Francisco Bay. Prepared for California State Water Resources Control Board.

Michelsen, T., M. R. Anderson. 2011. Description and Use of the RSET Floating Percentile Method Spreadsheets. Prepared by Avocet Consulting, for the Oregon Department of Environmental Quality Regional Sediment Evaluation Team. March 4, 2011

SAIC and Avocet Consulting, 2002. Development of Freshwater Sediment Quality Values for Use in Washington State: Phase I Task 6 Final Report. Prepared by SAIC, Bothell, WA and Avocet Consulting, Kenmore, WA for the Washington State Department of Ecology, Olympia, WA. Publication Number 02-09-050

SFBRWQCB. 1992. Sediment screening criteria and testing requirements for wetland creation and upland beneficial reuse. Interim Final. Public Notice No. 92-145.

SFBRWQCB. 2000. Beneficial reuse of dredged materials: sediment screening and testing guidelines. San Francisco Bay Regional Water Quality Control Board Draft Staff Report, 35 pp.

SFEI. 2015. 2015 Pulse of the Bay: The State of Bay Water Quality - 2015 and 2065. SFEI Contribution No. 759. San Francisco Estuary Institute: Richmond, CA.

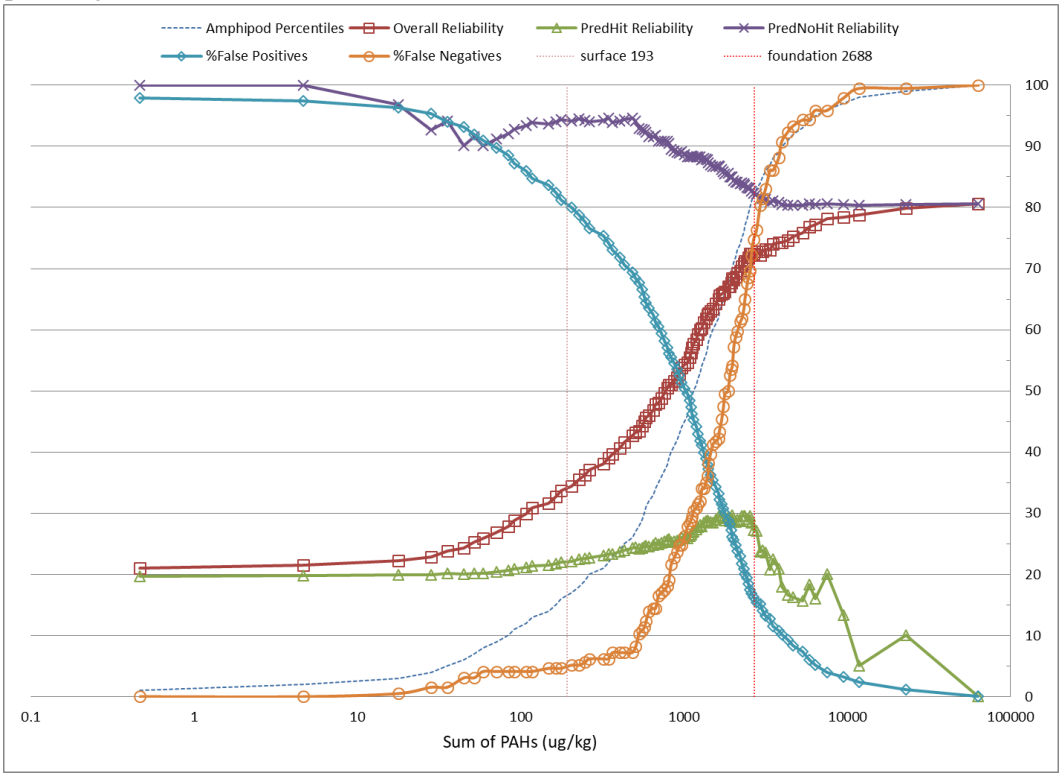
USACE, USEPA, San Francisco Bay Conservation and Development Commission, and San Francisco Bay Regional Water Quality Control Board, 2001. Long-Term Management Strategy for the Placement of Dredged Material in the San Francisco Bay Region Management Plan.

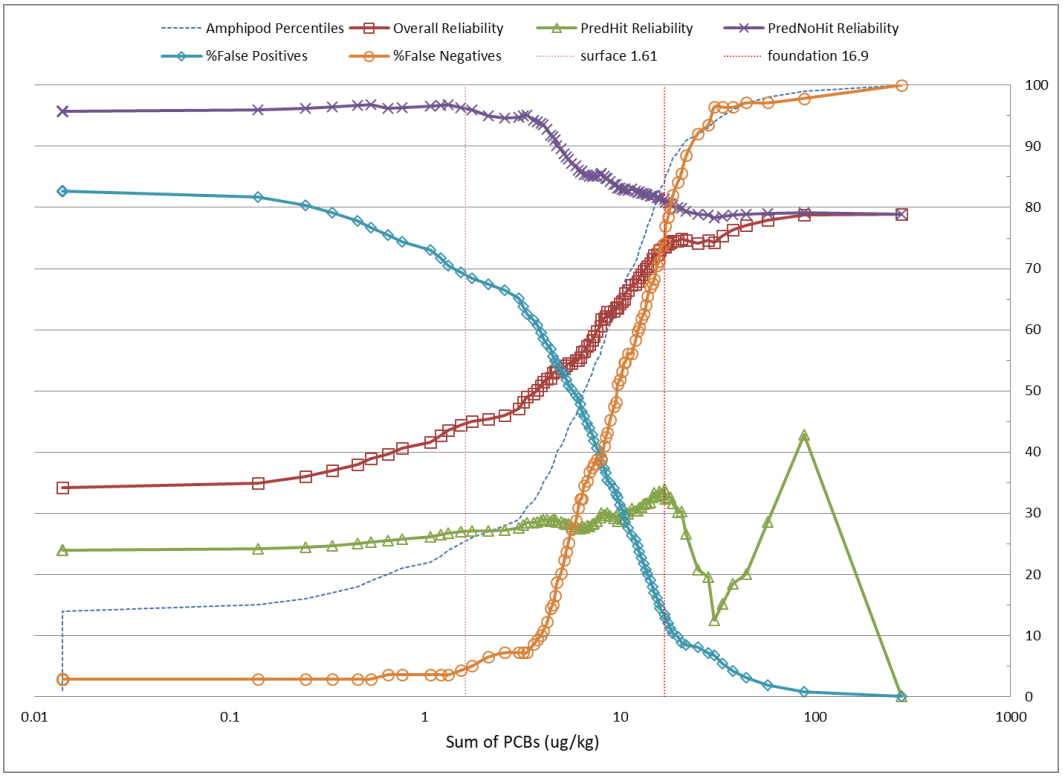
USEPA. 1986. Guidelines for the Health Risk Assessment of Chemical Mixtures. United U.S. Environmental Protection Agency, Washington, DC. EPA/630/R-98/002 Federal Register 51(185):34014-34025

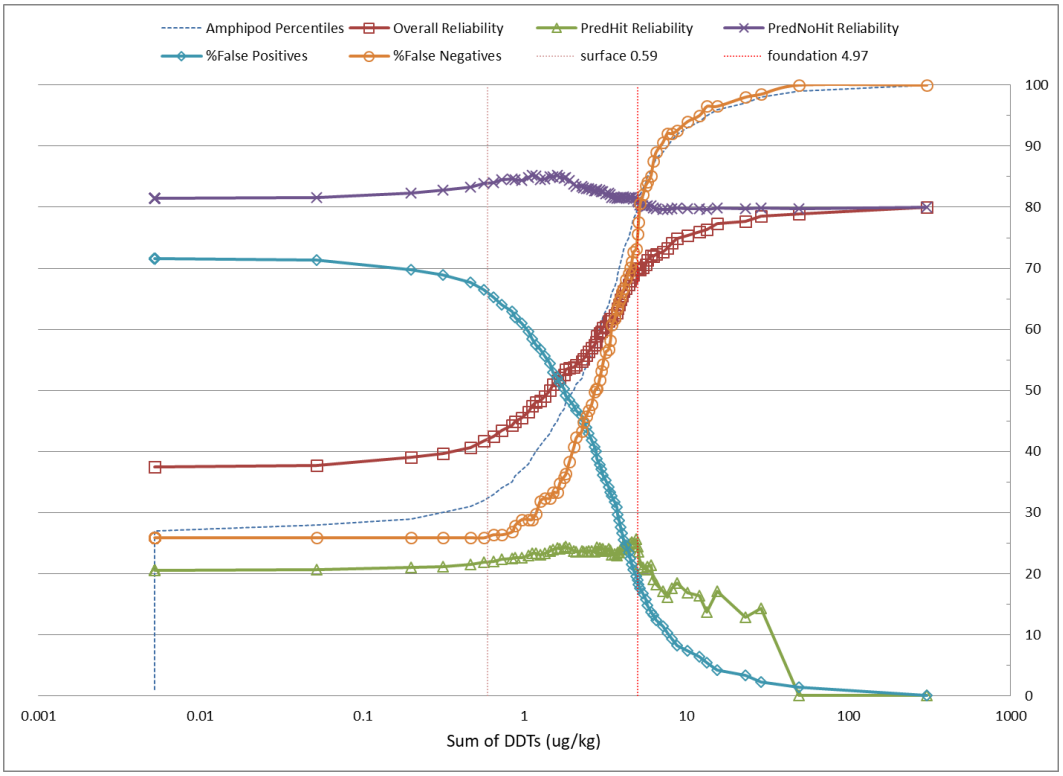
Appendix

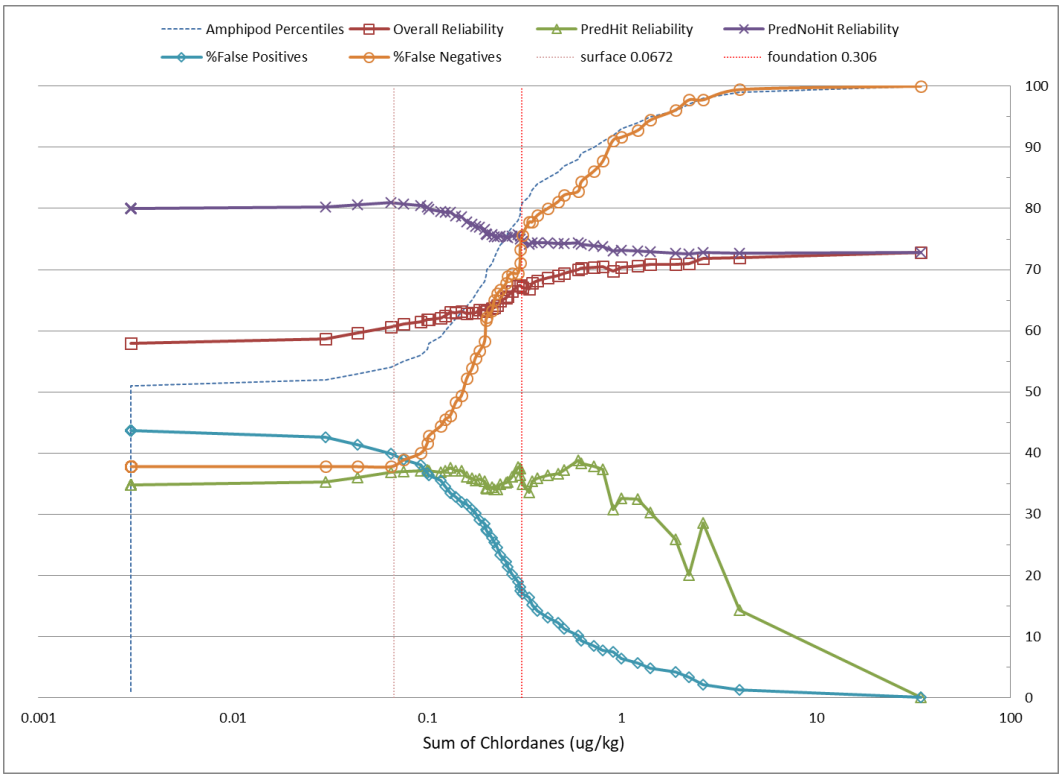
FPM derived charts for individual analytes. The y-axis for all graphs is expressed in units of percent (%) for each metric.

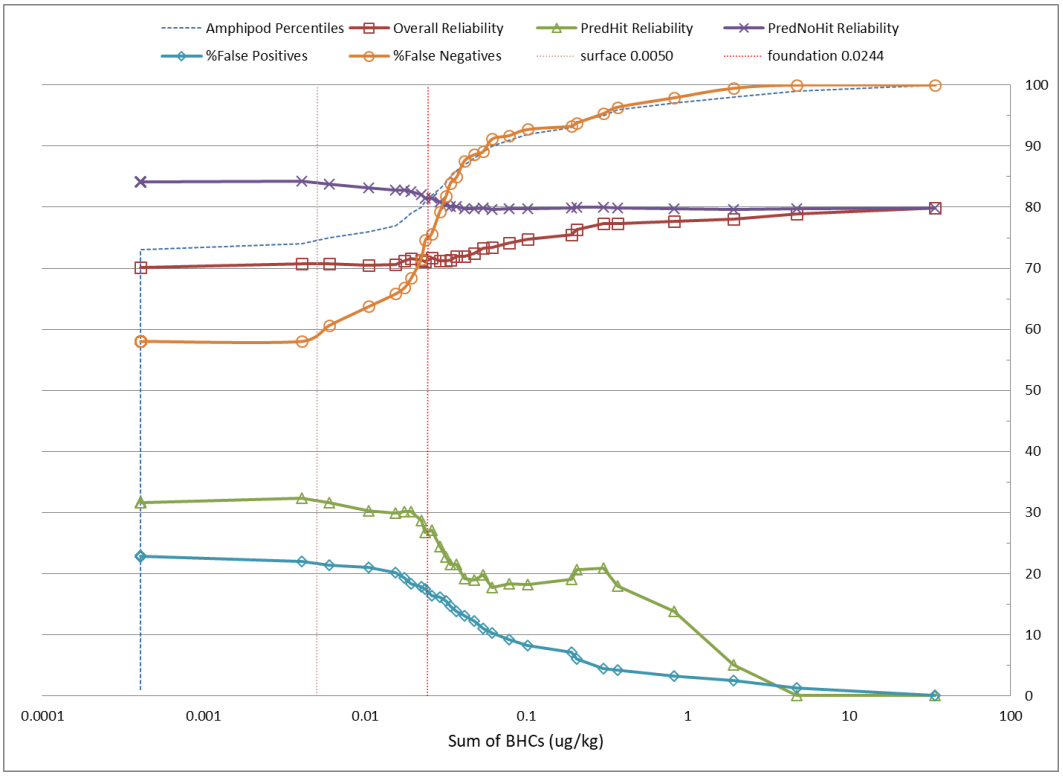
Amphipod organics charts



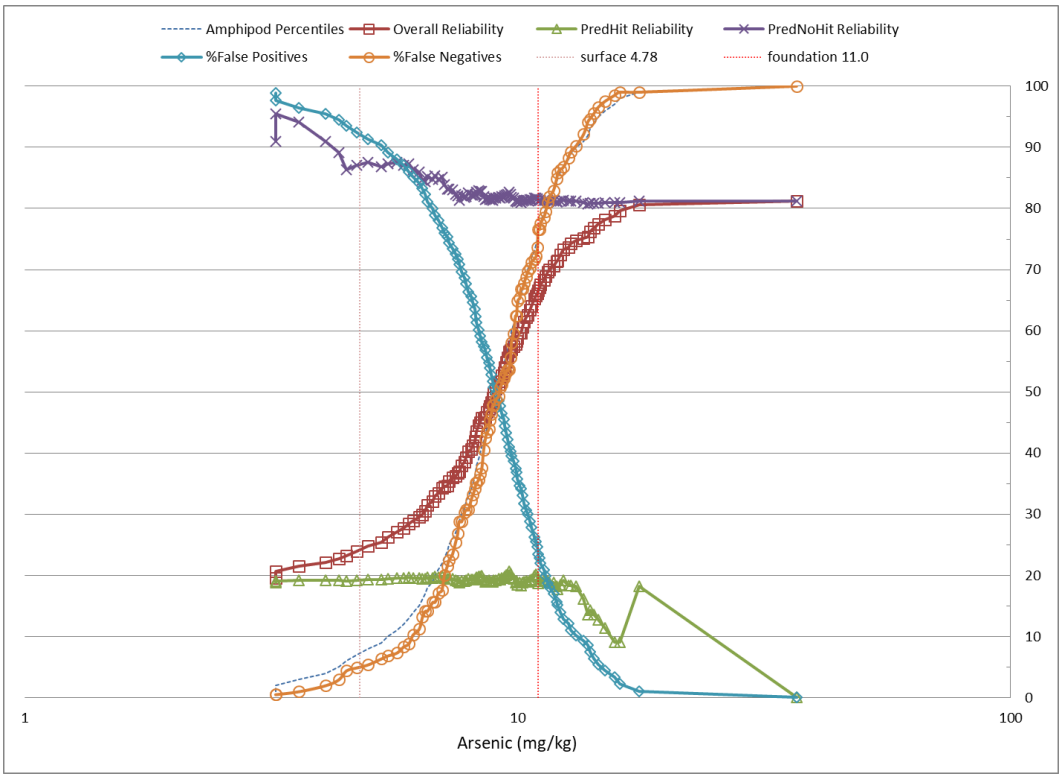


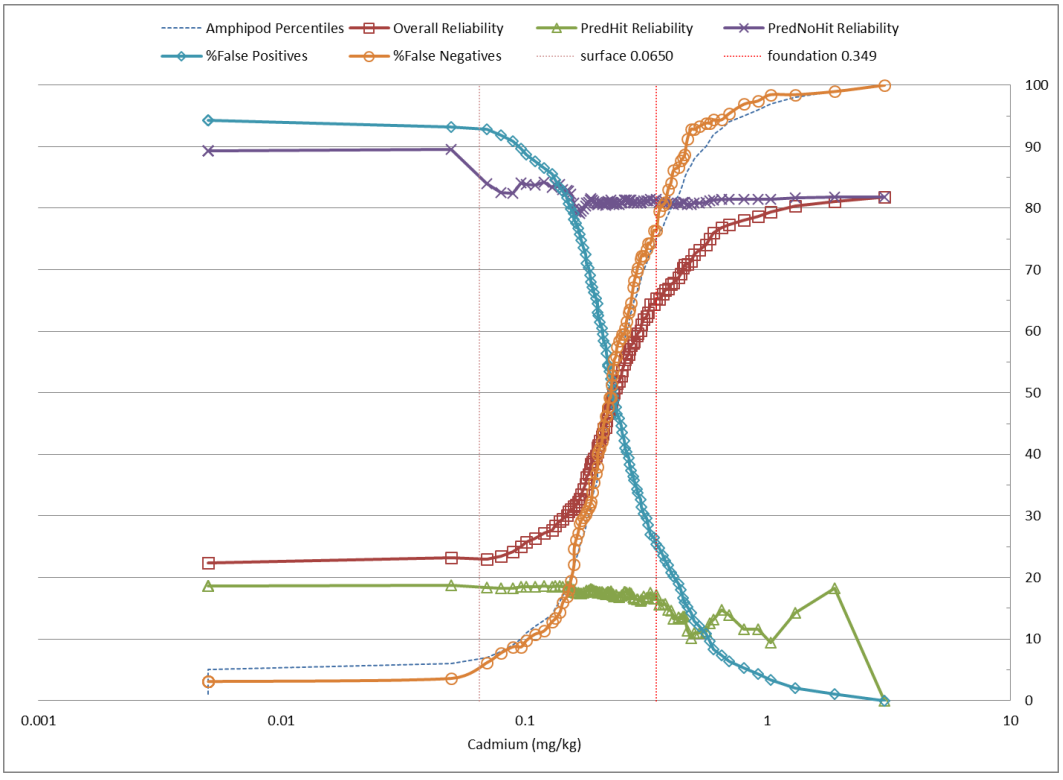

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.59
foundation 4.97
100
90
80
70
60
50
40
30
20
10
0
0.001
0.01
0.1
1
10
100
1000
Sum of DDTs (ug/kg)

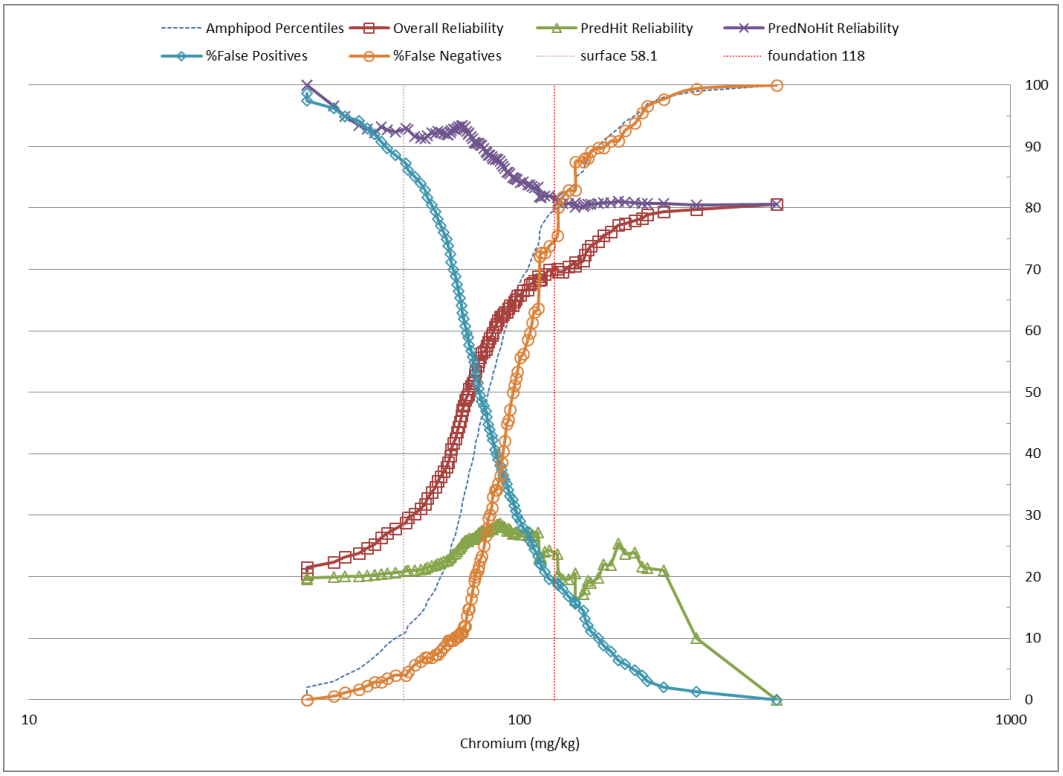

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0672
foundation 0.306
100
90
80
70
60
50
40
30
20
10
0
0.001
0.01
0.1
1
10
100
Sum of Chlordanes (ug/kg)

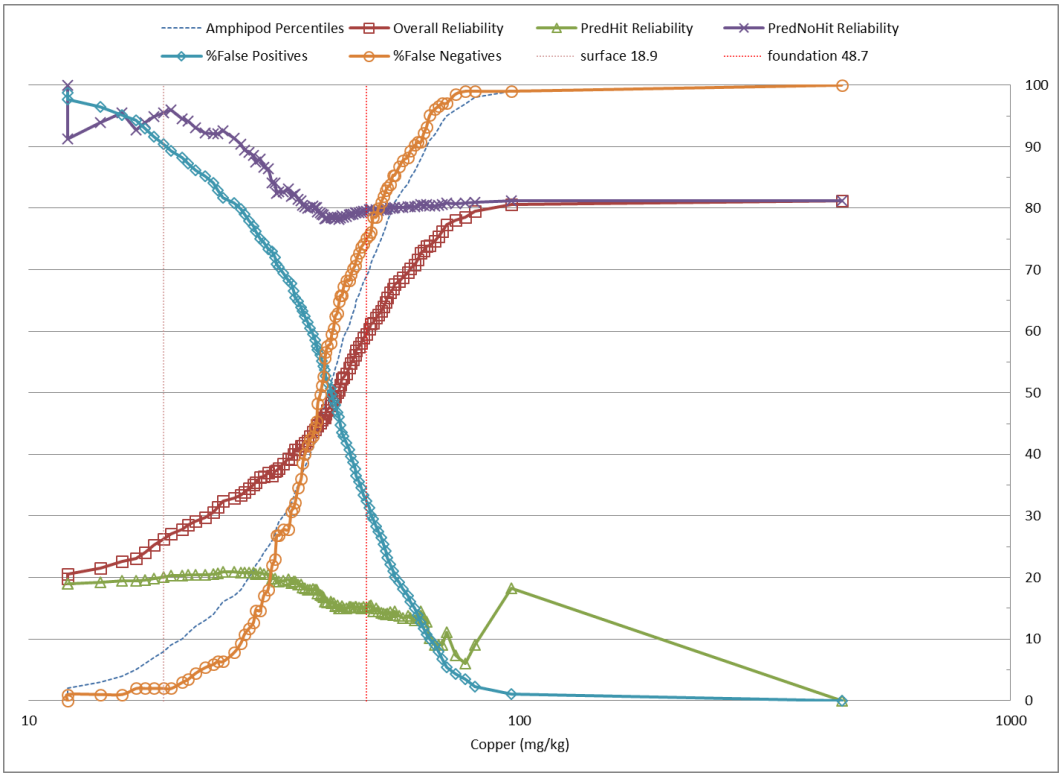

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0050
foundation 0.0244
100
90
80
70
60
50
40
30
20
10
0
0.0001
0.001
0.01
0.1
1
10
100
Sum of BHCs (ug/kg)

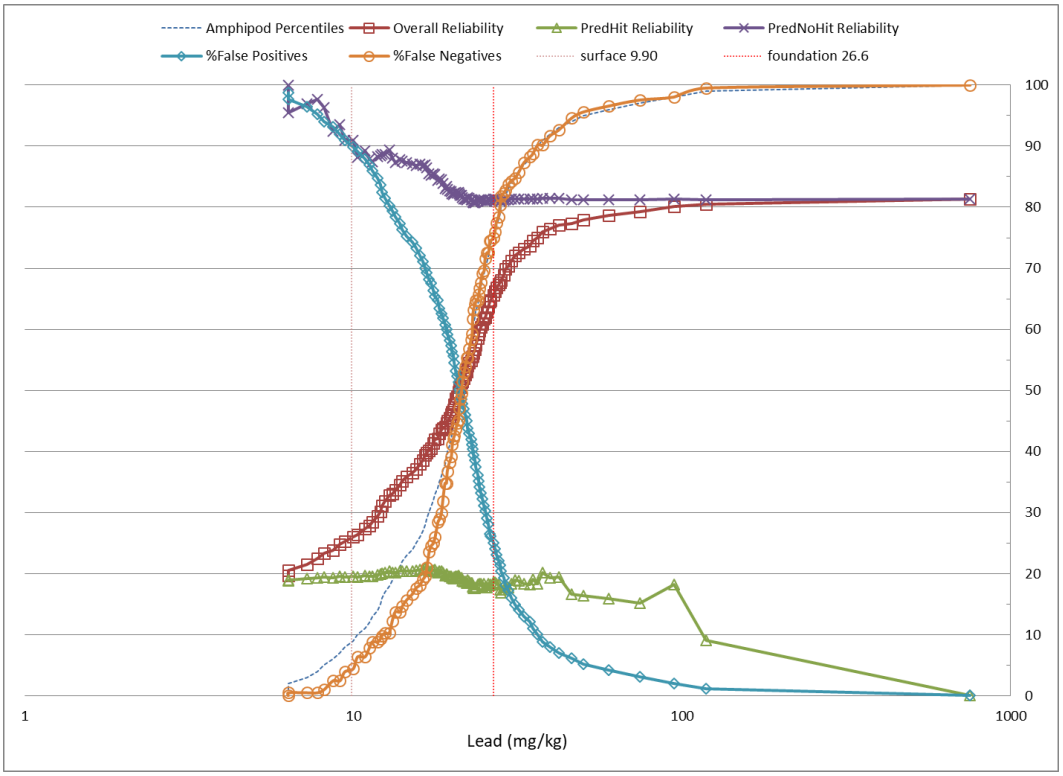
Amphipod metals charts

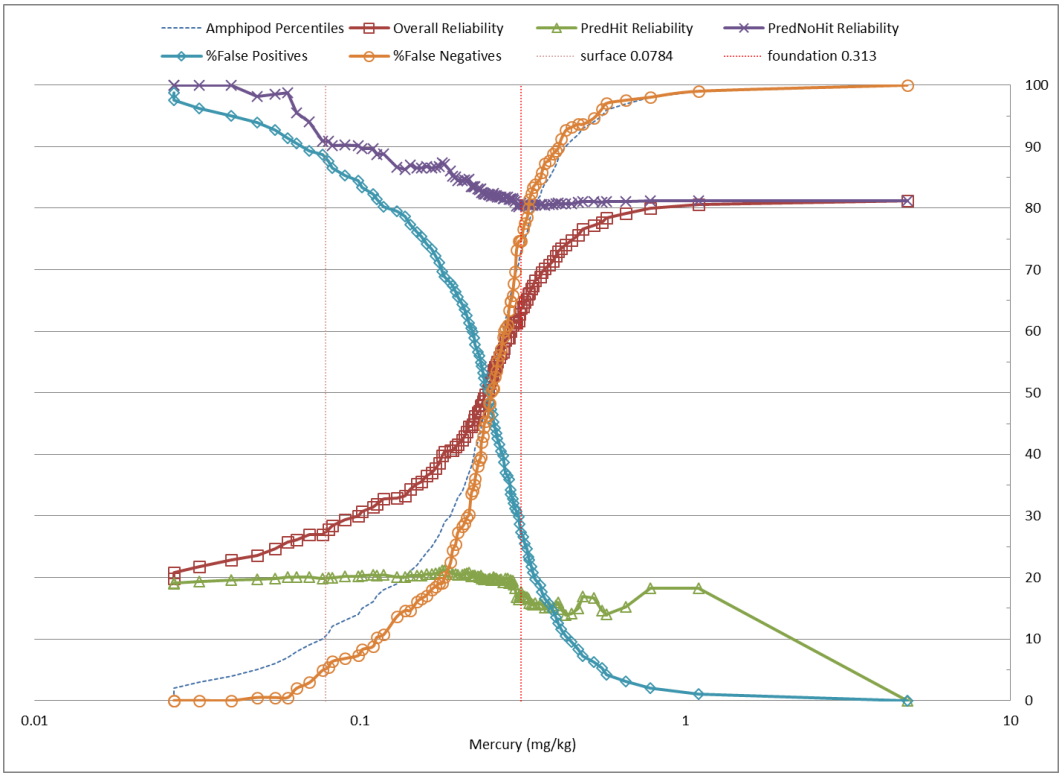


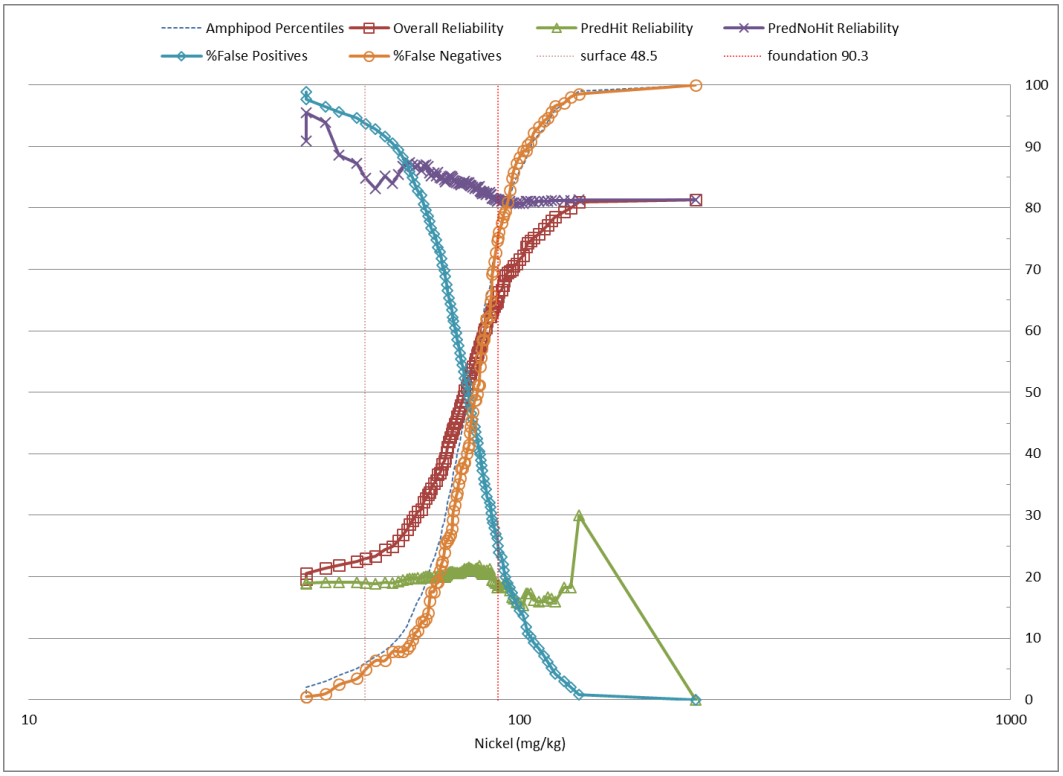


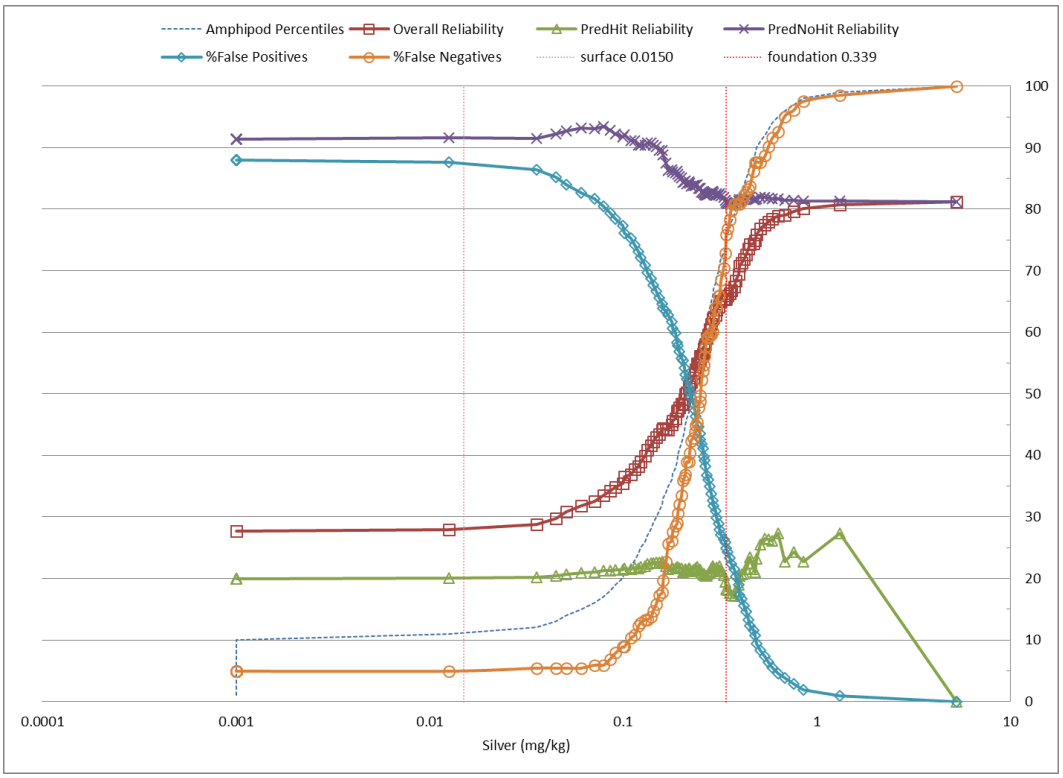

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 58.1
foundation 118
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Chromium (mg/kg)

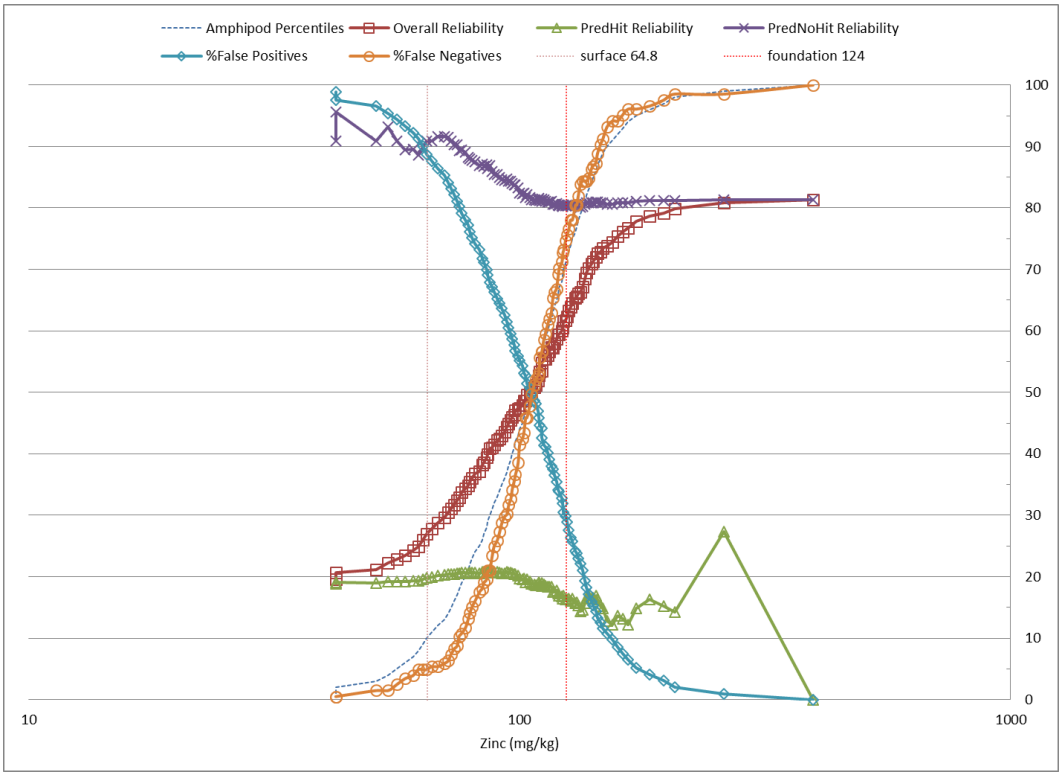

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 18.9
foundation 48.7
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Copper (mg/kg)


Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 9.90
foundation 26.6
100
90
80
70
60
50
40
30
20
10
0
1
10
100
1000
Lead (mg/kg)

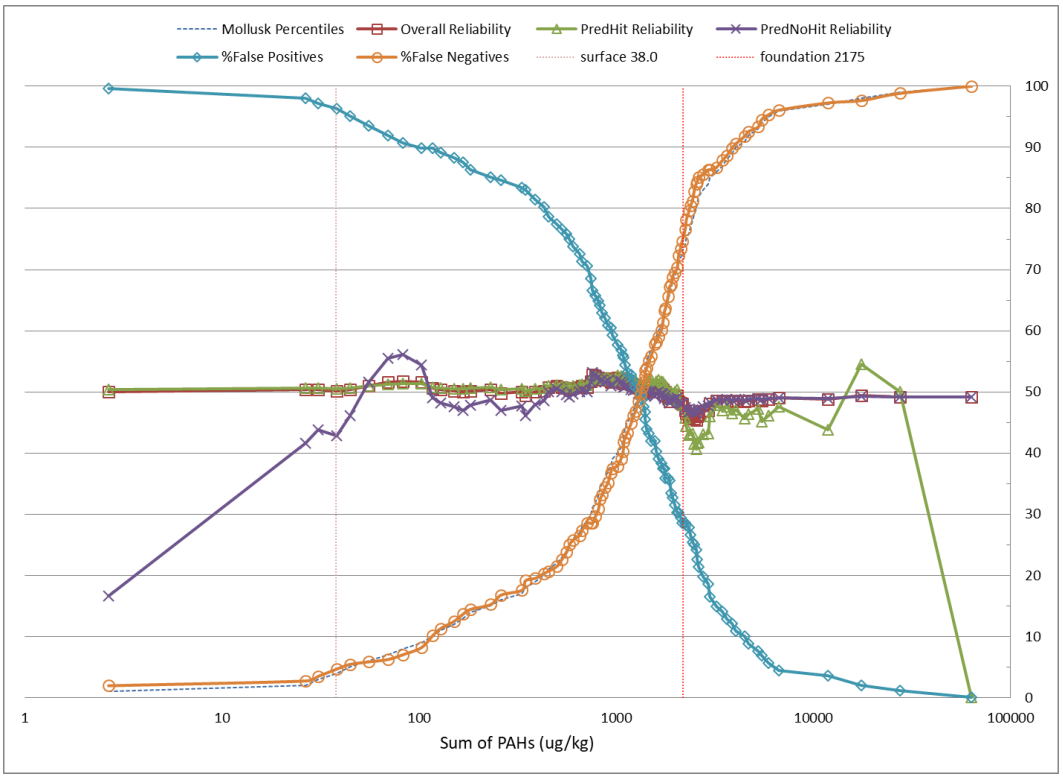

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0784
foundation 0.313
100
90
80
70
60
50
40
30
20
10
0
0.01
0.1
1
10
Mercury (mg/kg)

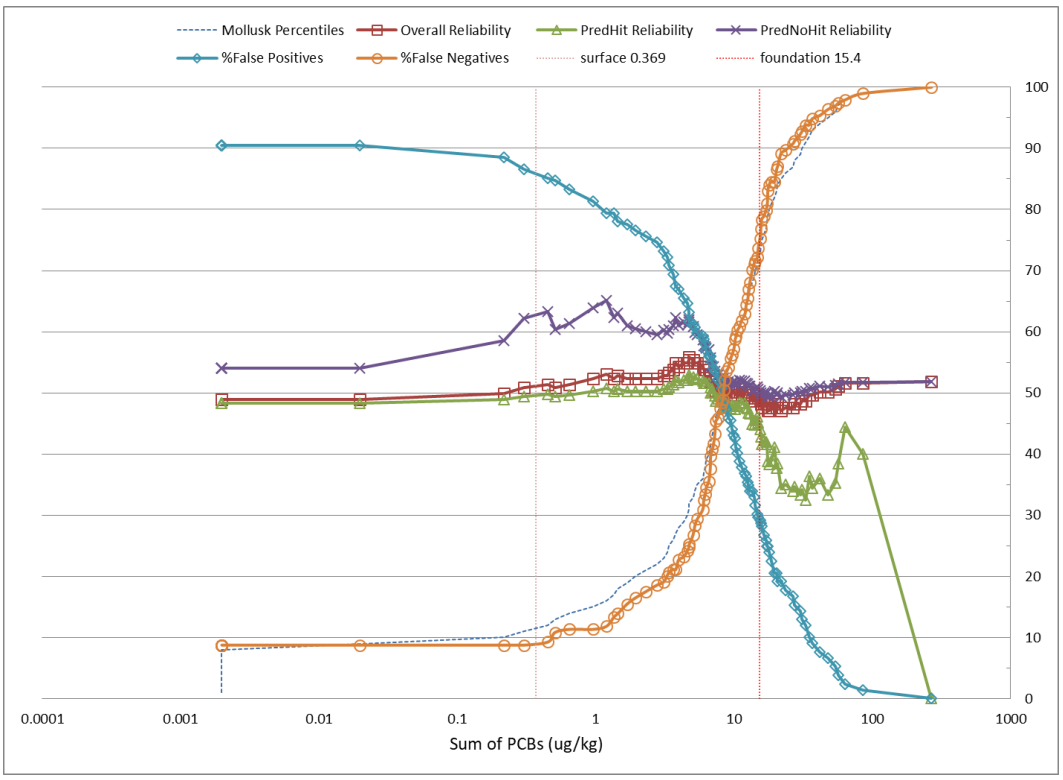

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 48.5
foundation 90.3
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Nickel (mg/kg)

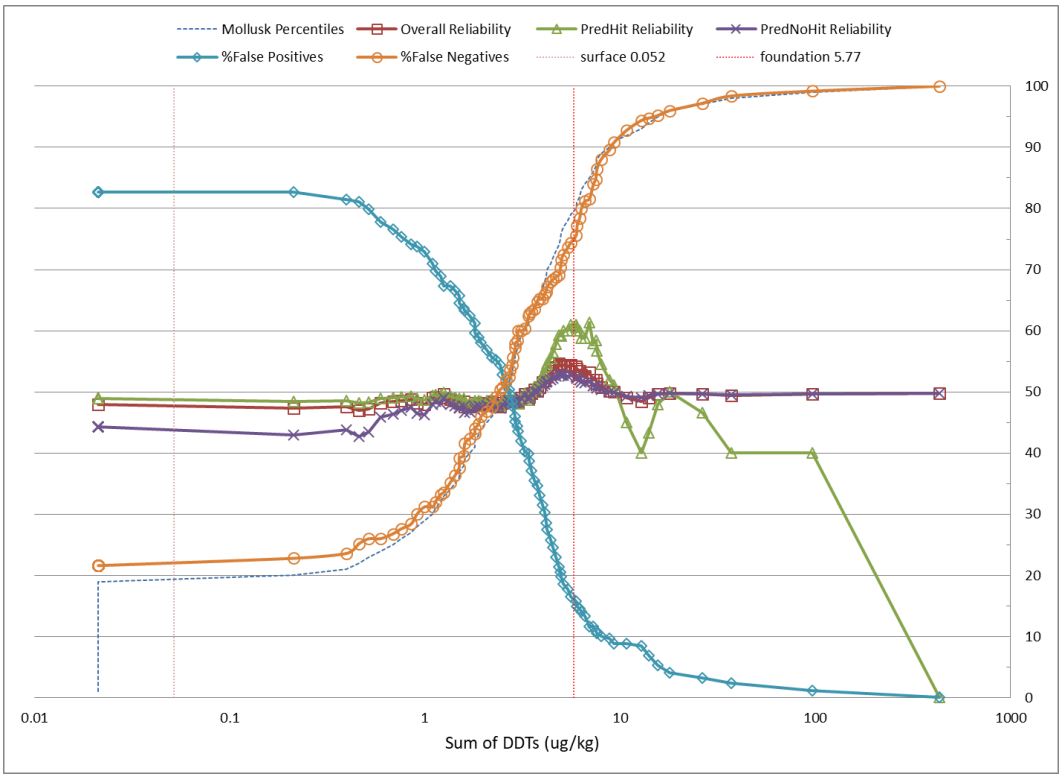

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0150
foundation 0.339
100
90
80
70
60
50
40
30
20
10
0
0.0001
0.001
0.01
0.1
1
10
Silver (mg/kg)

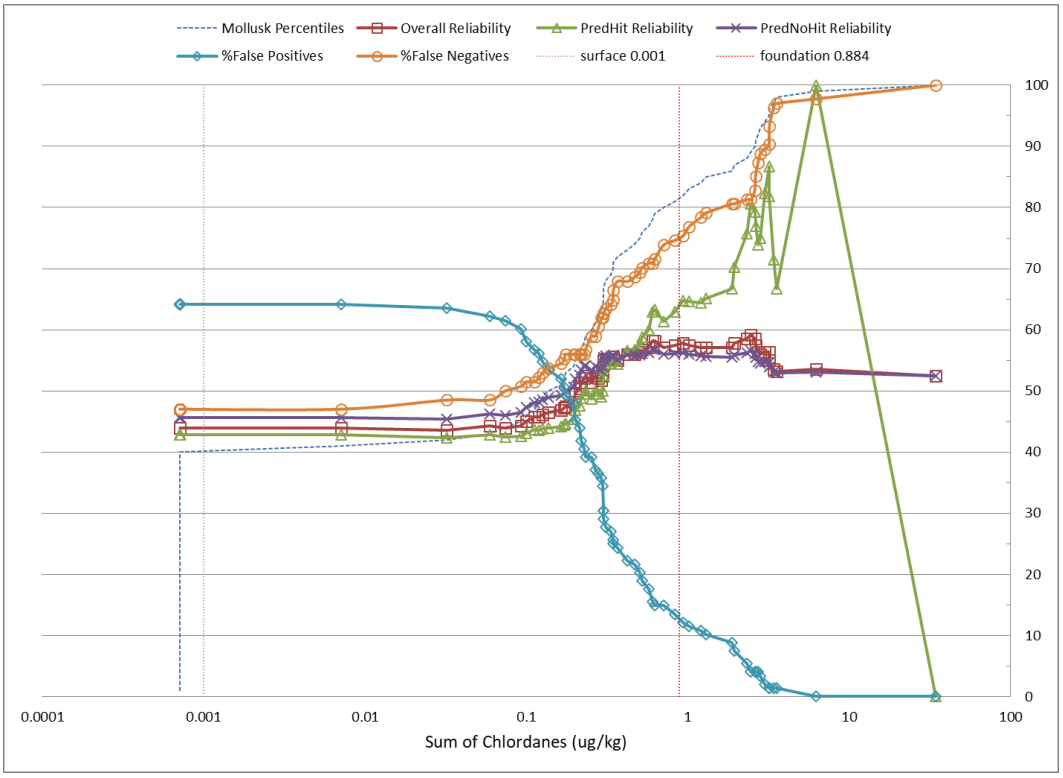

Amphipod Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 64.8
foundation 124
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Zinc (mg/kg)

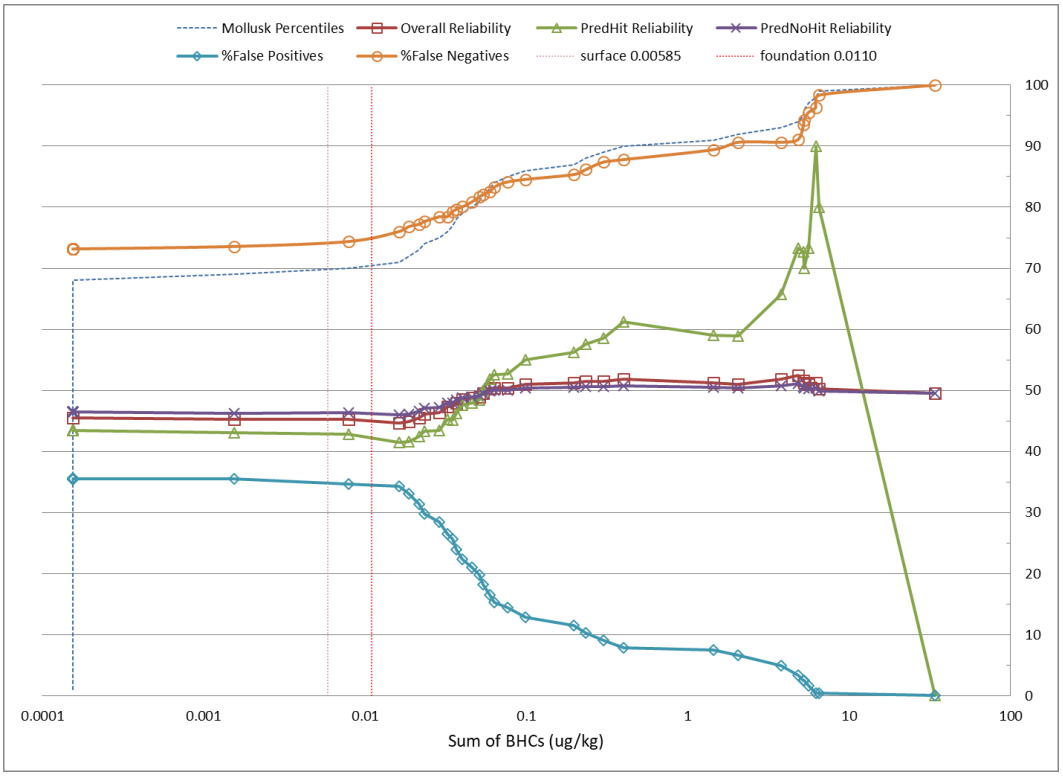
Mollusk organics charts



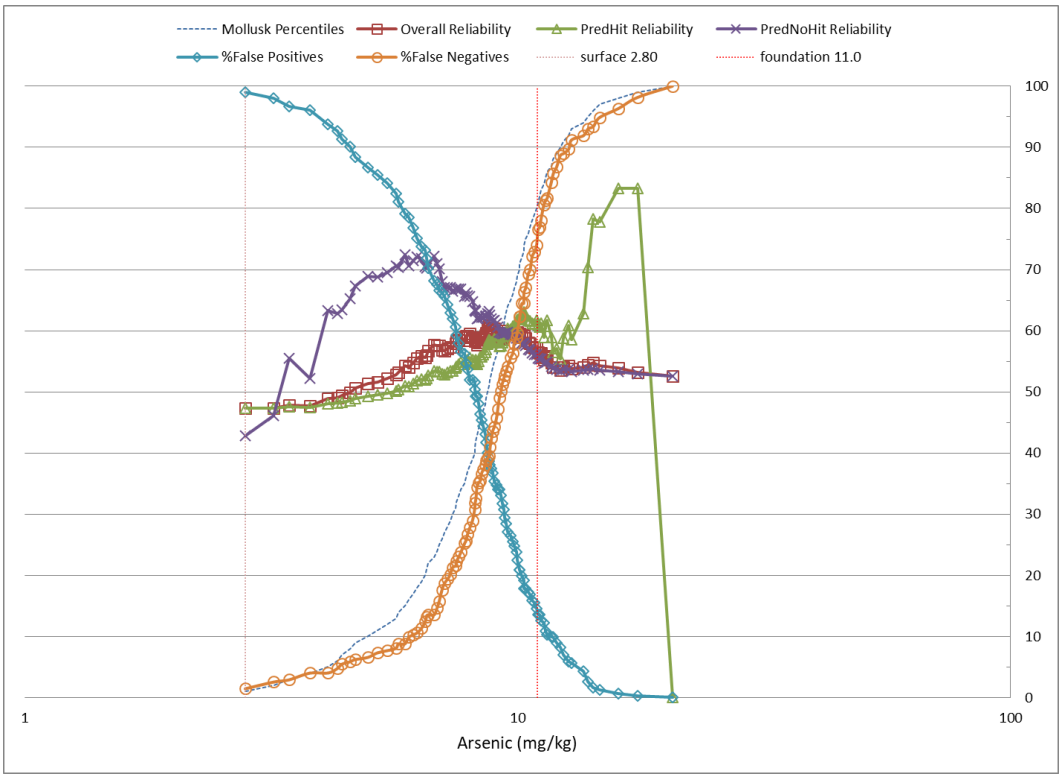


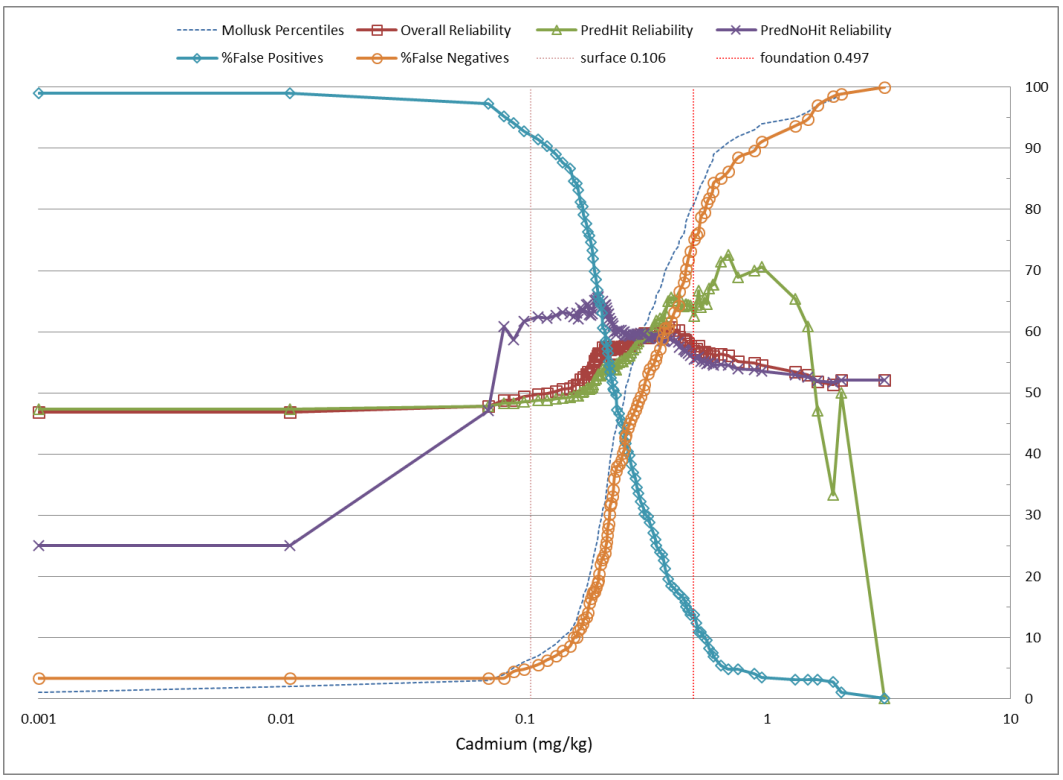

Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.052
foundation 5.77
100
90
80
70
60
50
40
30
20
10
0
0.01
0.1
1
10
100
1000
Sum of DDTs (ug/kg)

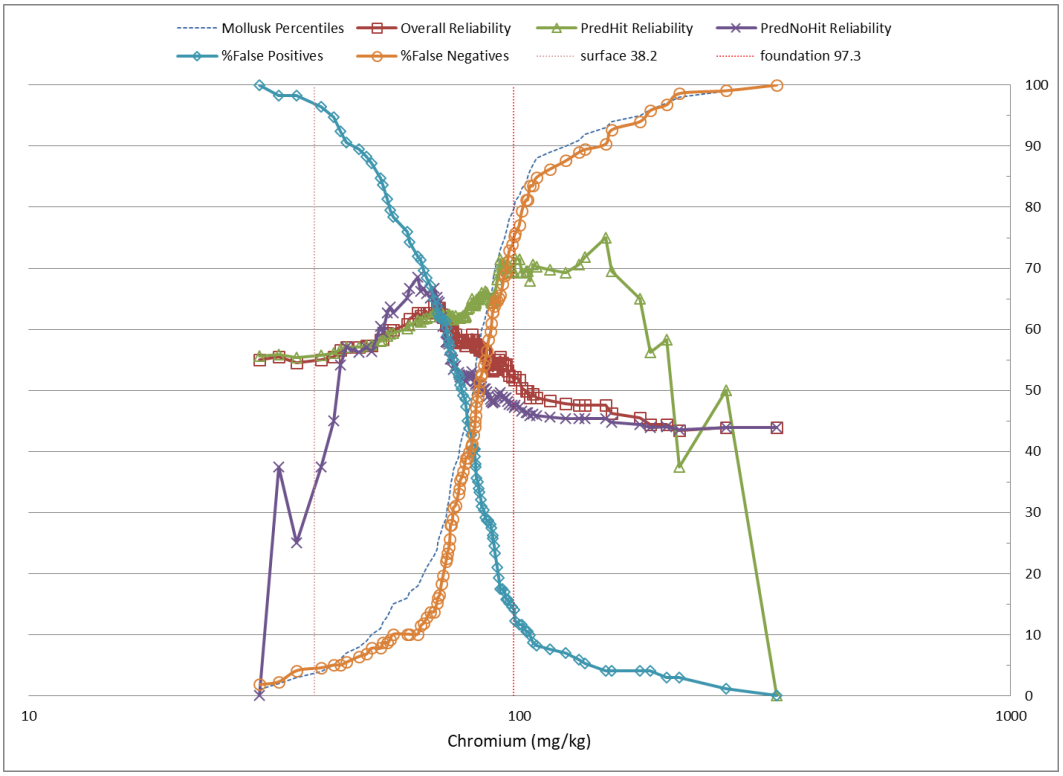

Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.001
foundation 0.884
100
90
80
70
60
50
40
30
20
10
0
0.0001
0.001
0.01
0.1
1
10
100
Sum of Chlordanes (ug/kg)

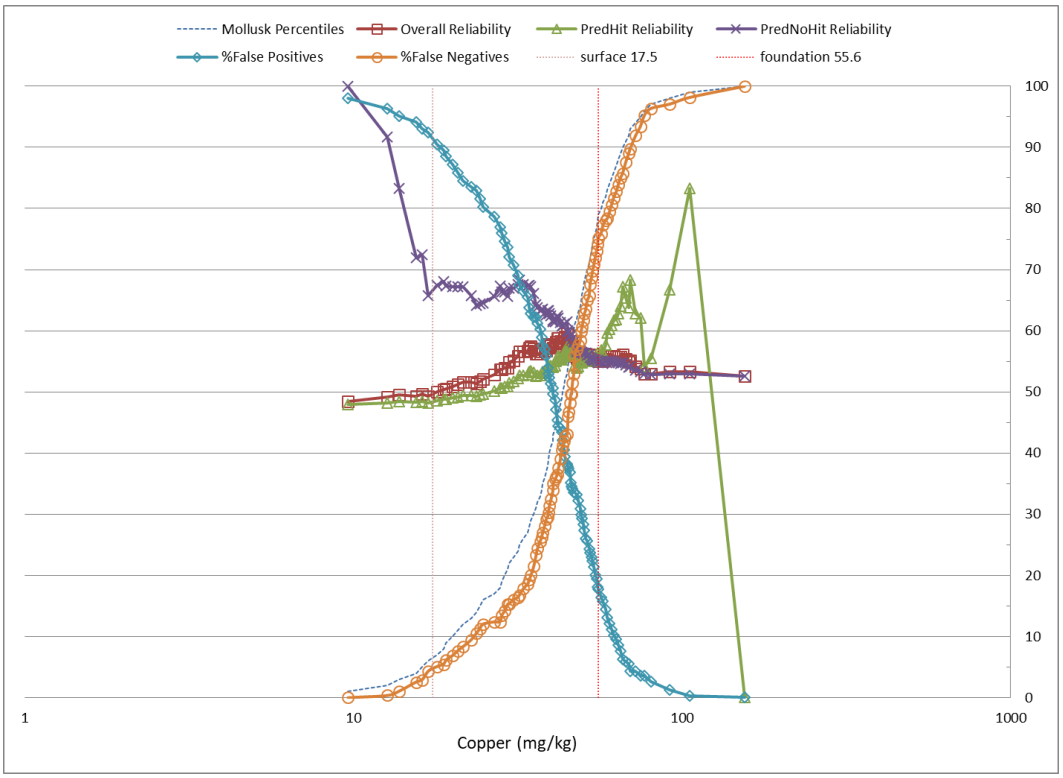

Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.00585
foundation 0.0110
100
90
80
70
60
50
40
30
20
10
0
0.0001
0.001
0.01
0.1
1
10
100
Sum of BHCs (ug/kg)

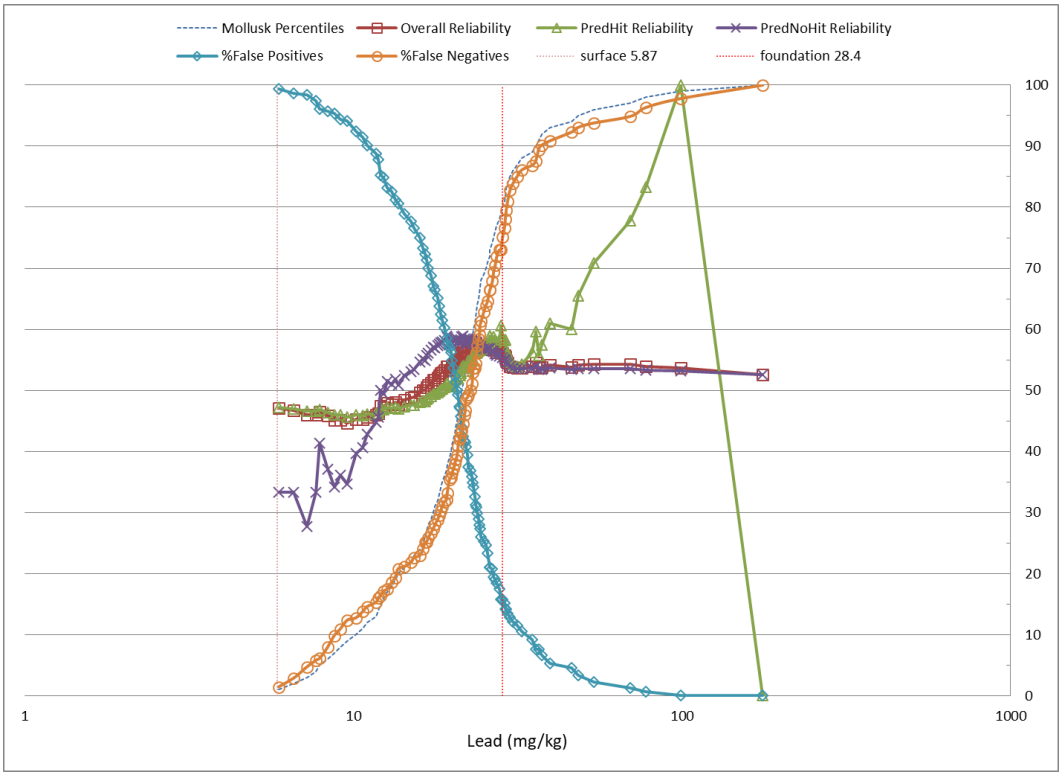
Mollusk metals charts

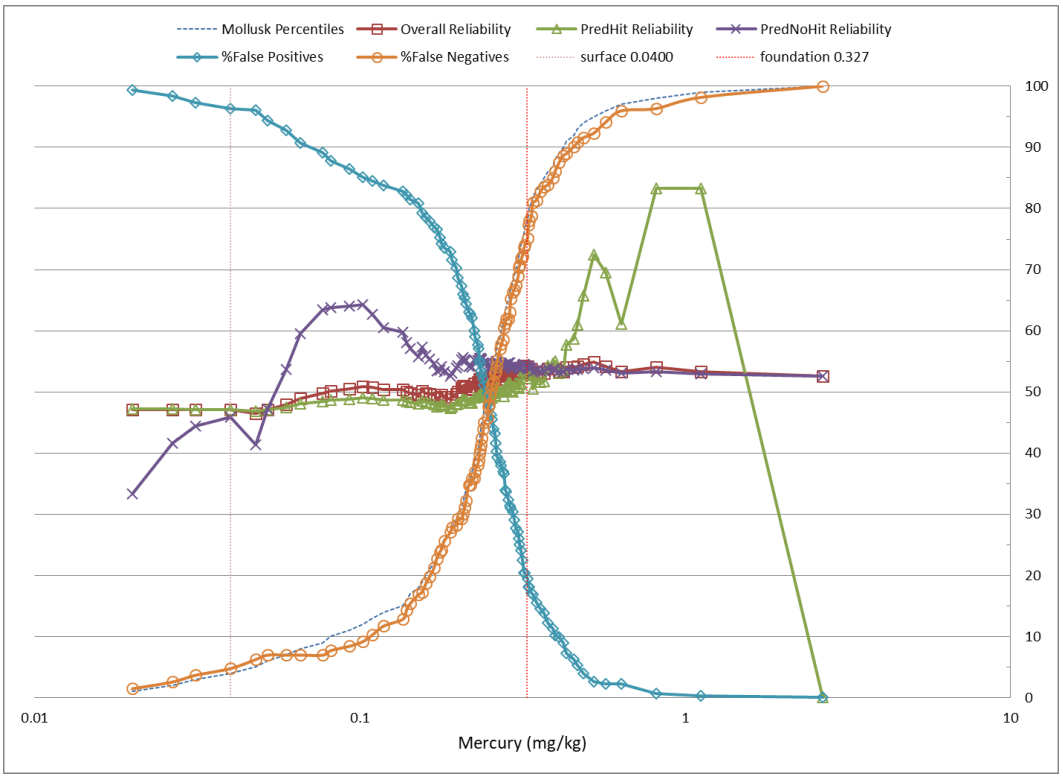


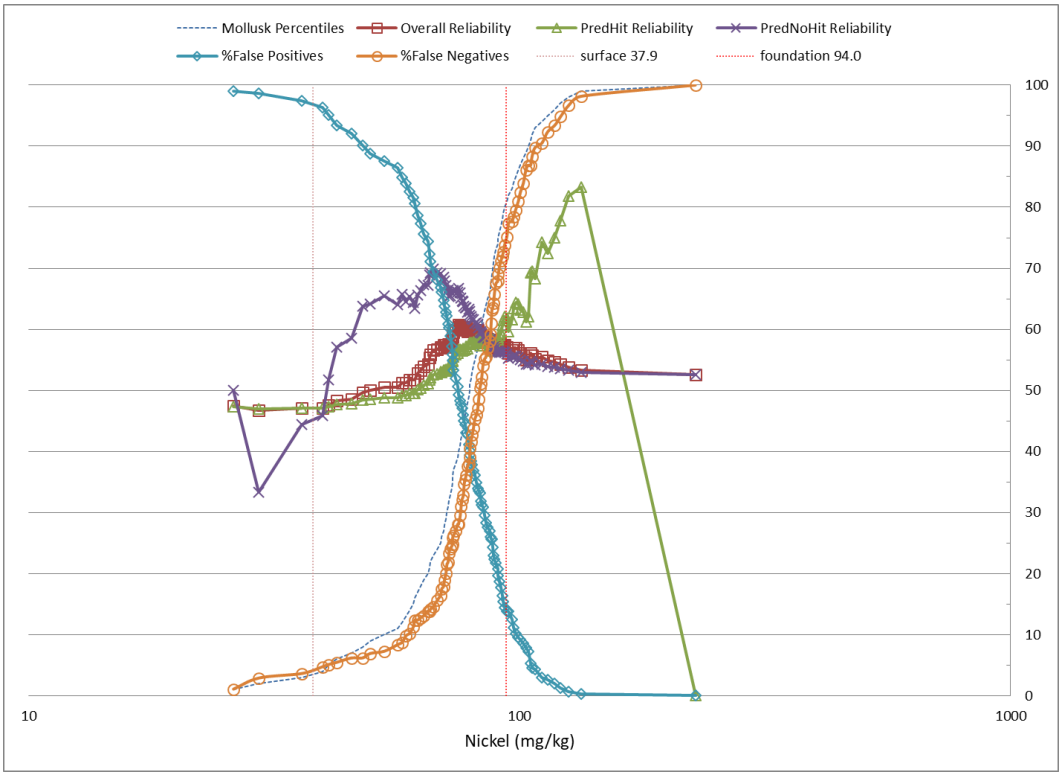


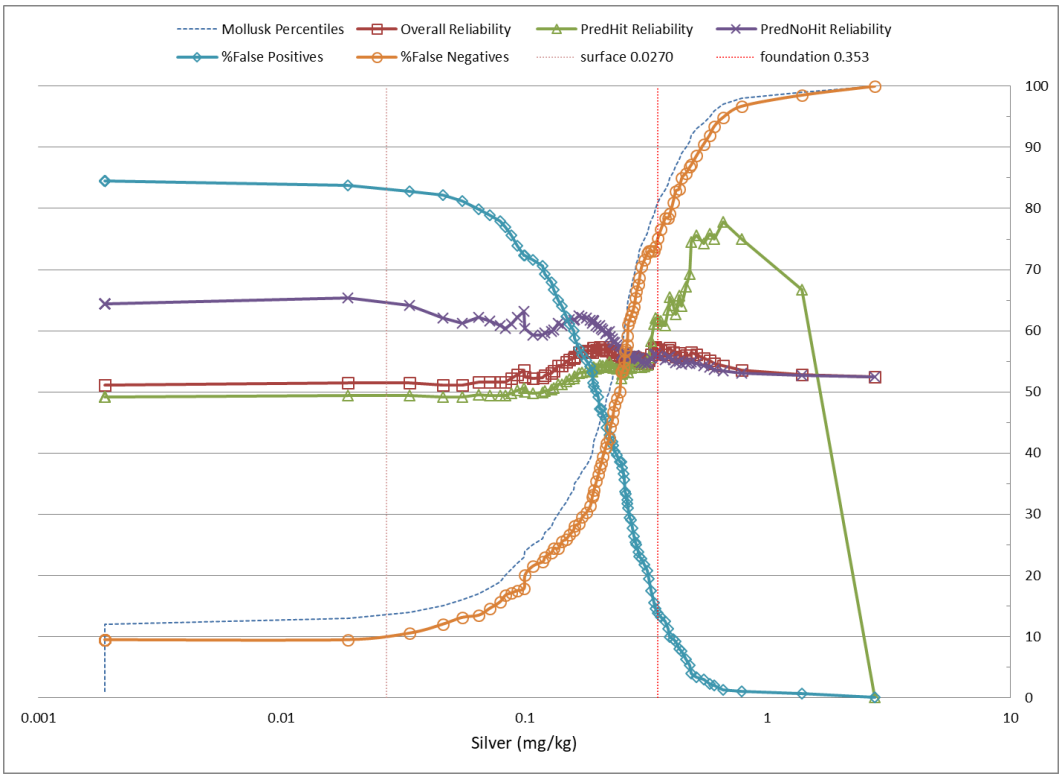

Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 38.2
foundation 97.3
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Chromium (mg/kg)

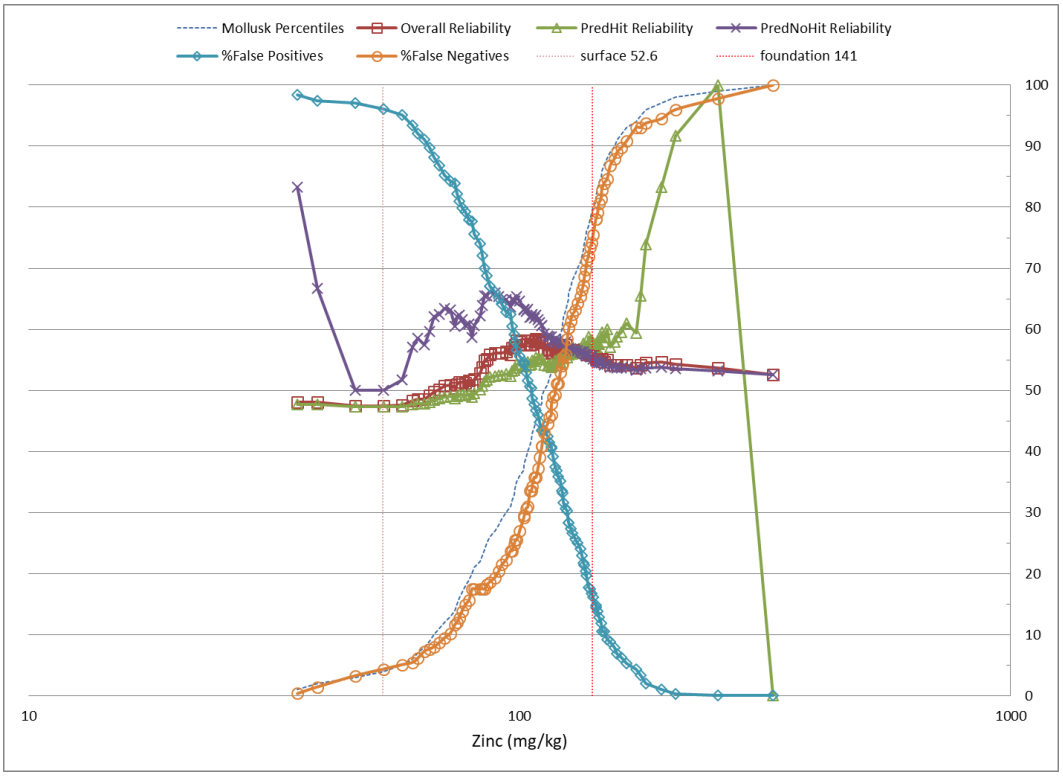

Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 17.5
foundation 55.6
100
90
80
70
60
50
40
30
20
10
0
1
10
100
1000
Copper (mg/kg)


Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 5.87
foundation 28.4
100
90
80
70
60
50
40
30
20
10
0
1
10
100
1000
Lead (mg/kg)


Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0400
foundation 0.327
100
90
80
70
60
50
40
30
20
10
0
0.01
0.1
1
10
Mercury (mg/kg)


Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 37.9
foundation 94.0
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Nickel (mg/kg)


Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 0.0270
foundation 0.353
100
90
80
70
60
50
40
30
20
10
0
0.001
0.01
0.1
1
10
Silver (mg/kg)


Mollusk Percentiles
Overall Reliability
PredHit Reliability
PredNoHit Reliability
%False Positives
%False Negatives
surface 52.6
foundation 141
100
90
80
70
60
50
40
30
20
10
0
10
100
1000
Zinc (mg/kg)