General Assistance Program (GAP) Evaluation of the Sediment Contamination in Lake Manistee Using Exploratory Data Analysis

---

Prepared for:

Frank Beaver, GAP Project Manager
GAP Grant #GA97546201-0
Little River Band of Ottawa Indians
Manistee, MI

and

United States Environmental Protection Agency

Prepared by:

Robert M. Powell, Principal Scientist

Powell & Associates

3125 S. Ortonville Road
Clarkston, MI

29 June 2004

Table of Contents

List of Figures

List of Tables

Executive Summary and Conclusions

This report was developed as a deliverable to the Little River Band of Ottawa Indians under their USEPA General Assistance Program (GAP) grant (Grant #GA97546201-0). It presents the results of an exploratory data analysis (EDA) and evaluation to further investigate the chemical and biological information provided in the manuscript "Preliminary Investigation of The Extent of Sediment Contamination in Manistee Lake," (Rediske et al., 2001).[1] The investigators and authors of that sediment investigation report did a thorough job of acquiring data from both sediment surface samples and sediment cores along the length of Manistee Lake. This EDA investigation and report sought to better understand the nature and impacts of some of the sediment contaminants that were quantified during the Rediske et al. (2001) study and to discern whether additional insights could be gained from the data accumulated in that report for better understanding of the Lake Manistee environment. An additional purpose of this report was to provide the basis of a strategy for acquiring further samples from Lake Manistee and to assist in the development of a conceptual model of the lake.

All data exploration and any statistical charts, tables, displays and values presented in this report have been generated using either the program DataDesk 6.2 (Data Description, 2003, http://www.datadesk.com) or Aabel 1.5.7 (Gigawiz, 2004, http://www.gigawiz.com). Data explorations focused initially on the metal and metalloid contaminants, total organic carbon (TOC), and hexane extractable materials with regard to depth in the sediment and sampling location. A continued evaluation explored the organism studies and looked for trends in these results versus location and potential chemical predictors that were measured in the uppermost sediments as well as with regard to potential pollutant sources.

This report is divided into two Parts. Part 1 investigates certain of the chemical contaminants in the sediments and their concentrations relative to sampling depth, relative to the current General Chemical site (the proposed location of Tondu Corporation's Northern Lights Power Plant), relative to sampling location along the length of Lake Manistee, and further evaluates the contaminant concentrations in the context of Probable Effects Concentrations (PEC).

Part 2 addresses several biological studies and additional contaminants from Rediske et al. (2001) and considers the chemical concentrations and locations relative to the potential sources and source types. Organism mortality and counts are also evaluated versus sampling locations and with correspondence to the contaminant concentrations. A detailed evaluation of the chemicals versus four organism studies using stepwise multiple

---

[1] Dr. Richard Rediske, Principle Investigator and Dr. John Gabrosek, Dr. Cynthia Thompson, Carissa Bertin, and Jessica Blunt Annis of the Water Resources Institute, Grand Valley State University, One Campus Drive Allendale, MI 49401; Dr. Peter G. Meier , University Of Michigan School Of Public Health I, Ann Arbor, MI 48106. AWRI Publication # TM-2001-7, Great Lakes National Program Office #985906-01; U. S. Environmental Protection Agency; National Oceanic and Atmospheric Administration. PROJECT OFFICER: Dr. Marc Tuchman U. S. Environmental Protection Agency, Great Lakes National Program Office, 7 West Jackson Boulevard, Chicago, IL 60604-3590, July 2001July 2001.

regression analyses was conducted and the most likely chemical causes of organism mortality were identified.

The results and conclusions of this EDA are neither intended to supplant nor fully reiterate the results of the Rediske et al. (2001) sediment investigation report but are intended to further examine certain of the data therein. Based on the results of the EDA the specific findings contained in this report are:

1) The measured constituents of concern (COCs) are primarily found in the uppermost sediments in the lake, indicative of the fact that they do not have a natural mineralogical occurrence in this environment but are of anthropogenic origin, i.e., these are introduced contaminants.
2) Many important contaminants in these upper sediments show an increase in concentration from the southern end to the northern end of the lake, i.e., as the water flows past the various industries
3) Contrary to the Rediske et al. (2001) report, the consensus based Probable Effects Concentrations (PEC's) are exceeded in the uppermost sediment for two contaminant metals, chromium and copper, and at a slightly greater depth by lead.
4) The PECs for some of the other individual contaminants are very closely approached even if not exceeded.
5) The PECs for organic contaminants, such as PAH compounds, were exceeded at 10 locations and oil is fairly ubiquitous in the lake sediments (from Rediske et al., 2001).
6) Because of 3, 4 and 5, it is likely that utilizing individual PECs for such a complex contamination scenario is an inadequate means to characterize the degree of toxicity to benthic organisms and the fish that feed upon these organisms. Multiple contaminant effects might be important and should be considered.
7) Box and whiskers plots of the biological study results for *H. Azteca* mortality, *C. Tentans* mortality, total organism count and species number count versus potential sources based on location (Rediske et al., 2001) and versus general contaminant type (control, chemical, salt) show that only the control locations have medians and 95% confidence intervals significantly different from the other categories. That is, the impacts upon the organisms cannot be distinguished from one another statistically based upon these categories (except for the controls).
8) Area plots of the four sets of biological study results versus location show extremely good correspondence. The acute toxicity studies on *H. Azteca* and *C. Tentans* correspond closely both with one another and with the organisms and species count data. Mortality of both test species increases dramatically upon exposure to the Lake Manistee sediments (relative to the control sediments) and organism counts plummet. Mortality/disappearance worsens by factors of 1.6 to 5.5 immediately within the Lake.
9) Stacked area charts of the contaminants overlain by *H. Azteca* and *C. Tentans* % Mortality lines show excellent correspondence between mortality and locations of very high total contamination, but do not elucidate which contaminants or subset of contaminants are responsible for the mortality. These plots also show that the

upper sediments of Lake Manistee are contaminated throughout the length of the Lake.

10) Simple linear regressions between the biological study results and individual contaminant concentrations do not adequately describe the results of the studies.
11) Stepwise multiple regression analyses on the results of biological studies strongly indicate that certain sediment contaminants are well correlated with the dearth of organisms in Lake Manistee. These contaminants are arsenic, hexane extractable compounds, chromium, PAH, mercury and possibly selenium.

General conclusions that can be reached based upon the results presented in this and the Rediske et al. (2001) report include:

1) The upper sediments in Lake Manistee are very contaminated and inhibiting the establishment of a normal benthic ecosystem that, in turn, negatively impacts the aquatic ecosystem in the Lake.
2) Lake Manistee cannot support additional contaminant loading without further endangerment of the remaining fish populations, human health, and the overall lake environment.
3) No new discharges to Lake Manistee or the rivers discharging to Lake Manistee should be permitted at this time and currently permitted discharges should be re-evaluated.
4) Dredging of the contaminated upper sediments to re-form the shoreline or deepening the Lake for various industrial/recreational activities is likely to release currently sediment-bound contaminants to the lake water and should be avoided at this time.
5) Additional study is needed prior to any dredging activities or alteration of the current shoreline usage. These studies could also serve as precursors to a potential remediation of the lake to return it to a higher state of beneficial use.
6) Lake Michigan is probably suffering added contamination due its connection as the outlet of Manistee Lake and this connection should be evaluated.

Planned and needed activities to enhance understanding and better manage the Lake Manistee contamination issues include:

1) Developing a sampling strategy based on the results of this report and Rediske et al. (2001) to further define and characterize the lake quality in critical areas.
2) Assess current pollutant inputs to Lake Manistee via an inventory, especially with regard to the contaminants implicated in this report as exerting the most significant toxic effects, and seek to control or eliminate their discharge.
3) Develop a conceptual model of Lake Manistee including the sediments and immediate shoreline.
4) Estimate the fate of the contaminants that are impacting sediment biota in terms of concentration versus time.
5) Assess the need and potential for remediation of contaminant outfalls and “hot spots” based on the results of these studies.

Approach to the Exploratory Data Analysis

The EDA was conducted on data resulting from the sampling and analysis of lake bottom surface sediments and core samples. Figure 1 depicts the coring/sampling locations per Rediske et al. (2001). Tables 1 and 2 provide the data that were subjected to EDA by DataDesk6.2. Part 1 of this report primarily addresses investigations of contaminant concentrations relative to sampling locations and Part 2 focuses mainly on exploration of the organism studies relative to contaminant concentrations.

The data of Table 1 were all extracted directly from a .pdf file of the Rediske (2001) report using Adobe Acrobat 6.0 Professional (Adobe Systems Inc., 2003), with the exception of the columns "Location#", "Depth Level", and "NLPP Relative Location" that were added for statistical purposes. The location number column simply assigns a numeric value to the "Station" from the Rediske report, which corresponds to a sediment sampling/coring location. The depth level S is for surface samples acquired using a Ponar sampler whereas T, M, and B refer to top, middle and bottom core sections, respectively. Details of the sampling design and methodology are presented elsewhere (Rediske et al, 2001). A variable (data column) was created for the NLPP Relative Locations. These refer to the locations of the sampling points with respect to the proposed location of the Northern Lights Power Plant and with regard to the flow direction in the lake; the water flows northwesterly from the south (Rediske et al, 2001). Upgradient points are south of the NLPP location (currently General Chemical) and downgradient locations are to the north. Background locations are those samples taken immediately at the inputs of the two rivers that flow into Manistee Lake, the Manistee River in the north and the Little Manistee in the south. All numerical values are in mg/kg except Mercury and ModMercury (µg/kg or ppb) and %TOC.

Table 2 repeats the locations information columns from Table 1, additional contaminants, potential sources and source types, two laboratory studies of species (*Hyalella Azteca* and *Chironomus Tentans*) exposed to collected surface sediment samples and counts of native species and organism totals in these samples (Rediske et al., 2001). Information on the performance of these studies and the test organisms *H. Azteca* and *C. Tentans* is provided in Rediske et al. (2001). For the purposes of the EDA, the "Mean Survival" values were recoded to be % Mortality for these two species. The "Species #" and "Organisms (Total)" values for each location were obtained from Table 4.8.1 in Rediske et al. (2001)[2]. It should be noted that with regard to location and depth characteristics the organism counts and testing data are only available for the Ponar-collected samples (M-1P through M-14P) and are therefore all "surface-sample" data. Because of this the EDA and statistical comparisons on the organism studies relate only to data from these samples and not to the data from the core samples that were collected. A caveat here is that no data were presented for chloride analysis of the Ponar samples. Because of this, along with awareness that the Ponar samples were collected very near the core samples and that chloride was potentially important, chloride values for the core "top" level samples were

[2] Note: There appears to be some discrepancies between Table 4.8.1 and Figure 4.8.1 in Rediske et al. (2001) with regard to the numbers of taxa counted (referred to as species in the EDA). Table 4.8.1 values were used for the EDA.

carried into the EDA. For example, the chloride value for M-1 Top was used as the chloride value for M-1 P. Due to this transposed usage, the column of chloride data was labeled as "ApproxCl$^-$" (Table 2).

The treatment within the EDA of values reported as less than their analytical reporting limits (< values) warrants mention. With the exception of mercury there were only a few values tabulated as less than reporting limits. These types of values have always been problematic in statistical analyses and there is still controversy about how best to manage them. The choice for this EDA was to err on the side of conservatism and use the reporting limit as the value for those sample data (rather than have "missing values" in the analyses or making some arbitrary decision about using half the reporting limit value or some other equally unsupportable formula). The maximum use of this approach occurred with the mercury data, which were reported at µg/kg (ppb) levels. Because of this being the "worst case scenario" Table 1 presents both the raw mercury data and the data modified to remove the "less than" values (ModMercury). Because the other contaminant results had so few of these reporting limit values they are not displayed in Tables 1 or 2 but can be observed in the Rediske et al. (2001) report if desired.

A variety of statistical exploration techniques were implemented to graphically display and quantify the data. DataDesk6.2 is not a traditional statistics application, although it uses statistical methodology at its core and generates the appropriate numerics associated with the particular analysis. It is designed to allow rapid comparison of multiple variables with graphical display so trends and patterns can be observed. Aabel 1.5.7 functions in a somewhat similar manner but with some differing capabilities. This report will not attempt to explain in any detail the statistics being used beyond what is needed to support the observations that result in some of the conclusions. Details of these methods can be found elsewhere should interest be sufficient.[3]

3 A) Velleman, P. F. (1997). DataDesk Version 6.0, Handbook, Volumes 2 and 3. Ithaca, N. Y., Data Description, Inc. B) J.W.Tukey, "Exploratory Data Analysis", 1977, Addisson Wesley

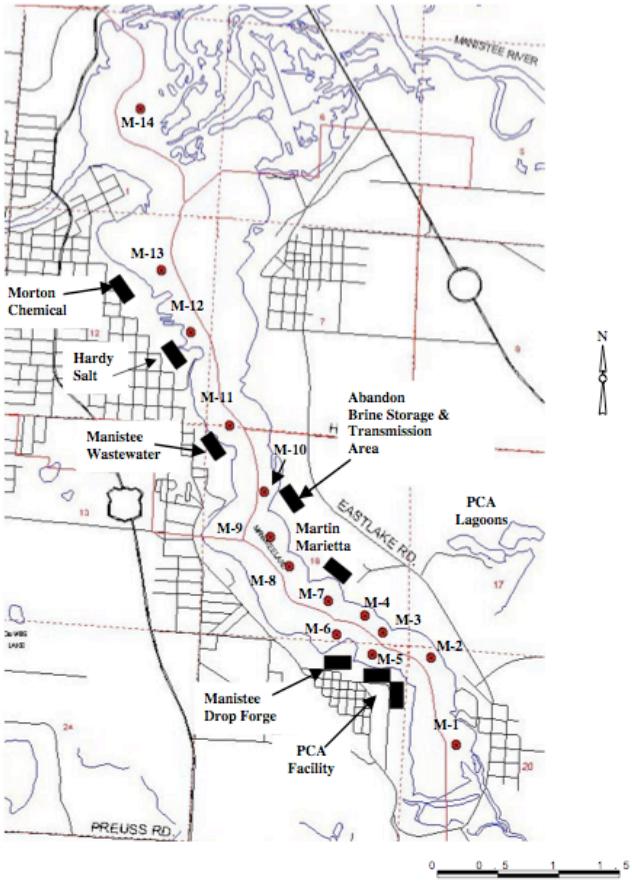

FIGURE 2.1 MANISTEE LAKE SAMPLING STATIONS (NOVEMBER 1998).

12

Figure 1 Manistee Lake Sediment Sampling Locations (Rediske et al., 2001)

Table 1 Data Used for EDA

Station	Location#	Depth Level	NLPP Relative Location	Barium	Selenium	Mercury	Mod Mercury	Arsenic	Cadmium	Chromium	Copper	Lead	Nickel	Zinc	% TOC
M-1 Top	1	T	Background	51	0.52	48	48	2.3	0.78	25	20	23	8.4	76	5.3
M-1-Mid	1	M	Background	62	0.5	27	27	0.24	0.47	20	27	16	9.8	53	7.4
M-1 Bot	1	B	Background	72	1.1	<25	25	0.33	0.54	30	12	4.8	9.5	53	16
M-2 Top	2	T	Upgradient	120	0.33	45	45	9.2	1.7	44	53	78	20	200	12
M-2 Mid	2	M	Upgradient	150	0.3	22	22	11	3.8	110	120	160	21	300	12
M-2 Bot	2	B	Upgradient	94	0.62	<25	25	8.6	0.74	78	18	15	14	64	14
M-3 Top	3	T	Upgradient	100	0.46	<25	25	8.4	0.85	37	29	24	16	92	5.2
M-3 Mid	3	M	Upgradient	120	0.73	<25	25	8.4	0.47	39	16	8.5	17	59	10
M-3 Bot	3	B	Upgradient	120	0.76	<25	25	7	0.49	41	16	7	18	60	11
M-4 Top	4	T	Upgradient	110	0.79	<25	25	6.7	0.41	40	17	73	19	60	9.4
M-4 Mid	4	M	Upgradient	110	0.71	<25	25	6.5	0.43	36	16	7.6	18	58	11
M-4 Bot	4	B	Upgradient	130	0.71	<25	25	6.3	0.47	35	16	8.1	20	210	8.4
M-5 Top	5	T	Upgradient	110	0.35	<25	25	2.2	2.5	72	75	88	22	60	12
M-5 Mid	5	M	Upgradient	100	0.7	123	123	7.3	0.5	34	16	10	18	57	12
M-5 Bot	5	B	Upgradient	120	0.91	<25	25	6.9	0.52	36	16	8.2	19	110	11
M-6 Top	6	T	Upgradient	93	0.44	27	27	8.1	1.8	56	30	26	23	57	10
M-6 Mid	6	M	Upgradient	110	0.72	<25	25	8.2	0.45	36	15	8.4	20	56	9.1
M-6 Bot	6	B	Upgradient	120	0.74	<25	25	6.9	0.5	34	17	7.8	21	56	11
M-7 Top	7	T	Upgradient	110	0.22	48	48	9.6	2.3	100	60	64	24	170	10
M-7 Mid	7	M	Upgradient	95	0.6	<25	25	5.4	0.63	33	17	12	22	67	9
M-7 Bot	7	B	Upgradient	120	0.68	<25	25	7.6	0.42	37	16	8.5	24	60	8.8
M-8 Top	8	T	Upgradient	110	0.36	95	95	17	2.6	130	100	91	26	230	4.8
M-8 Mid	8	M	Upgradient	110	0.52	<25	25	8.9	0.61	50	21	16	21	79	7.5
M-8 Bot	8	B	Upgradient	120	0.6	<25	25	7.8	0.35	39	16	8.2	19	61	7.5
M-9 Top	9	T	Upgradient	110	0.43	62	62	3	3.4	140	100	83	29	230	6.5
M-9 Mid	9	M	Upgradient	110	0.46	<25	25	7.7	0.37	39	19	12	21	71	2.9

Station	Location#	Depth Level	NLPP Relative Location	Barium	Selenium	Mercury	Mod Mercury	Arsenic	Cadmium	Chromium	Copper	Lead	Nickel	Zinc	% TOC
M-9 Bot	9	B	Upgradient	130	0.51	<25	25	6.2	0.3	36	15	8	20	56	6.4
M-10 Top	10	T	Upgradient	120	0.44	55	55	15	2.5	85	120	87	34	330	7.7
M-10 Mid	10	M	Upgradient	100	0.3	<25	25	6.4	0.36	36	21	20	24	30	5.3
M-10 Bot	10	B	Upgradient	120	0.38	<25	25	7.6	0.36	39	17	9.8	23	64	5.7
M-11 Top	11	T	Downgradient	110	0.39	150	150	14	1.3	48	150	67	33	190	4.7
M-11 Mid	11	M	Downgradient	110	0.35	<25	25	6.3	0.31	33	21	15	22	66	4.4
M-11 Bot	11	B	Downgradient	110	0.42	<25	25	4	0.44	29	16	9.5	22	63	4.9
M-12 Top	12	T	Downgradient	110	0.33	53	53	9.4	1.1	40	98	81	30	200	5.8
M-12 Mid	12	M	Downgradient	320	0.2	152	152	17	1.4	44	140	85	29	240	4.7
M-12 Bot	12	B	Downgradient	67	0.2	27	27	3.7	0.22	20	16	15	14	56	1.7
M-13 Top	13	T	Downgradient	88	0.29	48	48	11	0.82	35	180	58	35	150	6.2
M-13 Mid	13	M	Downgradient	94	0.21	188	188	9.4	0.57	34	84	30	24	120	4.6
M-13 Bot	13	B	Downgradient	96	0.25	<25	25	5.2	0.23	28	18	13	23	58	3.1
M-14 Top	14	T	Background	46	0.23	<25	25	2.1	0.14	8.6	7.1	6.1	7	20	2.5
M-14 Mid	14	M	Background	25	0.2	<25	25	1.6	0.16	6.8	5.7	5.8	8	15	1
M-14 Bot	14	B	Background	63	0.22	27	27	3.5	0.34	20	16	20	16	51	3.8
M-1 P	1	S	Background	8	0.2	29	29	0.63	0.05	2	2	1.5	4	4	1
M-2 P	2	S	Upgradient	110	0.65	39	39	9.1	1.7	38	45	54	18	160	9.3
M-3 P	3	S	Upgradient	110	0.62	33	33	10	2.6	38	49	54	19	160	8.8
M-4 P	4	S	Upgradient	120	0.58	39	39	9.9	1.4	36	42	43	17	130	13
M-5 P	5	S	Upgradient	110	0.51	230	230	9.1	3.1	38	72	85	16	190	15
M-6 P	6	S	Upgradient	84	0.52	44	44	13	3.1	68	71	71	19	160	13
M-7 P	7	S	Upgradient	83	1.2	<25	25	9.4	3.2	87	42	38	16	150	11
M-8 P	8	S	Upgradient	110	0.5	50	50	12	2.6	43	64	63	24	170	7.6
M-9 P	9	S	Upgradient	120	0.49	36	36	10	1.6	46	81	69	25	180	7.5
M-10 P	10	S	Upgradient	120	0.58	58	58	15	1.1	40	100	66	28	200	6.5
M-11 P	11	S	Downgradient	110	0.49	89	89	12	1.3	35	140	77	30	190	8.1
M-12 P	12	S	Downgradient	110	1.5	86	86	7.8	0.99	31	78	69	24	170	5.6

Station	Location#	Depth Level	NLPP Relative Location	Barium	Selenium	Mercury	ModMercury	Arsenic	Cadmium	Chromium	Copper	Lead	Nickel	Zinc	% TOC
M-13 P	13	S	Downgradient	120	0.72	52	52	7.9	0.82	34	95	56	34	150	4.7
M-14 P	14	S	Background	38	0.2	<25	25	2.7	0.18	12	9.6	8.9	9.6	25	2.3

All data are in mg/kg except Mercury and ModMercury (µg/kg or ppb) and %TOC.

Table 2 Additional Data for the EDA

Station	Location#	Depth Level	Rediske Potential Source	Source Type	Hexane Ext	PAH	Resin Acids	ApproxCl-	*H. Azteca* % Mortality	*C. Tentans* % Mortality	Species #	Organisms (total)
M-1 Top	1	T	Control	Control	130	0.33	3	16				
M-1-Mid	1	M	Control	Control			2	16				
M-1 Bot	1	B	Control	Control			1	25				
M-2 Top	2	T	PCA	Chemical	2800	8	8	120				
M-2 Mid	2	M	PCA	Chemical			11	180				
M-2 Bot	2	B	PCA	Chemical			4	180				
M-3 Top	3	T	PCA	Chemical	2300	1.95	10	190				
M-3 Mid	3	M	PCA	Chemical			6	260				
M-3 Bot	3	B	PCA	Chemical			2	300				
M-4 Top	4	T	PCA	Chemical	1200	0.33	12	210				
M-4 Mid	4	M	PCA	Chemical			7	250				
M-4 Bot	4	B	PCA	Chemical			2	320				
M-5 Top	5	T	PCA	Chemical	2900	1.79	11	172				
M-5 Mid	5	M	PCA	Chemical			18	250				
M-5 Bot	5	B	PCA	Chemical			3	300				
M-6 Top	6	T	MDF	Chemical	15000	2.41	13	270				
M-6 Mid	6	M	MDF	Chemical			7	430				
M-6 Bot	6	B	MDF	Chemical			2	530				
M-7 Top	7	T	PCA+MM	Chemical	6400	5.57	9	300				

Station	Location#	Depth Level	Rediske Potential Source	Source Type	Hexane Ext	PAH	Resin Acids	ApproxCl-	*H. Azteca* % Mortality	*C. Tentans* % Mortality	Species #	Organisms (total)
M-7 Mid	7	M	PCA+MM	Chemical			6	460				
M-7 Bot	7	B	PCA+MM	Chemical			3	640				
M-8 Top	8	T	PCA	Chemical	5700	5.81	9	100				
M-8 Mid	8	M	PCA	Chemical			4	260				
M-8 Bot	8	B	PCA	Chemical			3	380				
M-9 Top	9	T	PCA	Chemical	6700	6.57	5	160				
M-9 Mid	9	M	PCA	Chemical			3	380				
M-9 Bot	9	B	PCA	Chemical			1	390				
M-10 Top	10	T	ABW+P	Salt	2900	16.13	5	2500				
M-10 Mid	10	M	ABW+P	Salt			3	470				
M-10 Bot	10	B	ABW+P	Salt			3	550				
M-11 Top	11	T	MWTP+H	Salt	6500	11.71	8	98				
M-11 Mid	11	M	MWTP+H	Salt			4	230				
M-11 Bot	11	B	MWTP+H	Salt			1	360				
M-12 Top	12	T	Hardy	Salt	5400	13.53	7	980				
M-12 Mid	12	M	Hardy	Salt			3	2400				
M-12 Bot	12	B	Hardy	Salt			2	3500				
M-13 Top	13	T	Morton	Salt	9800	4.15	5	96				
M-13 Mid	13	M	Morton	Salt			4	260				
M-13 Bot	13	B	Morton	Salt			2	350				
M-14 Top	14	T	Control	Control	90	0.33	5	19				
M-14 Mid	14	M	Control	Control			4	69				
M-14 Bot	14	B	Control	Control			1	44				
M-1 P	1	S	Control	Control	100	0.33	4	16	11.3	5	23	5614
M-2 P	2	S	PCA	Chemical	1900	3.63	10	120	30	11.3	14	2870
M-3 P	3	S	PCA	Chemical	3200	4.81	9	190	30	10	12	2807
M-4 P	4	S	PCA	Chemical	2600	3.01	8	210	28.8	10	8	1127
M-5 P	5	S	PCA	Chemical	4300	4.8	10	172	53.8	6.3	14	798

Station	Location#	Depth Level	Rediske Potential Source	Source Type	Hexane Ext	PAH	Resin Acids	ApproxCl-	*H. Azteca* % Mortality	*C. Tentans* % Mortality	Species #	Organisms (total)
M-6 P	6	S	MDF	Chemical	26000	17.58	11	270	56.3	27.5	8	329
M-7 P	7	S	PCA+MM	Chemical	4000	11.89	7	300	36.3	12.5	10	511
M-8 P	8	S	PCA	Chemical	8800	9.61	7	100	41.3	7.5	7	392
M-9 P	9	S	PCA	Chemical	3300	8.8	6	160	27.5	12	10	1763
M-10 P	10	S	ABW+P	Salt	6600	15.01	7	2500	43.8	11.3	6	373
M-11 P	11	S	MWTP+H	Salt	8300	15.24	6	98	40	6.3	11	3311
M-12 P	12	S	Hardy	Salt	7200	12.01	6	980	20	8.8	7	1835
M-13 P	13	S	Morton	Salt	12400	29.37	11	96	40	27.5	6	1204
M-14 P	14	S	Control	Control	50	0.33	3	19	16.3	5	20	3770

Part 1: Exploratory Data Analyses of Contaminants

Constituents vs. Depth Level

The first analysis performed was to determine whether any of the constituents that were determined in the sediments (Table 1) were preferentially found at any particular depth(s) in the sediments. This was done by creating both dotplots and box and whisker plots of each of the constituents versus depth level (S, T, M, or B). Because box and whisker plots are not commonly used by environmental scientists but are heavily relied upon in this report, an explanation of their interpretation has been included in Appendix A of this report.

Figure 2 displays 12 boxplots of the COCs versus the depth at which the samples were acquired. Barium seems to be an example of reasonable consistency in concentration with depth, possibly indicating that a large fraction of Ba has a natural occurrence in minerals in this area. Se, Cr, Ni, and %TOC also seem to be relatively evenly distributed throughout the sampling depth but these data exhibit a much broader overall range of measured concentrations. With the possible exception of %TOC it is unlikely that the bulk of these four constituents is of natural origin in this environment. Their fairly ubiquitous distribution throughout the sediment column might result from downward transport in the sediments, older releases, or both.

Zn, Pb, Cd, and Cu strongly correlate to surface samples (S) and top sediment cores (T), indicating a likely, and possibly recent, anthropogenic origin. To a lesser extent As shows a tendency to be found in the uppermost sediments; the medians are higher in the arsenic S and T samples but there is some overlap of the 95% confidence intervals with the B and M samples. Although not clearly indicated from the plot, Hg is also a surface contaminant, the lower core sections are not well depicted in the Hg plot because they are mostly <25 µg/kg values, which results in a null value in the statistical analysis. Of 56 possible sampled values for Hg, only 29 exceeded the <25 µg/kg value; the next lowest was Se, with 51 of 56 possible values available. Of these 29 values for Hg, only two fell into the B (bottom core) group and only five into the M (middle core) group. Because of this the 95% confidence intervals are non-existent for the B group (not calculable for only 2 samples) and extremely broad for the M group (only 5 samples). Using a conservative approach to test whether there is a difference between the S or T groups and the B and M groups with 95% confidence, a new variable was created (ModMercury) and a value of 25 µg/kg was assigned to all samples that were designated as <25 µg/kg, eliminating the null value for these samples. Figure 3 shows that in fact the surface and top core samples are higher in Hg concentration than the bottom and middle core sections with >95% confidence, even though the middle core section does have some significant Hg contamination at sample locations 5, 12 and 13 (shown as location-labeled statistical outliers in Figure 3).

These plots have little relevance for the hexane extractable materials because data were only provided in the Rediske (2001) report for the S and T samples.

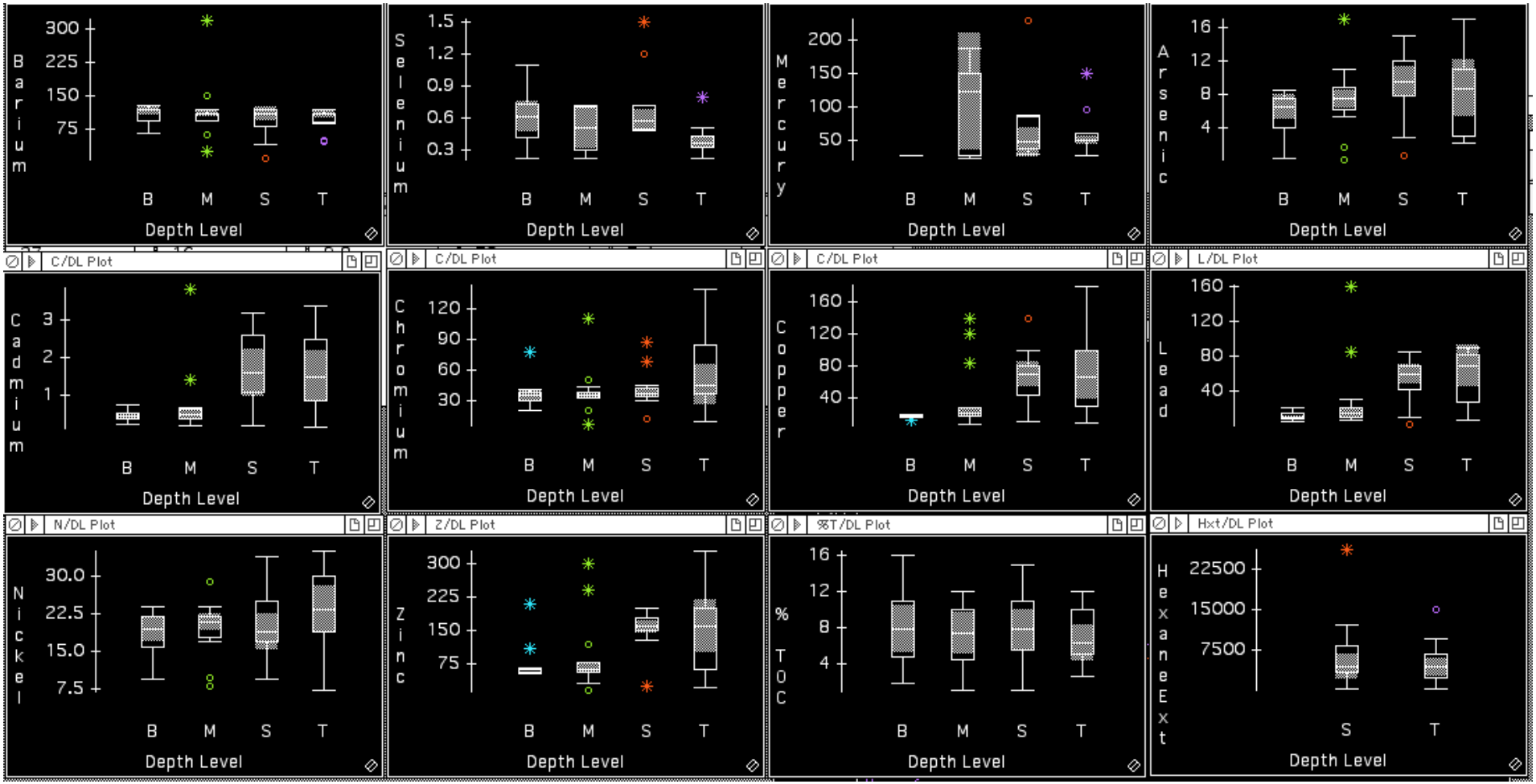

Figure 2 Box and Whisker Plots of COCs versus Depth

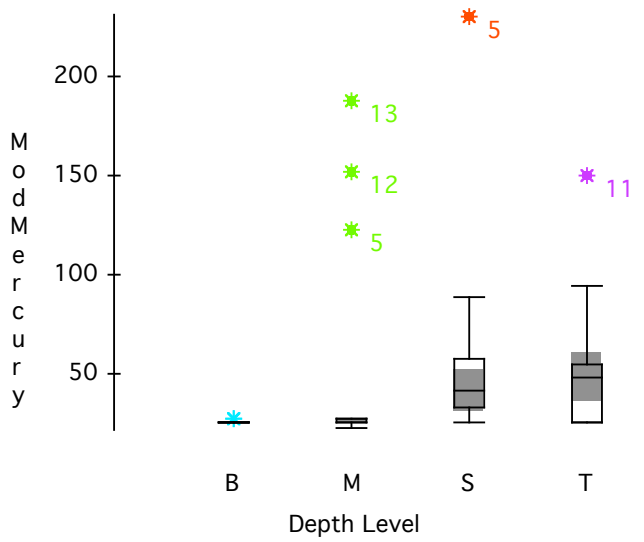

Figure 3 Box and Whiskers Plot of ModMercury versus Depth. Outliers Labeled By Location.

Constituents Versus Location

A) Relative to the NLPP Proposed Location

Figure 4 illustrates the sediment constituents of concern plotted versus background, upgradient, or downgradient locations relative to the proposed location for the Northern Lights Power Plant (currently General Chemical). Overall these plots do not show much in the way of distinguishing the upgradient and downgradient categories, even though the background samples are significantly lower in concentration for most of the COCs. This lack of distinction is probably because the upgradient and downgradient locations are artificially designated with respect to the location of the proposed NLPP, i.e., the proposed plant location currently creates no impact that is sufficiently strong at that single point to statistically distinguish downgradient impacts from the total of those upgradient. This does not mean that there is no impact from the old General Chemical site, it simply means that impacts at that point are not sufficient to statistically overwhelm those resulting from other industrial impacts further to the south, both historical and current. This information and approach might prove useful as a present-day baseline for evaluation of future contaminant additions, should the NLPP or other industrial activity be implemented at this location.

B) Relative to the Sampling Location Number

Figure 5 presents scatterplots with regression lines of the COC concentrations versus the location numbers from Table 1 (Location#). This generally follows the flowpath of the Lake. These scatterplots include concentrations at all depths (S, T, M, B) but exclude location 14 that was chosen at the mouth of the Manistee River as a background or

control (Rediske et al, 2001). Inclusion of Location 14 sediment samples that are not in the flowpath from the Little Manistee River past the industries and into Lake Michigan, and were deliberately selected as background controls, would have biased the scatterplots inappropriately to lower values at the higher (more northerly) locations.

Figure 5 displays the data on the scatterplots in numerous ways, in addition to visualizing whether there is a linear relationship between the concentrations and the location number. Background values are shown in the plot as circles, upgradient values (relative to NLPP) are short vertical lines and downgradient values are displayed as x's. The colors also help describe the data with blue being the B (bottom) core depth, green the M (middle) core depth, purple the T (top) core depth and red the S (surface Ponar) samples.

A fairly complex mixture of behaviors is evident in Figure 5 with the regression lines showing some constituents staying about the same or even decreasing (e.g., Ba, Se) as the sampled sediments proceed up the length of the lake while others show a pronounced increase (e.g., Ni, Hexane Extractables). Others (e.g., Cu, Zn) show that another modification to the plots might be needed and informative. We know from the previous analysis of constituents versus depth level that certain of the constituents were preferentially sequestered in the S and T sediment depths. Plots such as the scatterplot for Cu show all the bottom (B) core samples, in blue, hovering at very low concentration values near the x-axis while the uppermost sediment samples, in red and purple, are much higher in the plot and appear to be increasing in concentration with location. The regression line, when using all these points for the analysis, is biased low for such plots due to the fact that the COCs in these sediments are virtually all in the S and T sediments yet the very low B and M values are being included in the regression analysis.

Figure 6 shows the same scatterplots but with the B and M values removed from the analyses. This should provide a more accurate description of the behavior of the COCs that are predominantly in the S and T samples. From Figure 2 these include Cd, Cu, Pb, and Zn; As may also be primarily in the S and T samples but with reduced certainty. Based on the results presented in Figures 5 and 6, fully eight of the 12 COCs evaluated are increasing in the sediments from the most upgradient (Location 1) to the most downgradient location (Location 13). Three of the 12 seem to increase then decrease, while one of the 12 remains about the same throughout. Table 3 distills the information on these COCs from Figures 2, 4, 5 and 6.

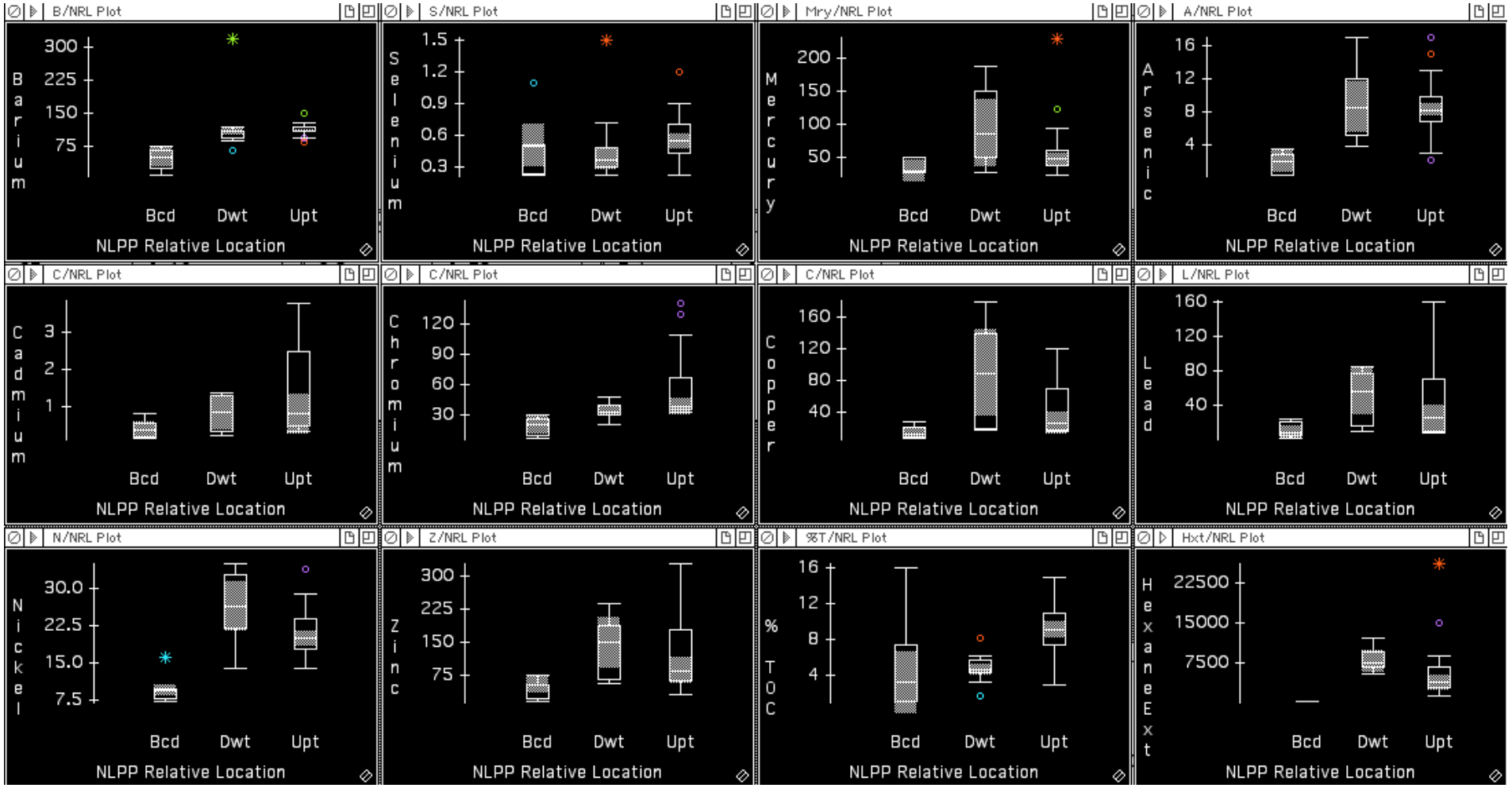

Figure 4 Box and Whisker Plots of COCs Relative to the General Chemical (NLPP Proposed) Location

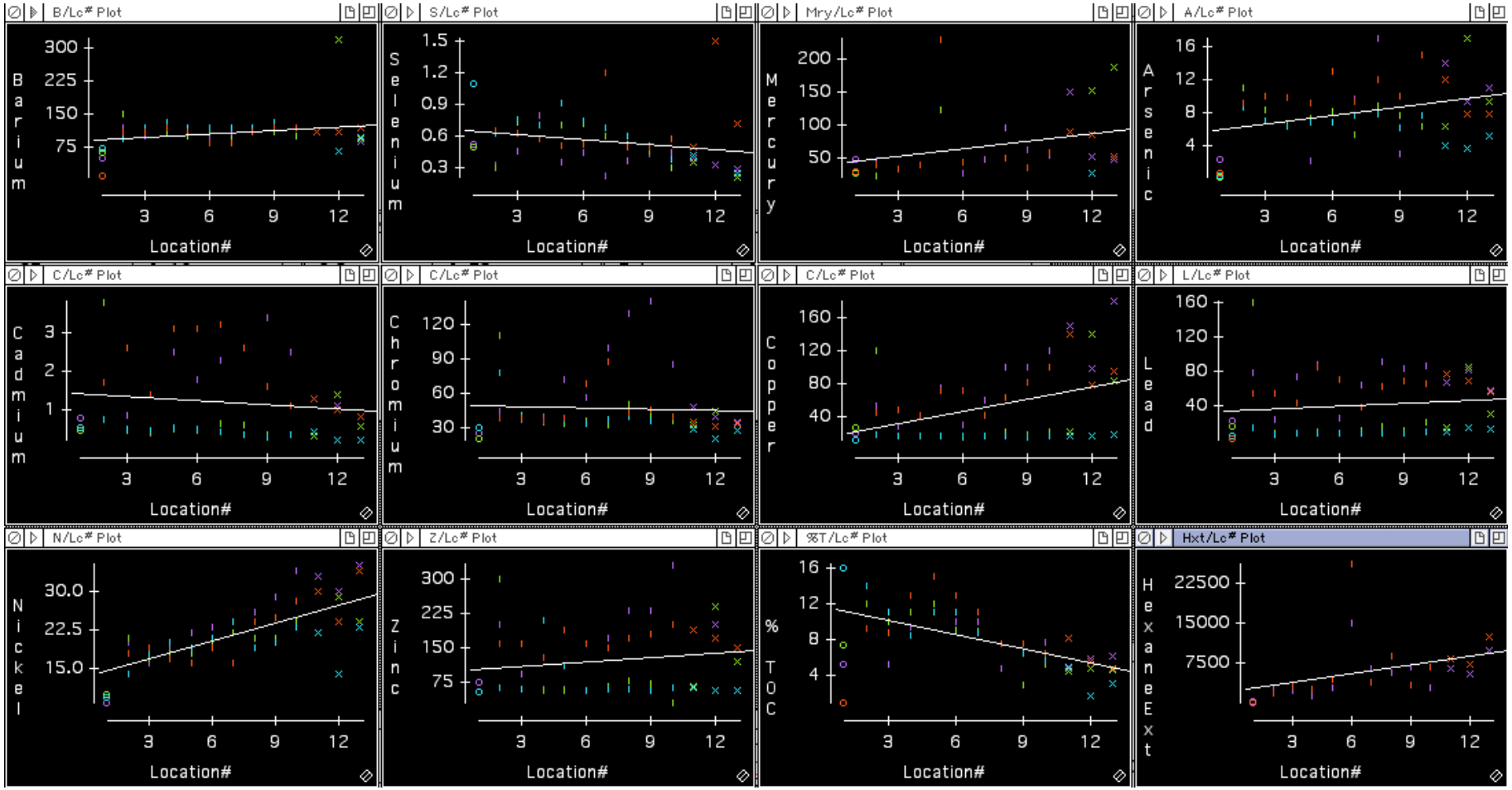

Figure 5 Scatterplots with Regression Line of COCs versus Location Number at All Sampled Depths

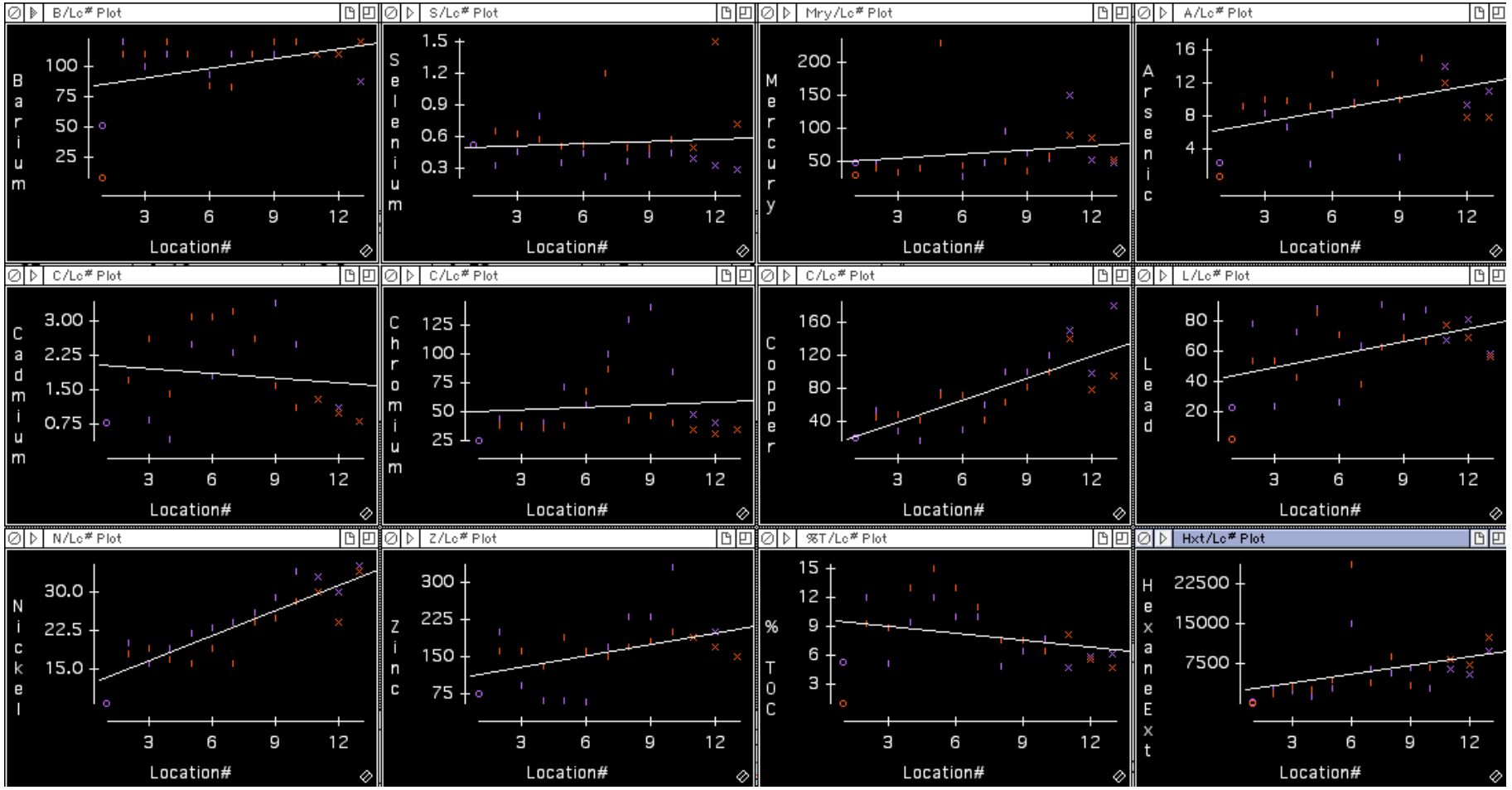

Figure 6 Scatterplots with Regression Line of COCs versus Location Number at S & T Depths

Table 3 Summary of COC Information from EDA and Statistical Analyses

	Statistical Differences*				
Sediment Constituent	**Depth Differences? (B, M, T, S)**	**Background Differs from NLPP UG?**	**Background Differs from NLPP DG?**	**UG & DG Differ?**	**Concentration Increases with Location#?****
Barium	No	Yes	Yes	No	Yes
Selenium	Approx. equal	No	No	No	No
ModMercury	Upper (S&T > B)	No	Yes	Yes	Yes
Arsenic	Upper (slight, S > B)	Yes	Yes	No	Yes
Cadmium	Upper (S&T > M&B)	Yes	Yes	No	No
Chromium	Approx. equal (T > B)	Yes	Yes	Yes	No
Copper	Upper (S&T > M&B)	Yes	Yes	Yes	Yes
Lead	Upper (S&T > M&B)	Yes	Yes	No	Yes
Nickel	No	Yes	Yes	No	Yes
Zinc	Upper (S&T > B)	Yes	Yes	No	Yes
%TOC	Approx. equal	Yes	No	Yes	No
Hexane Extractables	Unknown, S&T only sampled	No	Yes	No	Yes

UG = upgradient, DG = downgradient (from NLPP or General Chemical)
*Statistical differences are described based on the results of pooled t-tests with individual alpha = 0.05 (95% confidence).
**Response is from S&T scatterplot only (Figure 6) if the Depth Difference column designates "Upper." Otherwise the response is from the scatterplot using all four depths (Figure 5).

PEC Observations and Evaluation

The Rediske et al. (2001) report tabulates several of the metals that were analyzed and provides their highest measured concentrations in the Ponar samples (S samples in this EDA) while displaying the consensus-based PEC (mg/kg) for the metals. They conclude that none of the metals exceeded its PEC value and that there is probably no effect due to the metals. They then proceed to display another table showing that PAH compound concentrations do exceed PEC guidelines at 10 sampling locations. The EDA evaluations of the data show, however, that the Ponar (S) samples exhibit the highest metal concentrations in only one of the eight cases, i.e. mercury, and that the PEC values are indeed exceeded in some of the top core and middle core samples for other metals, notably chromium, copper and lead (Table 4). It should be noted that the top (T) core samples (0 to 20 inches depth of the total core, Rediske et al, 2001) incorporate the interval sampled by the Ponar device (0 to 6 inches depth) and appear to be as relevant for evaluating surface chemical conditions as the Ponar samples. The middle core section is the 20 to 40 inch below sediment surface interval.

Table 4 Metals Concentrations Relative to the PEC, exceedances in bold typeface

Metals	Highest Concentration Measured in Ponar (S) Samples, mg/kg	Consensus-Based PEC, mg/kg	Actual Maximum Measured Concentrations, mg/kg	Actual Maximum Sediment Level
Arsenic	15	33	17	T
Cadmium	3.1	4.98	3.8	M
Chromium	87	111	**140**	T
Copper	140	149	**180**	T
Lead	85	128	**160**	M
Mercury	0.23	1.06	0.23	S
Nickel	34	48.6	35	T
Zinc	200	459	330	T

Table 4 shows that chromium, copper and lead all exceed PEC values in Manistee Lake, with exceedances of chromium and copper occurring in the uppermost sediment layers.

Table 5 displays the PAH data from the Rediske et al. Report. The combination of PEC value exceedance by 3 metals, with several other metals approaching their PEC values, exceedances of PEC values by organic contaminants at 10 sampling locations, and the excessively saline conditions in some locations near the bottom of the lake indicate that Manistee Lake has been badly damaged by historical and current industrial activities. These results show that no additional industrial waste burdens should be allowed upon this lake and that every effort should be made to reduce ongoing industrial impacts.

Table 5 PAH Exceedances Relative to PEC Guidelines

PAH Compound	Consensus based PEC Guidelines (mg/kg)	Manistee Lake Stations that Exceed PEC Guidelines
Anthracene	0.85	None
Fluorene	0.54	None
Naphthalene	0.56	None
Phenanthrene	1.17	M-3P, M-5P, M-6P, M-7P, M-8P, M-9P, M-10P, M-11P, M-12P, M-13P
Benz[a]anthracene	1.05	M-6P, M-10P, M-11P, M-13P
Benzo(a)pyrene	1.45	M-11P, M-13P
Chrysene	1.29	M-6P, M-7P, M-10P, M-11P, M-12P, M-13P
Fluoranthene	2.23	M-6P, M-10P, M-11P, M-13P
Pyrene	1.53	M-6P, M-7P, M-8P, M-9P, M-10P, M-11P, M-12P, M-13P
Total PAHs	22.8	M-13P

Part 2: Exploratory Data Analyses of Biological Studies

Rediske et al. (2001) performed two types of biological studies, the results of which are used in this EDA. One of these consisted of counting organisms in the collected sediment samples at each location. A value was assigned to the total number of organisms counted for each location and the species were noted and counted. This resulted in two biological values for each surface sampling location (Table 2) that are being used in this EDA:

1. Organisms (total), and
2. Species #

The second type of biological study used here consisted of independent laboratory tests of acute toxicity, via exposure to the actual sediments collected at each location, for two types of organisms. The organisms were:

1. *Hyalella Azteca* (amphipod), and
2. *Chironomus Tentans* (midge)

Eight replicate tests were done with each organism for each location. For the EDA the final average number surviving at each location was converted to a percent mortality value (Table 2).

Biological Studies Data Assessment

The four sets of biological study data were briefly evaluated, in particular with respect to whether the test results were normally distributed. This assesses (to some extent) whether certain of the subsequent data manipulations are appropriate for the datasets (many statistics assume a normal or near-normal data distribution). However, there are not very many data and it should be noted that these data are known to result from what might be considered separate treatments (for example, some were exposed to contaminated sediments, some to control sediments) and the differing impacts of those treatments can skew the distributions. Determining whether there is any predictability in these impacts is the goal of the EDA. Normal probability plots of the data versus nscores (normal scores), regression of these data and correlation were used in the assessment of the results. A discussion of nscores and the interpretation of probability plots are beyond the scope of this report, but information can be found in various statistics texts and manuals[4]. In general, a near-normal distribution of a dataset versus its nscores will result in a straight line from the lower left to the upper right in a plot. Deviations from normalcy tend to result in sigmoid shapes in the plots. High R^2, F value, and correlation tend to support the hypothesis of a normal data distribution.

H. Azteca Mortality

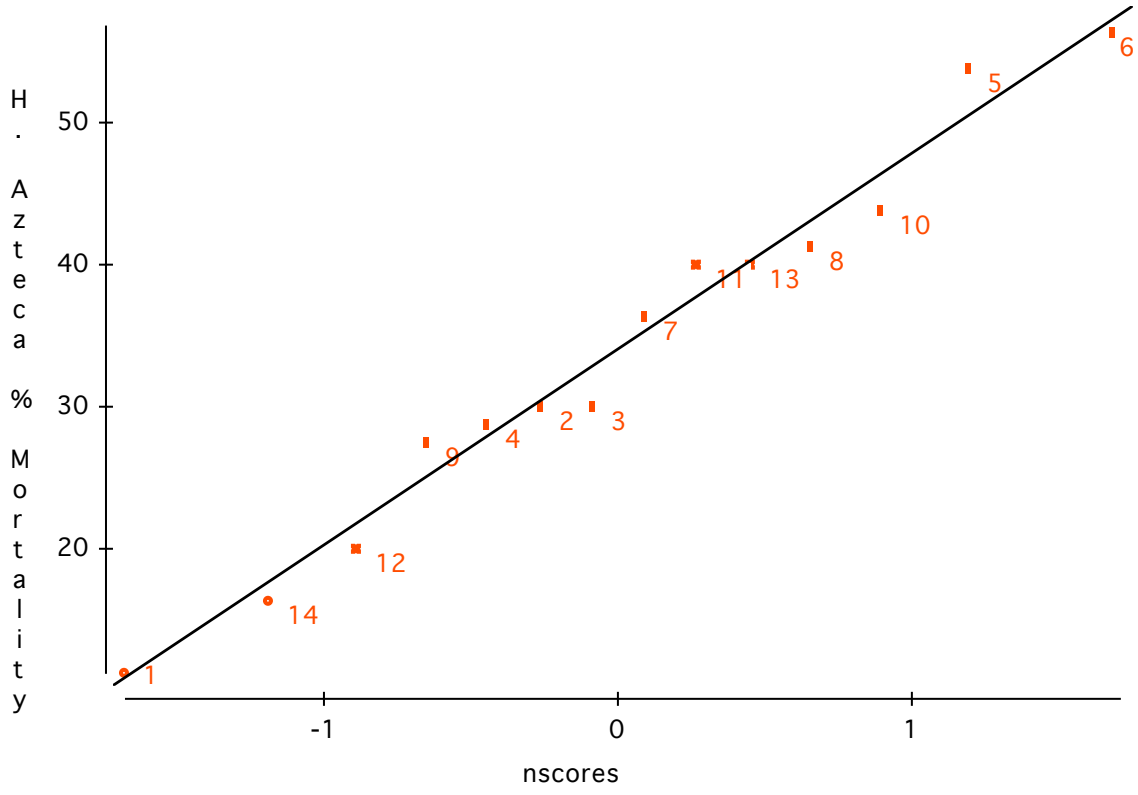

Figure 7 Probability Plot for *H. Azteca*. Labels are Location#.

[4] Ibid

Table 6 Regression of *H. Azteca* and nscores

Dependent variable is: H. Azteca % Mortality
No Selector
56 total cases of which 42 are missing
R squared = 97.9% R squared (adjusted) = 97.7%
s = 2.001 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	2193.76	1	2193.76	548
Residual	48.0343	12	4.00286	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	33.9571	0.5347	63.5	≤ 0.0001
nscores	13.8040	0.5897	23.4	≤ 0.0001

Table 7 Correlation of *H. Azteca* and nscores

Pearson Product-Moment Correlation

No Selector

	H. A...	nsco...
H. Azteca...	1.000	
nscores	0.989	1.000

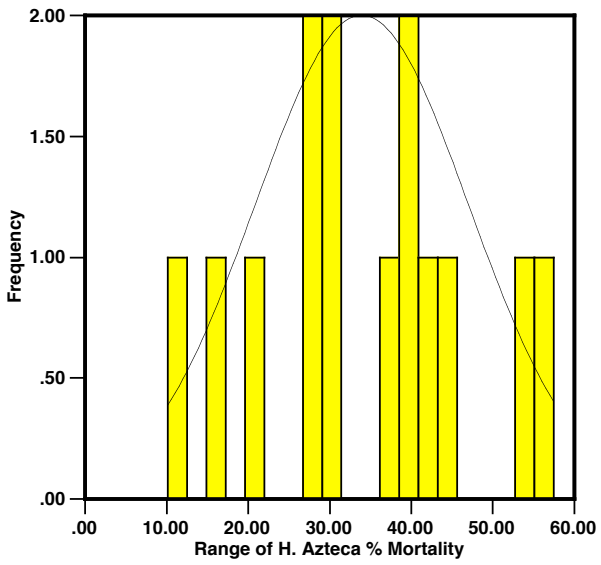

Figure 8 Histogram and Normal Curve for *H. Azteca*

The normal probability plot for *H. Azteca* % Mortality (Figure 7) indicates a normal distribution of these data, with all points falling very near the regression line and showing virtually no "S-shape" that would indicate the data are more long-tailed (right and left parts would point up and down), short-tailed (right and left parts pointing right and left), or skewed (only one side of the plot would bend away from the line). The high regression R^2 and F-ratio and, in particular, the 0.989 correlation support the normal distribution that is illustrated by the histogram and normal curve in Figure 8. This is the best fit of any of the biological study datasets.

C. Tentans

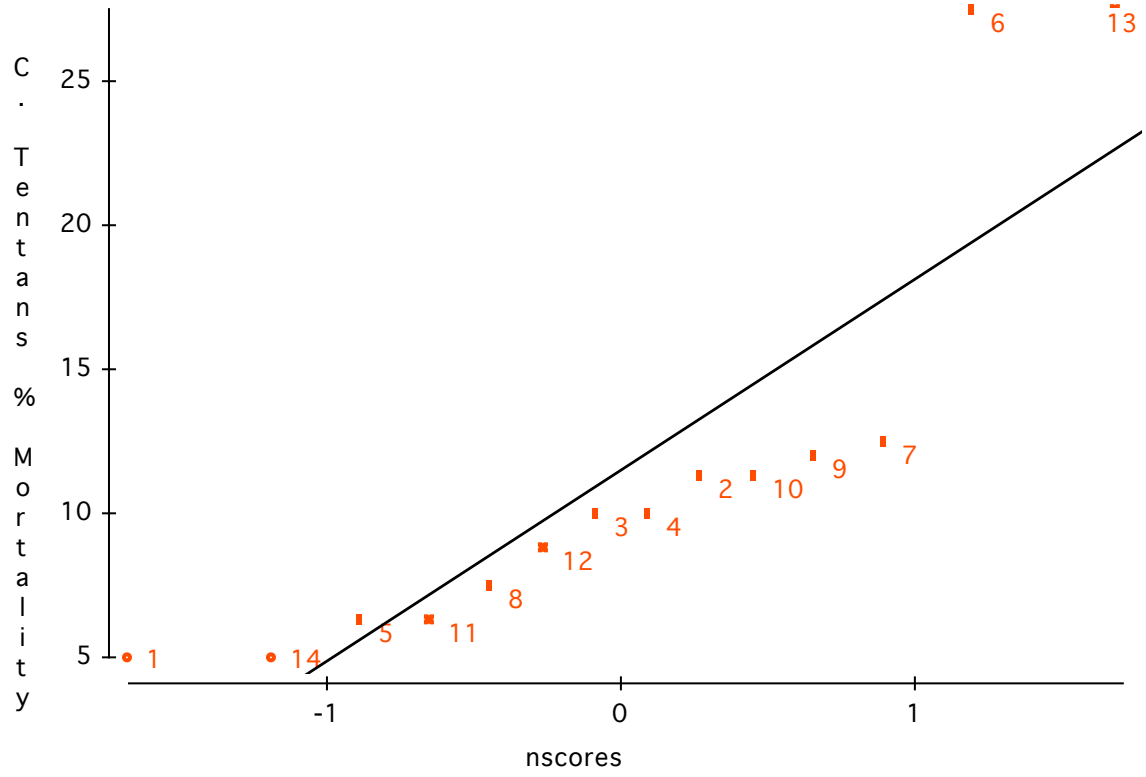

Figure 9 Probability Plot for *C. Tentans*. Labels are Location#.

Table 8 Regression of *C. Tentans* and nscores

Dependent variable is: C. Tentans % Mortality
No Selector
56 total cases of which 42 are missing
R squared = 74.4% R squared (adjusted) = 72.3%
s = 3.807 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	505.798	1	505.798	34.9
Residual	173.902	12	14.4918	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	11.5000	1.017	11.3	≤ 0.0001
nscores	6.62827	1.122	5.91	≤ 0.0001

Table 9 Correlation of *C. Tentans* and nscores

Pearson Product-Moment Correlation

No Selector

	C. T...	nsco...
C. Tentan...	1.000	
nscores	0.863	1.000

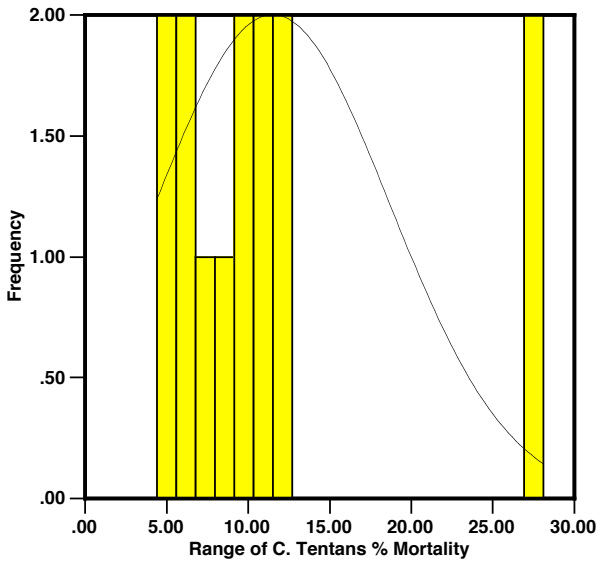

Figure 10 Histogram and Normal Curve for *C. Tentans*

The assumption of a normal distribution for *C. Tentans* % Mortality is not as well defended as it is for *H. Azteca*, because the high and low points on the probability plot (Figure 9) trend away from the line to the right and left (indicating a distribution that is more short-tailed than a fully normal distribution. It might be noted, however that there is a very large gap in the % mortality between two of the locations and the remainder of the data. These locations (6 and 13) might be outliers due to actual unaccounted probability (making them "apparent" outliers) or, more likely, these locations are subjected to some potential additional "effect" (in statistics the impact of a "predictor") that increased the mortality, or possible testing errors. Figure 10 shows the impact of these locations on the shape of the histogram and the normal curve. These issues will be addressed later in this report when exploring the causes of the mortality.

When locations 6 and 13 are removed from the plot the linear fit is greatly improved (the R^2 improves from 72.3% to 94.3% and the correlation improves from the still reasonable 0.863 (Table 9) to 0.974.

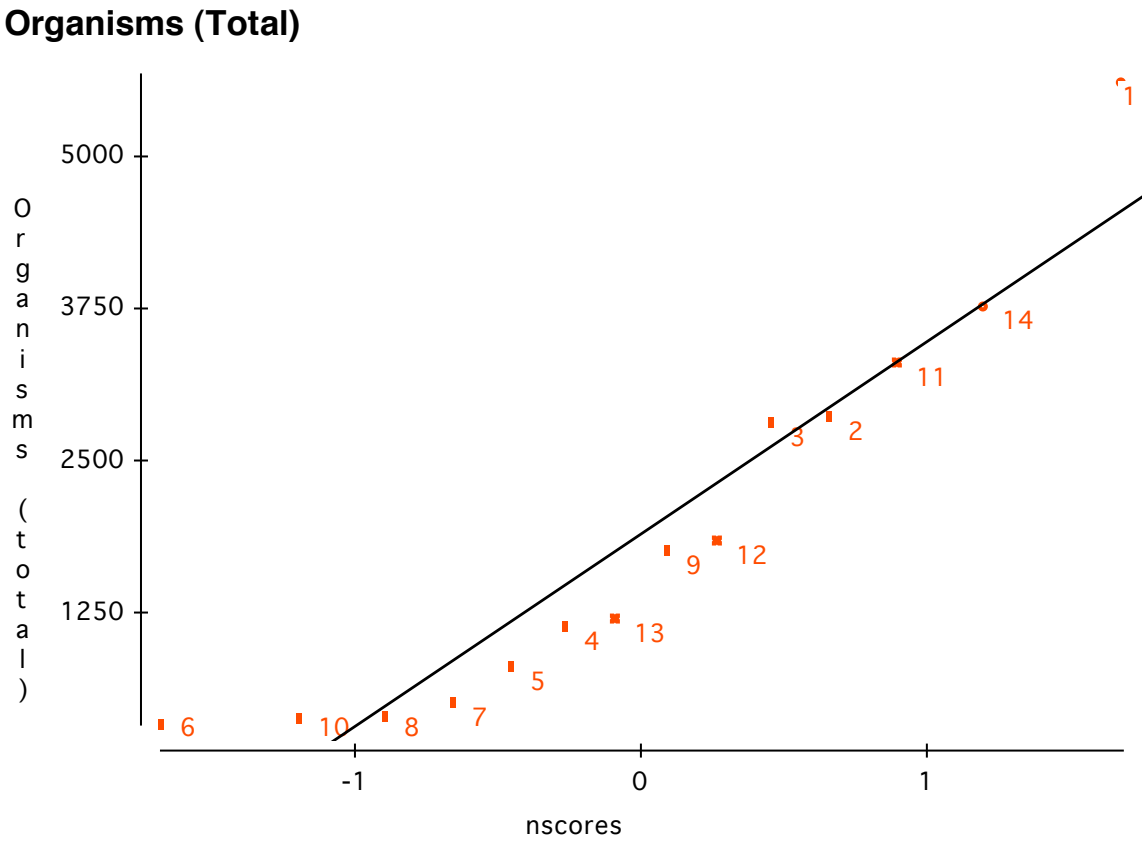

Figure 11 Probability Plot for Organisms (Total). Labels are Location#.

Table 10 Regression of Organisms (Total) and nscores

Dependent variable is: Organisms (total)
No Selector
56 total cases of which 42 are missing
R squared = 89.2% R squared (adjusted) = 88.3%
s = 539.8 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	28870274	1	28870274	99.1
Residual	3496618	12	291385	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1907.43	144.3	13.2	≤ 0.0001
nscores	1583.57	159.1	9.95	≤ 0.0001

Table 11 Correlation of Organisms (Total) and nscores

Pearson Product-Moment Correlation

No Selector

	Organis...	nscores
Organisms (tot...	1.000	
nscores	0.944	1.000

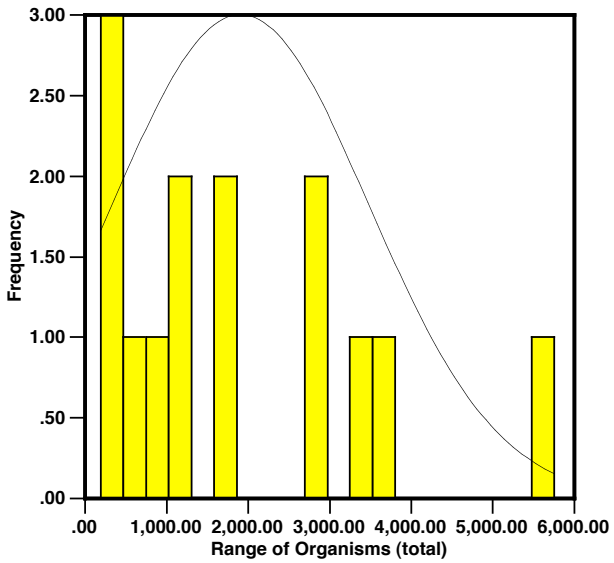

Figure 12 Histogram and Normal Curve for Organisms (total)

Of note in Figure 11, the probability plot for Organisms (Total), is that the dependent variable is reversed in the plot relative to the *H. Azteca* and *C. Tentans* plots. This is because in the acute toxicity tests higher percentage values represent a greater loss of organisms whereas in the count of organisms for Figure 11 (and for Species# in the next section) lower counts indicate a greater loss of organisms.

The distribution of Organisms (Total) appears to be relatively normal based on linearity of the variable versus nscores (Figure 11), which yield an R^2 of 88.3% (Table 10) and a correlation of 0.994 (Table 11). There are points (locations 1 and 6) that appear as possible outliers, however, with location 6 possibly being notable because it was also noticeable in the plot for *C. Tentans* (Figure 9).

Species

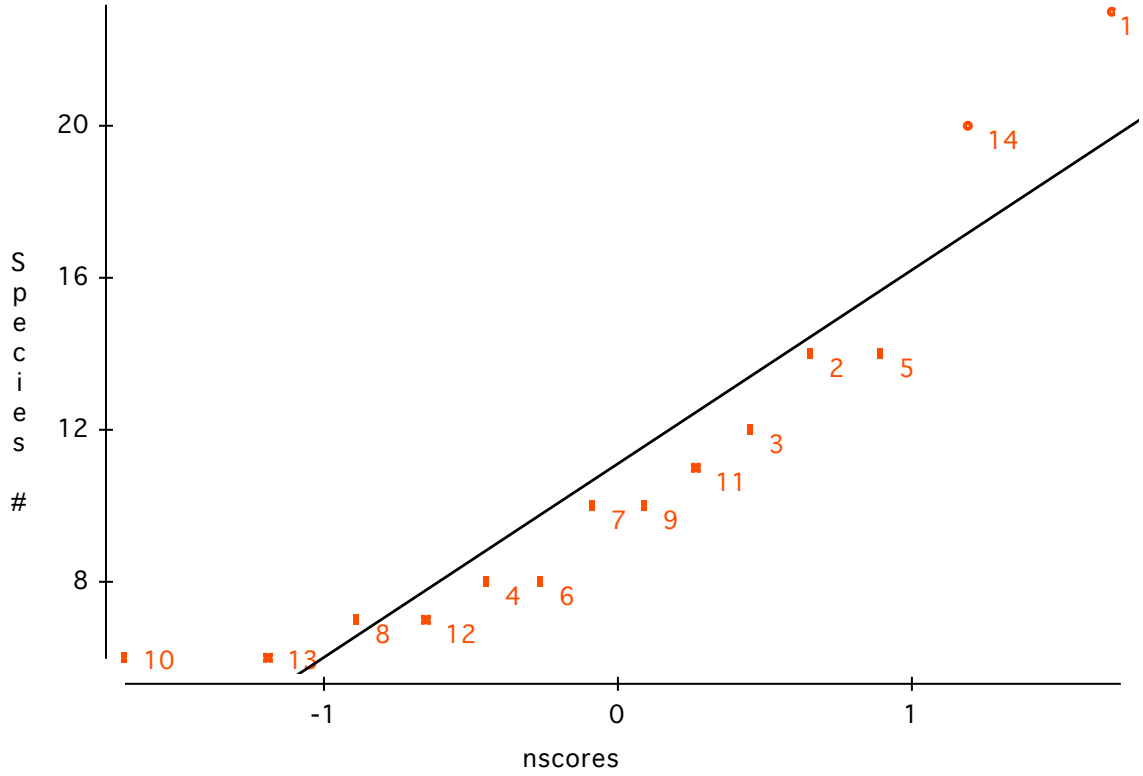

Figure 13 Probability Plot for Species #. Labels are Location#.

Table 12 Regression of Species # and nscores

Dependent variable is: Species #
No Selector
56 total cases of which 42 are missing
R squared = 86.6% R squared (adjusted) = 85.5%
s = 1.964 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	299.440	1	299.440	77.7
Residual	46.2745	12	3.85621	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	11.1429	0.5248	21.2	≤ 0.0001
nscores	5.09995	0.5788	8.81	≤ 0.0001

Table 13 Correlation of Species and nscores

Pearson Product-Moment Correlation

No Selector

	Spec...	nsco...
Species #	1.000	
nscores	0.931	1.000

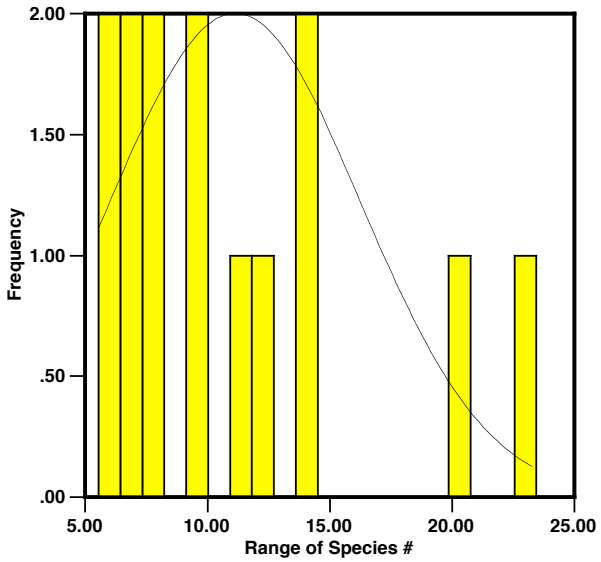

Figure 14 Histogram and Normal Curve for Species

The results for Species # appear from the statistics to be fairly normal, with a correlation of 0.931 and a few points on the probability plot that appear to be slight outliers due to differences in the treatments (i.e., differing environmental impacts). The histogram and normal curve in Figure 14 appear to skew to high values, because of the control samples having a high number of species while the other sampling locations, from within Lake Manistee, were greatly reduced in numbers of species.

Overall Organism Results Versus Sampling Location and Contaminant Loading

The Rediske et al. (2001) report creates a classification for the locations within Manistee Lake based on the potential sources near or upgradient from the sampling locations. From their report, these include:

Table 14 Sample Station Numbers Relative to Potential Sources

Station Number	Potential Source
M-1	Control at Little Manistee River Mouth
M-2 to M-5 and M-8 to M-9	PCA Superfund Site
M-7	PCA Superfund Site and Martin Marietta Chemical
M-6	Manistee Drop Forge
M-10	Abandoned Brine Wells and Pipeline
M-11	Manistee Wastewater Treatment Plant/Hardy Salt
M-12	Hardy Salt
M-13	Morton Chemical
M-14	Control at Manistee River Mouth

EDA was carried out to determine whether these categories were useful discriminators of organism mortality. Figure 15 provides an example box and whisker plot of the *H. Azteca* % Mortality results versus these source categories. There are essentially too few datapoints to be plotted versus so many categories, therefore these plots are not very revealing and those for the other organism studies are not presented.

Figure 16 attempts to reduce the number of categories to three from the nine displayed in Table 14 and Figure 15, based upon what might be considered the primary environmental impact at that location based on Rediske et al. (2001). These categories are control, chemical and salt (ignoring that salt is obviously a chemical, the chemical category refers to the other industrial chemicals such as those from the PCA site).

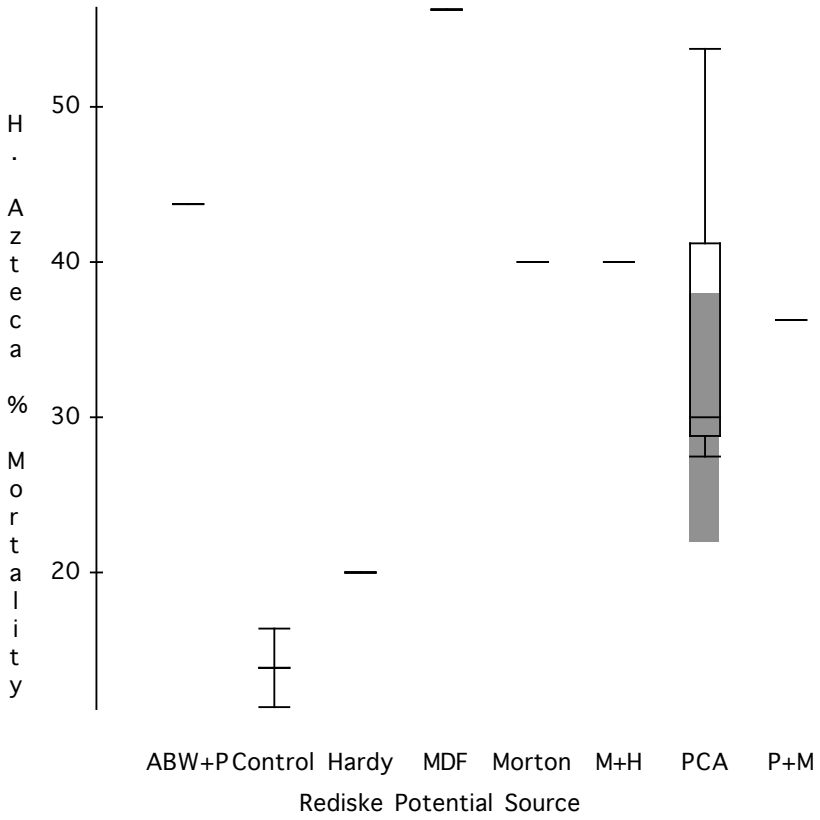

Figure 15 Boxplot of *H. Azteca* % Mortality versus the Rediske et al. (2001) Potential Source Categorization

Figure 16 again indicates that the control location mortalities are very different from the others but that within the lake itself it is difficult or impossible to discriminate based on these categorizations. This seems to indicate that Lake Manistee sediments are highly contaminated with respect to organism mortality by the first non-control sampling point and remain very toxic throughout its length.

Area charts of the four sets of biological study results are displayed versus sampling location in Figure 17. The charts for mortality to *H. Azteca* and *C. Tentans* are stacked whereas the magnitudes of the values for Species # and Organisms (Total) were too different for stacking without applying a data transformation. The acute mortality % values track one another very well and appear to be almost perfect counterparts to the organism and species counts, with increased % mortality corresponding very closely with reductions in the numbers and species of organisms natively present in the sediments. These plots again illustrate the extreme toxicity of the shallow Lake Manistee sediments just beyond the river mouths within the lake. Locations 1 and 14 were the control sample sediments and by the time one arrives at Locations 2 and 13, mortality/disappearance is worsening by a range of factors from 1.6 to 5.5.

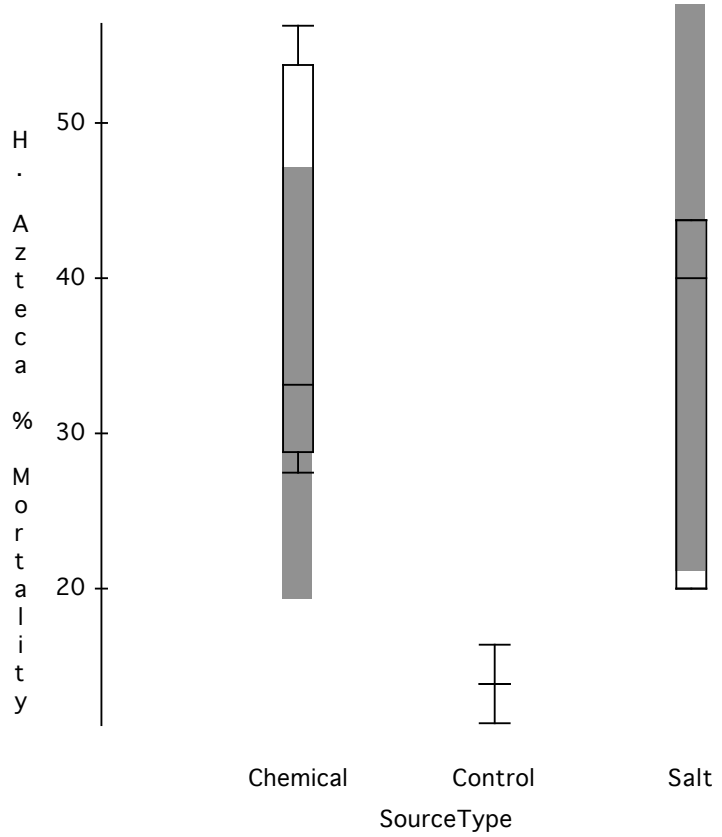

Figure 16 Boxplot of *H. Azteca* % Mortality versus a Source Type Categorization

Accounting for the specific environmental factors that contribute to the sudden increase in organism mortality upon entering the Lake environment is important for a variety of reasons, including:

1. The levels of these toxic chemicals need to be reduced to increase biodiversity in Lake Manistee
2. Identifying the primary contaminants of concern might also identify the sources, creating an opportunity to control their discharge to the lake
3. Predictions of residence time and fate in the sediments might be possible based on the contaminant chemistry and its geochemical milieu
4. Additional sampling with increased focus on those toxic parameters can be planned
5. In combination with other factors a conceptual model of the lakebed and its immediately overlying waters can begin to be developed.

A major goal of this report is to provide the underpinnings for accomplishing these five objectives. Figures 18 and 19 illustrate the complexity of understanding the specifics of toxicity to these organisms, however, by depicting stacked area charts of a variety of selected constituents. Virtually every measured constituent in these surface sediment

(Ponar) samples increases dramatically in concentration after leaving the river mouth locations (control Locations 1 and 14) and entering the lake (Locations 2 through 13). The % Mortality lines for *H. Azteca* and *C. Tentans* overlying these area graphs show the extremely close correspondence between the most highly contaminated zones and the acute toxicity of the sediments to these organisms. Cumulative, rather than individual, toxic impacts may be implied by these charts.

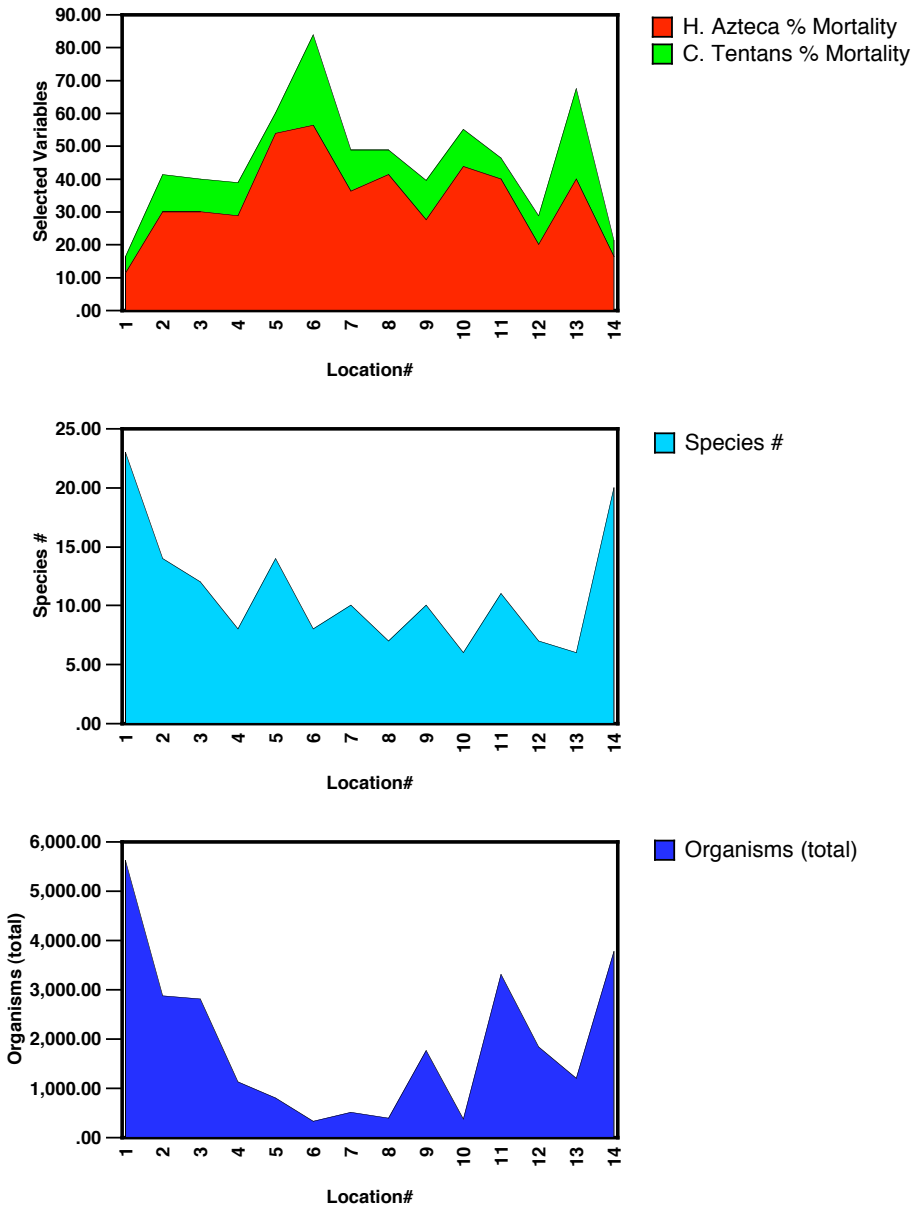

Figure 17 Area Charts of Organism Study Results Versus Location (*H. Azteca* and *C. Tentans* are stacked).

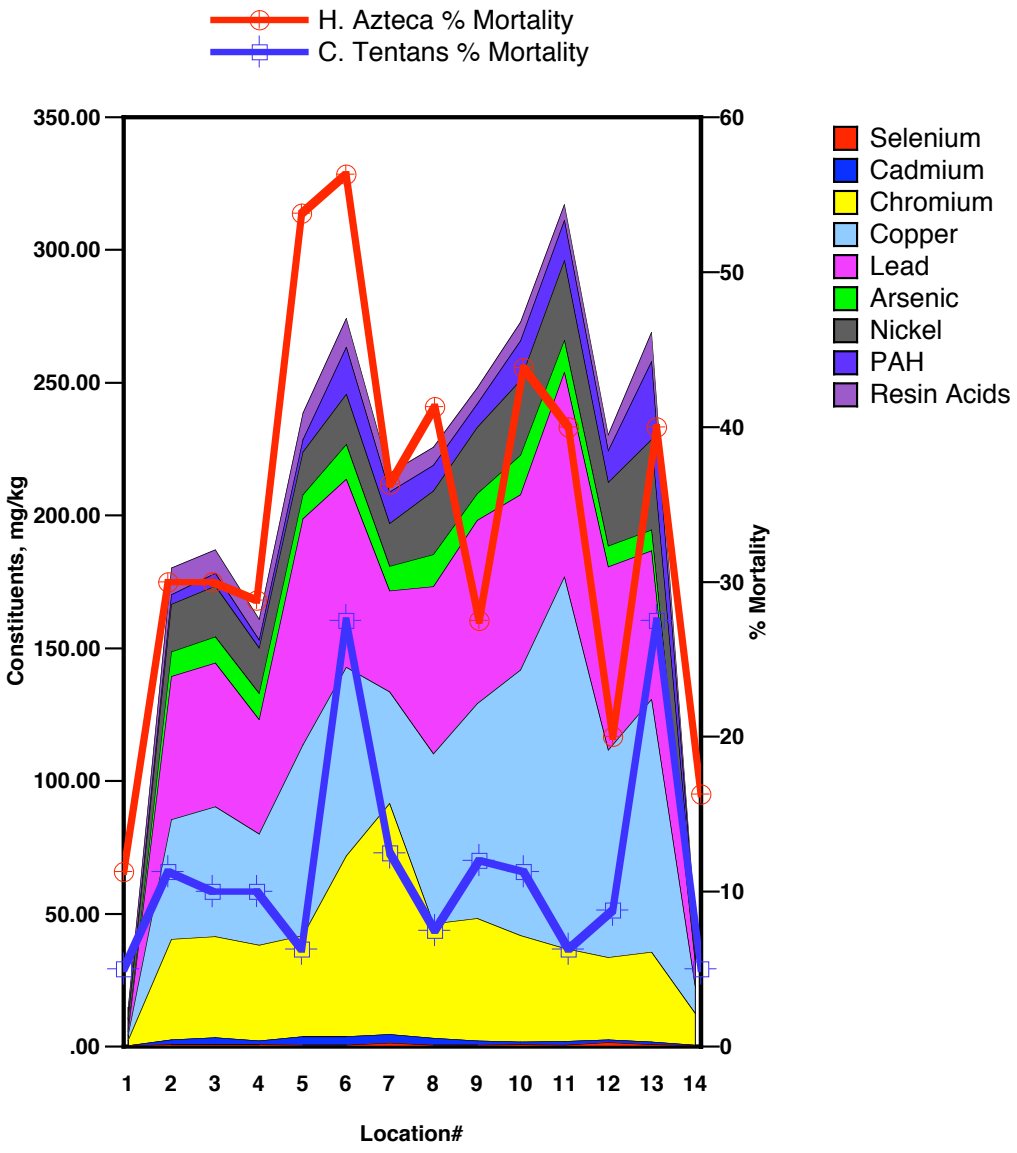

Figure 18 Stacked Area Charts of Selected Contaminants with *H. Azteca* and *C. Tentans* Mortality Rate Lines

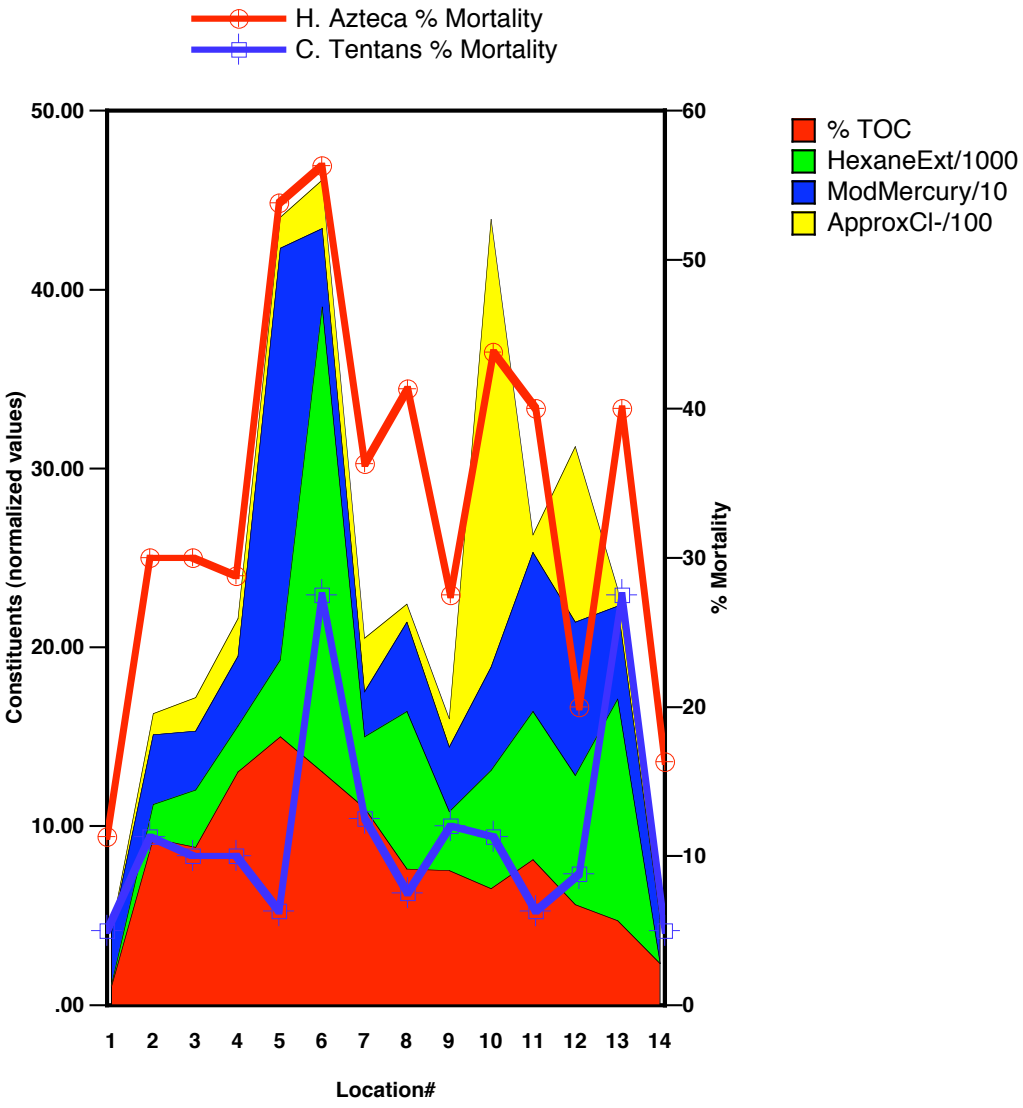

Figure 19 Stacked Area Charts of Additional Selected Contaminants (units normalized for scale) with *H. Azteca* and *C. Tentans* Mortality Rate Lines

Exploration of Toxic Effects on Organism Mortality

An intensive series of analyses were conducted to explore the organism acute mortality and native count data versus the concentrations of contaminants at the various sampling locations. The intent was to determine whether one or a few measured constituents were primarily responsible for the increased deaths of organisms within the Lake Manistee environment. The primary tool used for these determinations was stepwise multiple regression analysis using DataDesk 6.2 to develop a "model" for the impacts of the various toxins on the organisms. The familiar simple linear regression describes the relationship between a response variable (y) and a predictor variable (x) and the data can be plotted as a scatterplot that shows the datapoints, a regression line (based on the regression equation), and confidence interval boundaries. A multiple regression expands the regression equation to include more than one predictor variable to account for the response (dependent) variable values. The results of a multiple regression become difficult to visualize. With only two predictor variables the straight line of the simple regression becomes a flat surface. Further addition of predictors adds even more dimensions. Because of this, numerical values must be used to explain the model.

To understand whether predictor variables (such as arsenic, lead, hexane extractables, etc.) are significant and predict response values (such as *H. Azteca* % Mortality, Species #, etc.) requires the interpretation of tables containing t-ratios, probabilities, and R^2 fit values. A discussion of these values and their interpretation is beyond the scope of this report but can be found in standard statistics textbooks and the references previously cited in this report.

A brief description of the technique of stepwise multiple regression analysis is provided in Appendix B and can be used as a reference for the following sections. Additional detail on the approach is included in the Toxicity Exploration of *H. Azteca* % Mortality section, below, but is excluded from the other biological studies in the interest of brevity.

Toxicity Exploration of *H. Azteca* % Mortality

Table 15 contains the Pearson Product Moment Correlation values for *H. Azteca* versus all the possible predictors. The first column depicts the residual correlations used in the stepwise regression analysis before adding any predictors to the regression. When a predictor variable is added to the regression its residual correlation goes to 0.000 in the table. Regression began with the highest residual correlation, in this case arsenic at 0.756. Predictors were added to the regression analysis until the t-ratio probability of the last added predictor was > 0.05, indicating that it was no longer significant at the 95% confidence level. Table 16 provides the regression model, showing that arsenic, mercury, hexane extractable materials, and selenium were significant predictors of the percent mortality of *H. Azteca*. The regression model also includes a predictor variable labeled "12." This was an outlier in the plot of *H. Azteca* residuals versus HexaneExt (Figure 20). It was significant when in the regression and was retained (Appendix B). In short, statistical regression indicates that four contaminants, As, Hg, hexane extractable organic compounds, and possibly Se were probably most responsible for the mortality of *H. Azteca* species when exposed to the Lake Manistee sediments.

Table 15 Pearson Product-Moment Correlation for *H. Azteca* (y) and Predictors (x's)

Pearson Product-Moment Correlation

No Selector

	resid...	Bari...	Sele...	Arse...	Cad...	Chro...	Copp...	Lead	Nickel	Zinc	% TOC	Hexa...	PAH	Mod...	Resi...	Appr...
residuals(...	1.000															
Barium	0.516	1.000														
Selenium	0.018	0.073	1.000													
Arsenic	0.756	0.613	0.006	1.000												
Cadmium	0.706	0.238	0.018	0.474	1.000											
Chromium	0.611	0.295	0.014	0.439	0.759	1.000										
Copper	0.579	0.414	-0.162	0.668	0.556	0.435	1.000									
Lead	0.733	0.394	-0.115	0.577	0.796	0.607	0.800	1.000								
Nickel	0.461	0.560	-0.039	0.669	0.301	0.409	0.742	0.515	1.000							
Zinc	0.695	0.502	0.008	0.717	0.728	0.601	0.791	0.831	0.614	1.000						
% TOC	0.702	0.216	0.494	0.166	0.415	0.330	0.001	0.205	-0.025	0.182	1.000					
HexaneExt	0.680	0.214	0.099	0.405	0.354	0.242	0.368	0.262	0.420	0.222	0.281	1.000				
PAH	0.528	0.448	0.315	0.511	0.202	0.150	0.544	0.430	0.642	0.585	0.039	0.552	1.000			
ModMercury	0.484	0.300	-0.097	0.427	0.277	0.093	0.527	0.383	0.299	0.418	0.042	0.087	0.184	1.000		
Resin Acids	0.713	0.073	0.092	0.357	0.510	0.326	0.286	0.503	0.186	0.292	0.378	0.411	0.114	0.265	1.000	
ApproxCl-	0.169	0.357	-0.078	0.308	-0.017	0.005	0.207	0.136	0.267	0.274	-0.202	0.022	0.397	0.108	-0.159	1.000

Table 16 Multiple (Stepwise) Regression Model for *H. Azteca* % Mortality

Dependent variable is: H. Azteca % Mortality
No Selector
56 total cases of which 42 are missing
R squared = 98.6% R squared (adjusted) = 97.7%
s = 2.011 with 14 - 6 = 8 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	2209.43	5	441.887	109
Residual	32.3603	8	4.04503	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	5.84876	1.776	3.29	0.0110
Arsenic	1.10933	0.1975	5.62	0.0005
ModMercury	0.120536	0.0110	11.0	≤ 0.0001
HexaneExt	9.65144e-4	0.0001	9.73	≤ 0.0001
12	-27.6116	3.458	-7.98	≤ 0.0001
Selenium	10.5299	2.676	3.94	0.0043

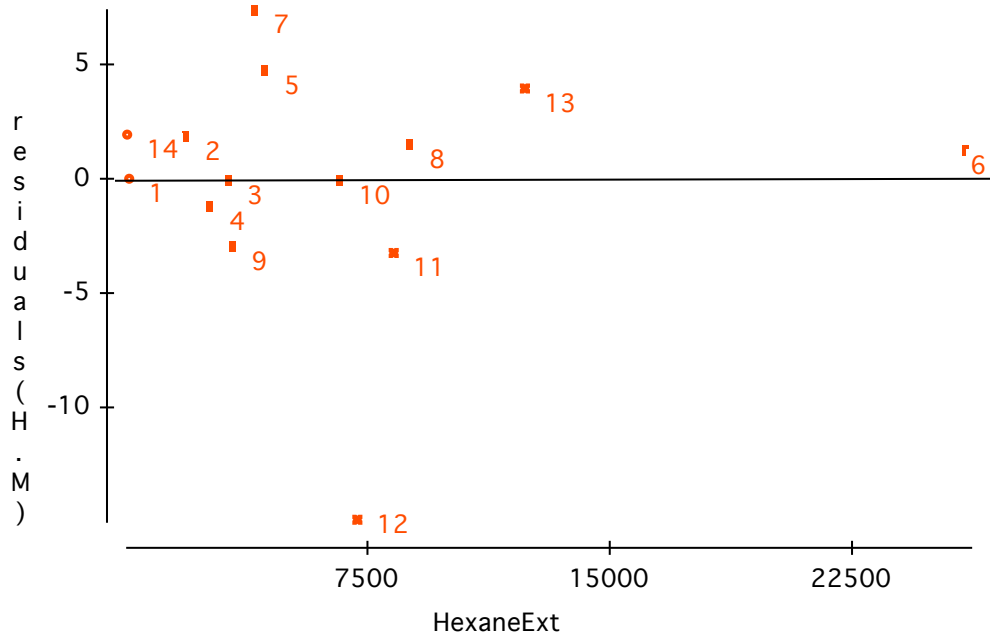

Figure 20 Scatterplot of *H. Azteca* % Mortality Residuals versus HexaneExt Concentrations. Location 12 is an Outlier.

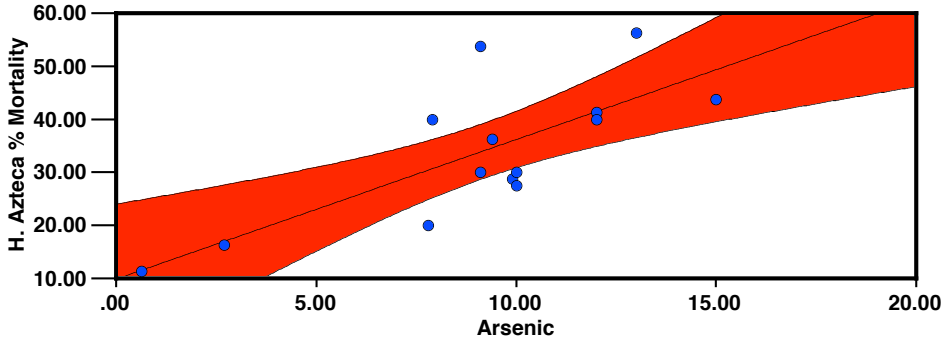

Figure 21 Regression and 95% Confidence Band of *H. Azteca* % Mortality versus Arsenic

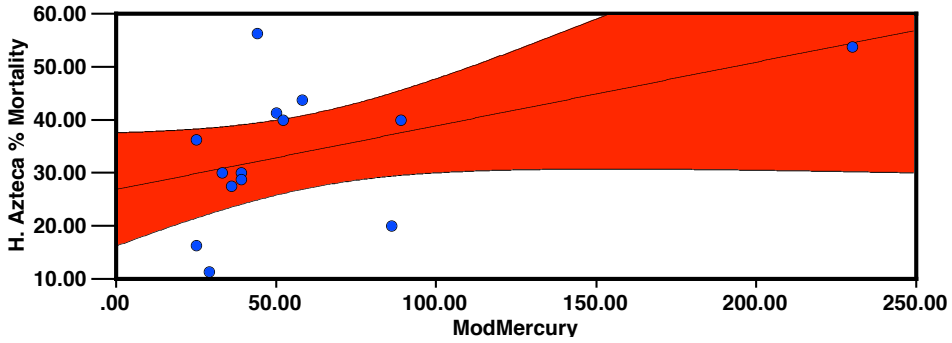

Figure 22 Regression and 95% Confidence Band of *H. Azteca* % Mortality versus ModMercury

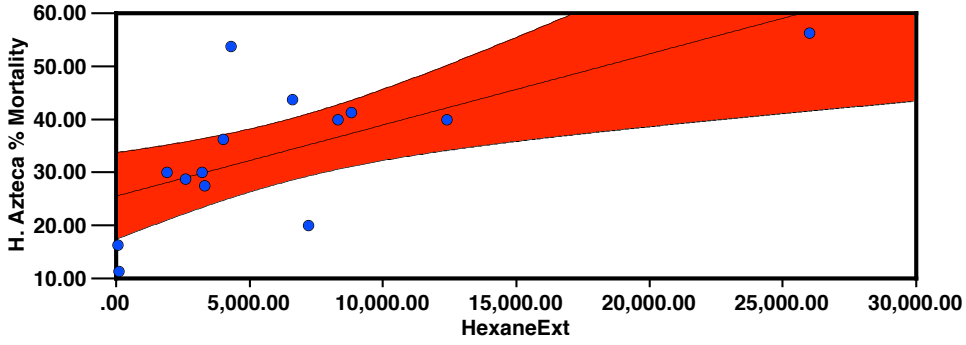

Figure 23 Regression and 95% Confidence Band of *H. Azteca* % Mortality versus HexaneExt

The difficulty explaining the mortality behavior of *H. Azteca* using simple linear regressions against one predictor variable at a time is illustrated in the three plots of Figures 21 through 23. Although all three of the plots exhibit an overall increase in mortality as the concentration increases, none of them individually exhibits really strong predictability over the full range of values. Figure 24 is 4-D depiction of the numerical data, with the *H. Azteca* % Mortality represented by the colors of the data points, that allows us to visualize the relationship between these factors to some extent.

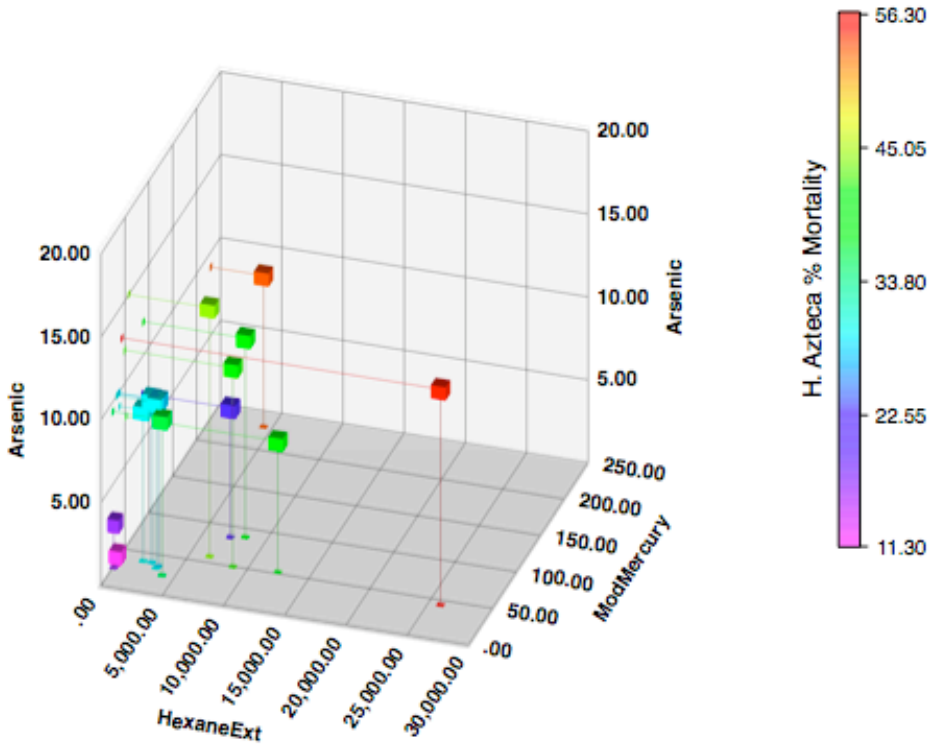

Figure 24 4-D Representation of *H. Azteca* % Mortality versus Arsenic, ModMercury, and HexaneExt Concentration

Figure 24 seems to show that mortality is highest when the concentrations of hexane extractable materials and/or mercury are moderate to very high and arsenic is around 10 mg/kg or greater in the sediments. Of note is a cluster of five datapoints where the mortality is moderate to slightly high while hexane extractables are < 5000 mg/kg and mercury is < 50 µg/kg with arsenic values hovering around 10 mg/kg. These points all reside above the leftmost grid in the floor of the chart (the background locations 1 and 14

are also located above this grid, but much lower in arsenic). Isolating these points and creating a table allows the observation that these samples (locations) are similar to one another in most analyzed characteristics, which likely explains the very similar mortality numbers for *H. Azteca*. Table 17 provides these data. There are a few instances, however, where major differences in constituents occur, one notable instance occurring in the Organisms (total) column where Location 7 has a much lower value (far fewer organisms present) than might be expected when compared to the other biological columns (*H. Azteca* % Mortality, *C. Tentans* % Mortality and Species #). Figure 25 uses scattered pie charts to graphically display similarities and differences for these datapoints (locations) relative to the total number of organisms in the sediment samples. The pie chart at Location 7 is very similar to those at the other locations with the exceptions of chromium (upper chart) and PAH (lower chart). This finding will be addressed in the section Toxicity Exploration of Organisms (total).

Toxicity Exploration of *C. Tentans* % Mortality

Regression analysis of *C. Tentans* % Mortality implies that the primary cause of acute toxicity for this organism was the hexane extractable materials. Table 18 shows the regression. Based on a confidence level of 95%, hexane extractable materials were the only significant impact on *C. Tentans*, but the R^2 value was only 57.8%. Location 13 was an apparent outlier in the plot of HexaneExt versus *C. Tentans* residuals and was included in the regression model due to its significance. This improved the R^2 value to 77.7%, which still indicates that only about 78% of the variability in *C. Tentans* % Mortality is being captured by the regression. If the confidence level is changed to 90% (t-ratio probability < 0.10) then chromium also becomes significant (Table 18) and the R^2 value improves slightly to about 82%. Figures 26 and 27 show simple linear regressions of *C. Tentans* % Mortality against the predictor variables HexaneExt and Chromium, respectively, with 95% confidence bands. As for *H. Azteca* % Mortality, these single predictor regressions show increasing *C. Tentans* deaths with increasing concentration of either hexane extractable materials or chromium but predictability using a single predictor regression equation is fairly noisy.

A 4-dimensional depiction for *H. Azteca* mortality was used in Figure 24 because three predictors were significant. Figure 28 uses a bubble chart (for depicting one less predictor dimension) to help us visualize the predictors versus dependant variable relationships for *C. Tentans*. Larger bubbles indicate higher % mortality, which is also indicated by color. This bubble chart shows that concentrations of hexane extractable materials above about 12,000 mg/kg resulted in the highest mortality levels for *C. Tentans*. Chromium above about 40-45 mg/kg may slightly increase mortality (Locations 7 and 9) but this is difficult to say with certainty. More data would be useful.

Table 17 Variables Data for the Five Points Selected from Figure 22

Station	Location#	Barium	Selenium	ModMercury	Arsenic	Cadmium	Chromium	Copper	Lead	Nickel	Zinc	ApproxCl-
M-2 P	2	110	0.65	39	9.1	1.7	38	45	54	18	160	120
M-3 P	3	110	0.62	33	10	2.6	38	49	54	19	160	190
M-4 P	4	120	0.58	39	9.9	1.4	36	42	43	17	130	210
M-7 P	7	83	1.2	25	9.4	3.2	87	42	38	16	150	300
M-9 P	9	120	0.49	36	10	1.6	46	81	69	25	180	160
	Mean =	108.6	0.708	34.4	9.68	2.1	49	51.8	51.6	19	156	196
	Std. Dev. =	15.16	0.28	5.81	0.41	0.77	21.59	16.57	11.97	3.54	18.17	67.31
	% RSD =	13.96	39.77	16.90	4.22	36.58	44.06	32.00	23.20	18.61	11.64	34.34

Station	Location#	% TOC	HexaneExt	PAH	Resin Acids	*H. Azteca* % Mortality	*C. Tentans* % Mortality	Species #	Organisms (total)
M-2 P	2	9.3	1,900.00	3.63	10	30	11.3	14	2,870.00
M-3 P	3	8.8	3,200.00	4.81	9	30	10	12	2,807.00
M-4 P	4	13	2,600.00	3.01	8	28.8	10	8	1,127.00
M-7 P	7	11	4,000.00	11.89	7	36.3	12.5	10	511
M-9 P	9	7.5	3,300.00	8.8	6	27.5	12	10	1,763.00
Mean =		9.92	3000	6.428	8	30.52	11.16	10.8	1815.6
Std. Dev. =		2.13	790.57	3.79	1.58	3.39	1.14	2.28	1033.63
% RSD =		21.47	26.35	59.03	19.76	11.11	10.23	21.11	56.93

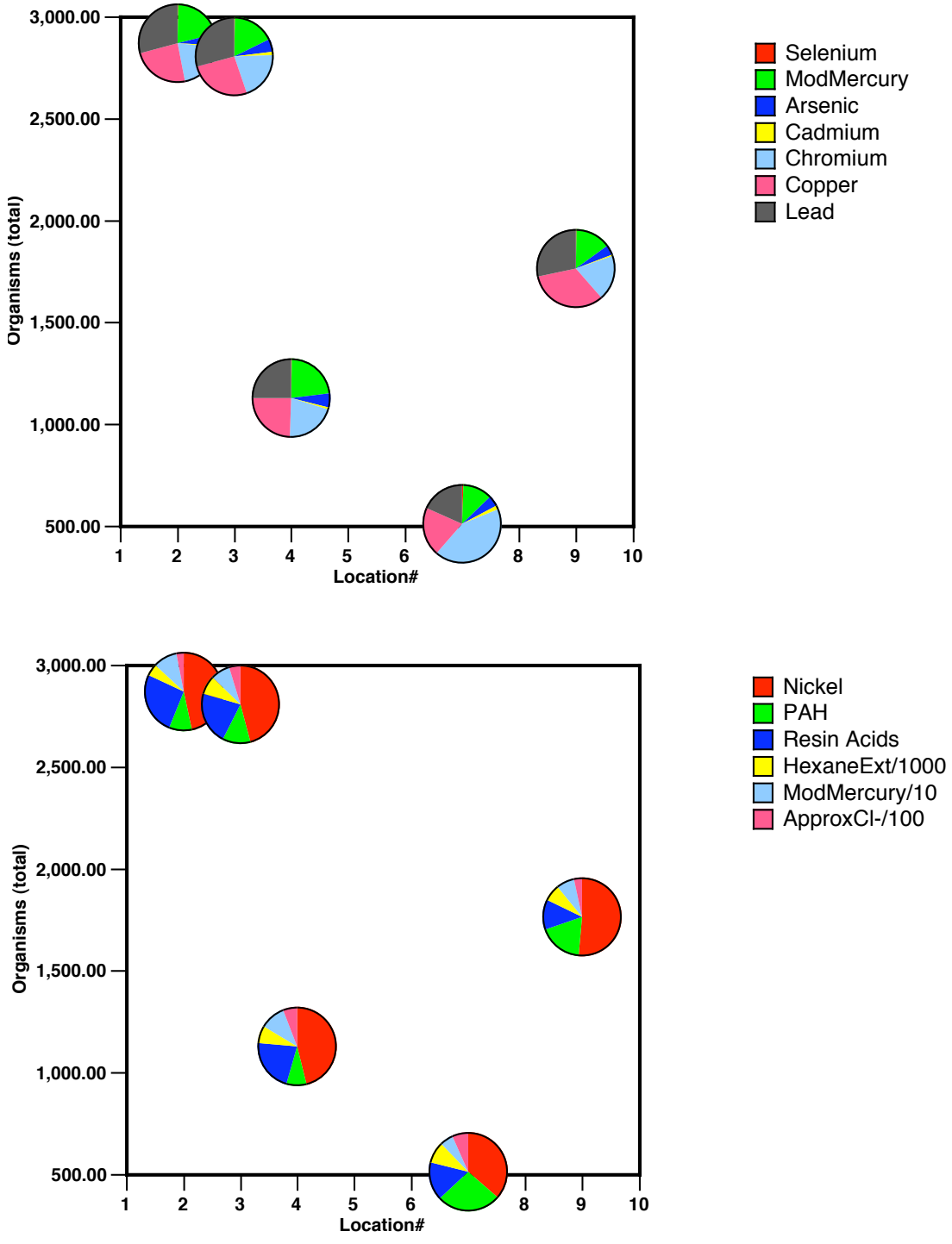

Figure 25 Scatterplot of Pie Charts Depicting Chemical Differences by Location and Organism (total) for Five Selected Datapoints.

Table 18 Multiple (Stepwise) Regression Model for *C. Tentans* % Mortality

Dependent variable is: C. Tentans % Mortality
No Selector
56 total cases of which 42 are missing
R squared = 85.9% R squared (adjusted) = 81.7%
s = 3.094 with 14 - 4 = 10 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	583.998	3	194.666	20.3
Residual	95.7018	10	9.57018	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.33785	1.874	1.78	0.1052
HexaneExt	5.73176e-4	0.0002	3.70	0.0041
13	13.9891	3.417	4.09	0.0022
Chromium	0.090167	0.0488	1.85	0.0941

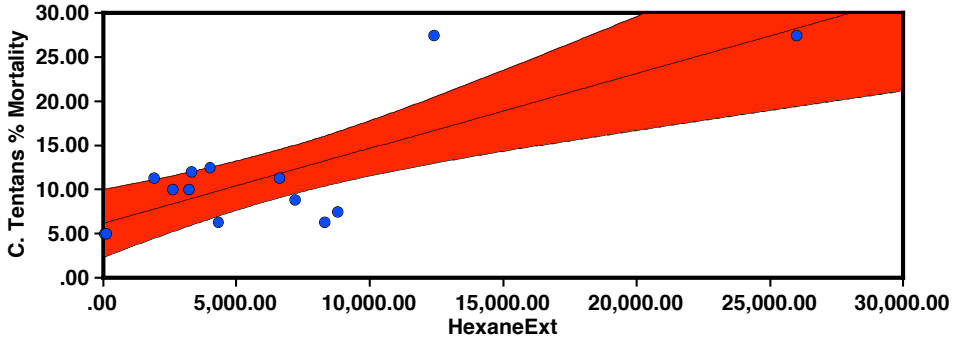

Figure 26 Regression and 95% Confidence Band of *C. Tentans* % Mortality versus HexaneExt

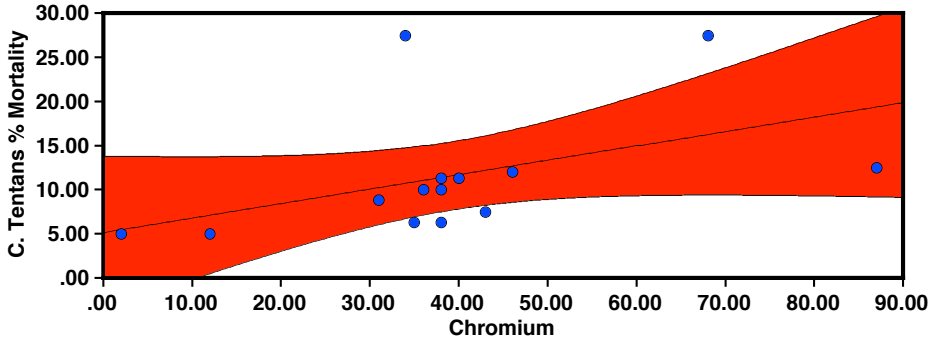

Figure 27 Regression and 95% Confidence Band of *C. Tentans* % Mortality versus Chromium

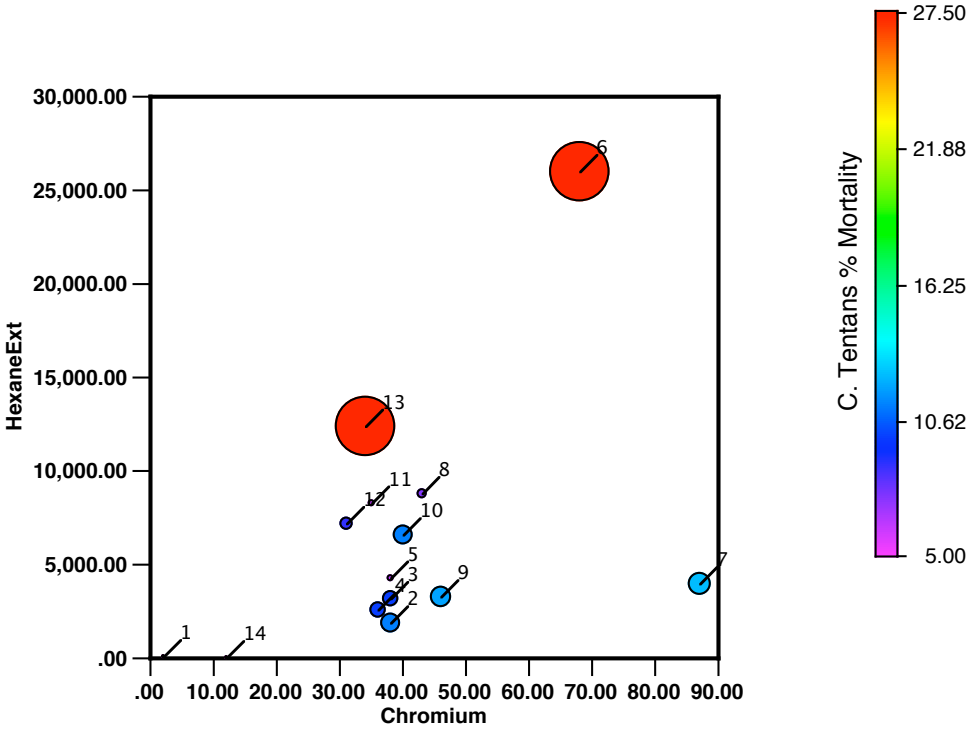

Figure 28 Bubble Chart Relating HexaneExt and Chromium with *C. Tentans* % Mortality. Bubble Size and Colors Indicate Increasing Mortality; Labels are the Sampling Locations.

Toxicity Exploration of Organisms (total) Counts

The Pearson Product Moment Correlation residual values for Organisms (total) versus all the possible contaminant predictors were extremely close for arsenic and chromium, -0.753 and -0.737, respectively. These values are negative because the contaminant impacts reduce the total number of organisms (as opposed to increasing the % mortality per the previously discussed studies). It might be recalled from Figure 25 and the discussion of the five-point cluster for *H. Azteca* that chromium was implicated as a possible factor in the low Organism (total) value for Location 7. High PAH concentration at that location was also a potential causal factor but this is not indicated by the correlation residual (-0.498) or the regression. Table 19 displays the regression analysis

for Organisms (total) with only arsenic as a predictor, Table 20 with only chromium as a predictor and Table 21 using both as predictors. Table 21 is probably the best fit based on the R^2 value (which is still not as good as those for the *H. Azteca* and *C. Tentans* % Mortality regressions) even though the t-ratio probability has exceeded 0.05 (0.078) and the confidence is now > 90% but < 95%.

Figures 29 and 30 display the simple linear regressions for Organisms (total) versus arsenic and chromium, respectively. As for the previously considered biological studies, the trend is that increasing concentrations correspond to decreasing organism counts but the fits are fairly poor on a single contaminant basis. Generally, the Organism (total) counts data do not respond to the predictors statistically as well as do either the *H. Azteca* or *C. Tentans* acute toxicity data. In this case, however, the bubble chart (Figure 31) shows that the lowest organism counts are indeed at the higher concentrations of both arsenic and chromium. The relationship seems fairly consistent with the exception of sampling location 11, which appears to be somewhat higher in organism count than would be predicted.

Table 19 Multiple (Stepwise) Regression Model for Organisms (total) and Arsenic

Dependent variable is: **Organisms (total)**
No Selector
56 total cases of which 42 are missing
R squared = 56.7% R squared (adjusted) = 53.1%
s = 1080 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	18362894	1	18362894	15.7
Residual	14003997	12	1167000	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4800.40	784.4	6.12	≤ 0.0001
Arsenic	-315.114	79.44	-3.97	0.0019

Table 20 Multiple (Stepwise) Regression Model for Organisms (total) and Chromium

Dependent variable is: **Organisms (total)**
No Selector
56 total cases of which 42 are missing
R squared = 54.3% R squared (adjusted) = 50.5%
s = 1110 with 14 - 2 = 12 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	17571891	1	17571891	14.3
Residual	14795000	12	1232917	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4135.01	660.5	6.26	≤ 0.0001
Chromium	-56.9090	15.07	-3.78	0.0026

Table 21 Multiple (Stepwise) Regression Model for Organisms (total) with both Arsenic and Chromium

Dependent variable is: Organisms (total)
No Selector
56 total cases of which 42 are missing
R squared = 67.8% R squared (adjusted) = 62.0%
s = 973.1 with 14 - 3 = 11 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	21950876	2	10975438	11.6
Residual	10416015	11	946910	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	5048.41	717.9	7.03	≤ 0.0001
Arsenic	-199.782	92.90	-2.15	0.0546
Chromium	-33.3864	17.15	-1.95	0.0776

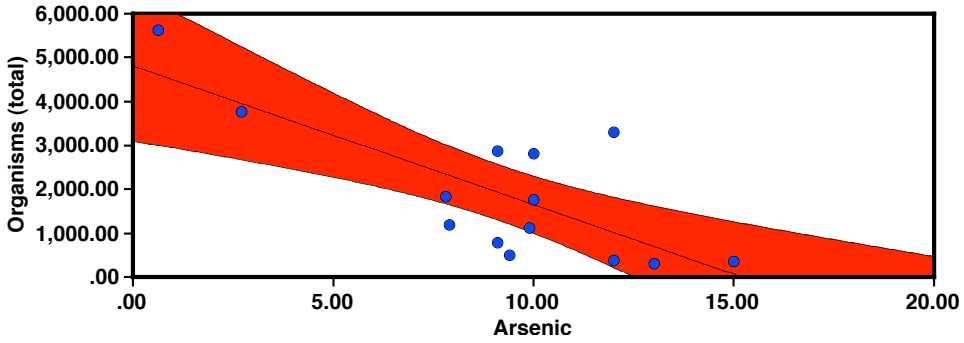

Figure 29 Regression and 95% Confidence Band of Organisms (total) versus Arsenic

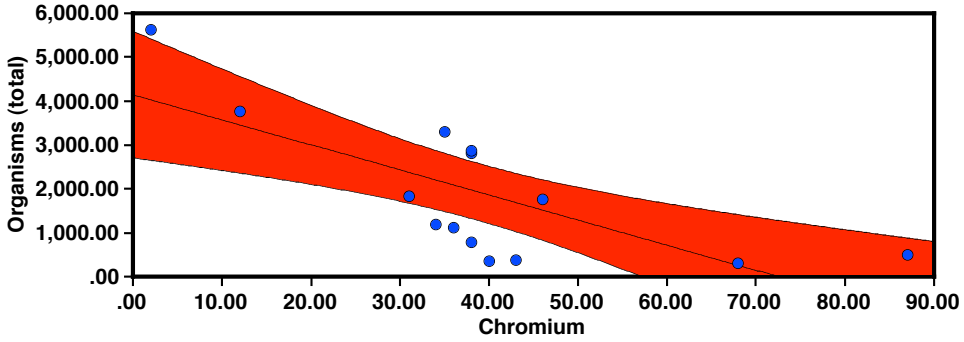

Figure 30 Regression and 95% Confidence Band of Organisms (total) versus Chromium

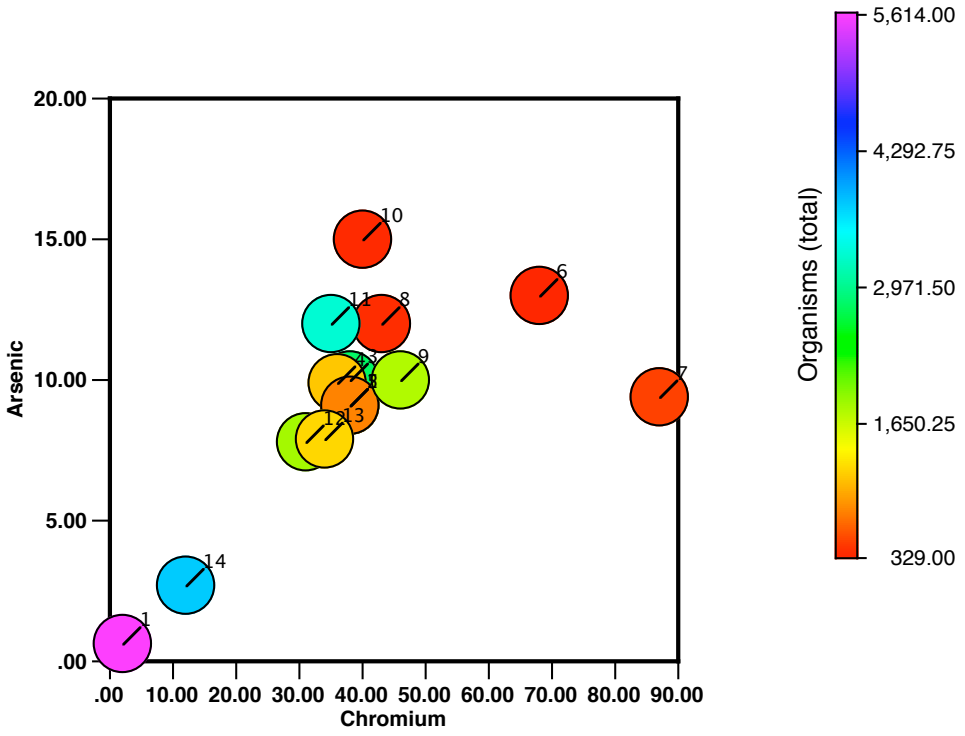

Figure 31 Bubble Chart Relating Predictor Variables Arsenic and Chromium with Organisms (total). Colors (not bubble size) Indicate the Organism Count; Labels are the Sampling Location#.

Toxicity Exploration of Species# Counts

The Pearson Product Moment Correlation for Species# regression residual versus all the possible predictors yielded arsenic as the toxic component with the highest correlation value. Table 22 provides the stepwise regression analysis for Species#. Arsenic was the only significant predictor when the outliers of Locations 12 and 13 were added to the regression, with the next most likely predictor (chromium) achieving a t-ratio probability of only 0.27. The R^2 value indicates that the regression accounts for 88% of the variability in the number of species counted. Locations 12 and 13 were indicated as

outliers in the plot of arsenic versus Species # correlation residuals and also appeared to be outliers in similar plots for other potential predictors that had high correlation residuals. Because only a single predictor variable (arsenic) was significant, the relationship is depicted only as a simple linear regression of Species# versus arsenic concentration in Figure 32. Locations 12 and 13 are the two points well outside the 95% confidence band in the lower part of the plot.

Table 22 Multiple (Stepwise) Regression Model for Species# with Arsenic

Dependent variable is: **Species #**
No Selector
56 total cases of which 42 are missing
R squared = 90.8% R squared (adjusted) = 88.0%
s = 1.786 with 14 - 4 = 10 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	313.825	3	104.608	32.8
Residual	31.8894	10	3.18894	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	23.2577	1.351	17.2	≤ 0.0001
Arsenic	-1.20617	0.1328	-9.08	≤ 0.0001
13	-7.72894	1.869	-4.13	0.0020
12	-6.84956	1.871	-3.66	0.0044

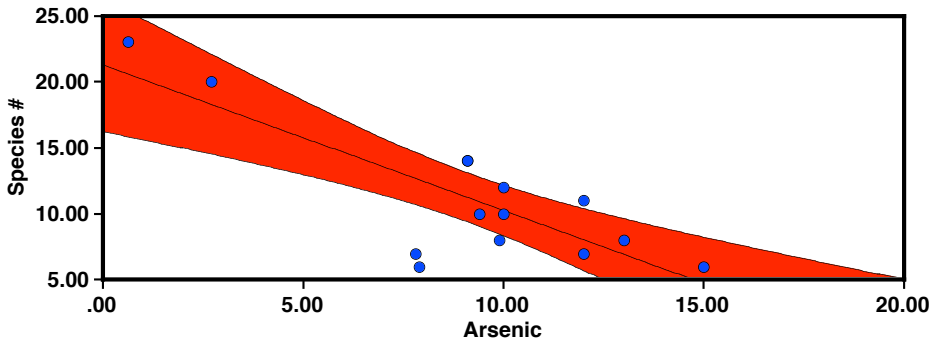

Figure 32 Regression and 95% Confidence Band of Species # versus Arsenic

Exploration of Outliers at Locations 12 and 13

Locations 12 and/or 13 were significant outliers for significant predictor variables in three of the four stepwise regression analyses (Tables 16, 18, and 22). It is important to understand the reason(s) why they are outliers to better understand the overall relationship between mortality and the contaminants at these locations.

Figures 33 through 36 present scatterplots to show the relationship of these two locations with each of the biological studies results versus the most significant predictor at that location. Arsenic was the predictor for *H. Azteca* [*Note*: 12 was an outlier for HexaneExt, the third most significant *H. Azteca* predictor], Species #, and Organisms (total) [although

neither location was a statistical outlier for that analysis], with HexaneExt for *C. Tentans* % Mortality. Therefore results from these studies are plotted versus the concentrations of these predictors with Locations 12 and 13 labeled.

In every case, even when not a statistical outlier (Figures 33 and 35), Location 13 represented an increased loss of organisms relative to the regression of the data versus the most significant predictor and always plots outside the 95% confidence bands. Figure 34 shows that the highest mortality rate experienced for *C. Tentans* (27.5%) occurred for sediments from Location 13 as well as for Location 6, which was well within the regression. To search for contaminant predictor variables that didn't follow the overall trend of contaminant concentrations by location, plots of all other possible predictor variables versus HexaneExt were created (not shown). Only PAH appeared abnormally high at Location 13 (Figure 37). Because Location 13 was also a significant outlier in the regression analysis for Species # and arsenic was the most significant predictor of reduced numbers of species, all possible predictor variables were also plotted versus arsenic concentration to look for something occurring outside the trend. In this case the trend was even clearer and the most extreme outlier was again PAH (Figure 38). The PAH value in this sample (29.4 mg/kg) exceeds the PEC value (Rediske et al., 2001). Therefore it seems reasonable to assume that the high PAH concentration is increasing the toxicity of this sediment to organisms and that is why Location 13 was an outlier that required addition to the regression analyses of *C. Tentans* and Species #.

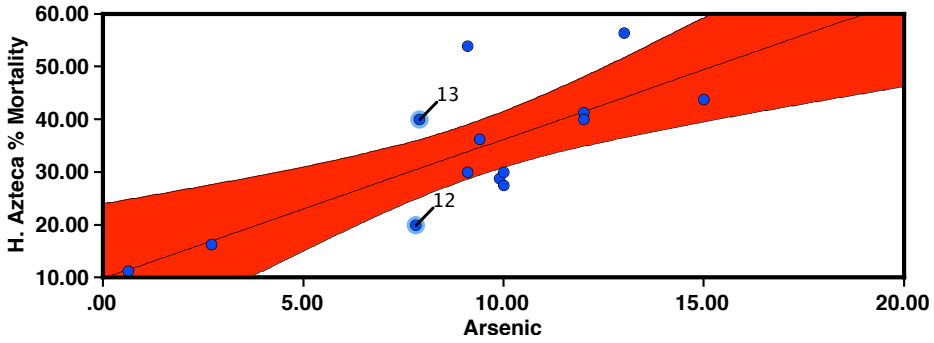

Figure 33 Scatterplot of *H. Azteca* and Arsenic Showing the Regression Outliers (12 only for *H. Azteca*)

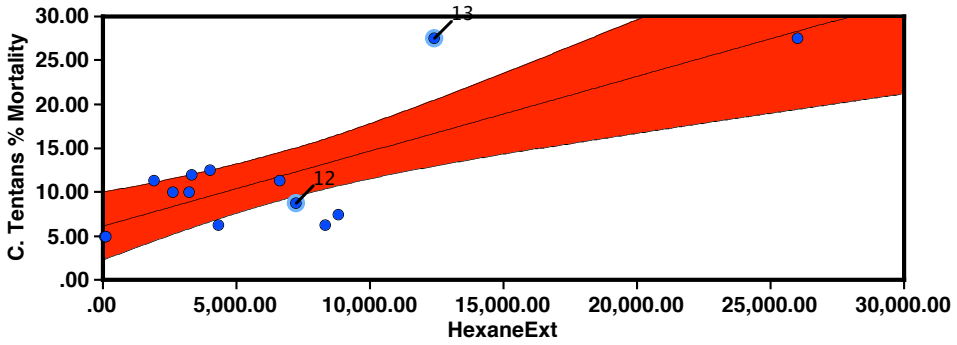

Figure 34 Scatterplot of *C. Tentans* and Hexane Extractables Showing the Regression Outliers (13 only for *C. Tentans*)

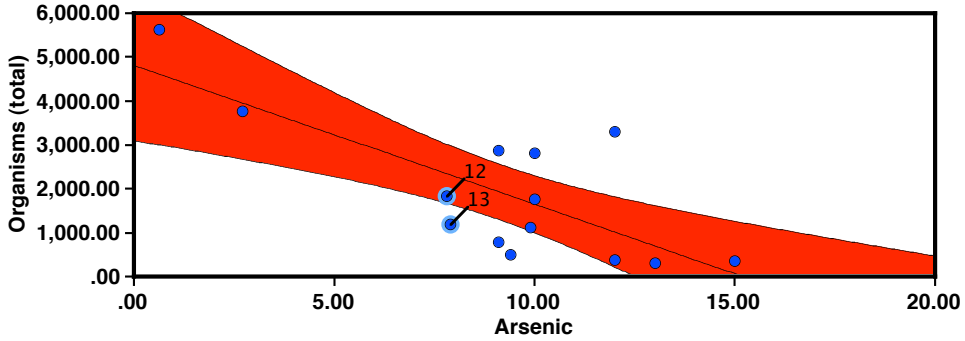

Figure 35 Scatterplot of Organisms (total) and Arsenic [Neither Location is a Regression Outlier for Organisms (total)]

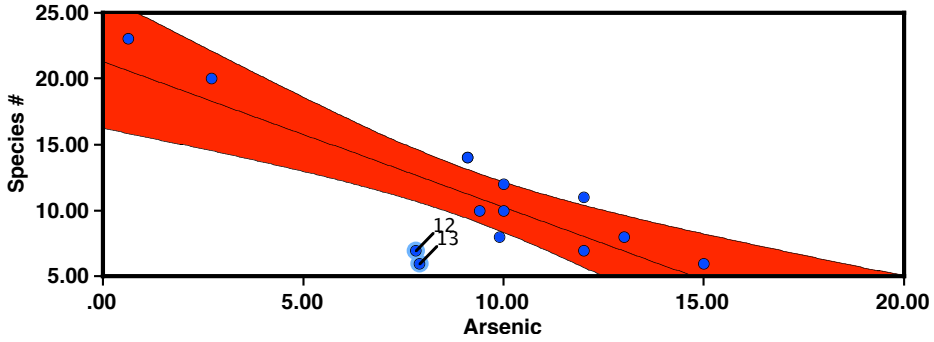

Figure 36 Scatterplot of Species # and Arsenic Showing Regression Outliers (Both 12 and 13)

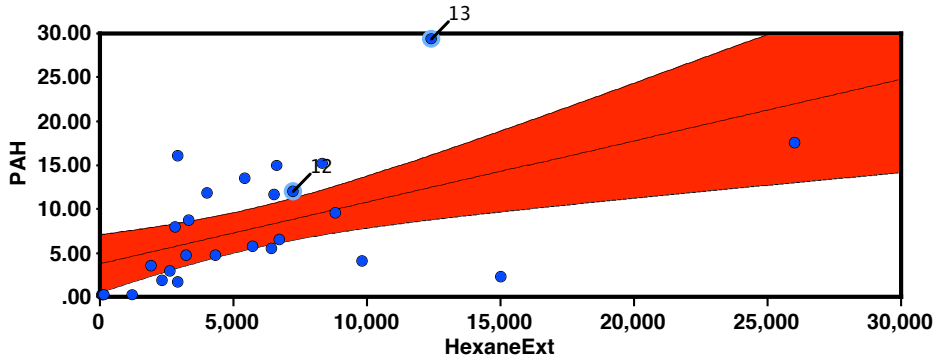

Figure 37 Scatterplot of HexaneExt and PAH Showing Location 13 as a High Outlier to the Trend

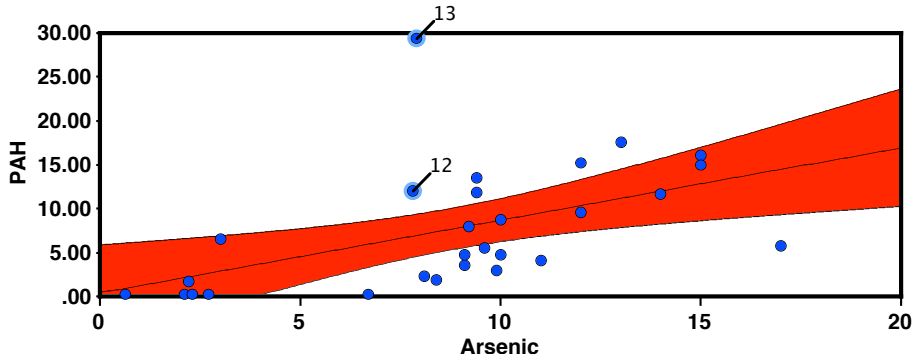

Figure 38 Scatterplot of Arsenic and PAH Showing Location 13 as a High Outlier to the Trend

Location 12 is not as easily understood as Location 13. It is a point of lower than predicted % Mortality for *H. Azteca* (Figure 33) but shows increased toxicity with regard to Species # in a way similar to Location 13 (Figure 36). For the other plots versus significant indicators it seems fairly unimportant (Figures 34 and 35) and it does not seem related to PAH concentration (Figures 37 and 38). The only measured constituent that appears as a slight trend outlier relative to the others at this location is selenium and the difference seems too small to be of much significance (Figure 39). It is possible that the significance of Location 12 is simply a result of random errors of some sort.

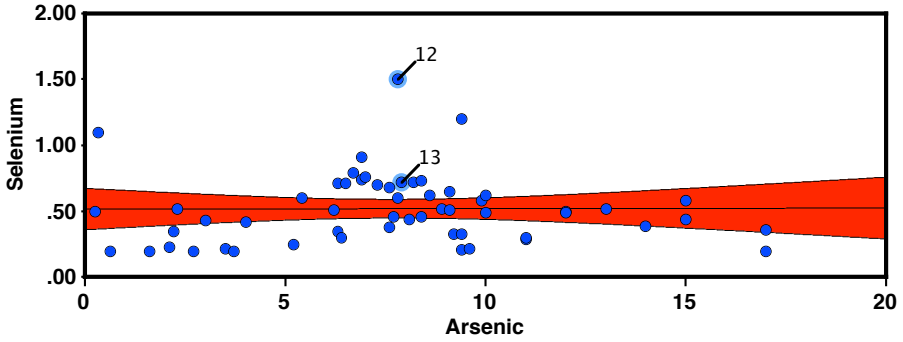

Figure 39 Scatterplot of Selenium and Arsenic, Showing Locations 12 and 13 Relative to the Trend of the Data

Summary Assessment of the Organism Studies EDA

Exploratory data analyses were performed on the results of biological studies that were conducted by Rediske et al. (2001) on Lake Manistee sediments from 14 locations. These studies included counts of the total number of organisms found in the sediments, a count of the number of species in each of the sediment samples, and two acute toxicity studies on organisms exposed to the sediments, including *Hyalella Azteca* and *Chironomus Tentans*.

Simple descriptive statistics were performed on the results of these four studies with awareness that the datasets were not very large and that treatment differences could impact the statistics. The four sets of data were found sufficiently normal given these caveats to proceed with the EDA without undue concern about this factor. The results from the study of the acute toxicity of sediments on *H. Azteca* were the most normally distributed and also tended to provide the most information on the remaining aspects of the EDA.

In order to assess whether specific contaminant outfalls or general contaminant types (that might be related to outfalls) would show a relationship to the results of the biological studies, box and whisker plots (Figures 15 and 16) of the results were constructed versus potential source categories from Rediske et al. (2001) and general (assumed) contaminant types. In both instances, only the control samples (Locations 1 and 14) had median values and 95% confidence intervals distinctly different from the other categories, all other categories had overlap of these parameters. Therefore these "bulk" identification procedures were unsuccessful.

Area plots of the results from the four biological studies versus location show very good correspondence between the four datasets (Figure 17). These plus additional stacked area plots of the contaminants versus location overlain by the mortality study results illustrate the extreme toxicity of these sediments for organisms in the bulk of Lake Manistee just beyond the river mouths (Figures 18 and 19).

Stepwise regression analyses of the biological study results versus the individual contaminants were done in an attempt to identify whether specific contaminants or groups of contaminants seemed to be most strongly implicated in the organism mortality. The purposes being to determine which chemicals need to be reduced in concentration and/or have their discharge controlled, to facilitate estimations of contaminant residence time and fate, to develop plans for additional sampling studies, and to help develop a conceptual model of the lake.

Table 23 provides information on the contaminants that were indicated to be most significant with respect to mortality for each of the organism studies.

Table 23 Summary of Regression Model Results for the Biological Studies

Study	Significant Predictors	Outlier Locations	Model Fit (R^2 value)	F-Ratio
H. Azteca % Mortality	As, Hg, Hexane Extractables, Se	12	97.7%	109
C. Tentans % Mortality	Hexane Extractables, Cr	13 (PAH)	81.7%	20.3
Organisms (total)	As, Cr		62.0%	11.6
Species #	As	12, 13 (PAH)	88.0%	32.8

A note of caution is warranted regarding these results. A high correlation between predictor and dependent variables (e.g., between a contaminant and a % Mortality) does not prove a causal relationship. As shown in Figures 18 and 19, as the biological mortality increased, concentrations for virtually all the contaminants also increased, so one must be cautious in interpreting these results.

Even though direct causality cannot be absolutely determined by these methods, is it probably important that arsenic turns up as significant in three of the four regression analyses and hexane extractable materials and chromium in two of the four when one considers the sheer number of analytes measured. A very high PAH value appears to have toxic significance in two of the four analyses as well. There is slightly less confidence that mercury is as important for these tests since it is indicated as significant in only one case and required some modifications to the values lower than the reporting limit to be usable at all. Selenium also achieved significance in only one case.

In terms of these biological study results for the Lake Manistee sediments, the contaminants, and possibly their order of importance, appear to be:

As > Cr, hexane extractables, PAH > Hg, Se.

A somewhat surprising result was that chloride did not seem to be a major factor in any of the studies based on the stepwise regression analyses, even though it is present in very large quantities at certain locations.

Conclusions

Please refer to the conclusions as presented in the Executive Summary Section.

References

Rediske, R., Gabrosek, J., Thompson C., Bertin C., Annis, J.B., Meier, P.G., and M. Tuchman. 2001. “Preliminary Investigation of The Extent of Sediment Contamination in Manistee Lake.” AWRI Publication # TM-2001-7.

J.W.Tukey, "Exploratory Data Analysis", 1977, Addisson Wesley.

Velleman, P. F. (1997). DataDesk Version 6.0, Handbook, Volumes 2 and 3. Ithaca, N.Y., Data Description, Inc.

Appendix A: Explanation of Dotplots and Box and Whisker Plots

Although box and whisker and dotplots are commonly used in data exploration, they are not as familiar to environmental scientists and will be briefly described using the following example figure.

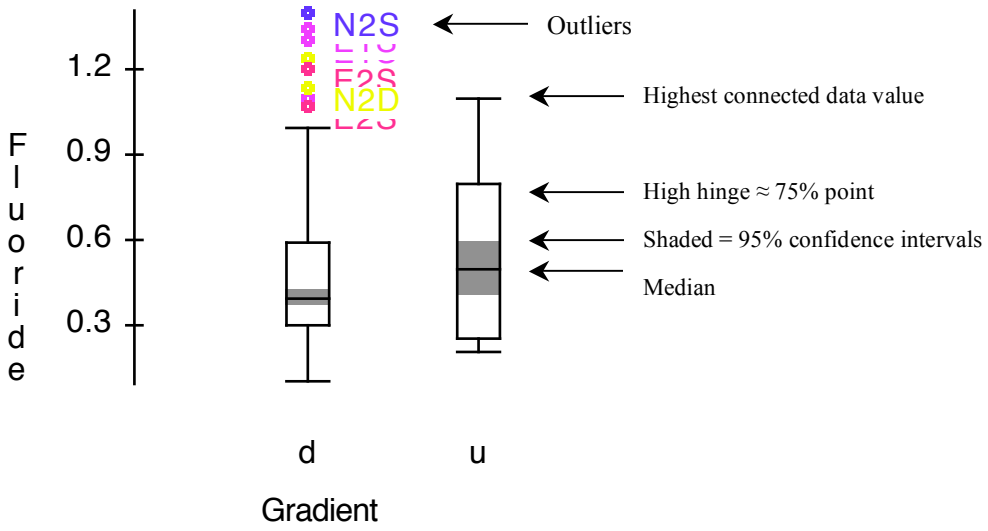

Figure 40 Example Box and Whisker Plot

This figure illustrates the distribution of fluoride (from a previous Powell & Associates project) grouped by upgradient and downgradient designation wells for all samples that have been collected, analyzed for fluoride, and reported in the spreadsheet. Boxplots have four components, more rigorous definitions of which can be found in Velleman, 1997 (Velleman, P.F., DataDesk Handbook Version 6, 1997, Data Description, Inc.). These are:

- The outlined central box depicts the middle half of the data between the 25th and the 75th percentiles.
- The horizontal line across the box marks the median.
- The whiskers extend from the top and bottom of the box to depict the extent of the main body of the data.
- In addition, extreme data values are plotted individually, usually with a circle. Very extreme values are plotted with a starburst (There are none of these in the example. In the example the outliers are also labeled with the well number at which the outlier occurred).

In addition to these four components the 95% confidence intervals of the data can be depicted with a shaded area. If these shaded areas for two or more groups do not overlap then there is 95% confidence that the difference between their medians is statistically

significant. If the shaded areas do overlap the medians might not be statistically different, at least not with 95% confidence. In the example figure the confidence intervals appear to have the slightest overlap, therefore, based on the available data, it cannot be said with 95% confidence that the means are statistically different between the upgradient and downgradient wells with regard to the concentration of fluoride.

For each variable selected, the boxplot shows:

- The overall level of values
- The overall variability or spread of the values
- Whether the main body of data values is distributed symmetrically around the median
- Any values that stray markedly from the rest.

In addition, a collection of boxplots (like the example figure) shows:

- How the levels of the variables compare
- How the spreads of the variables compare
- Relationships between the levels and spread. For example, do variables with a higher overall level tend to be more variable as well?

Dotplots are merely boxplots with the boxes removed. This allows observation of the number of data points that went into generating the statistics for the boxplots and yields some insight into whether there is data sufficient to draw sound conclusions. The dotplot example (Figure 2) is the same dataset as the previous boxplot with the boxes removed:

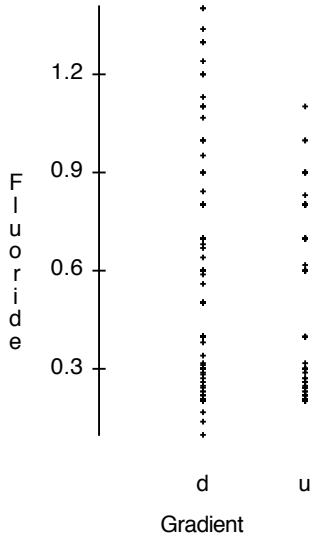

Figure 41 Example Dotplot

Appendix B: Approach to Stepwise Regression Analysis

Note: The stepwise regression process would be very difficult and time-consuming if manually calculated and plotted. All the following steps were carried out in DataDesk 6.2 that facilitates these analyses through automatically generating residual variables, correlation tables, drag and drop of predictors into the regression tables, etc.

1. Select the y-variable of interest (e.g., *H. Azteca* % Mortality) and perform a regression yielding only the constant term.
2. Compute the residuals of this regression and save these values as a variable.
3. Perform a Pearson Product-Moment Correlation of the residual variable versus all the candidate "predictor" variables (e.g., arsenic, PAH, etc.).
4. The predictor variable having the highest correlation with the residuals variable is added to the regression model and the t-ratio significance is checked (e.g., significant at the 95% confidence level or probability < 0.05).
    a. If the variable is not significant it is removed from the regression and the regression analysis is stopped.
    b. If the predictor variable is significant diagnostic plots should be generated to check for outliers and data patterns; plot a scatterplot of the residuals versus the contemplated predictor.
5. If there is no outlier, the predictor is left in the regression and its effect on the remaining residuals in the correlation is removed (this alters the values of the remaining correlations). The predictor variable with the new highest correlation is then tested in the same manner (i.e., add the predictor to the regression model).
6. If there is an outlier evident in the diagnostic plot(s) an "indicator variable" is created and added to the regression model and its significance is checked (Note: The indicator variable simply allows the outlier data point to be added to the regression independently of the predictor variable wherein it was indicated as an outlier).
    a. If the t-ratio for the indicator variable is not significant, remove it from the regression and proceed testing the next predictor variable.
    b. If the t-ratio for the indicator variable is significant and the significance of the preceding predictor variable (the one that showed the indicator point as an outlier) remains significant, leave the indicator in the regression and proceed with testing the next predictor variable.
    c. If the indicator variable is significant but having it in the regression made the t-ratio of the preceding predictor variable fall outside the significance criterion, remove the predictor from the regression and check the significance of the indicator variable.
        i. If the indicator is still significant, leave it in the regression and leave the predictor out, then proceed to add the next predictor variable to the regression and repeat these analyses.
        ii. If, after removing the preceding predictor the indicator variable is not significant then there is some kind of interaction between the predictor and the outlier and there is no right answer at this point.

Both can be left out of the model, or put one or the other may be put into the model. Leaving the predictor in may help with the understanding of the final model, or the outlier may be left in because there are other good reasons to suspect that the point is not correct.

7. Proceed with adding predictor variables and performing these assessments until the t-ratio significance criterion is exceeded (e.g., the probability is no longer < 0.05.