TIG environmental®
INVESTIGATIONS ▾ FORENSICS ▾ REMEDIATION

Evaluating the Challenges of Using Disparate Data Sets in Forensic Methods

Nicholas D. Rose, Timothy Negley, and Carlo Monti

February 13, 2019

Presented at the Tenth International Conference on Remediation and Management of Contaminated Sediments
New Orleans, Louisiana
February 11–14, 2019

A Member of TheIntelligenceGroup®
RESEARCH ▾ INVESTIGATIONS ▾ FORENSICS

Why Combine Disparate Data Sets for Forensic Investigations?

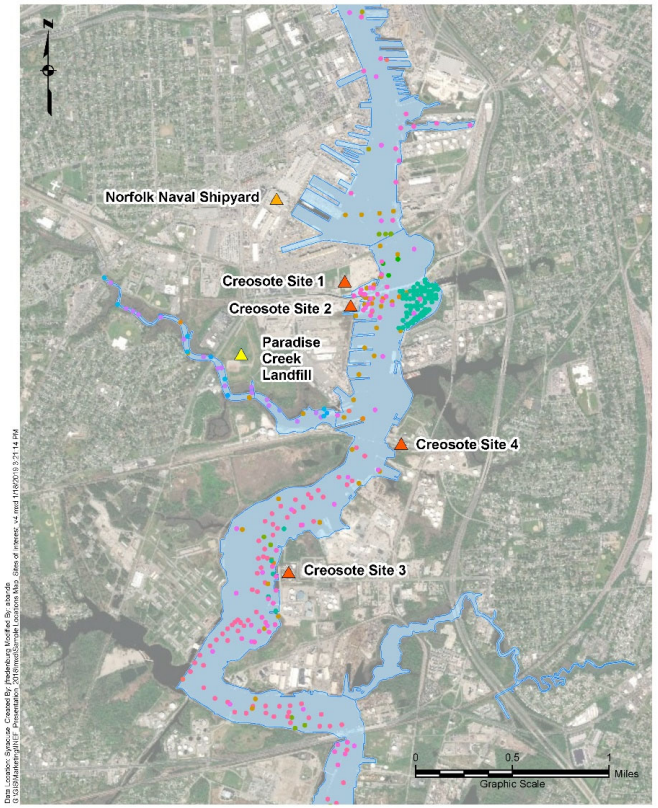
- Forensic methods are useful for identifying contaminant sources and assigning shares of costs in the allocation process
- The reasons for using and combining existing data sets include the following:
  - Providing a data set that is larger and has more coverage than any individual data set
  - Capturing data from historical sources that may have been removed or are no longer accessible
  - Avoiding the potentially prohibitive cost of additional studies/analyses
  - Informing future forensic sampling plans or investigations

Issues with Combining Data Sets for Forensic Investigations

- Available data are rarely collected with the aim of forensic analysis
- Some of the most common problems are:
  - **Varying contaminants analyzed**
  - **Censoring of data sets**
  - **Varying detection limits**
  - Variable units
  - Multiple contaminant names
  - Multiple analytical methods
  - Outliers
  - Distribution of sample locations due to study design
  - Lack of unique sample IDs within and between data sets
  - Inconsistent depth segmentation
  - Changes in surface elevation between studies
- **Combining data requires careful consideration of usability and harmonization**

Combining Data from Multiple Studies Provides Larger Data Set with Wide Coverage

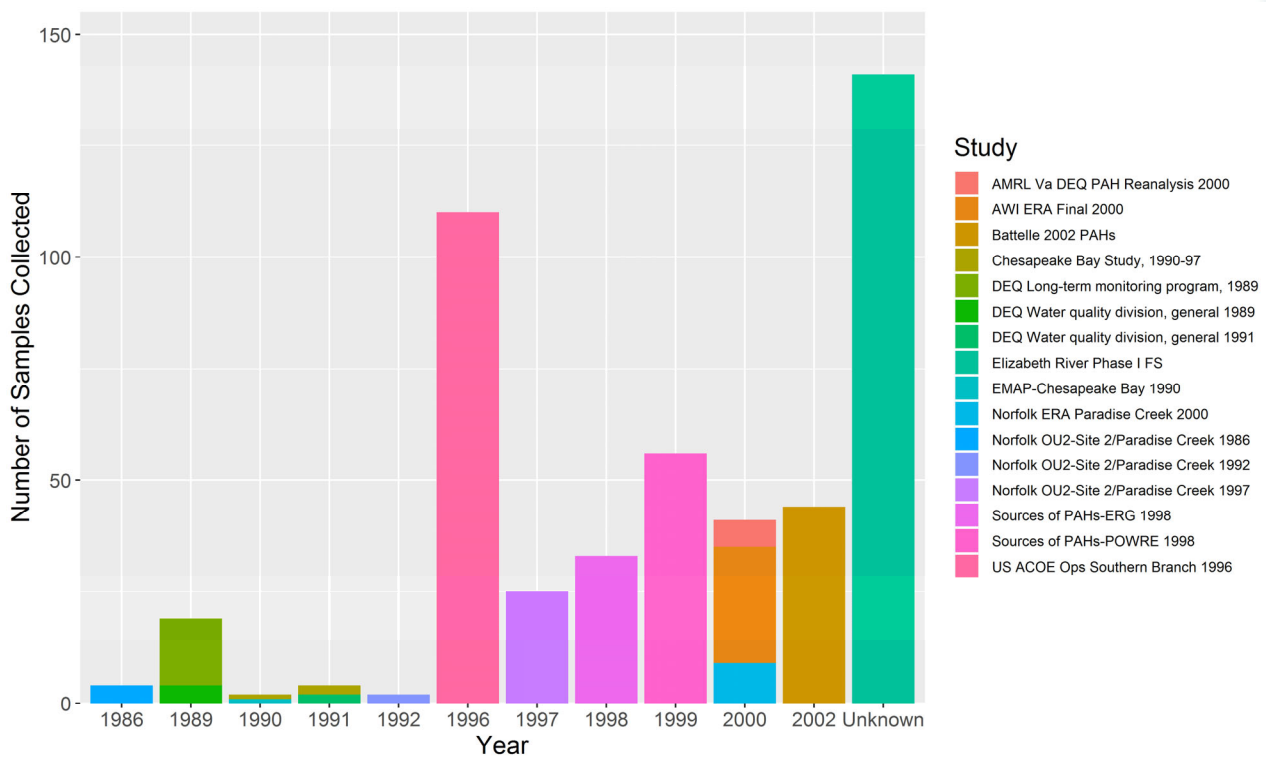
- Example data set from Southern Branch of Elizabeth River, VA
- All data came from National Oceanic and Atmospheric Administration's (NOAA's) Query Manager
  - NOAA's Query Manager addresses unit and analyte name consistency
- Multiple potential polycyclic aromatic hydrocarbon (PAH) sources identified in this area, so this example analysis was limited to PAHs



Multiple Studies Provide Extensive Historical Sampling

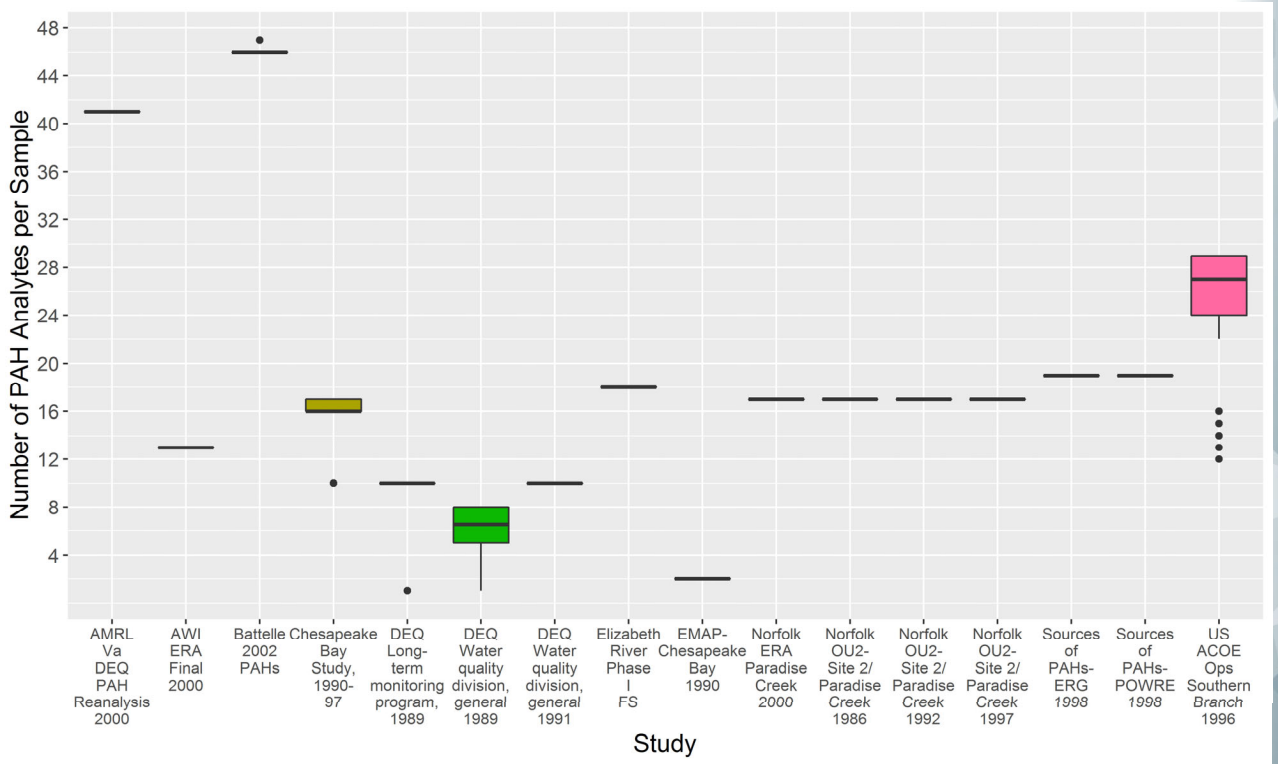
Elizabeth River PAH data

- Sampling events conducted over 16 years
- 16 studies
- 481 total samples
- 348 unique sample locations
- 10,246 results



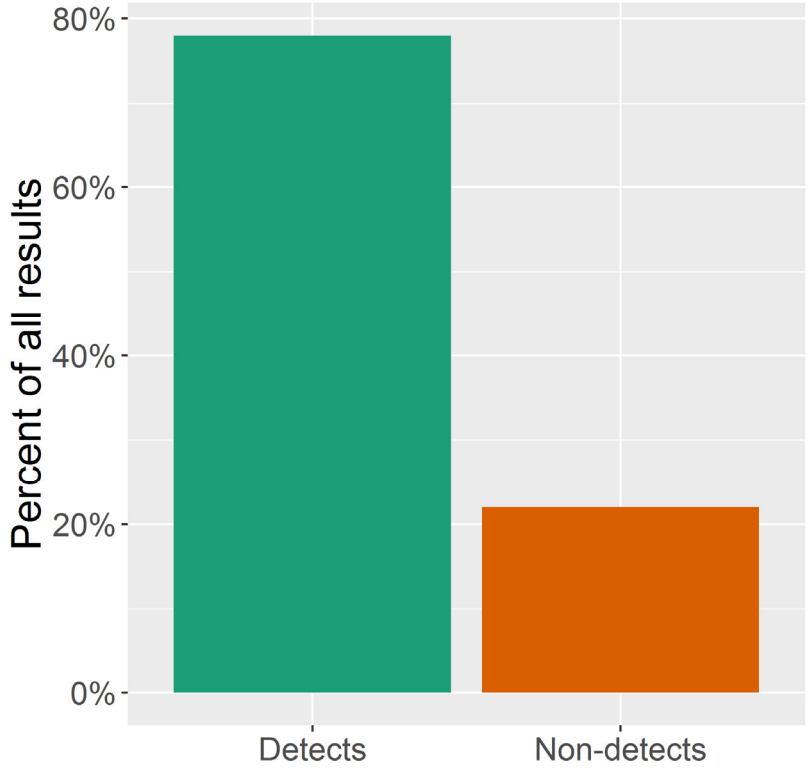
Different Studies Evaluate Different PAHs

- Number of PAHs analyzed per study ranges from 1 PAH (benzo[k]fluoranthene) to >40 PAHs (including methylated PAHs)
- Studies did not always analyze the same suite of PAHs



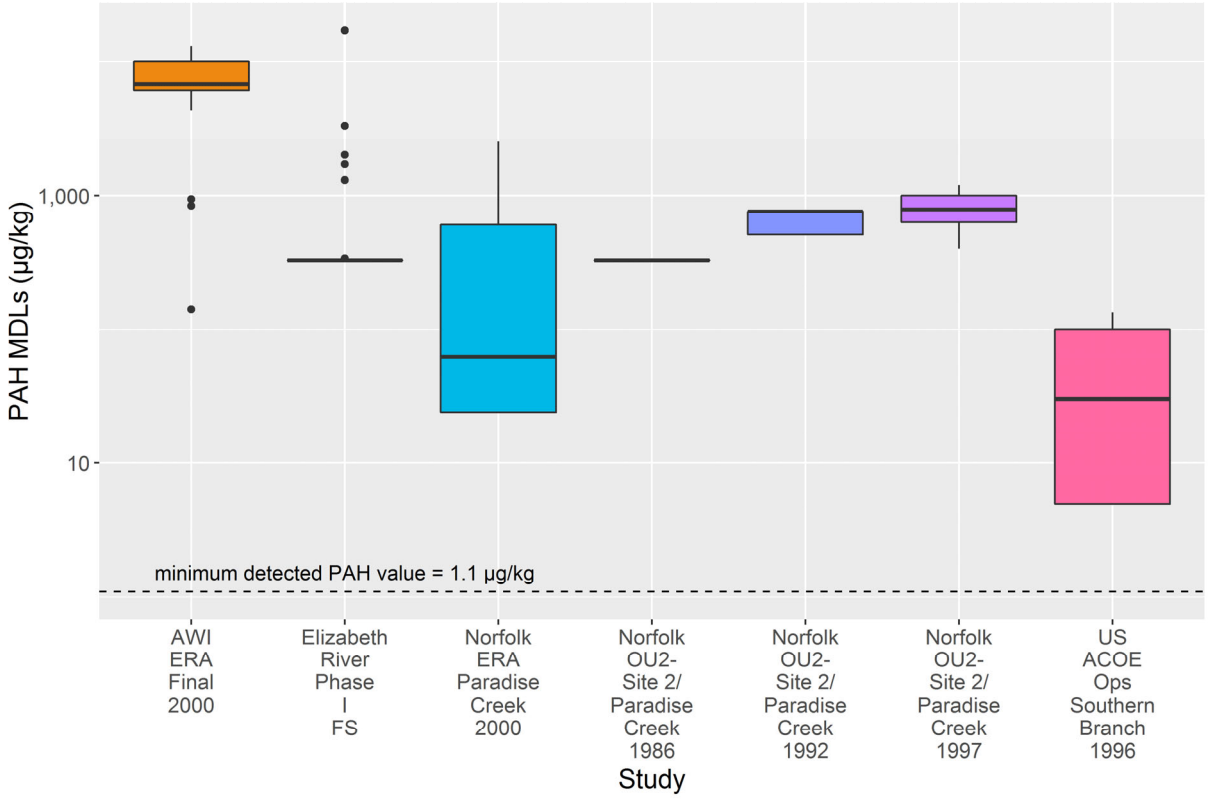
Combined Data Set Contains Censored Values

- More than one-fifth of all values are censored (below method detection limits [MDLs])
- Specific analytes may have:
  - High MDLs
  - No presence in sediments
- Specific studies may have:
  - High MDLs
  - Sampled non-contaminated locations



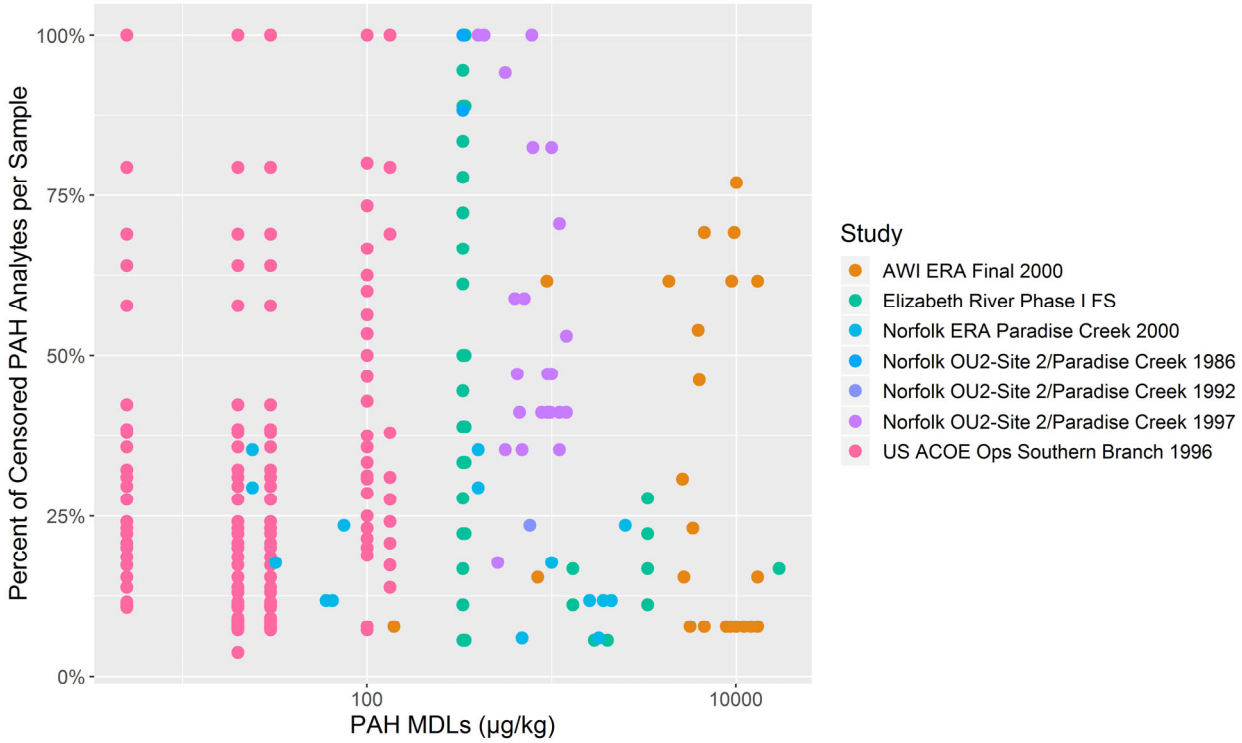
Studies Include Wide Range of MDLs

- MDLs only for non-detect results (7 studies)
- Wide range of MDLs potentially due to:
  - High concentrations
  - Lab quality/capabilities
  - Analytical method
- Does this affect censoring?



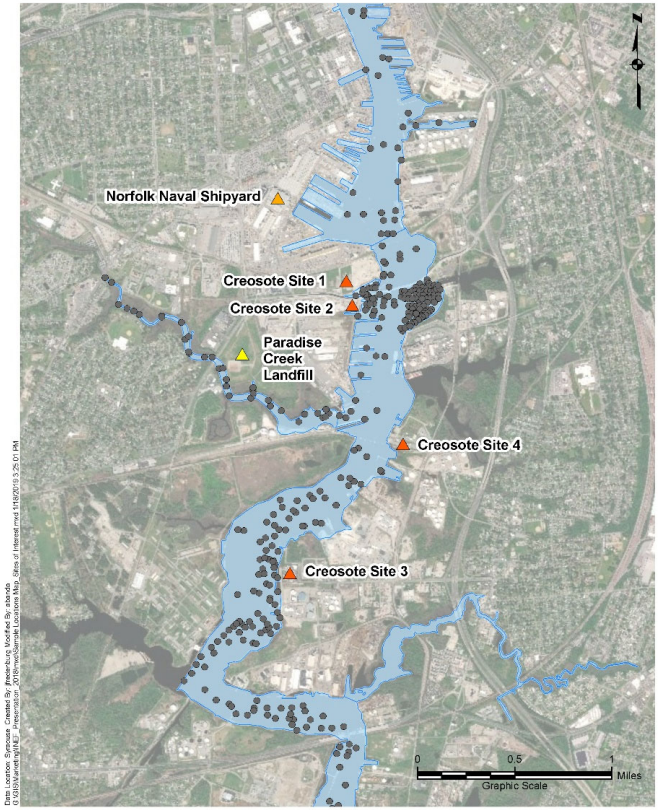
High MDLs Do Not Affect Censoring

- MDLs do not correlate with censoring
- Variability in censoring also potentially due to:
  - Matrix interference
  - Targeted study design



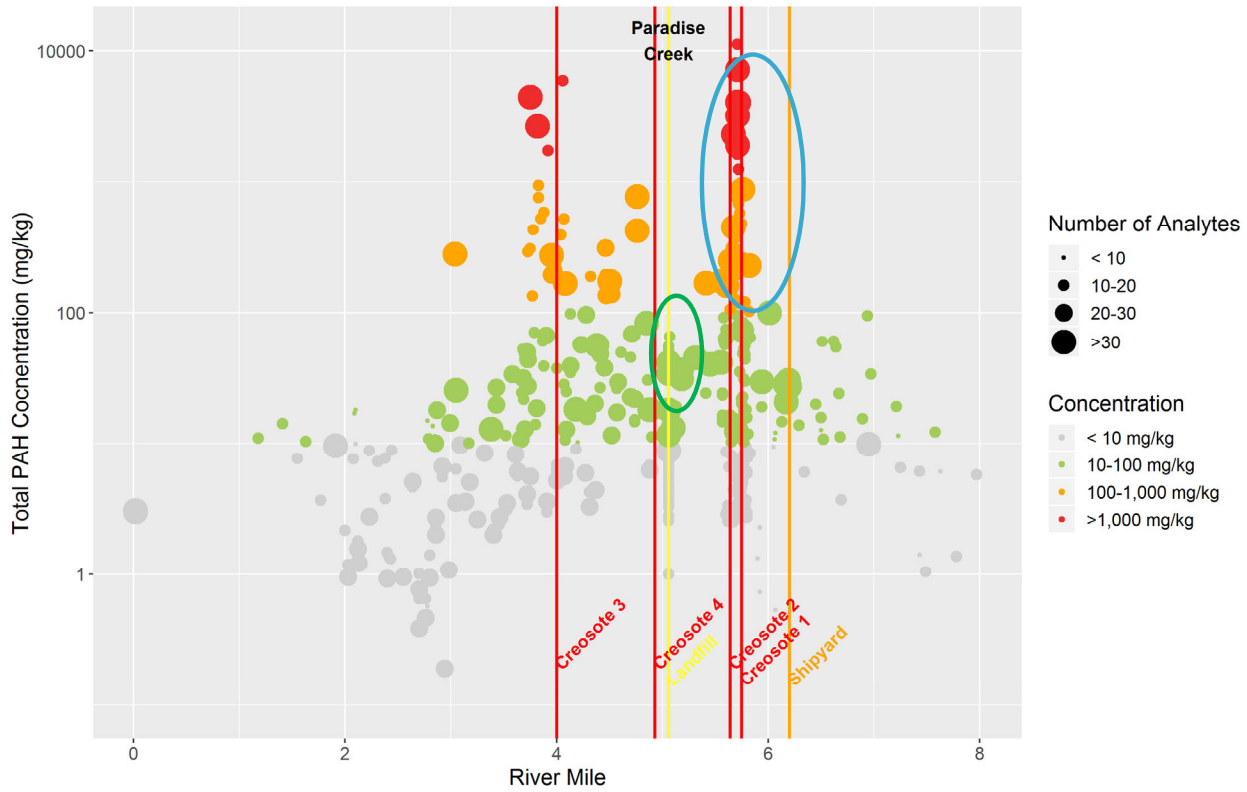
How Do Analyte and Censoring Issues Affect Forensic Analysis?

- There are multiple potential PAH contaminant sources in the Elizabeth River
- Analyte and censoring issues can result in drawing inappropriate connections between sources and analytical results



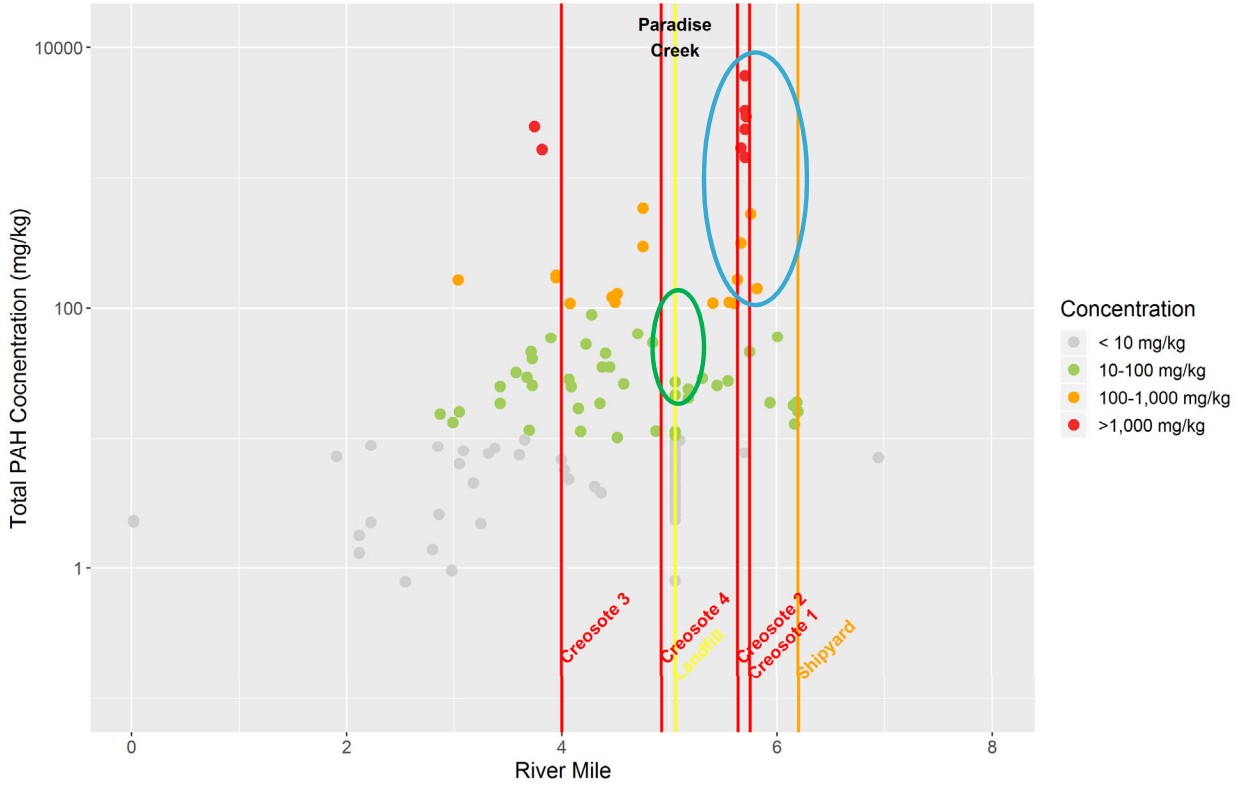
Larger Number of Analytes Results in Higher Total PAH Concentration

- Total PAHs calculated using all PAH analytes
- Often the larger the number of analytes measured, the larger the concentration of total PAHs



Standardization of Analytes Reduces Total PAH Concentration in Certain Areas

- Calculated total PAHs from U.S. Environmental Protection Agency's (EPA's) 16 priority pollutant PAHs
- Includes only samples analyzed for all 16 PAHs
- Elevated concentrations in Paradise Creek no longer present
- Samples near Creosote 1 and Creosote 2 sites with elevated concentrations still present

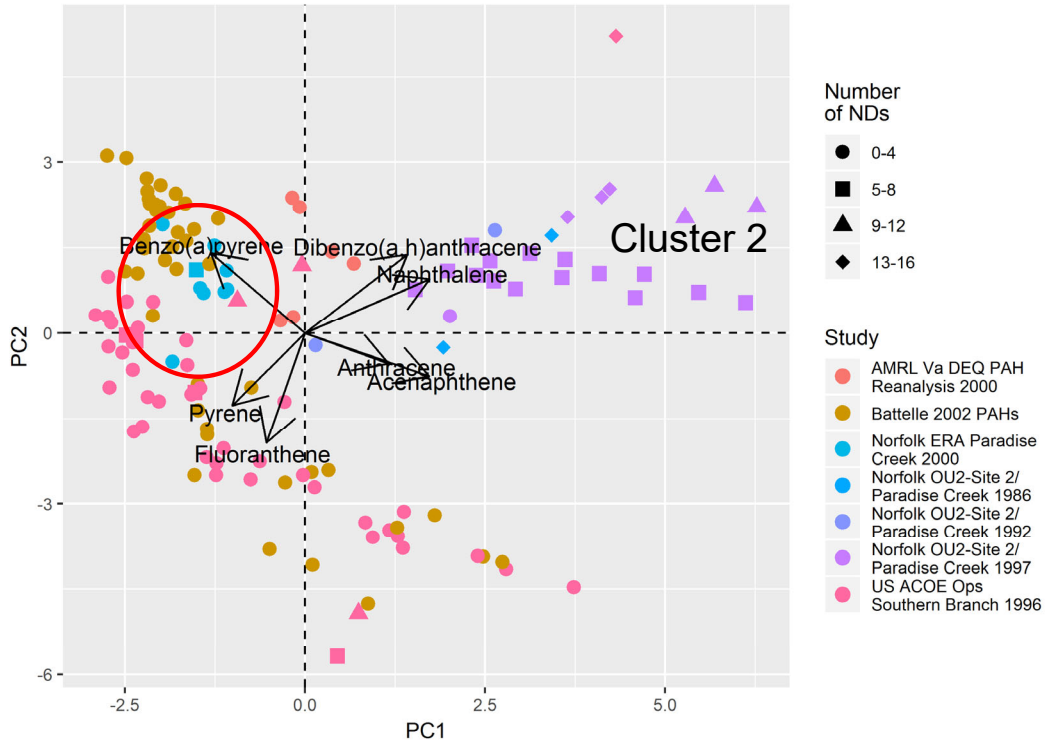


Highly Censored Samples Result in Samples Clustering Based on Number of Non-Detects (NDs)

- Principal component analysis (PCA):
  - 16 priority pollutant PAHs
  - Used half the MDL to represent NDs
  - Normalized by total PAH
- Clusters were identified using K-means clustering and the average silhouette method
- K-means analysis identified four clusters
- Results: Cluster 2 is mostly samples with high number of censored values

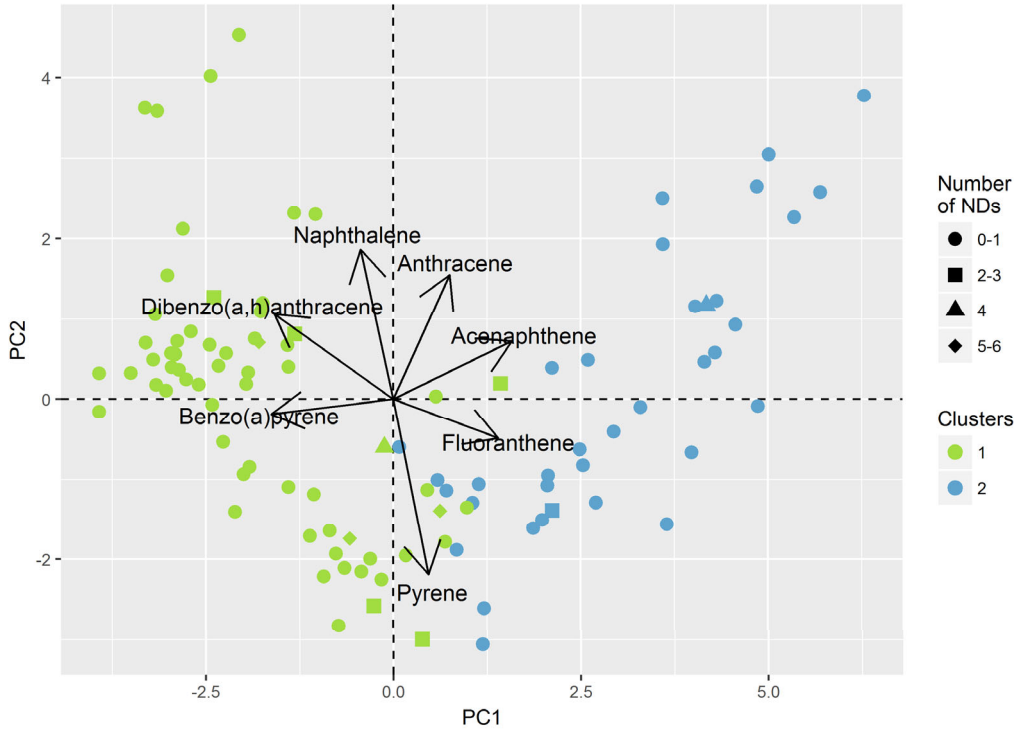
Cluster of Highly Censored Samples is Study Specific

- Cluster 2 is composed of 1986, 1992, and 1997 Paradise Creek studies
- 2000 Paradise Creek study doesn't cluster with other Paradise Creek studies
- 1986, 1992, and 1997 studies have high detection limits
- Cluster 2 likely represents differences in analytical methods and/or study design



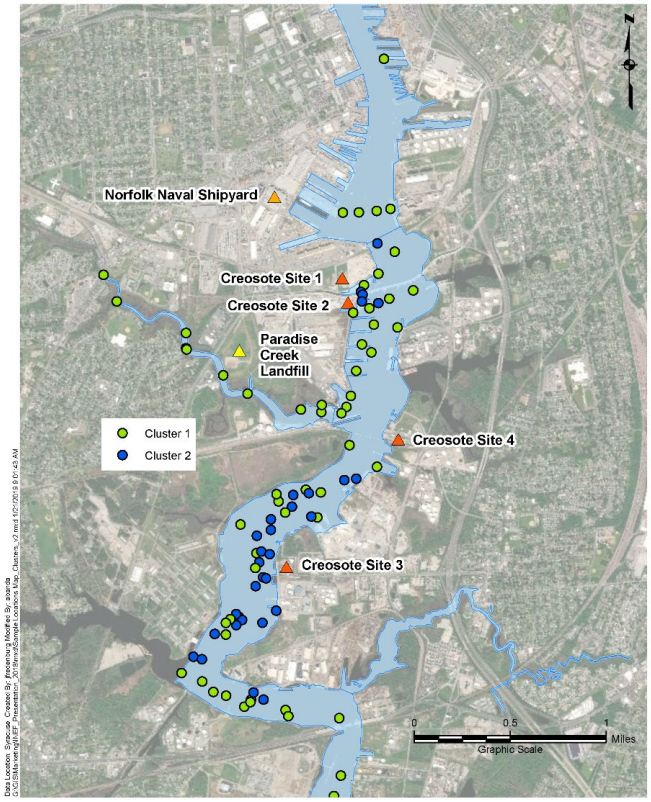
PCA and K-means analysis with Cleaned Data Set Identifies Two Clusters

- Updated PCA and K-means analysis use cleaned data
  - Excludes studies with high detection limits
  - Excludes samples with more than 7 analytes ND
- Results: Two clusters separated primarily by principal component (PC) 1



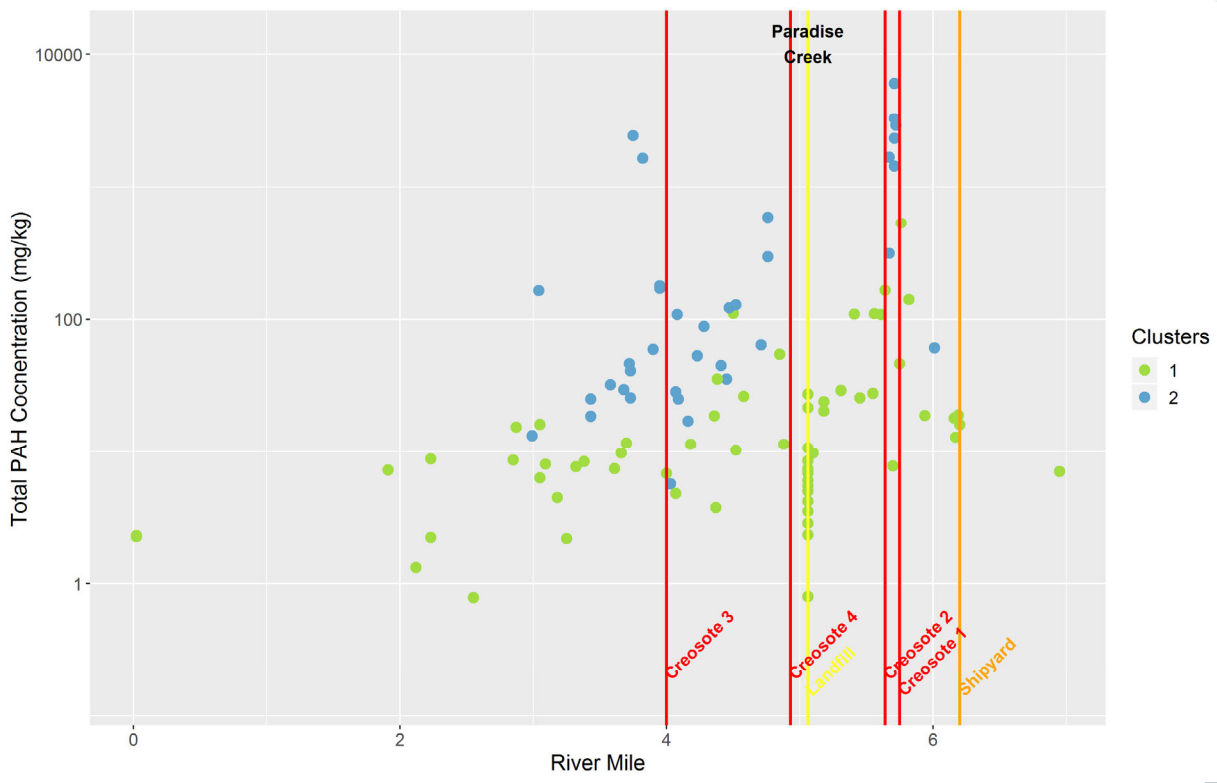
Samples in Cluster 2 Adjacent to Creosote Sites

- Cluster 1 samples in Paradise Creek are located near landfills
- Cluster 2 samples are often located near creosote treatment facilities
- PAHs in sediments near landfills appear distinct from PAHs in sediments near creosote treatment facilities



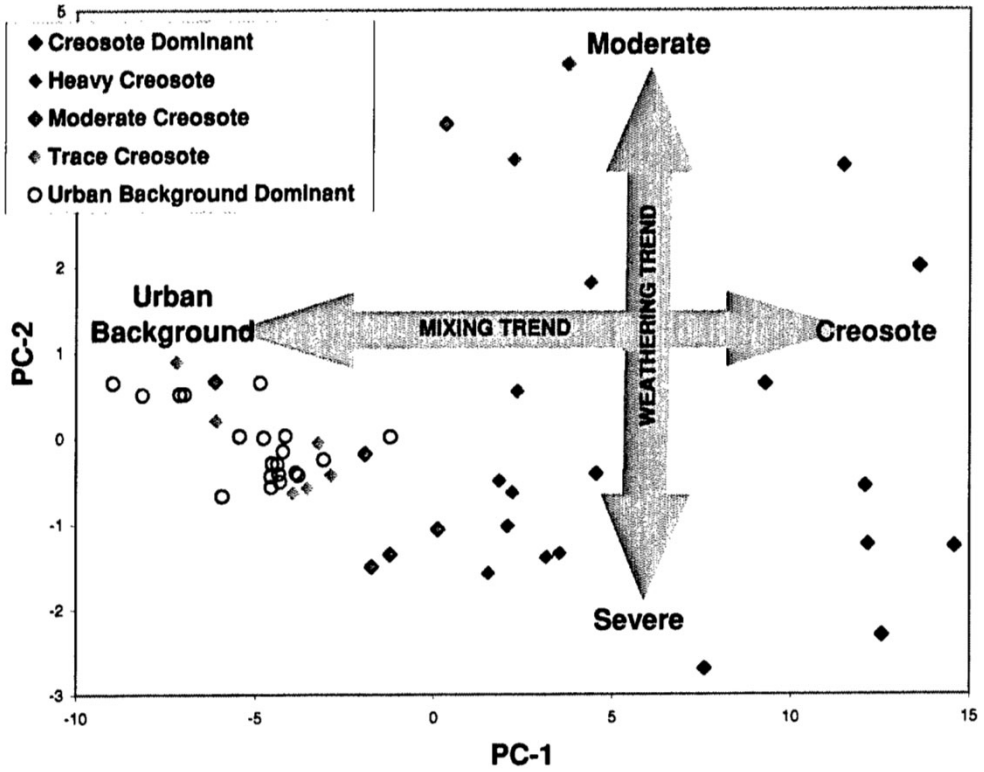
Cluster 2 Associated with Higher Total PAH Concentrations

- Highest total PAH concentrations are in Cluster 2
- Highest concentrations located near creosote sites
- Samples in Paradise Creek near landfill have low concentrations and are all in Cluster 1
- The landfill is likely not a major source of PAHs



Battelle Forensic Study also Identified Two Sources of PAHs

- Study focused on sources near Norfolk Naval Shipyard
- Identified creosote and urban background as primary inputs
- PCA of alkylated and non-alkylated PAHs
  - Background compared to creosote inputs
  - Weathering



Conclusions

- There are advantages to combining existing data sets when investigating sources and/or conducting allocation based on source identification:
  - Captures historical sources
  - Provides larger data set with more coverage
  - Cost of additional studies/analysis are prohibitive
- However, careful consideration must be taken to evaluate usability and ensure harmonization, including evaluating the following:
  - Issues arising from analysis of varying contaminants
  - Censoring of data sets
  - Variations in detection limits
  - Outliers

Thank You

Nicholas D. Rose

Senior Scientist

TIG Environmental

112 West 34th Street
18th Floor
New York, NY 10120

nrose@intell-group.com

+1 (212) 407-3405

www.intell-group.com

Timothy Negley

Principal Scientist

TIG Environmental

443 North Franklin Street
Suite 220
Syracuse, NY 13204

tnegley@intell-group.com

+1 (315) 401-7244

www.intell-group.com

Carlo Monti

Managing Director - Forensics

TIG Environmental

c/o BDO SA
Via G.B. Pioda 14
6901 Lugano
Switzerland

cmonti@intell-group.com

+39 0332 440905

www.intell-group.com