Evaluating the Challenges of Using Disparate Data Sets in Forensic Methods

Nicholas D. Rose (nrose@intell-group.com) (TIG Environmental, New York, NY, USA)
Timothy Negley (TIG Environmental, Syracuse, NY, USA)
Carlo Monti (TIG Environmental, Lugano, Switzerland)

Background/Objectives. Statistical methods are a valuable part of an environmental forensic investigation and can be useful in identifying sources of contamination or allocating costs. Due to removal of or lack of access to historic sources and/or cost limitations, forensic investigations may require the use of data sets that were not collected explicitly for this purpose. Combining data sets collected for different purposes, across multiple years, and by different entities presents challenges that must be carefully considered to ensure usability and to understand the limitations of the data. This work discusses some of the challenges involved in using disparate data sets for environmental forensic applications.

Approach/Activities. Sediment samples collected from the Elizabeth River in Virginia were obtained from the National Oceanic and Atmospheric Administration (NOAA) website. These sediment samples span much of the Elizabeth River and provide insights into potential source identification and design of future environmental forensic investigations. Because the samples were collected as part of 20 different studies between 1986 and 2002, they require a data usability evaluation and cleaning to harmonize the results. To investigate issues in the data set that could potentially skew the results of an analysis, the team used univariate statistical techniques and multivariate statistics, including principal component analysis and Kmeans clustering. This work highlights issues caused by multiple detection limits, censoring of data sets, and differences in the analytes tested for in different studies.

Results/Lessons Learned. The preliminary analysis highlights the multiple issues typically present in environmental data sets when attempting to use multivariate forensics methods. An analysis of polycyclic aromatic hydrocarbons (PAHs) in the data set, without an evaluation of usability, results in the identification of four clusters, one of which suggests a unique source of PAH contamination in Paradise Creek, a tributary of the Elizabeth River. However, a close evaluation of the data reveals that the cluster suggesting a unique source of PAH contamination is the result of studies where the detection limits for PAHs were high. Cleaning the data to account for this issue results in identification of only two clusters, which is consistent with other lines of evidence from investigations of the potential sources, including spatial location of known sources, river-mile plots, and previous forensic investigations. Results of the analysis of the cleaned data set, combined with the other lines of evidence, indicate that PAHs in Paradise Creek are the result of urban background that is also present in other parts of the Elizabeth River and not a separate source. Thus, analysis of the uncleaned data set could have resulted in inconclusive or erroneous conclusions regarding sources of PAH contamination in Paradise Creek. The results of this work emphasize that detailed evaluation of the data is required when using disparate data sets to ensure usability and harmonization.