PFAS Forensics through Applied Statistics: A Review of Case Studies in Chemometrics, Pattern Recognition, and Machine Learning

Zachary Neigh (zachary.neigh@aecom.com) and Julie McCurdy (julie.mccurdy@aecom.com)
(AECOM, Raleigh, NC, USA)
Rosa Gwinn PhD (rosa.gwinn@aecom.com) and Holly Brown (holly.brown@aecom.com)
(AECOM, Germantown, MD, USA)

Background/Objectives. To identify and characterize potential sources, the complex nature of per- and polyfluoroalkyl substance (PFAS) contamination requires a data-rich forensic fingerprinting approach, combining data science applications for pattern recognition with analytical chemistry techniques. We have developed and applied an exploratory, adaptive forensic investigation approach that maximizes the information extracted from PFAS analytical data using various statistical programming techniques from the fields of chemometrics, unsupervised pattern recognition, and machine learning. These powerful tools use the data at hand to reveal emergent patterns that are otherwise hidden by the limitations of our own perception.

Approach/Activities. The review is driven by a collection of PFAS investigative case studies, each highlighting the utility of various statistical techniques to differentiate PFAS sources and analyze their distribution in the environment. Exploratory data analysis is first performed on the PFAS analytical results to identify patterns in the data and to statistically define the relationships that differentiate key aspects of the PFAS samples. The exploratory results are used to further classify and group samples by their chemical signature, so that the entire dataset can be leveraged into powerful forensic conclusions supported by additional geospatial and statistical analyses. The specific methodology deployed depends on the types of data available, the results of the exploratory data analysis, and ultimately, the question that is intended to be answered. Statistical methods highlighted in this approach use anonymized case studies and public datasets include preprocessing and censoring, hierarchical clustering, k-means clustering, DBSCAN, principal component analysis, positive matrix factorization, multiple linear regression, k-nearest neighbors regression and classification, linear discriminant analysis, self-organizing maps, and kriging.

Results/Lessons Learned. Results are consistent with the growing body of research that demonstrates the utility of these methods for differentiating PFAS sources and extracting actionable insights on a wide range of applications. This data-driven approach allows the investigation to be more dynamic by refining the focus based on the impartial findings of the data analysis. For example, fingerprints identified in the targeted-PFAS analysis can then inform the sampling design for more specialized analyses, narrowing the focus to areas or types of PFAS mixtures of relevant forensic interest. This increases the value of the analytical data while optimizing the number of samples undergoing advanced laboratory analyses resulting in a cost reduction. The representative chemical compositions that form the signatures can be quantitively compared against established source signatures that are defined in the literature and run through AECOM's PFAS occurrence database for use in predictive machine learning algorithms.