EnviMetric: Findings from a Machine Learning Approach to Planning Site Characterization

Jason Dalton (jason.dalton@azimuth1.com) and Anna Harrington (anna.harrington@azimuth1.com) (Azimuth1, McLean, Virginia, USA)

Background/Objectives. In 2017 Azimuth1, with support from the National Science Foundation, researched the findings from thousands of contaminated site investigations. The purpose of this work is to leverage the knowledge gained from thousands of contaminated site investigations and remediation programs across the country and to determine if there are similarities and commonalities within subgroups of these sites that are meaningful and helpful for planning site characterization efforts and producing higher confidence results.

Approach/Activities. The data are compiled from public EPA and state agency data, and then augmented by the project team to include 3-D contaminant extent, soil, and groundwater characteristics. The data are then categorized according to contaminant type, soil conditions, groundwater conditions, climate, and age of the site. These categories of sites are then modeled using a machine-learning algorithm to determine the level of similarity among similar sites. This algorithm produces a statistical prior probability model automatically, which can be used to plan on-site data collection, or as a further line of evidence for site investigation. The statistical model is used much like an actuarial model would be used for insurance purposes. It does not intend to predict exactly the resulting contaminated zone, but it provides the weight of evidence from many previously observed cases, and serves as a starting point for further refinement. In practice, data from more than 100 prior sites are used to produce each new estimate, reducing uncertainty and providing confidence bounds on the extent or source of a known contaminant. Using many observations serves to filter out variation and determine if there are consistent migration patterns that can be used to guide a site investigation. This process reduces site revisits due to mischaracterized transport pathways, and ultimately will save time and money remediating the site.

Results/Lessons Learned. Our analysis has shown that using this machine learning approach reduces uncertainty and bias, creating an easier way to begin understanding a complex site. This presentation will show the savings in time and cost for collecting more ideally located sampling locations, and demonstrate the measures of internal consistency and back testing performed to validate the method and estimate levels of accuracy to be expected.