Applications of R Software and Machine Learning to Groundwater Data

Dustin Bytautas, P.E. (dustin.bytautas@aecom.com) (AECOM, Rocky Hill, CT)
Tomasz Kalinowski, Ph.D. (Syracuse, New York, USA)
Tauhirah Abdul-Matin, Matthew Panciera, PE, and Lucas Hellerich, Ph.D., PE (AECOM, Rocky Hill, Connecticut, USA)

Background/Objectives. Machine Learning is an approach to data analysis and model building where the details and parameters of the model are not explicitly specified by the analyst, but rather, trained by the computer through the use of computational power and large volumes of data. Machine Learning is commonly companion to 'Big Data,' a recently popular term describing datasets that require the use of specialized software because they are too large to analyze using conventional techniques. Driven by advancements in computers, internet technology and field sensors that allow for collection of datasets of unprecedented scale, there is currently a booming ecosystem of new ideas, tools, and techniques for conducting data analysis in the age of Big Data and Machine Learning. Here we discuss some of the benefits of these modern approaches to data analysis over conventional analysis toolchains, and present case-study examples where the tools have been applied to optimize remediation and monitoring programs at remediation sites.

Approach/Activities. R is a programming language and statistical data analysis environment where many Big Data and Machine Learning developments are occurring. Consequently, applications of Machine Learning to groundwater often involve the development of custom software in R. A few case studies will be presented of where R software was developed to process and analyze large datasets as part of environmental remediation and monitoring programs.

The approaches presented will focus on demonstrating how developing novel, quantitative metrics from analytical groundwater data can inform site-management decisions, including remediation technology selection based on past performance, predicting a site-specific likelihood of success of different remediation technologies, identifying opportunities for cost-savings through monitoring optimization, and prioritizing sites within a portfolio for remediation. Site metrics and statistics that allow for this must be unbiased by the overall size of the site, irregularities in the sampling frequency or well network distribution, destruction/creation of well locations over time, and other inconsistencies that typically inhibit cross-site comparisons in analytical groundwater data in environmental remediation applications.

In data-intensive efforts, interactive data visualizations are crucial aids that help both in conducting the data analysis and in communicating the findings. Some examples of interactive data visualizations of remediation-metrics built with R and Shiny are presented.

Results/Lessons Learned. The outcomes of the specific case studies presented were optimized remediation efforts. The lessons learned were that there are new capabilities to environmental remediation applications due to developments in the Big Data ecosystem. Modern development environments like R and Shiny allow for extremely rapid and low cost development of interactive data visualizations and other data-driven applications. The low-cost of development and deployment enables rapid and frequent rounds of iteration towards greater understanding, efficiency, communication, and automation.