Statistical Metrics for the Identification of Interdependant Analytes

Dr. Tiffany Downey (tiffany.downey@tetratech.com), Brian Caldwell, Ronnie Brito, Melissa Geraghty, and Rick Arnseth (Tetra Tech, Inc., Oak Ridge, Tennessee, USA)

Complex interactions between chemicals of concern (COCs) and geochemical species become critically important when evaluating the success of in-situ remediation techniques. However, these interactions are rarely obvious due to the multi-species correlations resulting from changes in pH or Oxidation-Reduction potential (ORP). Site investigations often rely on a broad suite of analyses to detect and characterize changes in groundwater quality as an indicator of remedial progress. Consequently, long-term remediation activities accumulate huge costs associated with monitoring activities, laboratory fees, and database maintenance. Many of the monitored analytes have little or nothing to do with the remedial activities and result in huge accumulated costs over the course of the project. Using well-established statistical analyses such as Pearson's linear relations and multi-variate Principal Components Analysis (PCA), the significance of each geochemical species can be evaluated to determine the overall contribution to site activities and the necessity of future monitoring. This study will present a simple iterative PCA process and standard metrics that can be employed at any site to assess the importance of each analyte in an effort to streamline future site activities.

Ongoing groundwater remediation efforts at the Iowa Army Ammunition Plant (IAAAP) provide an ideal case study to establish significant inter-species correlations at the field scale. In order to better understand the overall in-situ process and focus future sampling activities on relevant parameters, groundwater monitoring results were analyzed to determine if correlations exist between the site COCs and other geochemical analytes. For the purposes of this investigation, initial significance thresholds were set at $\alpha = 0.05$ for all Pearson's relations and eigenvalues greater than 1 for all PCA eigenvectors. Given uncertainties in temporal and spatial variability (ie the monitoring network included both source and downgradient indicator wells that could be expected to react differently and at different times), more rigorous significance criteria were not deemed appropriate. Datasets were initially screened for completeness, and all incomplete datapoints (i.e. unmeasured analytes) were eliminated from further inclusion. Results for analytes below detection limits were set equal to half the laboratory detection limit for all diluted samples, while the results for all undiluted analytes were set at one-half the lowest detection limit in order to minimize potential bias due to differences in instrumentation. Using a commercially available statistical analysis software, PCA was computed, including the calculation of squared cosine values for each analyte and eigenvector. Primary eigenvectors with eigenvalues greater than 1 were selected and the eigenvector best described by each analyte (i.e., eigenvector for which cosine squared is largest) were determined. Analytes whose highest squared cosine values were correlated with a single primary eigenvector were labeled as correlated analytes (thus establishing a dependency). Any vector best described by a single analyte or any analyte best described by a minor eigenvector were labeled as independent. The process was then iterated, excluding the independent analytes from previous iterations until all remaining analytes were correlated (i.e. dependent) within the primary eigenvectors. Results were compared to the Pearson's correlations to establish if any contradictions existed with simple uni-variate linear relationships. In concert, these statistical methods begin to highlight analytes that show evidence of interdependence.

Using well-established statistical techniques such as basic linear relations and PCA, complex interrelations between various chemicals of concern and geochemical parameters, as well as un-related or un-important species, can be established. Analogous statistical analyses at other sites can be used as a tool to eliminate superfluous sampling and focus future treatment and sampling efforts. Thus, project resources can be directed in the most productive and cost-effective manner.