AN EVALUATION OF METHODS FOR IMPUTATION OF MISSING TRACE ELEMENT DATA IN GROUNDWATERS

Bruce Dickson¹, Angela Giblin²,
Dickson Research Pty Ltd¹, CSIRO Honorary Research Fellow²

Introduction: Trace elements in groundwaters can be concentrated to unacceptable levels during evaporation of saline groundwaters into waste destined for disposal. In addition, saline waters provide a host for sulfate-reducing bacteria which can also lead to build-up of toxic metals in pond sediments. Because evaporation ponds are either in place or proposed as part of the solution to Australia’s groundwater salinity problems, we have been evaluating the potential for trace metal-related problems. For example, in the San Joaquin Valley, in California USA, such problems have resulted in the cessation of use of evaporation ponds due to high concentrations of Se, Mo and U in the sediments of the ponds and subsequent detrimental effects on wild fowl. In Australia, concentrations of U and Ra (radium) around salt-lakes have misled companies exploring for U into unwarranted expenditure. Groundwater data sets with the complete major cation and anion chemistry are widely available but data on toxic trace elements such as As, Cr, F, Se, U are much rarer. The study reported here investigated the use of data imputation methods to extend the small database of samples with multi-element analyses using the larger database of incomplete databases.

Methods: Two methods of imputation were evaluated in this study; self-organizing maps (SOM) and regularized expectation minimization (REM). SOM is a neural-net-like method that uses repeated learning to develop a “map” of the underlying trends in a data set. Once the map is obtained, replacements for missing values can be obtained from the best-matching map vectors for individual samples. The REM method, using an iterative regression, models between the missing data and available data, assuming an underlying Gaussian distribution in the data. Regularization assists in filtering the noise inherent in real data. SOM is also robust against noise and outliers in the data. In this study, a groundwater data set from NSW and Victoria was used. This set included U analyses that covered a wide range of U values (20% above 30 µg/L U). In addition, F analyses were included to indicate geochemical inputs from granites. Tests were made by setting a known fraction of U values as missing and then determining the accuracy with which they could be determined using the remaining data.

Discussion and Conclusion: Results show that neither method can accurately replace missing values in groundwater data. SOM-derived replacement values tend to fall within the range of the measured values in the original data whereas REM values can be widely scattered. For 25% of U data missing, SOM values were better by a number of measures than REM, but for 50% missing, this was reversed. We conclude that as U concentrations in groundwaters result from complex interactions involving groundwater, geology and mineralogy, the major composition of water alone is a poor indicator for U. There is no substitute for measuring U directly.

Poor imputation results found in this study for U should not be taken as a impugning imputation in general. Imputation is a valuable tool for solving various data problems and, through its model-free approach, is a useful addition to the array of imputation tools.