

**Ground-Water Status and Trends Analysis & Reports Meeting
San Antonio, TX
April 20-22, 2005**

EXTRAPOLATION SESSION GUIDANCE

Compiled by Bruce Lindsey, Wayne Lapham, Mike Rupert, Terri Arnold, and Tom Nolan

OBJECTIVES OF EXTRAPOLATION SESSION

- 1) Define extrapolation: **Prediction of water-quality conditions in areas not sampled.**
- 2) Provide examples of extrapolation at varying levels of complexity
- 3) Present potential pitfalls or things to watch out for and suggestions for presentation of extrapolation results to avoid misuse.

Issues to consider before considering extrapolation:

- How well do you understand the hydrologic framework where the sample came from?
- Is the area you are extrapolating to similar to the area where samples were collected: both lithologic/aquifer properties and hydrologic setting of sample.
- Do you know the temporal variation in water quality?
 - Relative to recharge events
 - Seasonal
 - Annual

Will your extrapolation represent susceptibility or vulnerability?

- Aquifer susceptibility: The ease with which a contaminant can reach the aquifer, determined by natural factors, and contaminant availability.
- Ground-water vulnerability: It is an estimate of the relative risk of ground-water contamination by a particular constituent, such as nitrate.
- Susceptibility map would be made only on basis of aquifer factors; vulnerability map would have to be made basis of aquifer factors and contaminant sources.

GUIDANCE ON EXTRAPOLATION

CLUSTER ANALYSIS (ALSO CALLED CLASSIFICATION ANALYSIS)

What is it: Cluster analysis is a statistical method used to define groups of objects with similar characteristics. Groups are formed such that objects within a group are more similar to each other than they are to objects in other groups. Most examples are from social or biological sciences.

Uses:

Cluster analysis could be used to explore relations between data points. It is best used to group data with similar characteristics. The clusters that are formed should be verified using as many different techniques as possible. If, through verification, the clusters are found to be meaningful, their meaning may be interpreted by assigning names or by identifying general characteristics that the groups represent.

Verification techniques include:

- 1) Run cluster analysis on random samples of the data and see if you get similar results
- 2) Assess predictive ability by selecting cluster-related variables that were not used to form the clusters to see if they vary across clusters.
- 3) Profile clusters using related data not used to form the clusters to interpret the meaning of the different groups. Techniques to profile the clusters include discriminant analysis or regression tree analysis.

- 4) See if similar clusters can be obtained using additional or fewer variables
- 5) See if similar clusters can be obtained using different clustering algorithms or different measures or dissimilarity.

Misuses:

Results of cluster analysis should not be used without verification of results. If you verify the clusters then you will know the proper way to use the information. Clusters can be formed using any combination of variables with as many or as few clusters formed as the user wants. Because clusters may not actually reflect the real structure of a data set, you may only be able to use results to describe your data.

Advantages:

- 1) Has graphical and numeric diagnostics to help select the best and most appropriate number of clusters
- 2) Easy way to explore the relations among data points using multiple variables.
- 3) Can be used as a predictive or as a descriptive tool
- 4) There are several different clustering algorithms and various methods of calculating dissimilarities between data points that can be tried to see which best fits the data.

Disadvantages:

- 1) Clustering solutions are not unique. This method is highly subjective. Clustering algorithms usually have several measures of how good the clusters are but the measures are all relative.
- 2) It takes many iterations and experimenting to get the appropriate variables and number of clusters.
- 3) You could have as many clusters as there are data points or as few clusters as 1. A cluster solution with only one group places all the data points in the same group and the group is not very homogenous. The more groups that are formed, the more similar the data points in each group are to each other.
- 4) Cluster analysis cannot be used to draw statistical inferences from a sample to a population. Clusters may change with the addition (or subtraction) of data.
- 5) Because of the nature of the clustering algorithms, clusters may not actually reflect a natural structure in the data; most clusters end up being round. Cluster solutions need to be verified.
- 6) Cluster analysis has no built in method for handling censored data

Additional References:

Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C., 1998, Multivariate data analysis: Prentice Hall, Upper Saddle River, New Jersey, 730p.

Everitt, Brian, 1980, Cluster analysis: John Wiley and Sons, New York, 136 p.

Romesburg, H. Charles, 1984, Cluster analysis for researchers: Lifetime Learning Publications, Belmont, California, 334 p.

Insightful Corporation, 2001, S-Plus 6 for Windows Guide to Statistics, Volume 2: Insightful Corporation, Seattle, Washington, 622 p

REGRESSION TREE ANALYSIS (ALSO CALLED DECISION TREE ANALYSIS)

What is it: Regression tree analysis is a statistical method used to explore structure in data using a single response variables and multiple predictor variables. It forms a binary tree of classification rules based on recursive partitioning.

Uses:

Regression tree analysis can be used to profile clusters as an alternative to discriminant analysis, logistic regression, or linear regression. It can be used to make prediction rules to assign data into groups.

Misuses:

Advantages:

- 1) Easy way to explore large multivariate data set
- 2) Can handle numerical as well as nominal data

- 3) Easy to interpret when predictor variables are both nominal and numerical
- 4) Can have missing values in predictor variables
- 5) Can prune and simplify the tree
- 6) At terminal nodes, the algorithm provides probabilities for each potential classification to show the how well the split predicted the response variable

Disadvantages:

- 1) Regression tree analysis has no built in method for handling censored data
- 2) If the tree is pruned (simplified) too much, misclassification rates will increase because variance increases
- 3) The recursive partitioning algorithm selects the “best” splits of predictor variables to get the most homogenous groups. There may be other splits that are possible and help the user to better understand the data. The user needs to use graphical tools to investigate splits.

Additional References:

Insightful Corporation, 2001, S-Plus 6 for Windows Guide to Statistics, Volume 2: Insightful Corporation, Seattle, Washington, 622 p.

GEOSTATISTICAL ANALYSIS—KRIGING

What is it: Kriging is a geostatistical analysis technique that interpolates values between data points based on the assumption that points nearer to each other are more similar than are points farther away.

Uses:

Ordinary kriging can be used to predict values for numerical data. Indicator kriging can be used to predict presence/absence of nominal data.

Misuses:

Kriging should not be used across large areas with no data points because the algorithm will try to predict values over the large area using very dissimilar points. This could create incorrect interpolations in areas with extremely sparse data points.

Advantages:

- 1) Kriging is good to use with sparse data points
- 2) Can be used to predict a value of one variable based on known relations with other variables
- 3) Can produce a standard error map to show how good the probability map is
- 4) There are good graphic tools like semivariograms to help select the best kriging model to interpolate the values
- 5) Kriging can be used to make contour maps as an alternative to inverse distance weighted (IDW) normal contours. IDW usually tends to create target shaped concentric rings around data points that don't fit well with their neighbors. Kriging smoothes the contours between data points and weights their values statistically by comparing the outside value to other nearest points.

Disadvantages:

- 1) does not have built in method for handling censored values
- 2) Kriging will predict values for any distance between points – it is to the user to realize when the data are too far apart. Distance depends on the scale of the mapping. If the distance between data points is great, then the lag search distance used by the kriging algorithm should also be pretty large

MULTIPLE LINEAR REGRESSION

- 1) The purpose of linear regression is to predict y given value of x. In the case of water quality, the purpose is to predict the concentration of a contaminant on the basis of explanatory variables.
- 2) Multiple linear regression is similar to linear regression, however the purpose is to predict y, given value of multiple explanatory variables.

Potential use of Multiple Linear Regression

- 1) Determine statistical relation between spatial data and water-quality variable.
- 2) Predict concentration of contaminant on the basis of explanatory variable(s)
- 3) Map predicted concentration, or establish procedure to predict concentration on the basis of explanatory variables at a given location.

Advantages of Linear Regression:

- 1) Linear regression produces a measure of 'goodness of fit' or r^2 value, which illustrates how well model fits data.
- 2) Plots can be visually assessed to determine model fit, and possible data problems.
- 3) Predicts specific value, model results can be compared to a threshold value if predicting a specific value is not desired.

Disadvantages of Linear Regression:

- 1) Producing a statistically significant model may be difficult, and may not explain much of the variation in the data.
- 2) Predicting a specific value may not be desired, or may imply more certainty to result than actually exists.
- 3) Linear regression cannot account for non-detects, which account for much of the data in NAWQA data sets.

Potential Misuses of linear regression:

Misuses would include violation of assumptions needed for linear regression:

- There is a linear relation between y and x
- Data are representative of population
- Variance is constant (doesn't change with values of x)
- Residuals are independent and normally distributed
- Outlier leverage may affect results

Multiple linear regression:

- Multi-collinearity of explanatory variables (2 variables account for the same information, such as land use and nitrogen load).

Summary:

- There are very few examples of using linear regression for extrapolation of ground water data
- Method doesn't handle non-detects well
- Assumptions may not be met for many of our data sets
- Useful for exploratory analysis

PRINCIPAL COMPONENT ANALYSIS

What is it:

PCA is a type of factor analysis, which is used to explain relationships among correlated, hard-to-interpret variables in terms of a few meaningful, relatively independent factors.

Uses:

PCA reduces large data sets to a few composite variables (principal components) that explain much of the variation in the original data. Extracted components are linear combinations of the original variables. Component loadings show the degree of correlation between the original variables and each principal

component. Variables that are highly correlated with a principal component are considered to represent a particular process influencing the data.

Misuses:

Advantages:

Data exploration

Data reduction

Extracted components can be used as variables in other methods, such as regression.

Disadvantages:

As an exploratory method, not well suited to confirming what is being measured by the data.

Interpreting principal components is a subjective process.

As the percent variance explained by successive extracted components decreases, they become more difficult to interpret.

Additional References:

Kleinbaum, D.G., L.L. Kupper, and K.E. Muller. 1988. Applied regression analysis and other multivariable methods. PWS-KENT Publishing Company, Boston.

Nolan, B.T., 1999, Nitrate behavior in ground waters of the southeastern USA: JEQ v. 28, p. 1518-1527.

Nolan, B.T., Ruddy, B.C., Hitt, K.J., and Helsel, D.R., 1998, Nitrate in ground waters of the United States: <http://www.nwqmc.org/98proceedings/Papers/63-NOLAN.htm>

LOGISTIC REGRESSION

What is it:

Logistic regression is a nonparametric statistical method that can be used to predict the probability of occurrence of an event of interest (such as detection of a contaminant in ground water) as a function of a set of independent variables (such as land cover and soils). Logistic regression is conceptually similar to multiple linear regression, but mathematically quite different.

Uses:

Originally developed to predict the probability of a patient getting cancer in epidemiologic studies. It is increasingly being used in environmental studies.

Misuses/caveats: (from Rupert, 2003, p. 31)

Probability maps developed by logistic regression portray the potential for contamination of ground water.

Probability maps do not show areas that are actually (currently) contaminated, but rather, the areas that have a potential (or likelihood) for being contaminated if a contaminant was released to the environment. More specifically, probability maps shows the probability of detection (in terms of a percent) of a particular chemical compound (the contaminant) in ground water. Probability is not the same as certainty; a well in a high probability area is not necessarily contaminated, because contamination can also depend on well depth and other local factors not taken into account by the models.

Probability maps are intended to be a first approximation at developing a consistent rating method for the entire Study Area. Additional site-specific data are needed before site-specific decisions are made, such as pesticide use restrictions. The most appropriate uses of these maps are to focus prevention programs in areas of greatest concern, to focus ground-water sampling programs in areas of greatest potential for contamination, and to assist educational programs for ground-water quality protection.

Advantages:

A major advantage of logistic regression over multiple linear regression is that the former is well suited for analysis of data sets with a large number of non-detections. Logistic regression is a nonparametric approach, so there is no need to transform data to a normal distribution. Also, logistic regression is insensitive to variations in constituent concentrations near the laboratory reporting limit because the data are transformed to detect-no detect prior to analysis by logistic regression. While building a logistic regression model the significant variables become apparent, and the insignificant variables can be discarded. The weighting and interaction of the independent variables is also calculated during construction of the models.

Disadvantages:

A disadvantage to logistic regression is that it is quite abstract and difficult to learn, and it can take a long time to construct a model because every possible combination of independent variables must be examined. It is rare to find a training class, and there are only a few textbooks that teach the subject (most focus on epidemiology).

Additional References:

Kleinbaum, D.G., 1994, Logistic regression, a self-learning text: New York, Springer-Verlag, 282 p.

Hosmer, D.W., and Lemeshow, Stanley, 1989, Applied logistic regression: New York, John Wiley & Sons, Inc., 307 p.

SELECTED REFERENCES ON APPROACHES TO EXTRAPOLATION

Overview of Ground-Water Vulnerability Mapping

Focazio, M.J., Reilly, T.E., Rupert, M.G., and Helsel, D.R., 2002, Assessing ground-water vulnerability to contamination: providing scientifically defensible information for decision makers: U.S. Geological Survey Circular C-1224, 33 p., <http://water.usgs.gov/pubs/circ/2002/circ1224/>.

Cluster Analysis (also called classification analysis)

Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C., 1998, Multivariate data analysis: Prentice Hall, Upper Saddle River, New Jersey, 730p.

Everitt, Brian, 1980, Cluster analysis: John Wiley and Sons, New York, 136 p.

Romesburg, H. Charles, 1984, Cluster analysis for researchers: Lifetime Learning Publications, Belmont, California, 334 p.

Insightful Corporation, 2001, S-Plus 6 for Windows Guide to Statistics, Volume 2: Insightful Corporation, Seattle, Washington, 622 p.

Regression Tree Analysis (also called decision tree analysis)

Insightful Corporation, 2001, S-Plus 6 for Windows Guide to Statistics, Volume 2: Insightful Corporation, Seattle, Washington, 622 p.

Geostatistical approach

- Hargrove, W.W., Luxmoore, R.J., 1997, A spatial clustering technique for identification of customizable ecoregions: ESRI User Conference Proceedings, paper 226, URL <http://gis.esri.com/library/userconf/proc97/proc97/to250/pap226/p226.htm> last accessed August 2004.
- Isaaks, E.H., and Srivastava, R.M., 1989, An introduction to applied geostatistics: Oxford University Press, New York, 561 p.
- Preston, Stephen D., 2000, Statistical identification of hydrochemical response units for hydrologic monitoring and modeling in Maryland: U.S. Geological Survey Water-Resources Investigations Report 00-4232, 8 p., <http://md.water.usgs.gov/publications/wrir-00-4232/html/>
- Wolock, D.M., Winter, T.C., McMahon, G., 2004, Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analysis: Environmental Management, On-line First (online only).

Linear Multiple Regression

- Battaglin, W.A., and Goolsby, D.A., 1996, Using GIS and Regression to estimate annual herbicide concentrations in outflow from reservoirs in the Midwestern USA, 1992-1993. AWRA Symposium on GIS and Water Resources, Sept 22-26, 1996, Ft, Lauderdale, FL.
- Evans, T.A., and Maidment, D.R., 1995, A Spatial and statistical assessment of the vulnerability of Texas groundwater to nitrate contamination, Center for Research in Water Resources Report 95-2, University of Texas at Austin, Austin, Texas, 358 p.
- Corwin, D. L., Carrillo, M. L. K., Vaughan, P. J., and Rhoades, J. D., 1998, Evaluation of a GIS-Based Model of Salt Loading to Groundwater, Proceedings of the 1998 ESRI user conference, 12 p.

Principal Component Analysis

- Kleinbaum, D.G., L.L. Kupper, and K.E. Muller. 1988. Applied regression analysis and other multivariable methods. PWS-KENT Publishing Company, Boston.
- Nolan, B.T., 1999, Nitrate behavior in ground waters of the southeastern USA: JEQ v. 28, p. 1518-1527.
- Nolan, B.T., Ruddy, B.C., Hitt, K.J., and Helsel, D.R., 1998, Nitrate in ground waters of the United States: <http://www.nwqmc.org/98proceedings/Papers/63-NOLAN.htm>

Wicoxon approach

- Rupert, M.G., 2001a, Calibration of the DRASTIC ground-water vulnerability mapping method: Ground Water, July-August 2001, vol. 39, no. 4, p. 625-630.
- Rupert, M.G., 1999, Improvements to the DRASTIC ground-water vulnerability mapping method: U.S. Geological Survey Fact Sheet FS-066-99, 6 p., <http://idaho.usgs.gov/public/factsheet.html/>.

Logistic Regression

- Erwin, M.L., and Tesoriero, A.J., 1997, Predicting Ground-Water Vulnerability to Nitrate in the Puget Sound Basin: USGS Fact Sheet FS-061-97, <http://wa.water.usgs.gov/pubs/fs/fs.061-97/>.
- Nolan, B.T., Hitt, K.J., and Ruddy, B.C., 2002, Probability of Nitrate Contamination of Recently Recharged Ground Waters in the Conterminous United States Environmental Science and Technology, v. 36, no. 10, p. 2138-2145. http://water.usgs.gov/nawqa/nutrients/pubs/est_v36_no10/.
- Rupert, M.G., 2003b, Probability of detecting atrazine/desethyl-atrazine and elevated nitrate concentrations in ground water in Colorado: U.S. Geological Survey Water-Resources Investigations Report 02-4269, <http://water.usgs.gov/pubs/wri/wri02-4269/>.
- Tesoriero, A.J., and Voss, F.D., 1997, Predicting the probability of elevated nitrate concentrations in the Puget Sound Basin-Implications for aquifer susceptibility and vulnerability: Ground Water, v.35, no.6, p.1029-1039.

GW flow modeling and particle tracking

- Snyder, D.T., Wilkinson, J.M., and Orzol, L.L., 1998, Use of A Ground-Water Flow Model With Particle Tracking to Evaluate Ground-Water Vulnerability, Clark County, Washington: USGS Water-Supply Paper 2488, 63 p., 1 table, 16 figures, 3 appendices, <http://pubs.er.usgs.gov/pubs/wsp/wsp2488>