Characterization of Groundwater Hydrology for Nitrate Contamination Analysis

Nathan Potratz[1]
Ming Ye[1],Huaiwei Sun[1],Richard W. Hicks[2]
[1]Department of Scientific Computing, Florida State University
[2]Florida Department of Environmental Protection, Tallahassee, FL, USA Tallahassee, FL, USA
E-mail: Nathan.Potratz@gmail.com, mye@fsu.edu, hsun4@fsu.edu, Richard.W.Hicks@dep.state.fl.us

Abstract One of the primary problems that scientists encounter when attempting to evaluate and ensure groundwater quality is the difficulty and expense of obtaining sufficient data for hydrological modeling. Consequently, scientists frequently desire to minimize the number of monitoring wells that are necessary to obtain sufficient data for their modeling purposes. They desire to increase their efficiency by extrapolating the data from one set to others that are similar. Frequently this is done subjectively based on field expertise. This study seeks to investigate more mathematical approaches to the process. In this study correlations between nitrate concentrations and hydrogeological conditions have been investigated. Quarterly data from 37 monitoring wells from 2003-2010 in Jacksonville, FL are considered. They are clustered using the K-means method based on characteristics of hydraulic conductivity, drainage type, septic tank density, and elevation difference within a 1000 ft. radius zone around each well. The overall purpose of the study is to determine how much nitrate contamination can be reduced by eliminating septic tanks in favor of a central sewage system in Jacksonville.

Background

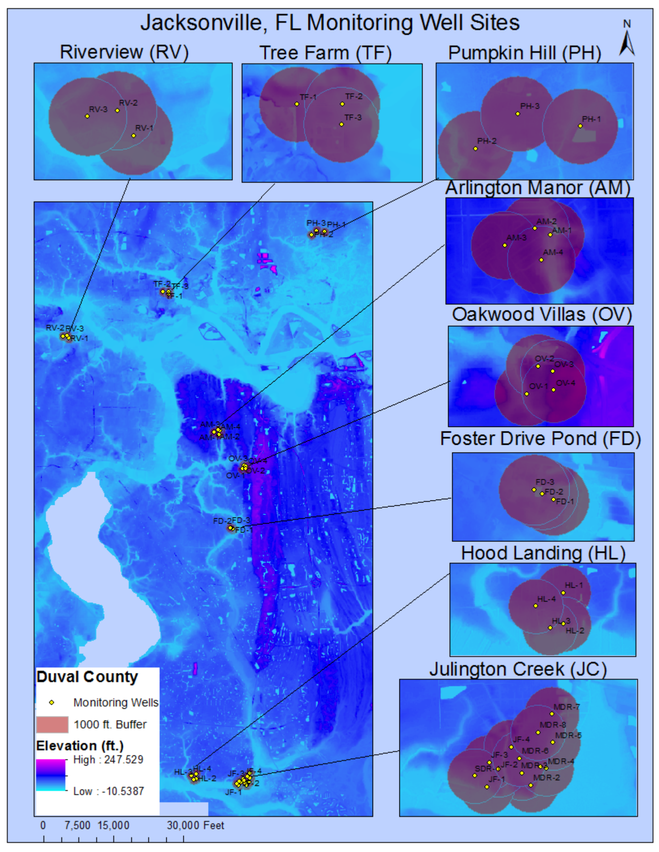
Nitrate is among the most prevalent groundwater contaminants in the United States. It has various adverse effects on both human health (primarily in infants) as well as the environment. Septic tanks are among the most significant sources and are easier to control than some of the other major sources such as fertilizer. The city of Jacksonville, FL is currently considering replacing many of the septic tanks in the city by a central sewage system in order to reduce the nitrate concentration in their water. They want to estimate how much they can expect the nitrate contamination to decrease if they do. The software ArcNLET, a Nitrate Load Estimator Toolkit is being used for the model. The city of Jacksonville desires to ensure accurate results for the nitrate modeling. It is difficult and expensive to obtain groundwater data, so the project is now in the process of selecting various sites at which to place monitoring wells in order to obtain data that can then be used for other similar areas. This study is concerned with characterizing sites into clusters with similar concentrations of nitrate and ammonium based on hydrogeological characteristics that are known all throughout the city. This clustering will help guide site selection process for building monitoring wells.

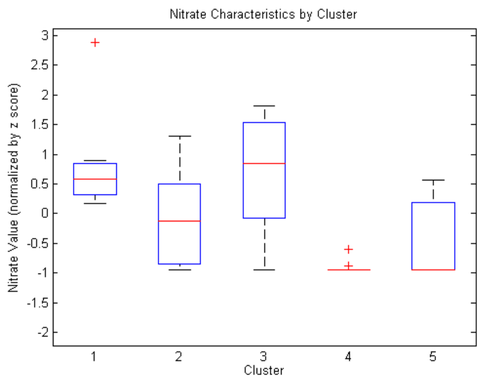
Groundwater Hydrology

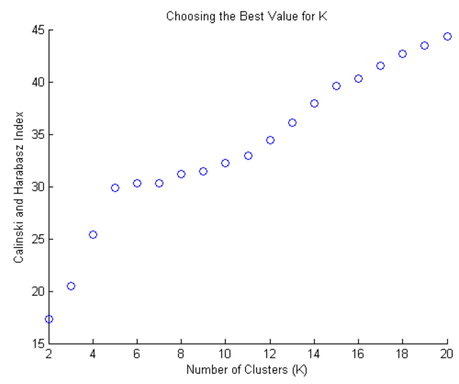
Hydrology is literally the science of water, and groundwater hydrology specifically encompasses the study of the occurrence and movement of water beneath the surface of the earth and its relationship with the living and material components of the environment. The primary governing equation for groundwater hydrology one needs to know to understand this study is[2]:

$$q = \frac{Q}{A} = -K\frac{dh}{dl} \tag{1}$$

where q is called the specific discharge (or Darcy velocity), Q is the total discharge of groundwater through a small section of porous space in a certain amount of time, A is the cross-sectional area of the porous space through which the groundwater flows, K, called the hydraulic conductivity, is a constant of proportionality, and $\frac{dh}{dl}$, called the hydraulic gradient, is the rate of change of the hydraulic head in the direction of flow. Note that The hydraulic head is the mechanical energy per unit fluid weight, and groundwater flows from a point of higher hydraulic head to a point of lower hydraulic head. Specific discharge is therefore defined as the discharge per unit cross-sectional area of flow through a porous media. Specific discharge provides a measure of how much groundwater flows through a particular region of soil in a certain amount of time.







Spearman Rank Correlation Coefficients

	Elev Min	Elev Max	Elev Diff.	Elev Median	Sept. Tank Dens	Hydraulic Cond.	Porosity	Drainage
Nitrate	0.23	0.65	0.61	0.52	0.55	0.42	0.21	0.33

Acknowledgment

This work is supported by contract WM993 with the Florida Department Environmental Protection (FDEP).

Clustering with K-Means

Clustering is a way by which one says that certain things are like each other and unlike certain other things (see Larose[3] and Xu[6]). For a given set of data points $\mathbf{x}_i \epsilon \mathbb{R}^d,\ i = 1, ..., N$, K-means places each data point into one of k clusters, where k is specified by the user beforehand. The algorithm starts with k initial points, called centroids (or means), denoted $\mathbf{m}_l \epsilon \mathbb{R}^d,\ l = 1, ..., k$. Each of the data points $\mathbf{x}_i$ are considered part of cluster C_l if they are closer to $\mathbf{m}_l$ than any other centroid. Euclidean distance is used to measure distance between a data point and a centroid. The following equation describes this step[6]:

$$\mathbf{x}_j \epsilon C_l,\ if\ \|\mathbf{x}_k - \mathbf{m}_l\| < \|\mathbf{x}_j - \mathbf{m}_i\| \quad \text{for } j = 1, ..., N,\ i \neq l, and\ i = 1, ..., k \tag{2}$$

The algorithm recalculates each $\mathbf{m}_l$ so that $\mathbf{m}_l$ is equal to the mean of all of the data points that are in cluster C_l. The algorithm continues until the difference between the means at the current step and the means at the previous step is less than a given tolerance.

Two issues need to be resolved when using K-means: choosing an initial partition for which the method converges to the global optimum, and choosing the number of clusters, k. K-means can be shown to converge to a local optimum clustering solution, but cannot be guaranteed to converge to the global optimum[5]. The choice of the initial partition directly affects whether the method converges to a local optimum or the global optimum. There are a various methods for generating an initial partition of the data assuming no prior knowledge of the structure of the data. However, no method can guarantee convergence to a global optimum[5]. Thus, it is common practice to run K-means many times for a given problem in order to ensure that one has obtained the global optimum. For this study, the clustering for a given k value was run 5000 times with generated initial partitions. Choosing a value for k is also nontrivial. Better clustering solutions have greater similarity between data points that are within the same cluster and greater dissimilarity between data points that are in different clusters. The most common criteria used to measure how good a particular k-means solution is, and the best-performing based on a study by Milligan and Cooper[4], is the Caliński and Harabasz (CH) index[1], which is defined as,

$$CH(K) = \frac{Tr(\mathbf{S_B})}{k-1} \Big/ \frac{Tr(\mathbf{S_W})}{N-k} \tag{3}$$

where N is the total number of objects, and $Tr(\mathbf{S_B})$ and $Tr(\mathbf{S_W})$ are the trace of the between and within-cluster scatter matrix, respectively. The between and within cluster scatter matrices are defined as[6],

$$\mathbf{S_B} = \sum_{i=1}^{k} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \text{ where } \mathbf{m} = \frac{1}{N}\sum_{i=1}^{k} N_i\mathbf{m}_i$$

$$\mathbf{S_W} = \sum_{i=1}^{k}\sum_{j=1}^{N} \gamma_{ij}(\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T, \text{ where } \gamma_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \epsilon C_i \\ 0 & \text{otherwise} \end{cases}$$

The trace of the within-scatter matrix is a value representing how similar values of each cluster are to one another. The trace of the between-cluster scatter matrix is a value representing how dissimilar values of each cluster are to values of different clusters. A larger CH index indicates a better clustering solution. Clearly a clustering solution that places each data point in its own cluster would result in the greatest amount of difference between clusters and the smallest difference within a cluster. However, this result is useless for the user. In fact a smaller value of k is often more desirable, but this does result in a decrease in the other qualities desired from the clustering solution i.e. the CH index. Thus, a value for k is sought that improves the CH index from lesser k values substantially but for which larger values of k do not improve the CH index very much[5]. The basic idea is to find the "elbow" point of a function that models the CH index as a function of k.

Results and Conclusions

It is worth mentioning that variables for each data point were normalized by using z-score transformations so that the results are independent of the scales for the variables[3]. K-means was run with 37 data points, one for each monitoring well, using four variables: hydraulic conductivity, septic tank density, drainage type, and the difference in elevation between the maximum elevation and minimum elevation. For each of the monitoring wells these variables are based on the area within 1000 ft. of each well. These variables and this distance were chosen after analysis of the Spearman rank correlation between the possible variables and the nitrate concentration at the following distances: 100 ft, 500 ft, 1000 ft, and 2000 ft. The Spearman rank correlation coefficients are provided in the table below, with a range of $[-1, 1]$. K-means was run for various values of k, and the CH index is plotted for the first 20 values of k. Based on this plot, the elbow point is found at $k = 5$, so this values was selected for the clustering results. The box plot indicates the characteristics of the nitrate concentration within each cluster. The clusters have some overlap in their range, but the median and range for each cluster is fairly distinct, indicating that the clustering based on the hydrogeological characteristics provides useful results.

Summary and Future Work

The results of the clustering have shown that sites can be grouped based on hydrogeological characteristics with the resulting clusters implying similarity in nitrate concentration for sites in the same cluster. In the coming stages of this study, the further implications of these results will be analyzed, and providing weights by which certain variables have a greater impact on the clustering process will be explored. Other clustering techniques will also be used on the same data to try to gain additional insight into the structure of the data.

References

[1] Caliński, R., and Harabasz, J. (1974), A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3, 1–27.

[2] Hornberger, G.M., Wiberg, P.L., Raffensperger, J.P., and Eshleman, K.N., *Elements of Physical Hydrology*, John Hopkins University Press, Baltimore, MD, 1998.

[3] Larose, D.T., *Discovering Knowledge in Data An Introduction to Data Mining*, John Wiley & Sons, Inc., Hoboken, NJ, 2005.

[4] Milligan, G., and Cooper, M. (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, 46, 159–179.

[5] Xu, R., and Wunsch, D. (2005), Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, 16(3), 645–678.

[6] Xu, R., and Wunsch, D. , *Clustering*, John Wiley & Sons, Inc., Hoboken, NJ, 2009.