Description: This course is designed to offer students an introduction to the management and mining of environmental data, primarily chemical and biological data collected at both contaminated sites and unimpacted locations. We will focus on datasets that are intermediate in size: large enough that they cannot be handled using a simple spreadsheet, but not so large that they require extensive computing power. Topics covered will include Microsoft Access, database structures, multiple linear regression, non-linear curve fitting, time series analysis, spatial relationships, and factor analysis. Students will complete a project analyzing and mining their own datasets.

Prerequisites: None, but students are expected to be proficient with Microsoft Excel.

Grading:
1. Attendance and Participation (10%)
2. Four short reports applying topics covered in lecture to a dataset of your choice (10% each)
   a. Apply what you learned in class to evaluate the robustness of various databases. What works, what doesn’t, what would you do better?
   b. After selecting a dataset from literature or your own research, describe the data, apply multiple linear or non-linear regressions or time series analysis. What trends or relationships are present, were these what you anticipated, what does this tell you about the nature of the dataset?
   c. Using topography or bathymetry data, create a map. Overlay data showing other parameters onto this map. Describe any observations you make from this overlay? Do observations correlate or cluster around geographic features? Discuss your observations.
   d. Using the same dataset from report 2, apply multivariate analyses (PCA, RDA, NMDS, PMF 5.0, cluster analysis, etc.). What trends or relationships are present, were these what you anticipated, were they consistent between techniques? Discuss your observations.
3. Paper review – read and provide a “News and Views” style response to a scientific paper (10%)
4. Abstract and Annotated Literature Review (10%)
   a. A paragraph detailing the topic of your paper, and an annotated literature review that will include at least 10 references with several full sentences describing each of the documents.
5. Final paper (30%)
   a. Write a research paper using the techniques discussed in class for creating, managing, and mining databases for insights into environmental process. This should include an introduction on your topic, a methods section, results from your data mining including figures and tables, and a discussion interpreting your results.

PLAGARISM: Plagiarism is taking credit for someone else’s work whether deliberately or unintentionally. This includes but is not limited to turning in all or part of an essay written by someone other than yourself (a friend, an internet source, etc.) and claiming it as your own, and including information or ideas from research material without citing the source. You will receive an automatic 0 for plagiarized work, and may be reported to the Office of Academic Integrity.
Schedule:

Course periods:

1. What is data? Introduction to the course, playing with MS Excel
2. Metadata
3. Querying online databases (STORET, NOAA Diver, USGS, IODP, and others)
4. Databases – example of a ‘good’ database and comparison to a ‘bad’ database
5. More on ‘good’ databases-querying and modifying databases
6. Making your own databases
7. Electronic Data Delivery
8. Multiple linear and non-linear regression and time series analysis
9. More on time series analysis
   Break to work on report #2, class resumes on October 28
10. Non-linear curve fitting with Excel
11. Non-linear curve fitting with R
12. Spatial data; making simple maps
13. Maps and maps with overlaid data
14. Factor analysis: PCA, RDA, NMDS, Cluster Analysis
15. Factor analysis: PMF 5.0
16. Remainder of semester - Work on independent projects – due Dec 10th