Objectives:
1. To introduce multivariate data analysis methods (e.g., graphic exploratory analysis, ordination, cluster analysis) commonly used in ecology and environmental studies.
2. To introduce R, a software used for statistical computation, graphics, and programming.

Summary of the course:
This course is designed for students who collect and analyze multivariate data (e.g., biological data, environmental data, or both). The datasets are usually complex, bulky and noisy with internal relationships among variables and usually with outliers. Multivariate data analyses will allow the students to effectively summarize complex data and detect underlying patterns. These analytical methods, often exploratory in nature, can generate hypotheses on complex systems in which experimental manipulations may not be feasible or not always the first option due to practical consideration (e.g., costs or scales). I expect that the students will understand why, when, and how each method can be used to address their research questions and eventually use some of these methods to their own researches.

Recommended reading materials:
- Legendre & Legendre 2012. *Numerical Ecology*. 3rd edition, Elsevier (If you will use numerical analyses for your research, this is a highly recommended reference book which provides a comprehensive coverage on the subject)

Recommended prerequisites:
ESM 556 Environmental Data Analysis or college-level statistics. A basic understanding of regression, especially multiple regression, and linear algebra will be very helpful.

Software:
- R (free downloadable from <http://cran.stat.ucla.edu/>). For the basic R tutorial to get a start with R, please go to <http://www.cyclismo.org/tutorial/R/> or You can go to Youtube <http://www.youtube.com/> and search for “R tutorial”

• **D2L**: an on-line learning system ([https://d2l.pdx.edu/](https://d2l.pdx.edu/)). You need to use your ODIN user name and password to log in. Class materials such as syllabus, homework assignments, lecture powerpoint presentations, and extra readings will be posted in “D2L”. Students are encouraged to use “D2L” to post questions, comments, and suggestions. **D2L’s email system works in a mysterious way and I strongly recommend that you don’t use D2L’s email system.**

**Approach:**
The term will be divided into three phases:
1. The first part of the term (6 weeks) will focus on introduction of commonly used multivariate methods. This part will include lectures, in-class exercises, and homework.
2. The 2nd part of the term (2 weeks) will emphasize on student-led research projects. The lecture will be condensed with no more homework. Each research group will have plenty time during the class to discuss their research ideas, analyze the data, and interpret the results.
3. The 3rd part of the term (2 weeks) will be largely on applications of some multivariate methods in environmental sciences. Each class period will be organized in the same way as a professional conference. Each group will present their group research projects followed by questions and discussion.

**Peer-evaluation:**
The class emphasizes tremendously on team-work and student-based learning. To be fair with every member of the team, each member will have a chance to evaluate their peers’ performance at the end of the term. The outcome of the peer evaluation will affect a student’s final grade.

**Grades:**
- Homework (3-4 homework exercises: 60%): Late homework will be accepted but will suffer a 10% per day grade reduction.
- Project (35%): You are required to formulate a study question and a conceptual model, collect/”borrow” data, analyze the data and interpret the results with relation to the study question, and write a professional research paper.
- Class participation (5%): Class participation includes class discussion, class presentations, and on-line discussion

**Research Project**: Each group is required to identify a dataset which is suitable for the multivariate data analysis covered by this course. It is preferred that the students use their own research datasets. Each group will then formulate research questions/objectives, construct a conceptual model, select appropriate multivariate data analyses, and perform
the analyses on the datasets. The evaluation of the project is based on (1) professional conference-style Powerpoint presentation (2) professionally written journal-style report.

Helpful Websites:

- An ordination website includes some useful information on ordination, software, and other useful links (<http://ordination.okstate.edu/>)
- Another website includes information on ordination and cluster analysis (both R scripts and examples in vegetation ecology) (<http://ecology.msu.montana.edu/labdsv/R/labs/>)
- There are many R-based multivariate methods available. We will introduce some of these methods during the class. You may find it informative if you go to this website (<http://cran.r-project.org/>, click “Task Views” on the left side, under the title of “Cran Task Views”, click “Multivariate”. Paul Hewson has kindly provided an overview of available statistical software which can be used by R. In addition, you may check “Graphics”, “Cluster” and “Machine Learning”.

On-line Resources:

- Excellent on-line lectures on linear algebra including eigenvalues and eigenvectors, taught by a MIT professor and the author for a linear algebra textbook (<http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/>)
- PCA: An good introduction of PCA with several video lessons (Chemometrics) (<http://www.youtube.com/playlist?list=PLBC24FD8C389FE9E4>)
- R-Bloggers: R news and tutorials contributed by 565 R bloggers (<http://www.r-bloggers.com/>)
- Quick-R (<http://www.statmethods.net/>)

Tentative Course Outline

Both lecture and workshop topics will be subject to changes depending on students’ interests and their data sets.

- Biological/environmental data and multivariate analysis
- Know your data: Data manipulation and summary using dplyr and graphic analysis using ggplot2
- Supervised learning vs. unsupervised learning
- Unsupervised learning method I: Ordination
  a. Eigenanalysis-based ordination: Principal Component Analysis (PCA)
  b. Distance-based ordination: Non-metric Multi-Dimensional Scaling (NMDS) and metric Multi-Dimensional Scaling (Principal Coordinate Analysis (PCoA))
- Supervised learning method I: Linking one data matrix (e.g., biota) to another (e.g., environment)
  a. Redundancy Analysis (RDA)
  b. Canonical Correspondence Analysis (CCA)
  c. Partial constrained analysis
- Compare two ordinations: Mental test, Procrustes analysis
- Testing differences between groups of samples based on distance measures: Analysis of Similarities (ANOSIM), Permutation Multivariate Analysis of Variance (PERMANOVA), and Multiple Response Permutation Procedure (MRPP)
- Unsupervised learning method II: Cluster analysis
  a. Hierarchical agglomerative cluster analysis
  b. k-means partitioning
  c. Self-organized map
- Supervised learning method II: Link clusters (categorical variable) to a data matrix (predictors)
  a. Classification trees
  b. Bagging classification trees (Random Forests)
  c. Boosting classification trees

Figure 1. A schematic diagram showing steps in an ordination analysis (modified from Clarke and Warwick 2001)

Figure 2. Using Self-Organized Map (SOM) to classify fish assemblages.