ESM 567 Multivariate Analysis of Environmental & Biological Data

(CRN: 44844, 4 credits)

SRTC B1-82

M & W: 11:30 to 13:20

Instructor: Yangdong Pan (SRTC Room 218, 5-8038, pany@pdx.edu)
Teaching assistant: Amy Ehrhart (email: aehrhart@pdx.edu)
Office hour: Monday and Wednesday 13:30-14:30 pm or by arrangement

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:

- Gotelli, N. J. and A. M. Ellison. 2013. *A Primer of Ecological Statistics*. Sinauer Associates, Inc. Publishers, Sunderland, MA. (2nd edition) (**required**)
- Legendre & Legendre 2012. *Numerical Ecology*. 3nd 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)
- Borcard, D., F. Gillet, and P. Legendre. 2011. *Numerical Ecology with R*. Springer. (A companion book for "*Numerical Ecology*". This book contains rich R scripts for multivariate analyses in ecology and environmental science)

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"

- *RStudio*: a text editor for R and others (free downloadable from <http://rstudio.org/download/desktop>). For more helpful documents on using R Studio, please go to <http://rstudio.org/docs/> and you can also watch a 2-minute Screencast on the RStudio website <http://rstudio.org/>.
- *D2L*: an on-line learning system (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.

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 (3 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 (1 week) 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.

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.

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

Peer-evaluations:

1. 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.

2. Each student will have a chance to evaluate each group's oral presentation during the week 10 using Google Form and all peer comments will be available so that the presenters can incorporate the peer comments/suggestions in their final written papers.

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 "*MachineLearning*".

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/>
- Two excellent on-line short videos on linear algebra with an emphasis on geometric view of linear algebra AND animation <https://www.3blue1brown.com/essence-of-linear-algebra-page><https://www.youtube.com/watch?v=ZKUqtErZCiU&list=PLHXZ9OQGMqxfUl0tcqPNTJsb7R6BqSLo6>
- PCA: A step-by-step introduction of PCA <https://www.youtube.com/watch?v=_UVHneBUBW0>
- R-Bloggers: R news and tutorials contributed by hundreds R bloggers <http://www.r-bloggers.com/>. You may sign up so that you will be informed about news and tutorials on R via email.
- 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.

Week	Topics
	Introduction
1	Biological/environmental data and multivariate analysis Know your data: Data manipulation and summary using *dplyr* Graphic analysis using *ggplot2*

Ordination: Put things in order (ch.12, p.406-428)

2 Eigenanalysis-based ordination: Principal Component Analysis (PCA) (**ch.12, p.406)**

3 Linking one data matrix (e.g., biota) to another (e.g., environment): Redundancy Analysis (RDA) and variation partition (**ch.12, p.438)**

4 Distance-based ordination: Non-metric Multi-Dimensional Scaling (NMDS) **(ch.12, p.425)**, Multi-Dimensional Scaling (MDS: Principle Coordinate Analysis)(**ch.12, p.418)**, and Linear vector fitting

5 Testing differences between groups of samples based on distance measures (**ch.12, p.387)**:
   a. Analysis of Similarities (ANOSIM)
   b. Permutation Multivariate Analysis of Variance (PERMANOVA)
   c. Multiple Response Permutation Procedure (MRPP)

Classification: Put things in groups (ch.12, p.429-437)

6 Cluster analysis
   a. Hierarchical agglomerative cluster analysis
   b. k-means partitioning

7 Self-Organized Map (SOM)

From a multivariate world back to a univariate world…

8 Supervised learning: numerical response variable (Y) and a data matrix (predictors): Bootstrap aggregation (Bagging, e.g., Random Forests) and Boosting trees

9 Supervised learning: categorical response variable (Y) and a data matrix (predictors): Support Vector Machine (SVM), Bagging, and Boosting classification trees

10 Group presentations

11 **Final paper due** (Wednesday by midnight in the finals week)

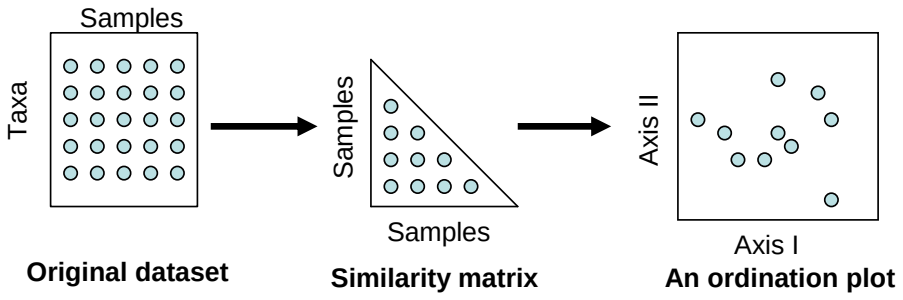

Figure 1. A schematic diagram showing steps in an ordination analysis (modified from Clarke and Warwick 2001)

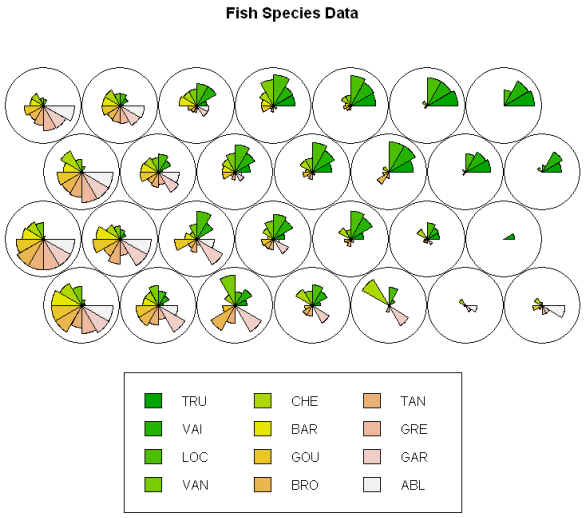

Figure 2. Using Self-Organized Map (SOM) to classify fish assemblages.

Statement on Academic Honesty

Plagiarism of any form will not be tolerated in this class and will result in failing grades for the assignment and course participation, and a referral to the Office of the Dean of Student Life. For more information, please see the Portland State University's Bulletin and how to avoid plagiarism.

PSU Student Resources

- Title IX reporting
- Disability accommodations and the Disability Resource Center
- Dean of student life
- Religious accommodations policy
- Library
- Writing Center
- Food assistance
- General PSU Policies (e.g., Student Conduct and Responsibility Policy)
- Student Resources and Centers (e.g., campus public safety, veterans resource center, etc.)
- Sanctuary campus information and resources
- DACA resources