ESM 567 Multivariate Analysis of Environmental & Biological Data

M & W: 2:00 to 3:50 pm

Instructor: Yangdong Pan (email: pany@pdx.edu)
Teaching assistant: Wendy Sangucho Loachamin (email: wen27@pdx.edu)

Student hours: Mondays 1:00 to 1:50 pm or by arrangement (Zoom meeting upon request)

Syllabus Content

Health Check, Illness, Exposure or Positive Test for COVID-19

- If you are feeling sick or have been exposed to COVID-19, do not come to campus. Call The Center for Student Health and Counseling (SHAC) to discuss your symptoms and situation (503.725.2800). They will advise you on testing, quarantine, and when you can return to campus.
- If you test positive for COVID-19, do not come to campus. SHAC will advise you on quarantine, notification of close contacts and when you can return to campus.
- Please notify me, (i.e. your instructor), should you need to miss a class period for any of these reasons so that we can discuss strategies to support your learning during this time.
- If I become ill or need to quarantine during the term, either I or the department chair will notify you via PSU email about my absence and how course instruction will continue.

Guidance May Change

Please note that the University rules, policies, and guidance may change at any time at the direction of the CDC, State, or County requirements. Please review the University's main COVID-19 Response webpage and look for emails from the University on these topics.

Course Learning Goals and Objectives:

1. To introduce multivariate data analysis methods commonly used in the field of ecology and environmental studies.
2. To develop skills in preprocessing and managing multivariate datasets.
3. To gain proficiency in using R to conduct multivariate analyses on environmental and biological datasets.
4. To better understand how to incorporate multivariate analysis results into research publications, including the proper interpretation and presentation of findings.

Summary of the Course:

This course is specifically tailored for students engaged in the collection and analysis of multivariate datasets such as biological species composition or environmental variables, or both. Such datasets are characterized by inherent complexity, voluminous dimensions, and noise, often manifesting internal inter-variable relationships and outliers. The application of multivariate data analyses is posited as a methodological approach to effectively summarizing multivariate datasets in a low-dimensional space, thereby elucidating patterns.

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 student-led research projects. The lecture will be condensed with no more homework. Each research group will have plenty of 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.

Recommended reading materials:

- Gotelli, N. J. and A. M. Ellison. 2013. *A Primer of Ecological Statistics*. Sinauer Associates, Inc. Publishers, Sunderland, MA. (2nd edition)
- 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/App:

- *R* (free downloadable from <https://cran.r-project.org/>. Please install the software to your computer.
- *RStudio*: a text editor for R and others (free downloadable from <http://rstudio.org/download/desktop>). Please install the software to your computer.
- We will primarily use RStudio in the course. Please watch this short video on how to get start.
- *Canvas*: an on-line learning system (https://www.pdx.edu/academic-innovation/canvas). 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 on Canvas. Students are encouraged to use Canvas to post questions, comments, and suggestions.

Approach:

1. **PowerPoint slides with audio**: Most of course contents will be delivered via PowerPoint slides coupled with audiotaped explanation by the instructor. The slides with audio will be loaded to Canvas weekly and students are encouraged to go through the slides prior to attending the class.
2. **Worksheets and in-class exercises**: A worksheet with several in-class exercises will be distributed weekly on Canvas. A large amount of each class period will be allocated for group-based hands-on in-class exercises. In other words, during each class period, students are encouraged to work as a group on each of the in-class exercises and discuss any issues with the instructor or GTA.
3. **Lectures**: The instructor will try to limit the lecture to 20-30 minutes each time. Each class will start with a short introduction and then follow by in-class exercises using either students' own laptops or desktops in B1-82. When students run into similar issues during the exercise, the instructor will start to lecture to address the issues to the entire class.
4. **Research Project**:
    a. This course emphasizes tremendously student-led and group-based learning. The class will be divided into groups, each with 2-3 members. Each group will work together on group research projects.

b. Each group will formulate research questions based on their shared research interest, construct a conceptual model, identify a dataset which is suitable for addressing their research question (it is preferred that the students use their own research datasets), select appropriate data analyses, and perform the analyses on the datasets.

c. Each group is required to present their work to the class following a conventional scientific meeting format (12 minutes presentation with PowerPoint slides and 5-10 minutes for Q&A) and then write a paper based on their research results following scientific format (more details will be provided).

5. **Peer-evaluation**: Since the class emphasizes tremendously on teamwork and student-based learning, 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.

Grading:

- **Homework (3 homework exercises: 60%):** Each homework assignment, its expectation and grading rubric, and due date will be posted on Canvas.
- **Group-based Research Project (35%):** Each group is 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.

Grading Scale (percent scores and grade break points for letter grades):

A: "excellent", comprehensive knowledge and understanding of subject matter;
B: "good", moderately broad knowledge and understanding of subject matter;
C: "satisfactory", reasonable knowledge and understanding of subject matter;
D: "inferior", minimum knowledge and understanding of subject matter
A: 94–100; A^-: 90–93; B^+: 87–89; B: 84–86; B^-: 80–83; C^+: 77–79; C: 74–76; C^-: 70–73
Pass: C^- or above
Incomplete: Departmental and university policies dictate that incompletes can be given only for verified medical reasons (through the Office of the Dean of Student Life).

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/>
- 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*".

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
6	(**ch.12, p.387)**: a. MANOVA, Discriminant function analysis b. Analysis of Similarities (ANOSIM), Permutation Multivariate Analysis of Variance (PERMANOVA), Multiple Response Permutation Procedure (MRPP)
	Classification: Put things in groups (ch.12, p.429-437)

7 Cluster analysis
   a. Hierarchical agglomerative cluster analysis (e.g., Nearest neighbor, Flexible beta)
   b. Non-hierarchical partitioning (e.g., k-means partitioning, Partitioning around medoids)
   c. Fuzzy clustering
   d. Self-Organizing Maps (SOM)

From a tree to a forest: Tree-based multivariate models

8 Regression and classification tree models
Multivariate Regression and classification tree models

9 Random Forest, Gradient Forest
Multivariate Random Forest

10 Group presentations

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

Group-based Research Project:

Group-based research projects allow students to work together and use what they learn from the class to solve a real-life environmental problem. Several research papers from past students using different statistical methods are posted on Canvas for your reference.

- *Research group*: By week three, 2-3 students will form a research group on your own.
- *Research*: Each group will then formulate a research question based on their shared research interest, construct a conceptual model based on the relevant literature, identify a dataset which is suitable for addressing their research question (it is preferred that the students use their own research datasets), select appropriate data analyses, perform the analyses on the datasets, interpret the results both statistically and scientifically. Each group is required to write a paper based on their research results following a scientific format. **This process is typically more time-consuming than we usually expected and thus act early.**
- *Final paper due*: The final revised paper will be due on **Week 11 (Wednesday)**

Grading Guidelines on the ESM 567 Final Project Paper:

Title & Abstract (2 pts)
Does the abstract concisely summarize the study with (1) purpose (2) research objective/question/hypothesis (3) study design (4) major findings and (5) conclusion(s)?

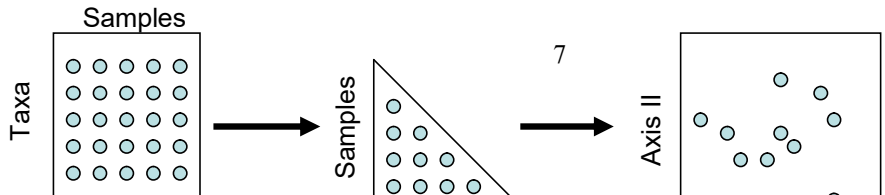
Introduction (5 pts)

Is the purpose of the paper clear in the introduction?
Does the author summarize the current knowledge on the subject using a conceptual model?
Does the author effectively use the conceptual model to provide enough background information that it is very clear to the reader on why this study is necessary?
Is the question well formulated and unambiguous?

Methods (10 pts)
Does the author clearly explain the study design (e.g., experiment/study unit, data collection, sample size, site selection criteria, etc. Hint: the best way is to use a diagram to illustrate the study design)?
Are the study design and methods appropriate for the research questions/hypotheses?
How does the author prepare the data for the analysis (e.g., standardization or transformation)? Is each decision on the data preparation defensible with enough rationale?
Does the author clearly explain how the data are analyzed (e.g., using Principal Component Analysis) and the rationale for selecting a particular method over the others (e.g., Non-metric dimensional scaling)? In other words, how this particular method is suitable for addressing the research question? Does the author make an explicit linkage between each analytic analysis and the study objective/question?
Is the analytical method appropriate for the study design? Does the author describe ALL methods used in this study?
Can the reader repeat the same analysis according to this section?

Results (10 pts, the number of figures in the text <4, the number of tables in the text <3)
Does the author adequately characterize the patterns and trends in the data prior using summary statistics such as mean and standard deviation or median and range)?
Does the author provide too many results which are not well related to the study objectives?
Does the author provide key findings with relation to the study question?
Does the author report key results from each key data analysis (e.g., eigenvalue and eigenvector in PCA)?
Does the author present the key finding effectively in each figure or table?
Is every figure or table included in the paper referred to in the text?
Does each figure/table have a complete caption, axis labels, so that each figure/table can stand alone?
Does the author provide any results from the analysis which is not described in the methods section?
Does the author present the results only in this section (with no discussion/explanations and methods)?

Discussion (8 pts)
Does the author discuss the major findings with relation to the study objective?
Is the conclusion fully supported by the results?
Does author provide any logical and meaningful interpretation of the results?
Has the author been objective in the discussion of the topic?
Is all of the discussion relevant to the study questions?
If the results are negative (contrast to the author's original expectations), does the author adequately discuss major possible reasons and provide any leads for further studies?
Does the author simply repeat all results in this section?



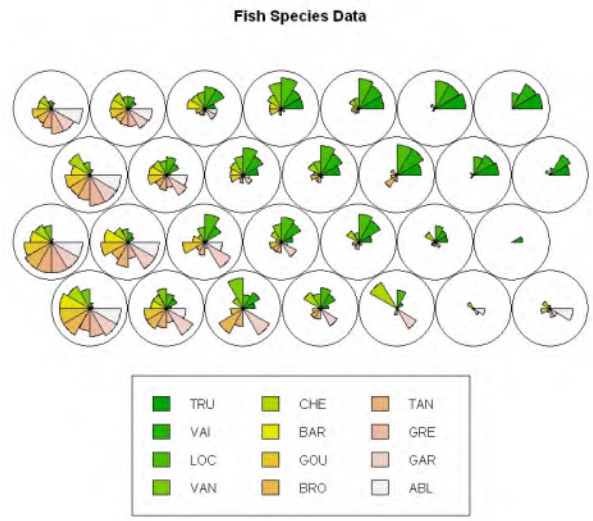


Figure 2. Using Self-Organized Map (SOM) to classify fish assemblages.

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
- PSU Prohibited Discrimination & Harassment Policy
- 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