**Statistical Analysis of Ecological Communities**

**Introduction to the course:**

Syllabus (course website): <http://myweb.ttu.edu/nmcintyr/Stat_Analysis_Ecol_Communities/SAEC.html>

Ecological communities are groups of multiple interacting species. They are thus characterized in terms of their biodiversity. Because biodiversity losses are accelerating, there is a need to be able to describe ecological communities quantitatively in order to be able to monitor any changes in them over time, evaluate any management activities conducted on their behalf, or compare different areas. This course focuses on some (but by no means all) of these methods.

Conservation biologists, wildlife biologists, natural resource managers, and community ecologists need to know how to answer certain questions about ecological communities, so I’m structuring this course around methods to address eight such questions:

* What are ecological communities, why do we study them, and what kinds of data are involved?
* How do I describe the structure of a community?
* How do I describe the structure of a taxonomic subset of a community?
* How do I summarize complex relationships and identify the most important patterns from a near-infinite number of possible arrangements of multivariate data?
* How do I winnow down a large number of correlated variables to an independent set?
* How can I determine which environmental variables are associated with species occurrences at different sites?
* How do I determine whether my data are grouped?
* If I have communities with *a priori* groupings, how do I determine whether those groups are valid?

My objective in this course is to teach you the currently best-established and most heavily used methods in analyzing site x species x environment data.

To do so, we will use the R statistical computing environment. R is freeware with a steep learning curve but is now perhaps the most heavily used statistical analysis software.

What this course is NOT (and what it is instead):

-a deep dive into R (R is being used as a tool)

-a help session for problems you are encountering in analyses for your research (it’s structured around only certain topics and is no substitute for consultation with a statistical expert who can spend time getting to know your data and your objectives)

-comprehensive (it’s an overview)

By the end of this course, you will have gained the following:

1. An understanding and appreciation of ecological communities and the kinds of analyses that are performed on biodiversity data;
2. Knowledge about which statistical method or family of methods is needed to address a particular question;
3. Experience with the R computing environment (including RStudio and R Markdown);
4. Example code that you may be able to use in future analyses involving your own research; and
5. An ability to interpret the output of such analyses in future.

Although this will not give you a comprehensive understanding of the statistical analysis of ecological communities, it should enable you to conduct some analyses on your own, and empower you to explore other forms of analysis.

**Structure of the course:**

On Tuesdays, I will give a lecture that brings up one of the eight questions, which methods are used to answer that question, and then will provide you with an example (data and R code). I will then discuss the output. You will then be provided with new data and be asked to analyze it and interpret the output. This is the only form of grading assessment that will be done in this class (no tests). You will turn in your assignment before the following Monday, using R Markdown.

Tuesday lectures will be narrated PowerPoint presentations available on YouTube; a Word document including additional information and references will be available on the course website for download. This Word document will also include your weekly assignment. You may watch the video and download the Word document at any time prior to the assignment being due, but I suggest that you do so on Tuesday or Wednesday because on Thursdays, I will be available live via Zoom to answer any questions you may have about the materials.

The narrated PowerPoint video will be an overview of the week’s topic; the Word document will contain very detailed notes about the topic.

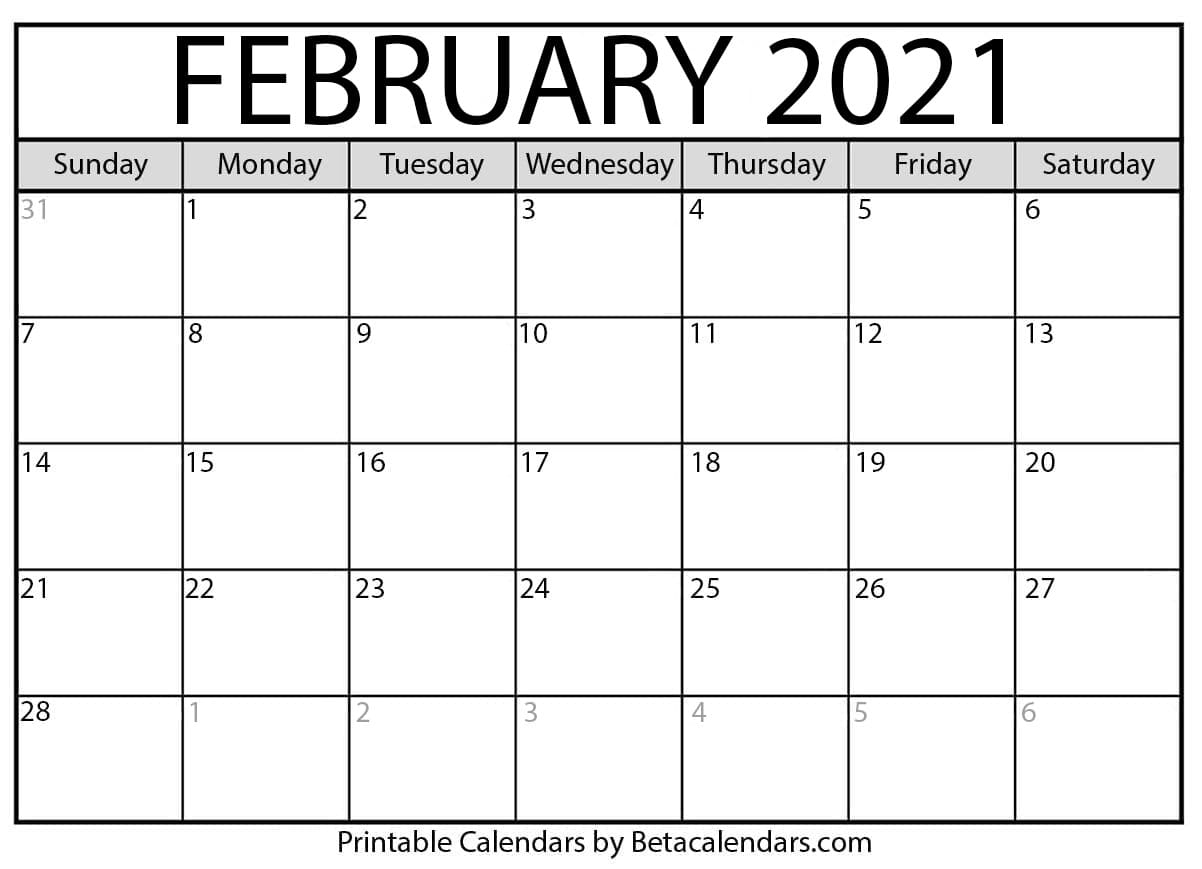
The Thursday sessions are meant to be informal opportunities for you to ask questions. Zoom links will be emailed to course participants only; please do not share those invitations with others.

I recommend that you interact with the course materials in the following order:

1. First, read the online notes.
2. Then watch the lecture video.
3. Then do the assignment.

**Assessment/grading:**

Each weekly assignment must be done in R Markdown as a Word document and should be turned in via email to the Teaching Assistant, Iroro Tanshi ([iroro.tanshi@ttu.edu](mailto:iroro.tanshi@ttu.edu)), before 8:00 a.m. on Monday of the next week. For example, each lesson below is color-coded to indicate when that lesson’s assignment is due:



Assignment due by 8:00 a.m.

Assignment due by 8:00 a.m.

Assignment due by 8:00 a.m.

Online

(Zoom)

Online

(Zoom)

Online

(Zoom)

Lecture; assignment given

Lecture; assignment given

Lecture; assignment given

Online

(Zoom)

Assignment due by 8:00 a.m.

Lecture; assignment given

In your email, please include the following as the Subject line:

Assignment on \_\_\_\_ (you will fill in the blank with the week’s topic; for example: Assignment on PCA)

Now that the course details are out of the way, on to the course subject!

**What is an ecological community?**

**community** ‑ an association of biotic populations of multiple species that are spatially delimited (i.e., they live in a particular area or habitat)

**Characteristics of a community:**

**richness** - number of species (or other pre-defined operational taxonomic units)

**abundance** - number of individuals/species; some species are common whereas others are scarce

**evenness** - relative numbers of individuals across species in a community; if every species is represented by the same number of individuals, that is perfectly even whereas if most individuals belong to only a few species, that community is not even and can be said to be dominated by those few species

**richness + evenness = diversity**

**environment** - both abiotic and biotic factors determine species occurrence and abundance

**Why do we study them?**

There has long been interest in why diversity differs in space and over time, what factors maintain diversity at certain levels (i.e., why are there only *this* many species here?), and how human activities affect diversity. Understanding these things is the purview of conservation biology, landscape ecology, natural resource management, and many other fields.

It is important to be able to take a rigorous, quantitative approach in examining ecological communities. A community is, by definition, a multispecies collective, with each species represented by different numbers of individuals. This collective exists in a discrete area that has environmental characteristics (e.g. soil types, canopy cover, soil moisture, depth, salinity, etc.) that may limit which species can be present. Thus, examining ecological communities necessitates multivariate data analyses.

**Multivariate data:**

Most community analyses involve, at minimum, two multivariate datasets:

1. At minimum, there will be data on species abundance or occurrence at various sampling locations 🡪 this is a **site x species** matrix and it may be binary (presence/absence of a species) or continuous (abundance of each species)
2. You may or may not also have measured environmental variables at those sampling locations 🡪 **site x environment** matrix

Typically our focus is on whether there are any patterns of species abundances/occurrences by sites, whether sites show patterns in environmental variables, and whether species are associated with environmental variables independent of site. Before any analyses can be done, however, some data screening should occur first.

**Data screening:**

**-Zeros:** In the site x species matrix, there are often lots of zeros; such matrices are called “sparse” and may require transformation for some analyses (more info about data transformation is given below).

Note that a zero is *not* the same as missing data! Having lots of cells with missing data is a different problem that takes different solutions:

**-Missing data:** Many analyses will delete entire rows if they contain missing values, which is rather draconian. If your missing values are few, dispersed throughout your dataset, and essentially random with respect to groups of data, then you can do some corrections to fill in values:

-Insert means or medians

-Use regression to predict a value

If you cannot reasonably insert a value as a stand-in, then denote missing data cells with NA or with -999. Those are standard designations for missing data.

**-Outliers:** Outliers have extreme values (typically defined as >3 standard deviations from the mean) and thus can have a large effect on the outcome of an analysis. They can typically be detected by examining plots of your data; once identified, then check for accurate data entry to make sure the outliers are real and not simply typos. If the value is indeed correct and not a mis-entry, then consider whether the outlier is part of your main focal group. If not, then it’s common to simply remove that outlier data point. If the outlier is part of your focal group, then consider a data transformation (see below), or analyses that are relatively insensitive to outliers.

**-Normality:** many statistical analyses assume normality, assessed via skewness (asymmetry) and kurtosis (peakiness). If the data are normal, then skewness = 0 and kurtosis = 0. Species data are almost always skewed, with a long tail (because most species are rare, few are common), so one rule of thumb is to try to have |skew| < 1. Log transformation can help alleviate skew (see below).

Include how to test for departure from normality and/or how to determine what kind of distribution your data have

**-Data transformation:** Transformation may be needed to meet assumptions of normality and homogeneity of variance, deal with a lot of 0’s, equalize the relative importance of common vs. rare species, and other reasons. In general:

-Square root transformations normalize variables with a Poisson distribution (e.g. counts over time).

-Root transformations (not just square root) are useful when your data are right-skewed (i.e., have a long right tail in their distribution); the root compresses the spread of values in the right tail.

-Power transformations have the opposite effect to root transformations.

-Proportions and percents should be arcsine-square root transformed to improve normality; this procedure flattens the tails of the distribution and compresses the middle.

-Log transformation is useful when there is skew, when your data are ratios, or when there is high variation within variables or when there is high variation among attributes within a sample. (These are commonly true for count data and biomass data.) A log transformation rescales the values such that very large or very small values are brought closer to the mean. To log transform a dataset containing 0’s, a small number must be added to all data points (e.g. 0.001). If the lowest non-zero value in the dataset is 1, then it is recommended to add 1 to all data values before log transforming.

-If your dataset has many 0’s (e.g. due to species being found only at certain sites but not others), then the Chord transformation or Hellinger transformation are recommended (see Legendre and Gallagher 2001 for details).

-Examine your residuals before and after a transformation: if the residuals are smaller and no longer form a pattern, then you can conclude that the transformation was effective.

Now that your data are tidy, you can begin analysis. In this course, we will cover analyses that address various objectives:

-Basic quantitative descriptions of ecological communities can be provided by various diversity and evenness indices as well as abundance patterns.

-To simplify multivariate datasets and examine associations between explanatory variables and response variables, various forms of ordination can be used. Most of the topics we will cover this semester will be different types of ordination techniques. These different techniques have different data assumptions or are used to address slightly different questions.

-**Data standardization:** Standardization is needed if you have variables that differ in range. For example, some organisms will be present in the hundreds in a community whereas others will be present in only single digits in the same community. Data standardization rescales data to make them more comparable to each other. This topic will be covered in more detail in the “Site x environment” lesson.

**I have structured this course to address a broad series of questions:**

-What are ecological communities, why do we study them, and what kinds of data are involved? 🡪 That was addressed in today’s lesson.

-How do I describe the structure of a community? 🡪 We will explore a variety of diversity and evenness indices as well as abundance patterns.

- How do I describe the structure of a taxonomic subset of a community? 🡪 We will examine assemblage structure along ecomorphological gradients.

- How do I summarize complex relationships and identify the most important patterns from a near-infinite number of possible arrangements of multivariate data? 🡪 This is the objective of the various forms of ordination. The remaining lessons in this semester will involve forms of ordination.

- How do I winnow down a large number of correlated variables to an independent set? 🡪 This question is addressed by a variety of techniques. We will cover Principal Components Analysis (to detect any variable in your data that explains species composition) and Redundancy Analysis (to detect the best explanatory variables).

- How can I determine which environmental variables are associated with species occurrences at different sites? 🡪 This question is addressed by a wide variety of techniques. We will be able to compare Nonmetric Multi-Dimensional Scaling (unconstrained ordination that identifies patterns in a dataset and the variables responsible for those patterns) and Canonical Correspondence Analysis (constrained ordination that identifies the effect of one dataset on another, related dataset).

- How do I determine whether my data are grouped? 🡪 This is the realm of Cluster Analysis.

- If I have communities with *a priori* groupings, how do I determine whether those groups are valid? 🡪 This is the focus of Discriminant Function Analysis; we will use a Classification and Regression Trees approach.

**R: the tool we will use to answer these questions**

The next few lessons will cover the statistical platform/computing environment we will use to address these questions (called R). R is perhaps the most widespread statistical analysis platform now used in the life sciences; it is the go-to for many governmental agencies and companies because it is free. It is extremely versatile but has a steep learning curve. It is definitely a useful skill in which to have some experience, proficiency, or mastery.

There are many other software packages that will do many if not all of the kinds of analyses we’ll do, but they aren’t free. So what do you get for the money? Some are easier to use than R because they have a graphical user interface. Some make beautiful output plots. And some have customer service hotlines/messaging/email when you run into problems. Here are some alternatives to R:

* [Brodgar](http://www.brodgar.com/) – especially useful for multivariate time series analyses
* [CANOCO](http://www.canoco.com) – for ordinations (especially constrained ordinations)
* [Genstat](https://www.vsni.co.uk/software/genstat) – does a wide variety of univariate and multivariate analyses, including generalized linear modelling, principal components analysis, canonical variates analysis, factor rotation, principal coordinates analysis, correspondence analysis, discriminant analysis, hierarchical and nonhierarchical cluster analysis, and multidimensional scaling
* [PC-ORD](https://www.wildblueberrymedia.net/pcord) – perhaps the best all-around community analysis package, covering various forms of ordination, canonical and detrended correspondence analyses, redundancy analysis, nonmetric multidimensional scaling, indicator species analysis, principal coordinates analysis, diversity indices, cluster analysis, and many others. Handles large datasets well (32,000 rows x 32,000 columns!).
* [SAS](https://www.sas.com) – the OG of powerful statistical software for the masses. Handles large datasets very well (unlimited rows/columns, and up to 9.2 quintillion observations!!!), better than any of these other packages, including R (which struggles in datasets larger than 1 million observations). But steep learning curve, pricey (annual license), notoriously poor graphics, and doesn’t do some analyses.

This week’s live session (Th, Jan. 28) will simply be a “getting to know you” session and experience in how the live Q&A sessions will go.

Next Tuesday’s lesson will be an introduction to R, RStudio, and RMarkdown. It will also introduce you to some of the datasets we will be using in this course.

**References:**

Legendre, P., and E.D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280.

**Assignment:** None (continue working through R primers if necessary)