**Tutorial - Morphological traits and non-metric multidimensional scaling (NMDS)**

This tutorial is an introduction for NMDS analysis in R to assess the similarity of the fossils to the extant taxa. This ordination approach, although most common in community ecology, has become increasingly used for analyzing fossil diversity and taxonomic affinities. In this tutorial, we provide prompts to help you check you understanding of the steps – feel free to ask whenever you are not sure!

1. Prepare the files you need to analyze the data and place them in the same folder. They consist of:
   1. R script
   2. CSV file with morphological traits (Sol\_traits.csv)
   3. CSV file with species and classifiers (Sol\_classifier.csv; i.e., additional information used to label taxa; e.g., fossil vs. extant taxa)
2. Open the CSV file with **morphological traits** (e.g., in Excel) and describe the data.
   1. How many characters are included?
   2. How many taxa are scored?
   3. What characters are not allowed in taxa and character names?
   4. What is the nature of the morphological characters?
   5. Why can’t I perform a PCA with these data?
   6. How are represented the missing data?
   7. The meaning of the names for each character is:

fruiting\_calyx\_LT = fruiting calyx length

fruiting\_calyx\_ratio = fruiting calyx lenght:width ratio

calyx\_lobe\_LT = length of calyx lobes

calyx\_lobe\_ratio = calyx lobe lenght:width ratio

calyx\_LT\_lobe\_ratio = ratio between calyx total length to lobes length

base\_inv = presence/absence of calyx base invaginated

angled = presence/absence of angled calyx,

lobe\_sinus = sinus rounded or angled,

teeth\_presence = presence/absence of calyx teeth,

widest\_veins = presence/absence of calyx widest veins distinctive of other vein orders that terminate in lobe tips,

secon\_veins\_dist = presence/absence of calyx secondary veins distinct from other vein orders that emerge from base,

secon\_veins\_fork = presence/absence of calyx secondary veins forking before lobe sinus,

fruit\_type = berry, capsule, diclesium, drupe, mericarp, pyxidium or non-capsular dehiscent fruit,

calyx\_inflated = presence/absence of inflated calyx,

venation\_pattern = calyx venation pattern (parallelodromous or other type).

* 1. Do you think any of these characters would be correlated?
  2. How would this correlation affect the analysis and results?

1. Open the NMDSscript in RStudio to analyze the data.
   1. Install packages required and complete the following table using the help option in R or Google.

|  |  |
| --- | --- |
| Package name | Description |
| vegan |  |
| cluster |  |
| ggplot2 |  |

* 1. Set your working directory using the function *setwd* or within the Session tab, select ‘Set working directory’; e.g., setwd ('C:/Users/rocio/Desktop')
  2. Read the CSV file.
     1. Why is the header set as True and row.names=1?
     2. Identify the function involved.
     3. How is now represented the missing data?
  3. Prepare the **dissimilarity matrix**.
     1. Identify the function called.
     2. According to the following table, why would you choose the gower metric for this dataset?

|  |  |
| --- | --- |
| Metric | Meaning |
| Euclidean | Root sum-of-squares of differences |
| Manhattan | Sum of absolute differences |
| Gower | Each variable (column) is first standardized by dividing each entry by the range of the corresponding variable, after subtracting the minimum value; consequently, the rescaled variable has range [0,1], exactly |

* + 1. Indicate the type of characters in the dataset

|  |  |
| --- | --- |
| Type of character (in R) | Meaning |
| symm |  |
| asymm |  |
| ordratio |  |
| logratio |  |
| ordtype |  |

* 1. Perform the **NMDS analysis**.
     1. Identify the function.
     2. What is the meaning of trymax and k?
     3. What happens if you reduce the trymax to 10?
     4. Keep the ‘trymax’ in 50 and change the k from 3 to 4. What happens? Why?
  2. Fit the morphological characters into the ordination.
     1. When you analyze this result, what does this tell you about your characters?
     2. Which is the least informative character?
     3. Which is the character with highest correlation to NMDS1?
     4. Which is the character with highest correlation to NMDS2?
  3. Calculate the scores corresponding to the length of the segments for each morphological character.
     1. Identify the character best correlated to NMDS1 and NMDS2
  4. Determine the **stress**, or the disagreement between 2-D configuration and predicted values from the regression.
     1. Observe the Shepard plot. What number of dimensions would be best to analyze our data?
  5. **Plot** the results. First, we’ll plot the results without discrimination between fossil and extant taxa. You’ll find the following warning ‘In arrows(at[1], at[2], vect[, 1], vect[, 2], len = 0.05, col = col) : zero-length arrow is of indeterminate angle and so skipped’, meaning that the arrow for the character ‘teeth\_presence’ is not plotted because its length is 0.
     1. How do you relate this graphic with the data obtained in 3.g?
  6. **Plot** the results using classifiers and the ggplot2 package.
     1. What are the differences between these two plots?
     2. Change font, colors, etc. to the plot.