Data types and import into R

Theory R functions Examples

Community ecology data sets

Data sets used in community ecology usually consist of one, two or three of the following matrices (see also Fig. 1):

- the matrix of species composition (*sample* × *species* matrix **L**),
- the matrix of environmental variables, or generally *sample attributes* (*sample* × *environmental variables* matrix **R**),
- the matrix of species characteristics (e.g. traits), or generally *species attributes* (*species* × *traits* matrix **Q**).

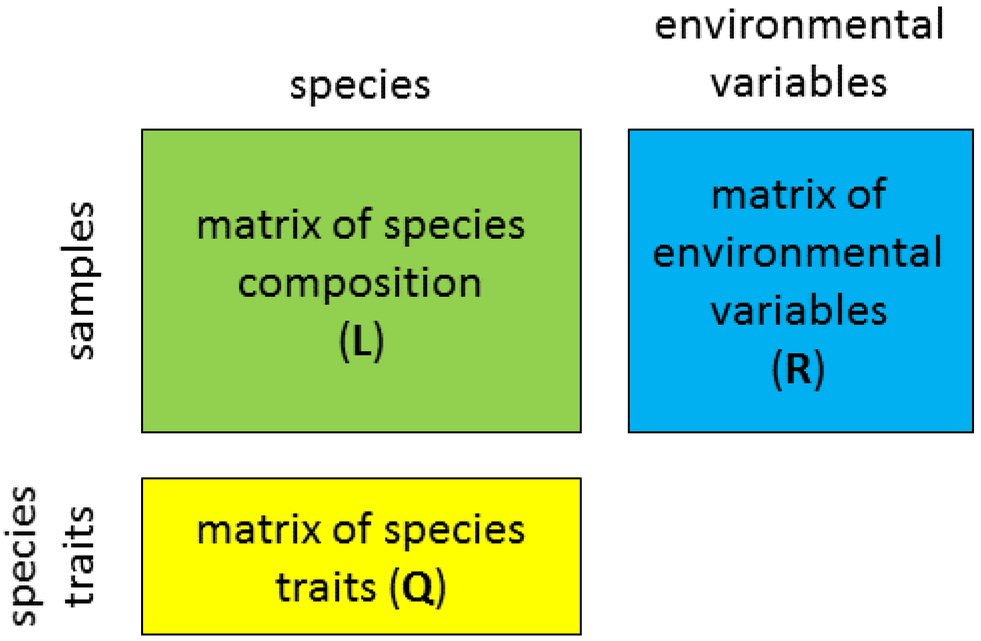
species
environmental variables
samples
matrix of species composition (L)
matrix of environmental variables (R)
species traits
matrix of species traits (Q)

Figure 1: Three matrices used in multivariate analysis of community data: matrix of species composition, environmental variables (or other sample attributes) and species traits (or other species attributes).

Types of variables

Variables (be it environmental variables, species attributes, or values in species composition matrix) fits into on of the three main categories:

- **Qualitative** (categorical, nominal): individual categories are unique (each observation belongs to only one of them), and they cannot be meaningfully ordered; e.g. geological type, soil type, binary (presence-absence) variables).

- **Ordinal** (semi-quantitative): arbitrary categories which can be ordered (from low to high, the least to the most etc.); e.g. abundance or cover scales, replacing real counting of individuals or estimating their percentage cover by few ordered scores (e.g. Braun-Blanquet dominance-abundance scale for estimating cover of vegetation: r, +, 1, 2, 3, 4, 5). Often used by ecologists to speed-up the sampling process.
- **Quantitative**: variables measuring quantities
  - *discrete* vs *continuous*: discrete variables are e.g. numbers of individuals, measurements taken with low precision; continuous variables are e.g. precise measurements.
  - variables on the *relative (or ratio) scale* or on *interval scale* (Fig. 2). The difference is in the meaning of zero: for variables on a relative scale (or ratio scale) zero is a meaningful value (usually it means that the measured variable is not present), while for variables on interval-scale zero is arbitrarily chosen. Variables on interval-scale don't allow for the relative comparisons (e.g. in case of temperature measured in degrees of Celsius we cannot say that this temperature is twice higher than that one; however, if we measure temperature in Kelvins, which is an example of the relative scale variable (since 0°K is absolute zero), we can make such comparison; similarly, e.g. wind speed is the example of the variable on the relative scale with meaningful zero - no wind).

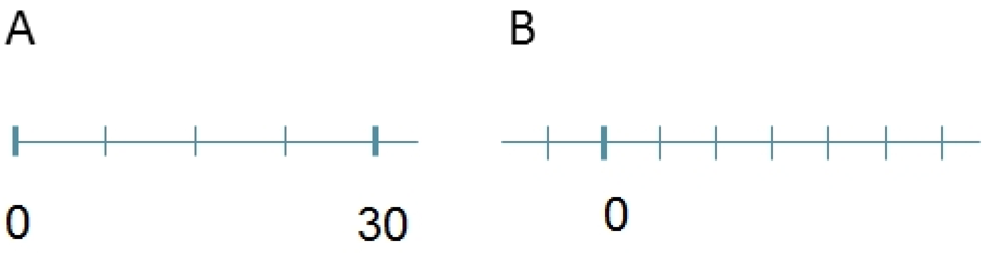

Figure 2: Variable on the relative scale (A), where zero has the real meaning (absence of the measured variable), and on the interval scale (B), where zero is chosen arbitrarily.

Primary data

Primary data are at the beginning of all analytical exercise. While collected into notebooks or newly also to some electronic devices like recorders, tablets or smartphones, they need to be retyped to a spreadsheet and archived. I found it useful to keep data in a multi-sheet Excel file, each matrix in a separate sheet, with one extra sheet containing **metadata** - a detail description of what kind of data each sheet contains, what is the meaning of abbreviations (if these are used for e.g. environmental variables or species names) and whether there were some changes or corrections done after the data were retyped (there are usually some). Such file can be used in future as a base for all further analyses, keeping all primary data in one place.

Long term storage of primary data is still a bit complex issue - it seems that the safest way is still to print them on acid-free paper, using the laser printer. The other option is to append data to published papers (usually as an online appendix) - some journals (like *Ecological Monographs*, *Journal of Ecology* etc.) requires attaching the data as a condition for accepting the paper. Another option is to store data in public and managed databases or electronic repositories (e.g. Dryad Digital Repository, www.datadryad.org), which (for free or some small payment) offers time-unlimited storage of your data (with the advantage of data being citable by assigning them DOI identifier).

Import of data into R

For analysis, the first step is always to get data into R. In most cases, R expects that community ecology data will have samples in rows and species (or other descriptors, like environmental variables) in columns[1]. For work in R, a good practice is to save each data matrix into one file (preferentially *.txt file with cells delimited by tabulator), save these files to a certain folder, and at the beginning of the script locate a set of lines reading the data from the files into R working space. This will make the script reproducible - just wrap the script together with the files into a zip file and the analysis (using the same data) can be fully reproduced. There are also other types of import (e.g. directly via the clipboard, from other format types like *.csv, directly from Excel's *.xls or *.xlsx file, etc.), and RStudio contains also a button functionality to load data. Still, I suggest that *.txt format loaded via script is the most stable solution.

Before importing the file into R, make sure that **all variable names are valid R names**. Valid name can contain letters, numbers, dots and underline characters; it should start with a letter (not a number) or dot not followed by a number. R is case sensitive, which means that uppercase and lowercase letters have a different meaning. Examples of **valid names**: `var_1`, `var.1`, `.var1`, `var_1.1`, `Var1`. **Not valid names** include `1var`, `1_var`, `.1vars`. There is also a list of **reserved words** which cannot be used as variable names, since they have fixed meaning in R: `if`, `else`, `repeat`, `while`, `function`, `for`, `in`, `next`, `break`, and also `TRUE`, `FALSE`, `NULL`, `Inf`, `NaN`, `NA`, `NA_integer_`, `NA_real_`, `NA_complex_`, `NA_character_`.

Species names of taxa, consisting of the genus, species and sometimes subspecific Latin name, may need to be **abbreviated**. This is not because R cannot handle long names of variables (it can, up to 255 characters), but because long species names may be difficult to display, e.g. in ordination diagram. Library `vegan` contains handy function `make.cepnames`, which abbreviates species names into up to 8 letters abbreviations; it takes four first letters from genus name and merges them with first four letters from species names.

Example section contains various examples of how to import data from other sources and formats into R.

[1] There are, however, notable exceptions, e.g. functions in the package `iNEXT` for analysis of diversity (iNterpolation/extrapolation) expect samples in columns and species (descriptors) in rows; this needs to be kept in mind when using it!

From:
https://www.davidzeleny.net/anadat-r/ - **Analysis of community ecology data in R**

Permanent link:
https://www.davidzeleny.net/anadat-r/doku.php/en:data_import

Last update: **2021/03/03 19:47**