# Multivariate statistics in R

Hannes PETER
Martin BOUTROUX
Zhe LIU

*This lecture benefits from teaching experience at different universities by Prof. Alexandre Buttler, Prof. François Gillet and Dr. Daniel Borcard. Yaniss Augot and Elie Roth helped with the implementation of exercises in Jupyter Notebooks.*

# Aims of the course

- Overview of principles of multidimensional data analysis
  - mostly for ecologists/environmental scientists
- Choice of statistical tools
- Learn how to use these tools
- Interpretation of the results
- Efficient communication with other experts
- Use 

# Course structure

- Wednesday 9:15 – ~13:00
  - ~1h theoretical lecture
  - ~1h practical work on noto (R on Jupyter Notebooks)
  - 1-2h hands-on practical work/group work
    - before you leave: briefly present your results to a teacher (~2-3 minutes)
- short paper discussions
  - read paper for discussion the following week

# Tentative schedule

- 10.09. **session 1**
- 17.09. **session 2**
- 24.09. **group work**
- 01.10. ***«modern R» with Martin*** (tidyverse)
- 08.10. **session 3**
- 15.10. **group work**
- 22.10. autumn holidays
- 29.10. **session 4**
- 05.11. **session 5**
- 12.11. **mid-term exam**, **group work**
- 19.11. **session 6**
- 26.11. **session 7**
- 03.12./10.12. **group work**
- 17.12. hand in report
- 07.01. group presentations 1
- 14.01. group presentations 2

# topics

- **data exploration**
  - summary statistics
  - visualization
- **transformations**
- **resemblance metrics**
  - dis/similarity, distance
- **unsupervised classification**
  - cluster analysis
- **supervised classification**
  - classification and regression trees, random forest classifier
- **unconstrained ordination**
  - PCA, CA, NMDS
- **constrained ordination**
  - RDA, CCA
- **auxiliary multivariate analysis**
  - LDA, mantel correlation, procrustes, etc…

# Evaluation

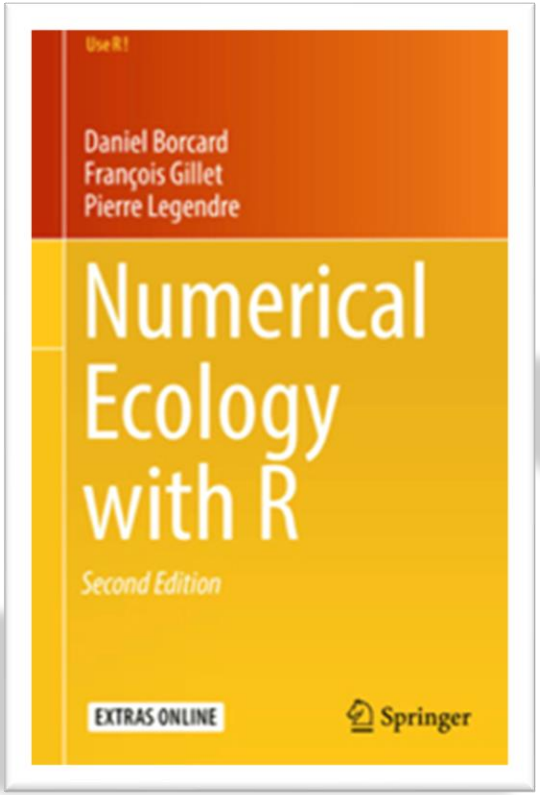
- **mid-term exam** (40%) (individual)
  - ca. 10 multiple-choice questions
- **oral presentation** (60%) (group)
  - in January 2025, 15 + 5 minutes
- **bonus** (+0.25 on final grade, details later this class)

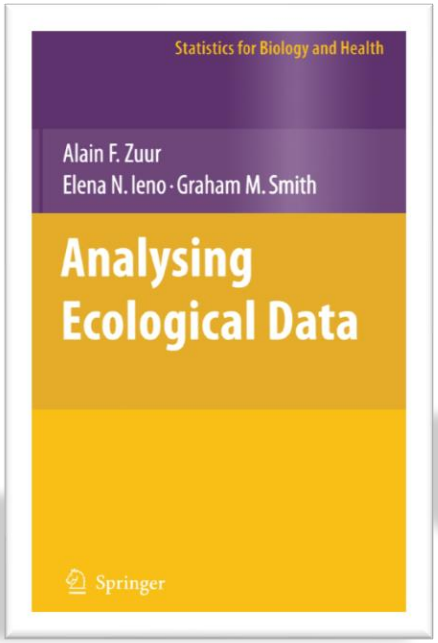
# Group projects

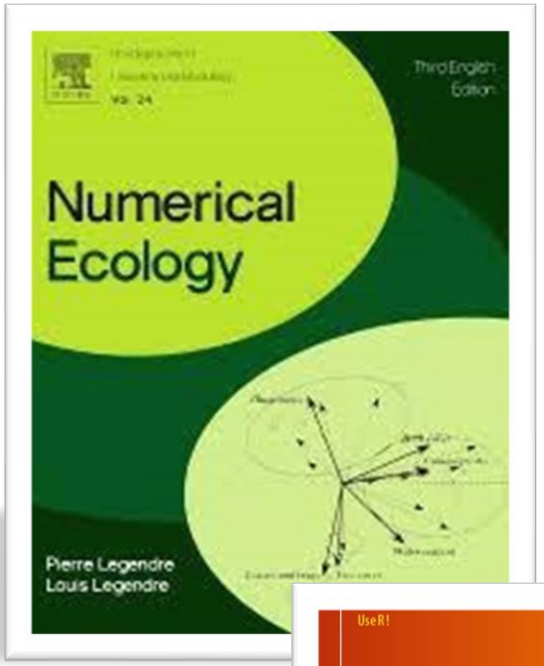
- 6-7 groups of 5 students
- find data, define a research question
- present research idea (1 page)
- perform multivariate analyses to address research question
- present results to the class

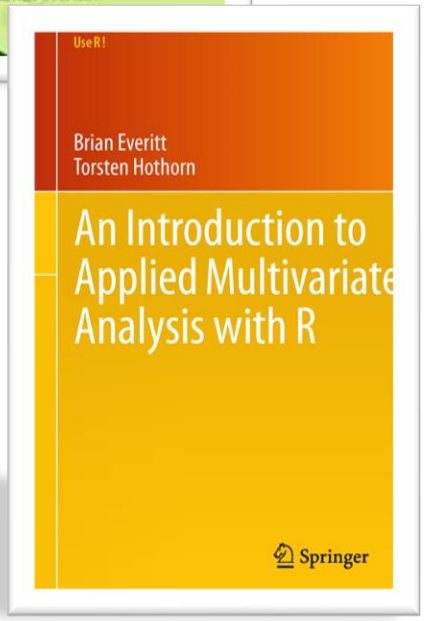
*> Check moodle for data sources*

# material




http://www.numericalecology.com/numecolR/index.html

available as pdf

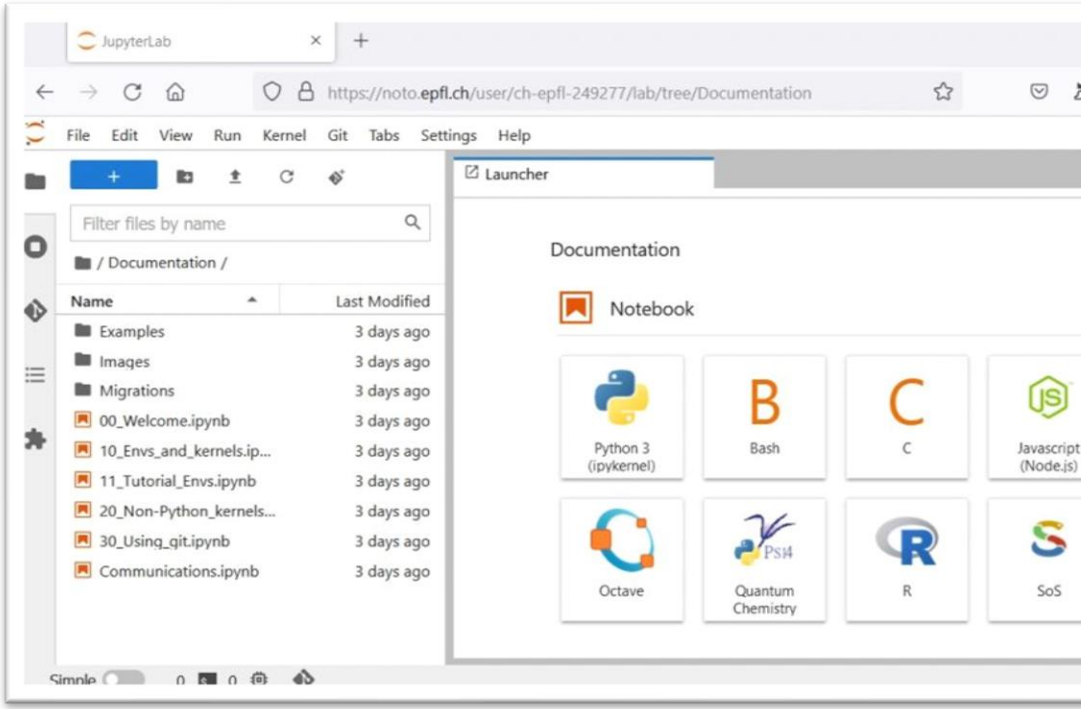
# Online resources

- https://ordination.okstate.edu/
- https://www.davidzeleny.net/anadat-r/doku.php/en:start
- https://environmentalcomputing.net/statistics/mvabund/
- https://sites.google.com/site/mb3gustame/

# practical part: Jupyter notebooks hosted on noto

- web-based interactive developer environment
  - supports different environments ("kernels") R, Python, ...
  - organized in «cells»
  - contains code, text, embedded objects,...
  - executed directly

  > *The Scientific Paper Is Obsolete*

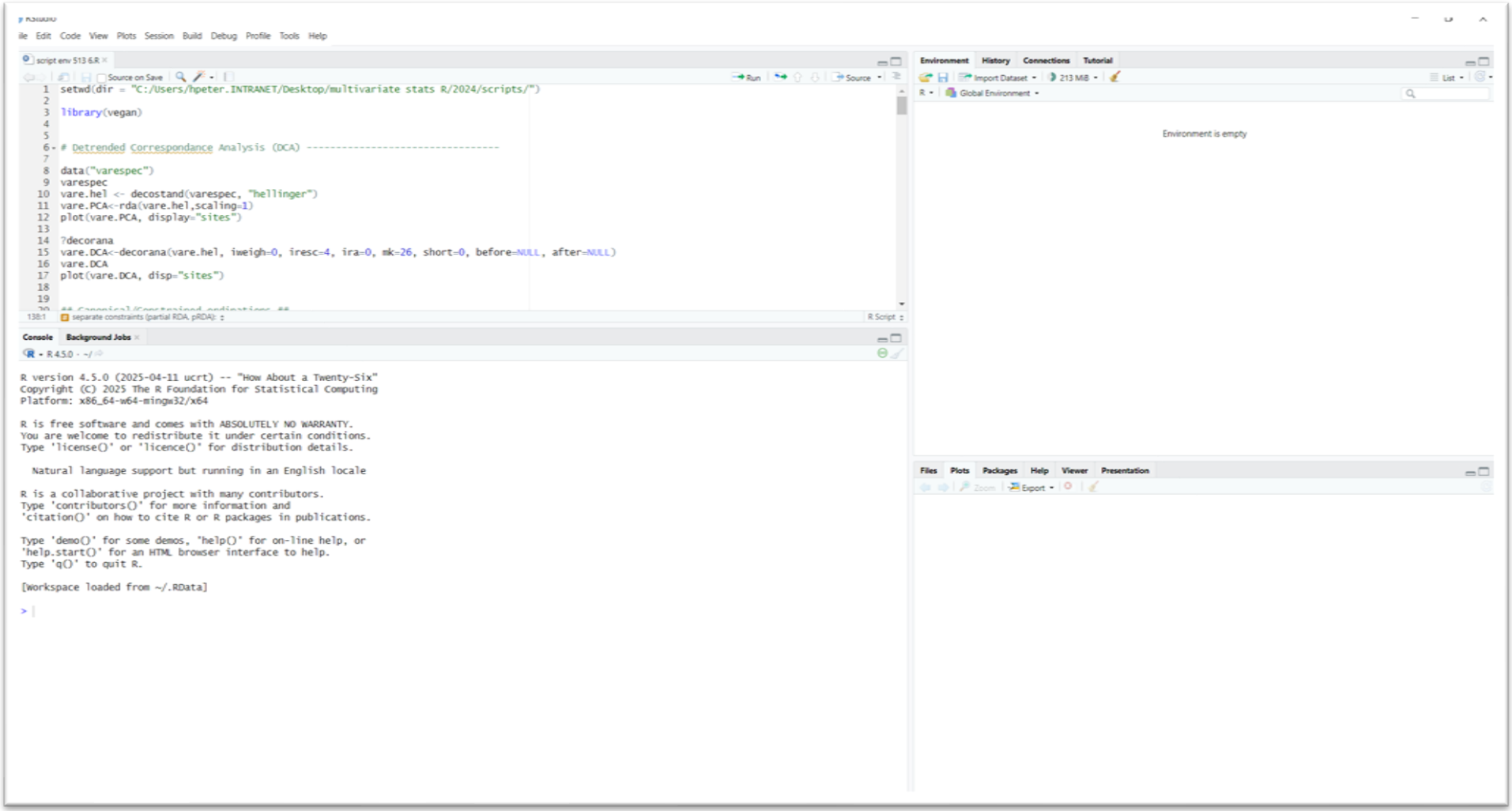



NOTO

- EPFL'scentralized JupyterLab platform
- use gaspar to connect
- notebooks accessible via moodle links

EPFL

# Rstudio


ile Edit Code View Plots Session Build Debug Profile Tools Help
script env 513 6.R
Source on Save
Run
Source
1 setwd(dir = "C:/Users/hpeter.INTRANET/Desktop/multivariate stats R/2024/scripts/")
2
3 library(vegan)
4
5
6 # Detrended Correspondance Analysis (DCA) ---------------------------------
7
8 data("varespec")
9 varespec
10 vare.hel <- decostand(varespec, "hellinger")
11 vare.PCA<-rda(vare.hel,scaling=1)
12 plot(vare.PCA, display="sites")
13
14 ?decorana
15 vare.DCA<-decorana(vare.hel, iweigh=0, iresc=4, ira=0, mk=26, short=0, before=NULL, after=NULL)
16 vare.DCA
17 plot(vare.DCA, disp="sites")
18
19
138:1
separate constraints (partial RDA, pRDA):
R Script
Console
Background Jobs
R 4.5.0 · ~/
R version 4.5.0 (2025-04-11 ucrt) -- "How About a Twenty-Six"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment
History
Connections
Tutorial
Import Dataset
213 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

# Multivariate statistics (in Ecology)

- « *Domain of quantitative ecology dealing with the numerical analysis of complex data* » (Legendre & Legendre 1998)
- Origin in the ecology of biological communities (synecology, community ecology)
- Original numerical methods, often developed by ecologists (e.g. diversity/similarity measures)

# Why is it important?

*You will try to answer difficult questions about the complex world we live in.*

*Which test should I apply?*

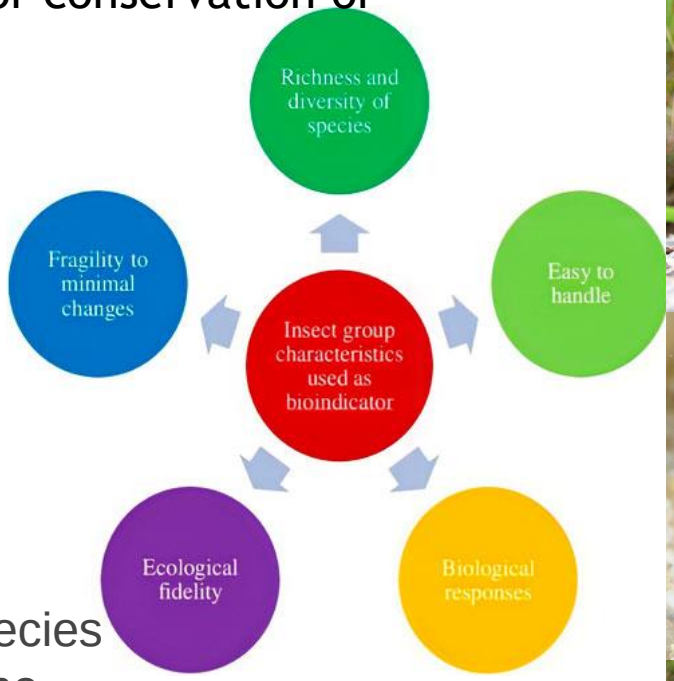
*There is not a single «test», but rather a series of analyses - data exploration, model formulation, evaluation and interpretation is required.*

# Practical Example

Indicator Species Analysis (ISA)

# Indicator Species

Monitoring the occurrence (or abundance) of indicator species is often used in long-term environmental monitoring for conservation or ecological management.




**Indicator species:**

1. reflect the environment (abiotic or biotic)
2. respond to environmental change
3. representative of other species (i.e. can be used to predict the diversity of other species, taxa or communities)

EPFL

Lawton & Gaston, Encyclopedia of Biodiversity, 2001

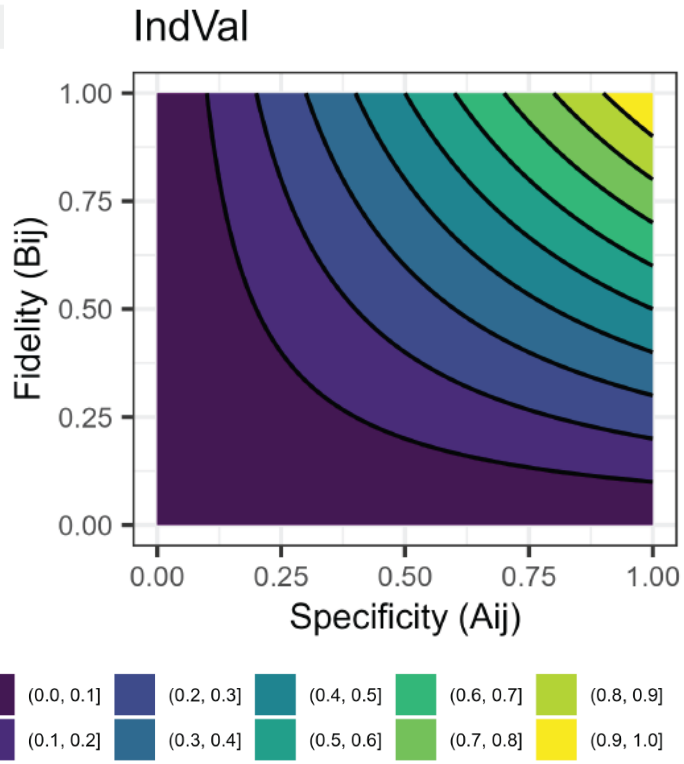
# Indicator Species Analysis (ISA)

ISA requires classification of sample units into groups.

ISA involves calculating the **specificity** (*A*ij; relative abundance) and **fidelity** (*B*ij; relative frequency) of species *i* in group *j*.
These values are then multiplied to yield the test statistic, the **Indicator Value (*IV*ij)**.

| | Formula | Verbal Interpretation | Range |
|---|---|---|---|
| Specificity / Relative Abundance: | $A_{ij} = \frac{\bar{x}_{ij}}{\sum_j \bar{x}_{i\bullet}}$ | Mean cover of species *i* in group *j* as a proportion of its mean cover in all groups | 0 to 1 |
| Fidelity / Relative Frequency: | $B_{ij} = \frac{n_{ij}}{n_{\bullet j}}$ | Proportion of plots in group *j* on which species *i* occurs | 0 to 1 |
| Indicator Value: | $IV_{ij} = A_{ij} \times B_{ij} \times 100$ | As proposed by Dufrêne & Legendre (1997) | 0 to 100 |
| | $IV_{ij} = \sqrt{A_{ij} \times B_{ij}}$ | As reported in `indicspecies::multipatt()` | 0 to 1 |

- $\bar{x}_{ij}$ is the mean cover of species *i* within group *j*
- $\sum_j \bar{x}_{i\bullet}$ is the sum of the mean cover of species *i* in all groups
- $n_{ij}$ is the number of plots in group *j* occupied by species *i*
- $n_{\bullet j}$ is the total number of plots in group *j*.




EPFL

Heft 85, 2019

**WSL Berichte**

ISSN 2296-3456

**Zustand und Entwicklung der Biotope von nationaler Bedeutung: Resultate 2011–2017 der Wirkungskontrolle Biotop-schutz Schweiz**

Ariel Bergamini, Christian Ginzler, Benedikt R. Sch
Angéline Bedolla, Steffen Boch, Klaus Ecker, Ulric
Helen Küchler, Meinrad Küchler, Oliver Dosch,
Rolf Holderegger

Eidg. Forschungsanstalt für Wald, Schnee und
CH-8903 Birmensdorf

info fauna karch, Bellevaux 51, 2000 Neuchât

2023 | Umwelt-Zustand

Biodiversität

## Biodiversität in der Schweiz

Zustand und Entwicklung

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesamt für Umwelt BAFU

ISSN 1470-160X

ELSEVIER

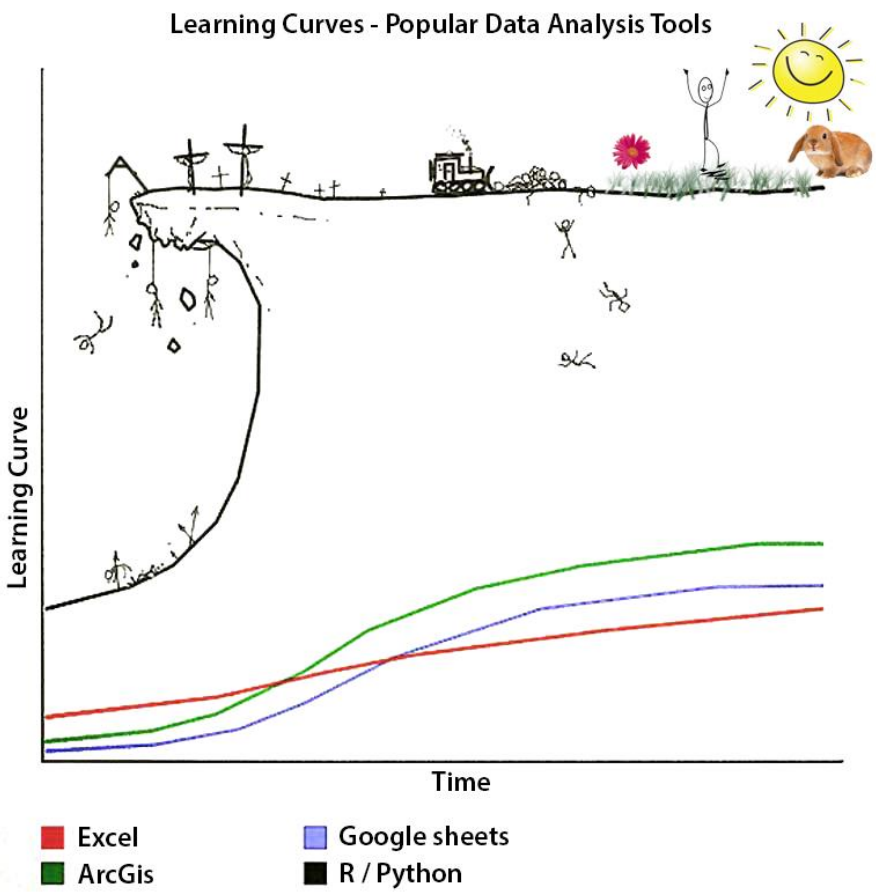
# ECOLOGICAL INDICATORS

INTEGRATING, MONITORING, ASSESSMENT AND MANAGEMENT


Co-Editors-in-Chief
**J.C. Marques**
**F. Müller**

EPFL

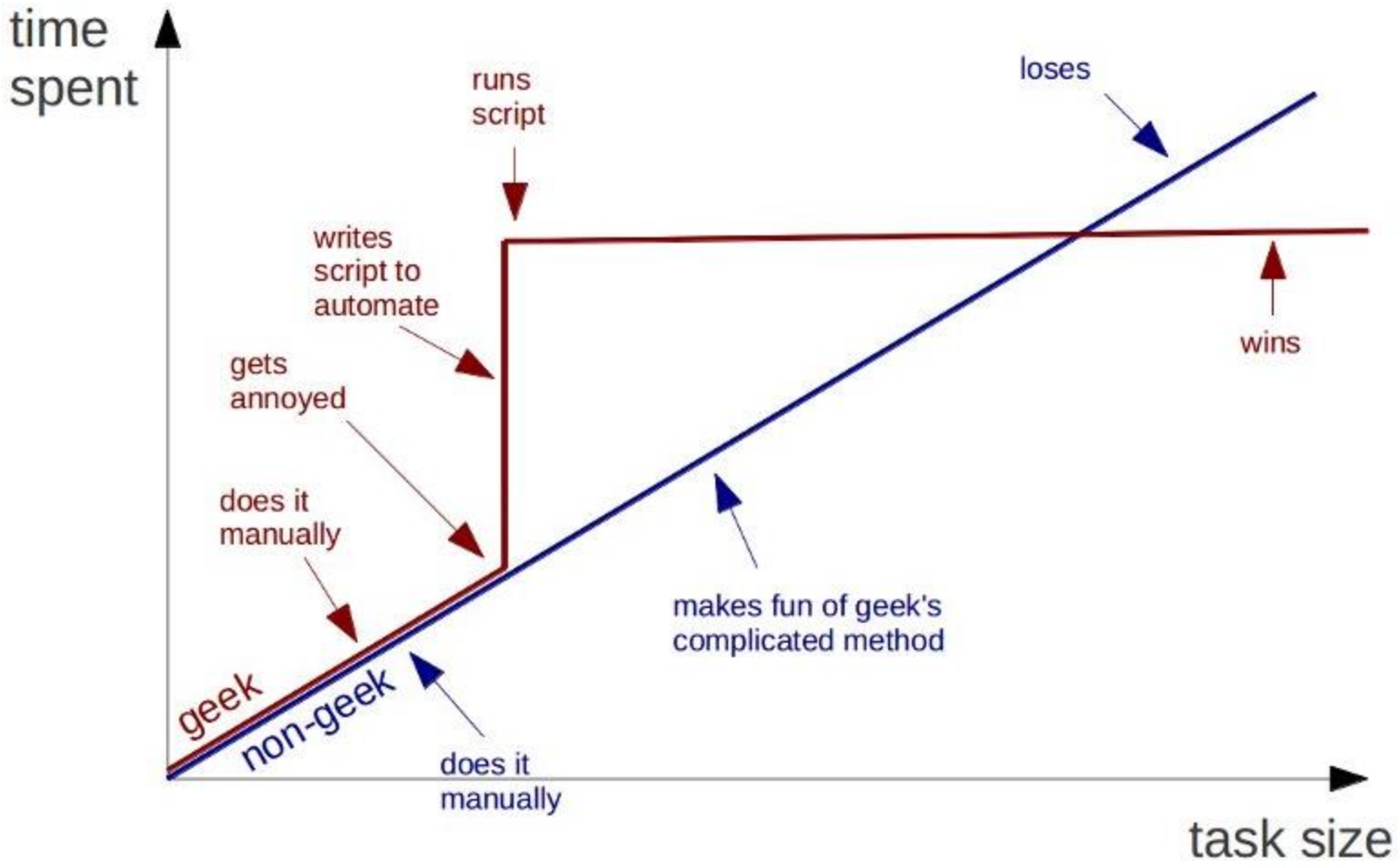
# Why R?




- freely available
- constant development
- professional visualization
- documentation and help
- large community (support)
- reproducability (code sharing)

https://nceas.github.io/oss-lessons/

# Geeks and repetitive tasks




https://nceas.github.io/oss-lessons/

EPFL

# Multidimensional ecological/environmental data

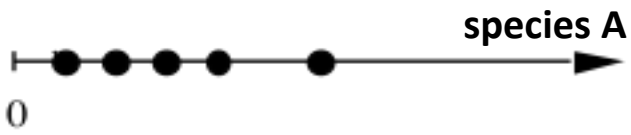
- **Univariate analysis -** one variable
- **Bivariate analysis -** two variables
- **Multivariate analysis** - more than two variables
  Every object (sample) is characterized by several descriptors
  Direct graphical representation is impossible beyond 3 dimensions

# Unidimensional data

Objects

samples

| Descriptors | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| species A | 1 | 4 | 2 | 1 | 6 | 3 |




# Multidimensional data

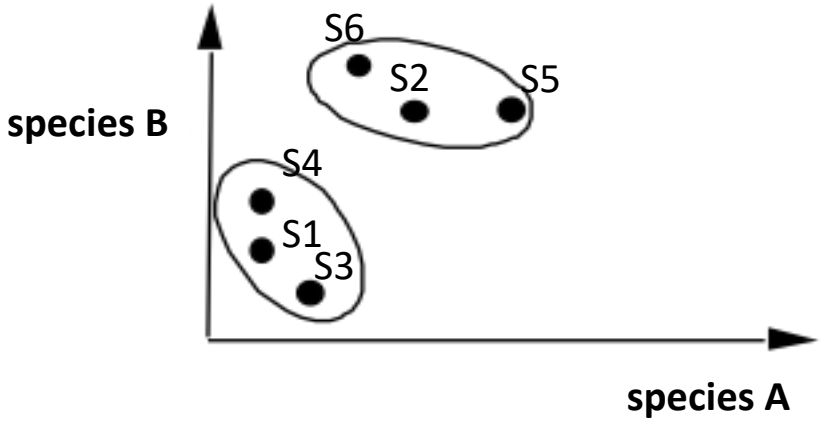
samples

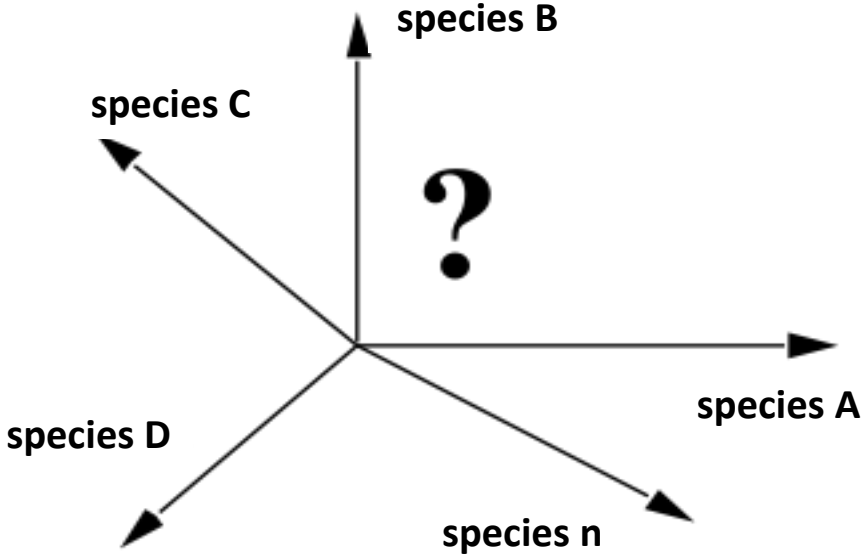
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| species A | 1 | 4 | 2 | 1 | 6 | 3 |
| species B | 2 | 5 | 1 | 3 | 5 | 6 |
| species C | 1 | 4 | 3 | 1 | 2 | 2 |
| species D | 3 | 1 | 6 | 5 | 6 | 2 |
| : | | | | | | |
| species n | 1 | 6 | 3 | 2 | 2 | 4 |

# Bidimensional data

samples

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| species A | 1 | 4 | 2 | 1 | 6 | 3 |
| species B | 2 | 5 | 1 | 3 | 5 | 6 |



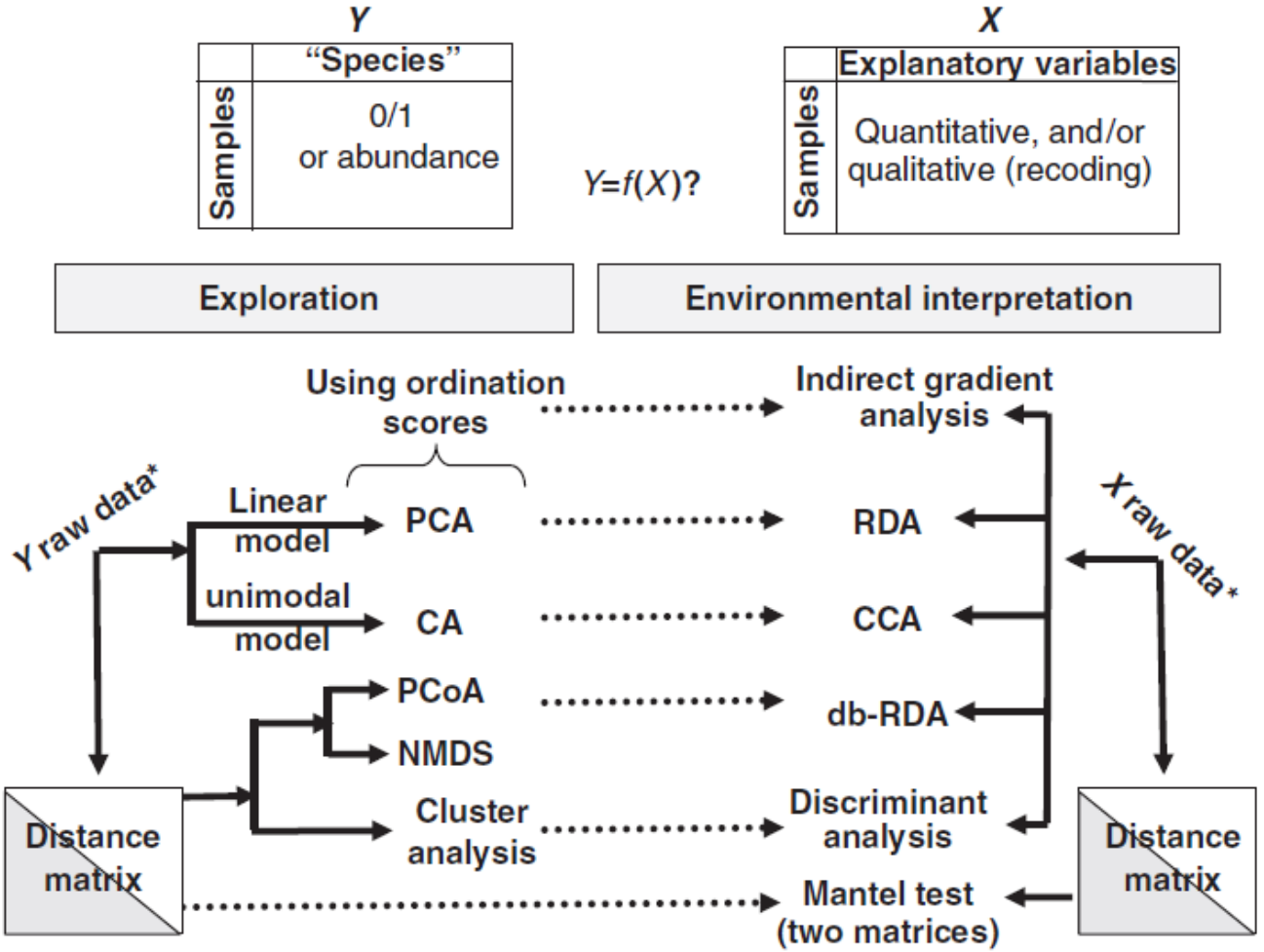

EPFL

# Types of scientific tasks

## most suited to the application of multivariate methods

- **Data reduction and simplification**
  - the summary of multiple variables via a small set of (synthetic) variables. High-dimensional patterns are presented in a lower-dimensional space, aiding interpretation.
  - *Principal Component Analysis*
- **Sorting and grouping**
  - Tasks concerned with the similarity of samples and their assignment to groups.
  - *Cluster analysis*
- **Investigation of dependence among variables**
  - Methods that detect dependence among variables are valuable in detecting influence or covariation.
  - *Redundancy Analysis*
- **Hypothesis construction and testing**
  - Exploratory techniques can reveal patterns in data from which hypotheses may be constructed.
  - *Mantel test, PERMANOVA*

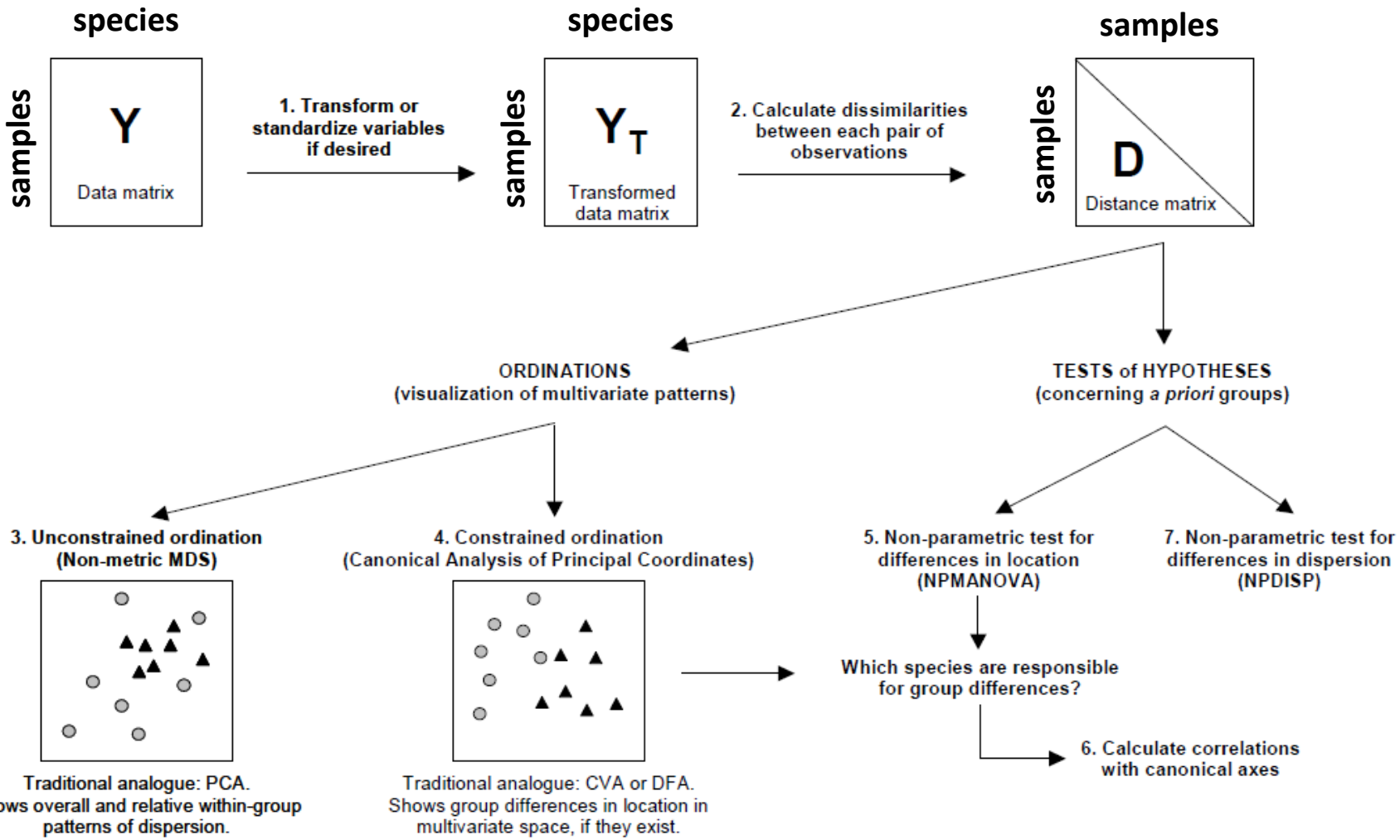
# Accurate choice of methods…






Ramette 2007

# Accurate choice of methods…



# Objects and descriptors

- Objects = observations
  - samples, field surveys, sites, experimental units

- Descriptors = measured variables
  - **Biological variables**
    - species, presence/absence, abundance
    - traits
    - evolutionary/phylogenetic relationships
    - activity measurements
  - **Environmental variables**
    - temperature, pH, soil type, nutrients…
  - **Spatial variables**
    - Geographical coordinates (x, y), island size, latitude, …

# Descriptor types

- Binary (boolean, qualitative with two modalities)
  - *Ex. : presence (1) or absence (0) of a species, terrestrial vs aquatic*
- Multiple
  - Unsorted (nominal, qualitative multiclass, categorial)
    - *Ex. : soil type, group affiliation (for instance following cluster analysis)*
  - Sorted
    - Semi-quantitative (ordinal, rank)
      - *Ex. : weak – medium – strong (coded 1 2 3)*
      - *Ex. : dominance code of a species (r + 1 2 3 4 5)*
    - Quantitative (cardinal)
      - Discreet (integer)
        - *Ex. : number of individuals of a species (abundance s.s.)*
      - Continuous (numerical)
        - *Ex. : biomass, altitude, activity rate measurement*
- Synthetic (complex)
  - *Ex. : relative abundance of a species*
  - *Ex. : C/N ratios of organic matter*

# Doubs

## abundance of 27 fish species along 30 sites in the river Doubs

| | Cogo | Satr | Phph | Neba | Thth | Teso | Chna | Chto | Lele | Lece | Baba | Spbi | Gogo | Eslu | Pefl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 5 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 5 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4 | 0 | 4 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2 | 2 |
| 5 | 0 | 2 | 3 | 2 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 | 2 | 4 | 4 |
| 6 | 0 | 3 | 4 | 5 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 1 | 1 |
| 7 | 0 | 5 | 4 | 5 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 1 | 4 | 4 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 1 | 0 | 0 |
| 11 | 1 | 3 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 12 | 2 | 5 | 4 | 4 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 2 | 5 | 5 | 2 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 3 | 5 | 5 | 4 | 4 | 3 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 15 | 3 | 4 | 4 | 5 | 2 | 4 | 0 | 0 | 3 | 3 | 2 | 0 | 2 | 0 | 0 |
| 16 | 2 | 3 | 3 | 5 | 0 | 5 | 0 | 4 | 5 | 2 | 2 | 1 | 2 | 1 | 1 |
| 17 | 1 | 2 | 4 | 4 | 1 | 2 | 1 | 4 | 3 | 2 | 3 | 4 | 1 | 1 | 2 |
| 18 | 1 | 1 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 3 | 3 | 3 | 2 | 1 | 3 |
| 19 | 0 | 0 | 3 | 5 | 0 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 4 | 1 | 1 |
| 20 | 0 | 0 | 1 | 2 | 0 | 0 | 2 | 2 | 2 | 3 | 4 | 3 | 4 | 2 | 2 |
| 21 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 2 | 2 | 2 | 4 | 2 | 5 | 3 | 3 |
| 22 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 2 | 3 | 4 | 5 | 1 | 5 | 3 | 4 |

## environmental parameter

| Variable | Code | Units |
|---|---|---|
| Distance from the source | dfs | km |
| Elevation | ele | m a.s.l. |
| Slope | slo | ‰ |
| Mean minimum discharge | dis | $m^3 \cdot s^{-1}$ |
| pH of water | pH | – |
| Hardness (Ca concentration) | har | $mg \cdot L^{-1}$ |
| Phosphate concentration | pho | $mg \cdot L^{-1}$ |
| Nitrate concentration | nit | $mg \cdot L^{-1}$ |
| Ammonium concentration | amm | $mg \cdot L^{-1}$ |
| Dissolved oxygen | oxy | $mg \cdot L^{-1}$ |
| Biological oxygen demand | bod | $mg \cdot L^{-1}$ |

## traits

| | sort | LatinName | Family | EnglishName | FrenchName | BodyLength | BodyLengthMax | ShapeFactor | TrophicLevel | omnivory |
|---|---|---|---|---|---|---|---|---|---|---|
| Cogo | 1 | Cottus gobio | Cottidae | Bullhead | Chabot commun | 80 | 120 | 3.13 | 3.10 | |
| Satr | 2 | Salmo trutta fario | Salmonidae | Brown trout | Truite fario | 280 | 800 | 5.03 | 4.00 | |
| Phph | 3 | Phoxinus phoxinus | Cyprinidae | Eurasian minnow | Vairon | 60 | 100 | 2.95 | 3.10 | |
| Babl | 4 | Barbatula barbatula | Nemacheilidae | Stone loach | Loche franche | 80 | 130 | 4.50 | 3.10 | |
| Thth | 5 | Thymallus thymallus | Salmonidae | Grayling | Ombre commun | 300 | 450 | 4.33 | 3.10 | |
| Teso | 6 | Telestes souffia | Cyprinidae | Vairone | Blageon | 90 | 180 | 5.26 | 3.40 | |
| Chna | 7 | Chondrostoma nasus | Cyprinidae | Common nase | Hotu | 280 | 480 | 6.50 | 2.00 | |
| Pato | 8 | Parachondrostoma toxostoma | Cyprinidae | South-west European nase | Toxostome | 160 | 250 | 5.00 | 2.00 | |
| Lele | 9 | Leuciscus leuciscus | Cyprinidae | Common dace | Vandoise | 180 | 360 | 2.81 | 2.57 | |
| Sqce | 10 | Squalius cephalus | Cyprinidae | European chub | Chevaine | 240 | 500 | 3.97 | 3.50 | |

## + spatial coordinates

EPFL

Verneaux J. 1973. - Cours d'eau de Franche-Comté (Massif du Jura).
Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie.

# paper to read for next class

*Annu. Rev. Ecol. Syst. 1990. 21:129–66*


## MULTIVARIATE ANALYSIS IN ECOLOGY AND SYSTEMATICS: PANACEA OR PANDORA'S BOX?

*Frances C. James*

Department of Biological Science, Florida State University, Tallahassee, Florida 32306

*Charles E. McCulloch*

Biometrics Unit, Cornell University, Ithaca, New York 14853

KEY WORDS: multivariate analysis, data analysis, statistical methods

---

### INTRODUCTION

Multivariate analysis provides statistical methods for study of the joint relationships of variables in data that contain intercorrelations. Because several variables can be considered simultaneously, interpretations can be made that are not possible with univariate statistics. Applications are now common in medicine (117), agriculture (218), geology (50), the social sciences (7, 178, 193), and other disciplines. The opportunity for succinct summaries of large data sets, especially in the exploratory stages of an investigation, has contributed to an increasing interest in multivariate methods.

The first applications of multivariate analysis in ecology and systematics were in plant ecology (54, 222) and numerical taxonomy (187) more than 30 years ago. In our survey of the literature, we found 20 major summaries of recent applications. Between 1978 and 1988, books, proceedings of symposia, and reviews treated applications in ecology (73, 126, 155, 156), ordination and classification (13, 53, 67, 78, 81, 83, 90, 113, 121, 122, 159), wildlife biology (33, 213), systematics (148), and morphometrics (45, 164,



0066-4162/90/1120-0129$02.00

# Quiz