R project: powerful tool for vegetation data analysis and visualization

David Zeleny

Masaryk Univerzity Brno, Czech Republic

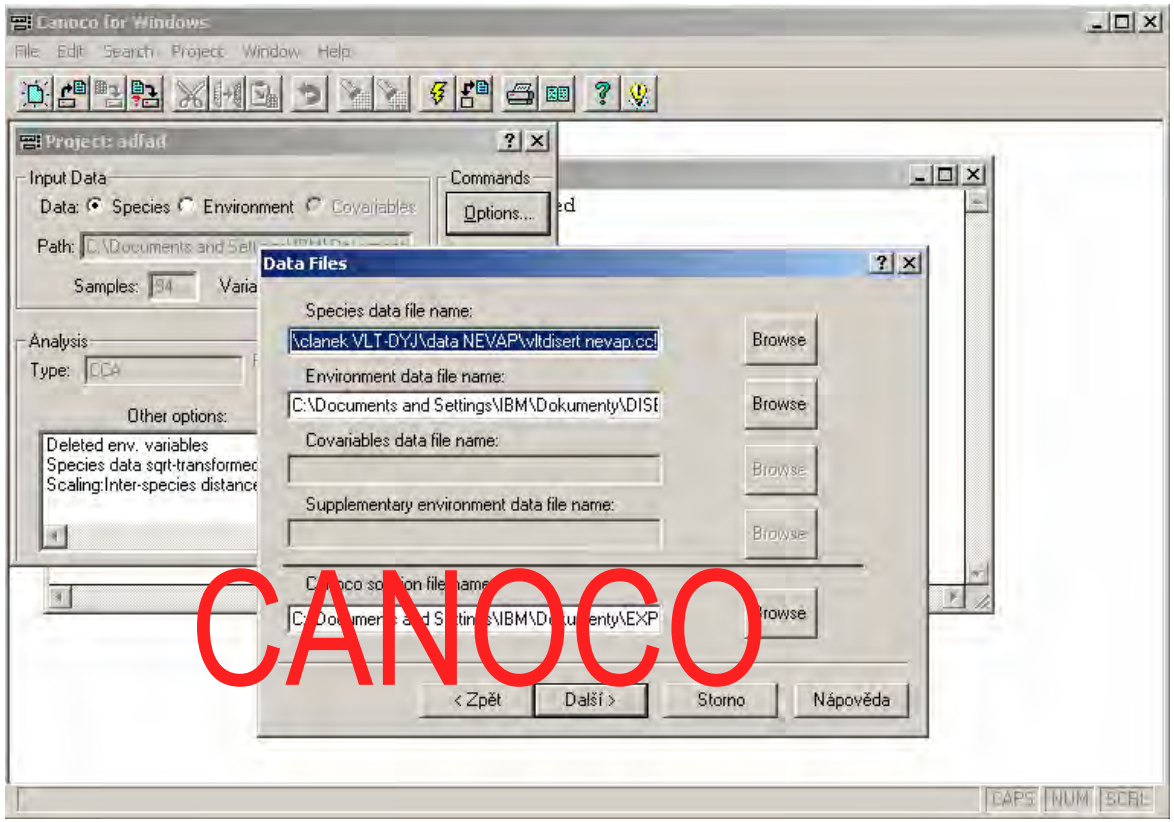
Overview of my talk

- Short introduction: what is R and what is R not?
- What says Web of Science about R?
- What R offers to vegetation ecologists?
- Examples of using R

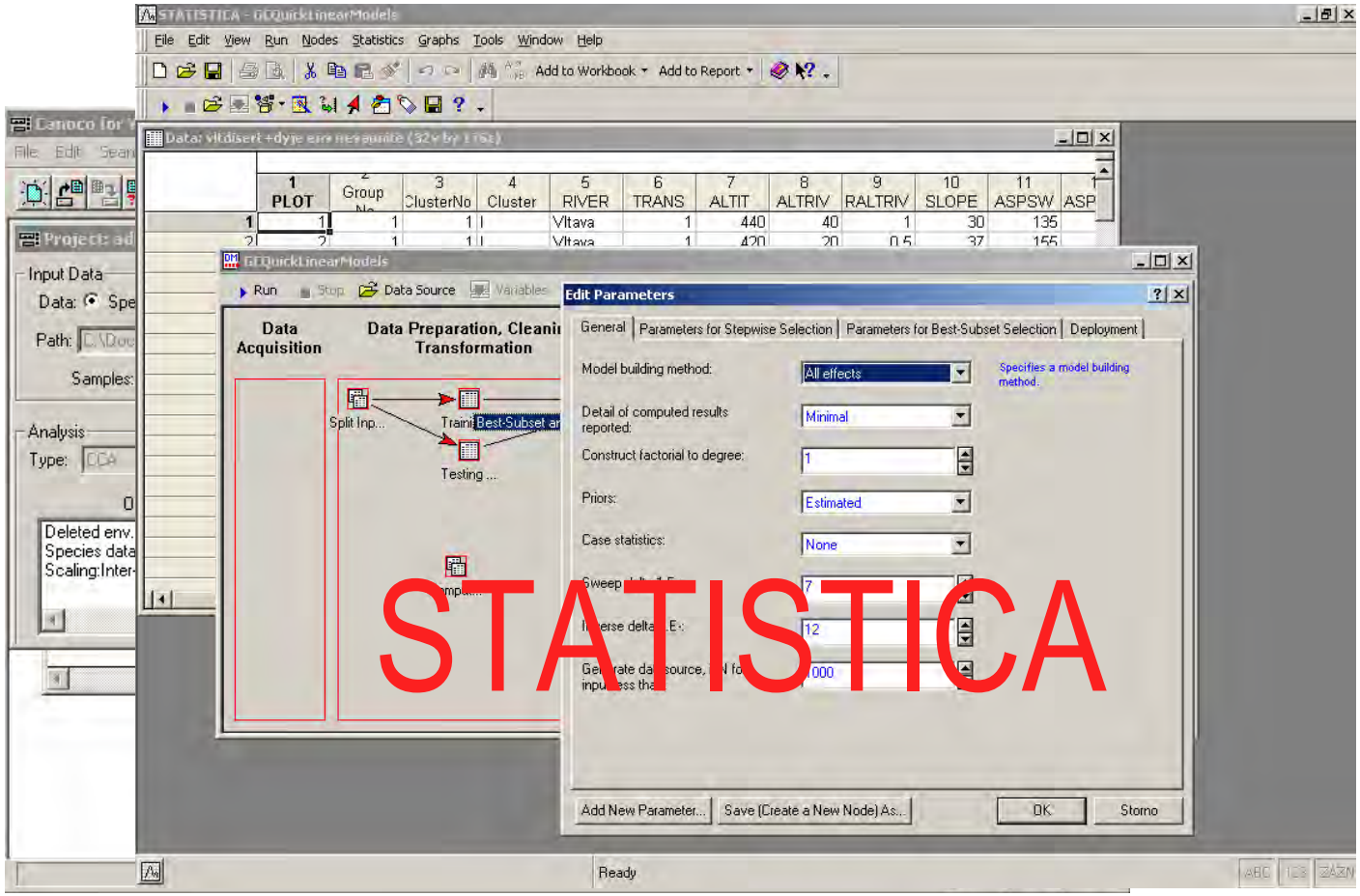
What is R project?

- Programming language based on S syntax
- Free software package, frequently updated
- Open community – everybody is welcome to contribute changes, improvement or new functions to this software
- People, who like to meet and discuss (yearly conferences useR! held in different parts of the world)

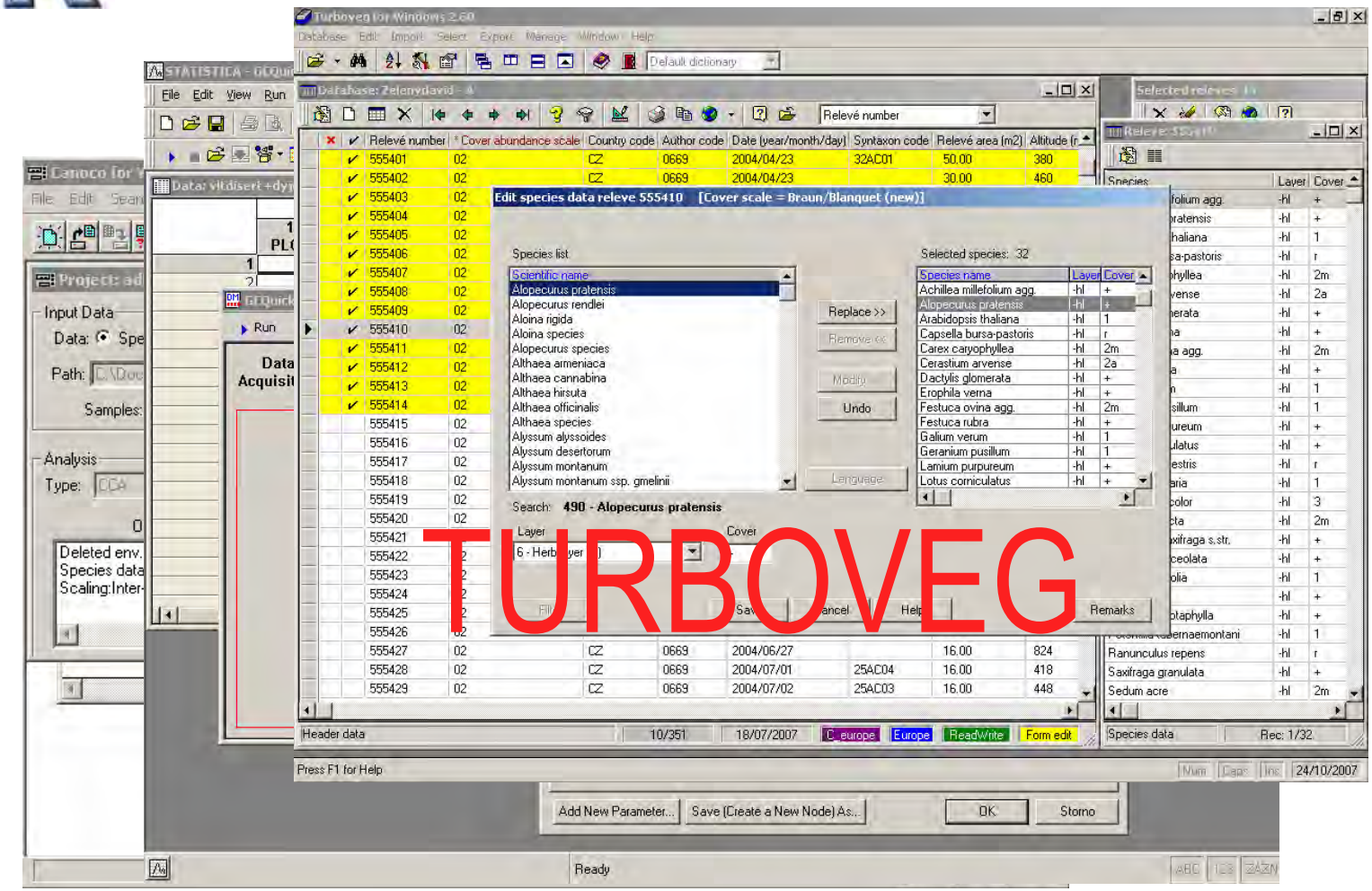
What R is not?


Canoco for Windows
File Edit Search Project Window Help
Project: adfad
Input Data
Data: Species Environment Covariables
Path:
Samples:
Analysis
Type: CCA
Other options:
Deleted env. variables
Species data sqrt-transformed
Scaling:Inter-species distance
Commands
Options...
Data Files
Species data file name:
\clanek VLT-DYJ\data NEVAP\vltdisert nevap.cc
Browse
Environment data file name:
C:\Documents and Settings\IBM\Dokumenty\DISE
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
Browse
Browse
< Zpět
Další >
Storno
Nápověda
CAPS
NUM
SCRL
CANOCO

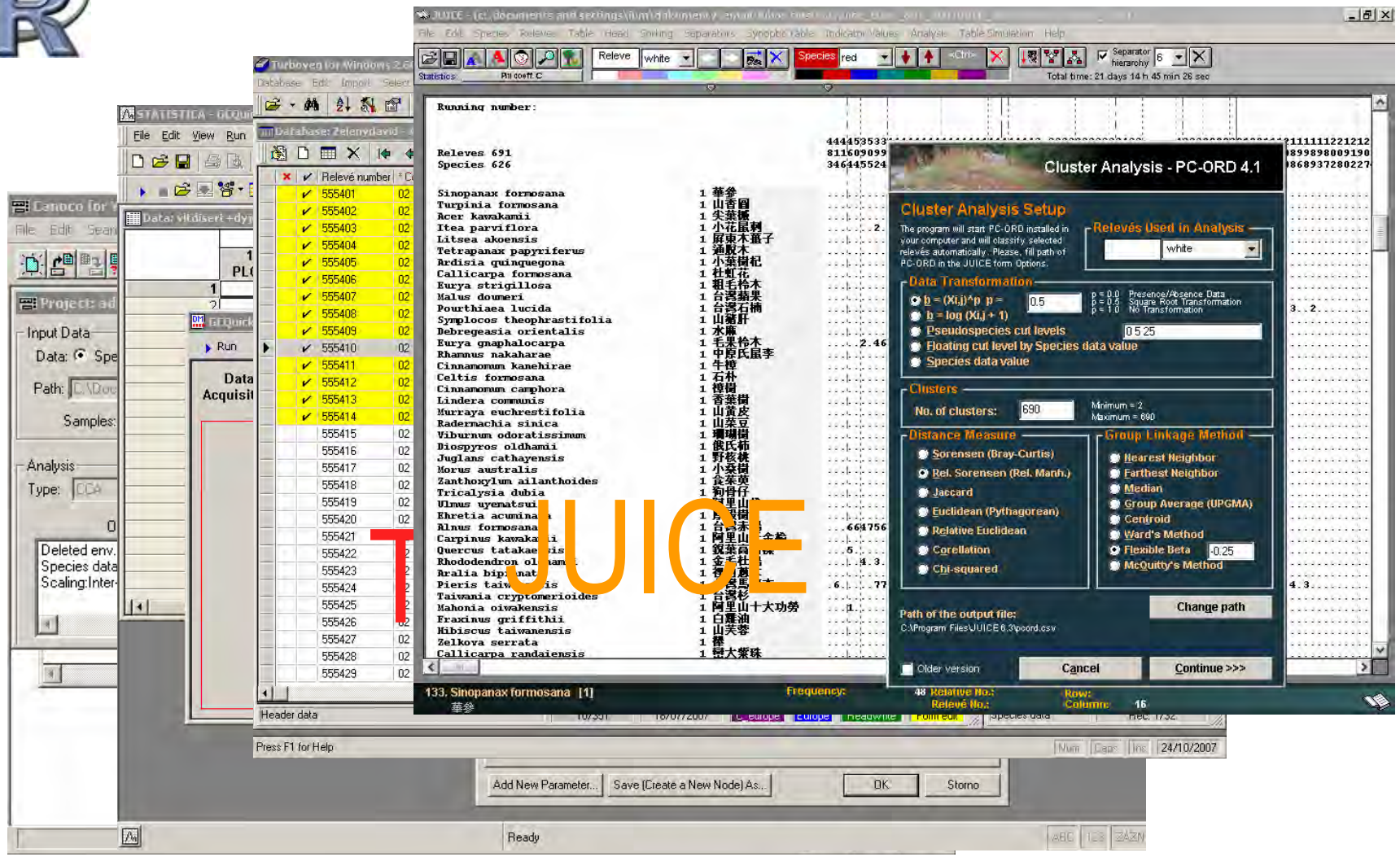
What R is not?


STATISTICA - GDQuickLinearModels
File Edit View Run Nodes Statistics Graphs Tools Window Help
Add to Workbook
Add to Report
Data: vltdisert +dyje sire nerasunite (32v by 1156)
1 PLOT
2 Group No
3 ClusterNo
4 Cluster
5 RIVER
6 TRANS
7 ALTIT
8 ALTRIV
9 RALTRIV
10 SLOPE
11 ASPSW
ASP
1 1 1 1 Vltava 1 440 40 1 30 135
2 2 1 1 Vltava 1 420 20 0.5 37 155
GDQuickLinearModels
Run
Stop
Data Source
Variables
Data Acquisition
Data Preparation, Cleani Transformation
Split Inp...
Testing ...
Edit Parameters
General
Parameters for Stepwise Selection
Parameters for Best-Subset Selection
Deployment
Model building method:
All effects
Specifies a model building method.
Detail of computed results reported:
Minimal
Construct factorial to degree:
1
Priors:
Estimated
Case statistics:
None
7
12
000
Add New Parameter...
Save (Create a New Node) As...
OK
Storno
Ready
Input Data
Data:
Path:
Samples:
Analysis
Type:
CCA
Deleted env.
Species data
Scaling:Inter-
STATISTICA

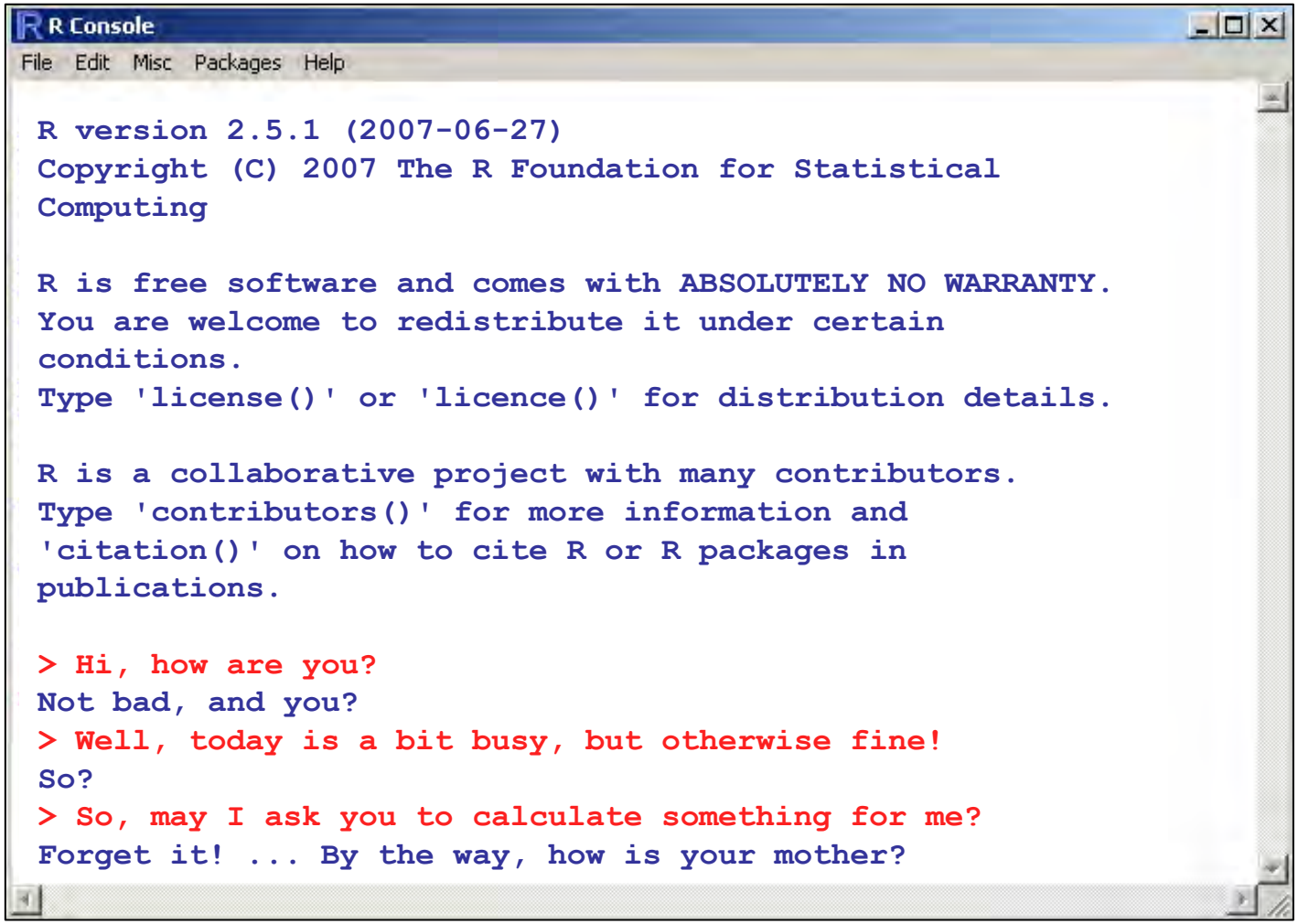
What R is not?



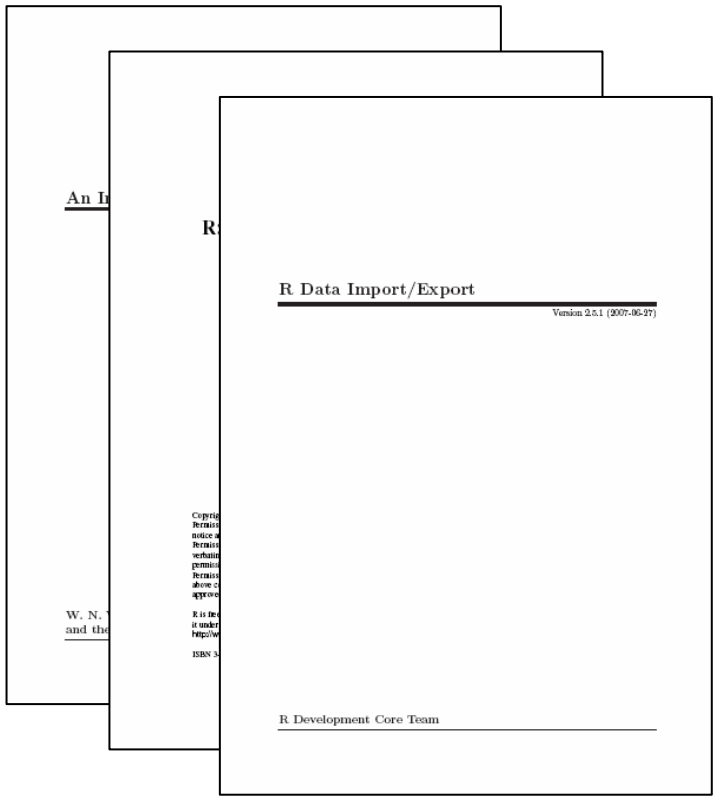
What R is not?


Cluster Analysis - PC-ORD 4.1
Cluster Analysis Setup
Data Transformation
Clusters
Distance Measure
Group Linkage Method
Change path
Cancel
Continue >>>
JUICE

R has very simple interface...



What comes with R?



+ so called “idiot’s guides”

What are the strong features of R?

- program structure is open – you can expand functionality simply by downloading and installing add-on packages from central repository (CRAN, recently around 1.300 packages)

- you can cook everything in one pot – import data, analyze them, export result and draw nice figures!

- there are extensive discussion forums on internet, where you can put your question dealing with R and you can be almost sure, that it will be answered

and (of course) some disadvantages...

- you need to spend some time and energy to learn it
- it could have some problems with analyzing large datasets
- Windows version works much less effectively due to Windows system limitations – so if you can, go for Linux version!
- ... and much more, if you start to use, you will see...

Web of Science Analysis

How does it work:

- I searched for papers published in SCI journals with reference to the R project

  R Development Core Team (2006 and older). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. – and similar references

- and sorted these papers according to the subject category (mathematic, statistic, ecology, plant science etc.), publication year, journal name and the country of authors.

Web of Science®

Web of Science Analysis

Which field of science most commonly publishes results analyzed using R project?

Field: Subject Category	Record Count	% of 2893	Bar Chart
ECOLOGY	549	18.9768 %	
STATISTICS & PROBABILITY	306	10.5773 %	
GENETICS & HEREDITY	216	7.4663 %	
ENVIRONMENTAL SCIENCES	178	6.1528 %	
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	164	5.6689 %	
ZOOLOGY	156	5.3923 %	
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	150	5.1849 %	
BIOCHEMISTRY & MOLECULAR BIOLOGY	140	4.8393 %	
PLANT SCIENCES	135	4.6664 %	
MATHEMATICAL & COMPUTATIONAL BIOLOGY	127	4.3899 %	

Web of Science®

Web of Science Analysis

How has the number of papers using R project increased during the last 6 years?

(papers from all subject categories are included...)

Field: Publication Year	Record Count	% of 2893	Bar Chart
2006	1172	40.5116 %	
2007	1069	36.9513 %	
2005	547	18.9077 %	
2004	93	3.2147 %	
2003	7	0.2420 %	
2002	4	0.1383 %	

Web of Science®

Web of Science Analysis

Which journals most frequently publish papers using R project?

(considered subject categories: Ecology and Plant Science)

ECOLOGY	42	6.2038 %
JOURNAL OF APPLIED ECOLOGY	28	4.1359 %
AMERICAN NATURALIST	26	3.8405 %
BIOLOGICAL CONSERVATION	25	3.6928 %
JOURNAL OF ANIMAL ECOLOGY	22	3.2496 %
OECOLOGIA	22	3.2496 %
MOLECULAR ECOLOGY	21	3.1019 %
ECOLOGICAL MODELLING	20	2.9542 %
MARINE ECOLOGY-PROGRESS SERIES	19	2.8065 %
OIKOS	17	2.5111 %
ECOLOGICAL APPLICATIONS	16	2.3634 %
ECOLOGY LETTERS	15	2.2157 %
EVOLUTION	15	2.2157 %
JOURNAL OF BIOGEOGRAPHY	15	2.2157 %
JOURNAL OF ECOLOGY	14	2.0679 %
JOURNAL OF EVOLUTIONARY BIOLOGY	14	2.0679 %
JOURNAL OF WILDLIFE MANAGEMENT	14	2.0679 %
THEORETICAL AND APPLIED GENETICS	14	2.0679 %
NEW PHYTOLOGIST	13	1.9202 %
AGRICULTURE ECOSYSTEMS & ENVIRONMENT	12	1.7725 %
BEHAVIORAL ECOLOGY AND SOCIOBIOLOGY	12	1.7725 %
GLOBAL ECOLOGY AND BIOGEOGRAPHY	11	1.6248 %
ECOLOGICAL RESEARCH	9	1.3294 %
PLANT AND SOIL	9	1.3294 %
BIODIVERSITY AND CONSERVATION	8	1.1817 %

Web of Science®

Web of Science Analysis

TOP 23 countries

using R project for publication by their scientists (all fields of science included)

Field: Country/Territory	Record Count	% of 2893	Bar Chart
USA	1005	34.7390 %	
GERMANY	345	11.9253 %	
ENGLAND	340	11.7525 %	
FRANCE	268	9.2637 %	
AUSTRALIA	219	7.5700 %	
CANADA	185	6.3947 %	
SWITZERLAND	177	6.1182 %	
ITALY	121	4.1825 %	
NORWAY	114	3.9405 %	
SPAIN	112	3.8714 %	
SWEDEN	111	3.8368 %	
NETHERLANDS	109	3.7677 %	
JAPAN	77	2.6616 %	
FINLAND	76	2.6270 %	
DENMARK	70	2.4196 %	
BRAZIL	66	2.2814 %	
NEW ZEALAND	65	2.2468 %	
AUSTRIA	61	2.1085 %	
SCOTLAND	60	2.0740 %	
BELGIUM	40	1.3826 %	
PORTUGAL	40	1.3826 %	
SOUTH AFRICA	30	1.0370 %	
CZECH REPUBLIC	29	1.0024 %	

Web of Science®

What does R offer to vegetation ecologists?

- Basic statistics – ANOVA & MANOVA stuff
- Multivariate analysis – ordination, cluster analysis
- Machine learning methods (supervised classification) – neural networks, CART, Random Forests etc.
- Species response modeling – GLM, GAM
- Diversity modeling
- Simulation studies and randomization procedures
- Analysis of spatial data (spatial autocorrelations etc.)
- Analysis of time series

Competition of DCA and NMDS

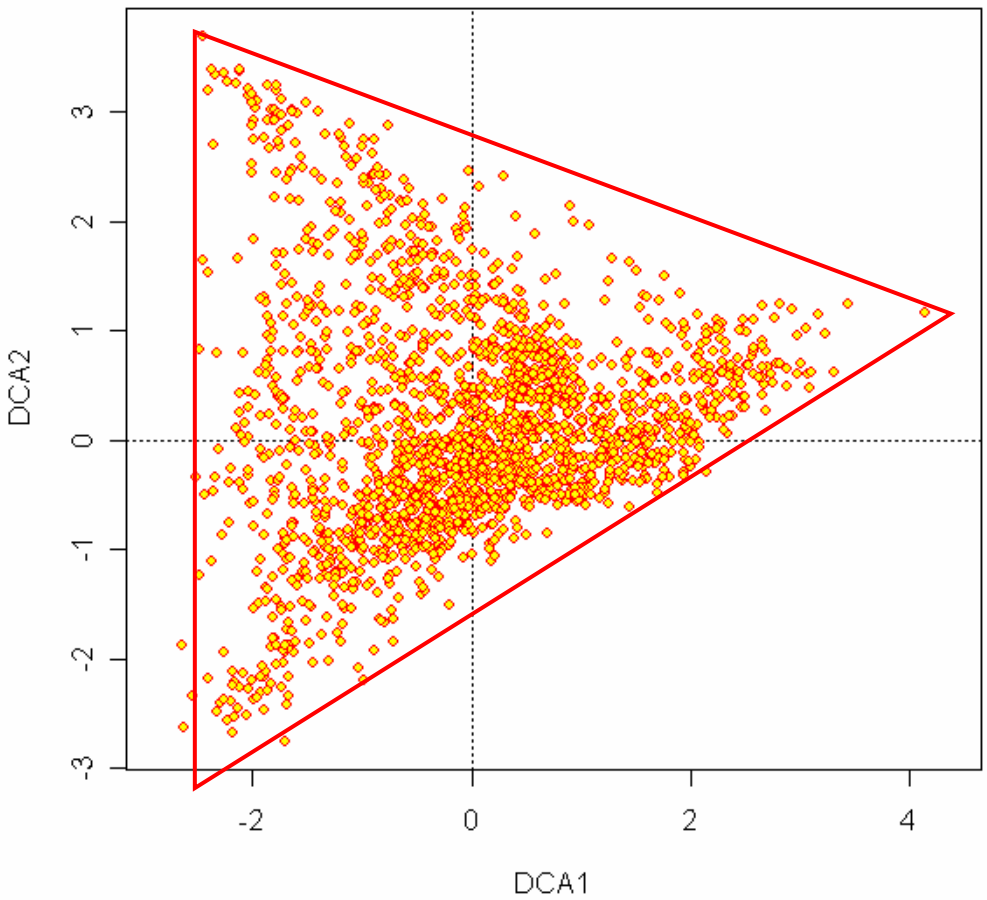
NMDS – non-metric multidimensional scaling

- non-metric alternative of unconstrained ordinations (PCA, DCA)
- the method tries to project samples into 2D figure, so as distances between these samples maximally correspond to the sample dissimilarities (measured by selected dissimilarity measure – Bray-Curtis etc.)

Why not DCA?

- DCA was originally developed by Mark Hill in order to correct some obvious problems of CA (Correspondence Analysis).
- This is done using two problematic steps – detrending along axes, which removes so called arch effect, and rescaling of axes, which removes packing of samples in the ends of the gradient.
- Detrending in reality works by twisting the ordination space, so it looks pretty in 2D, but not so in 3D and higher.
- Points will produce triangle or diamond shape – which is actually artifact of detrending!

Why not DCA?

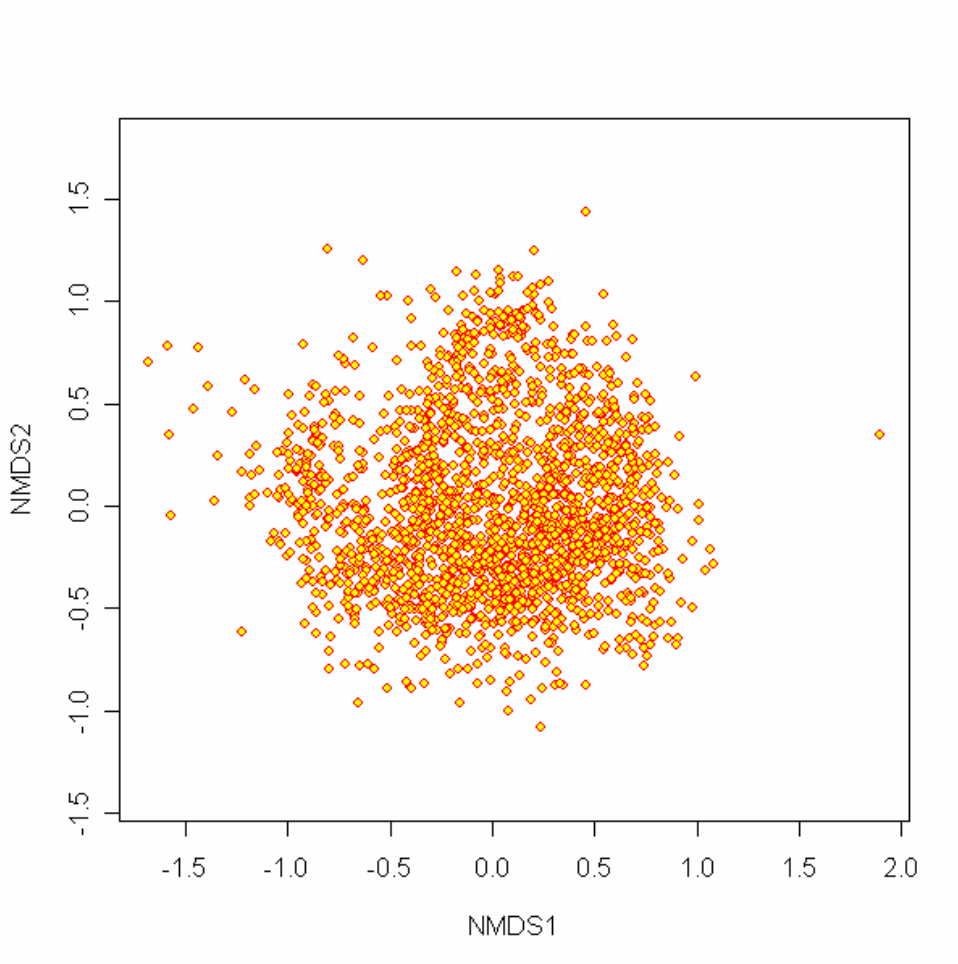


DCA analysis of 2000 forest releves randomly selected from Czech National Vegetation Database.

Why yes NMDS?

- It doesn’t suffer from the artifacts of DCA.
- According to Minchin (1987) it’s the most robust unconstraint ordination method in vegetation ecology.
- It’s non-metric method, and doesn’t assume the unimodal shape of species response curves.

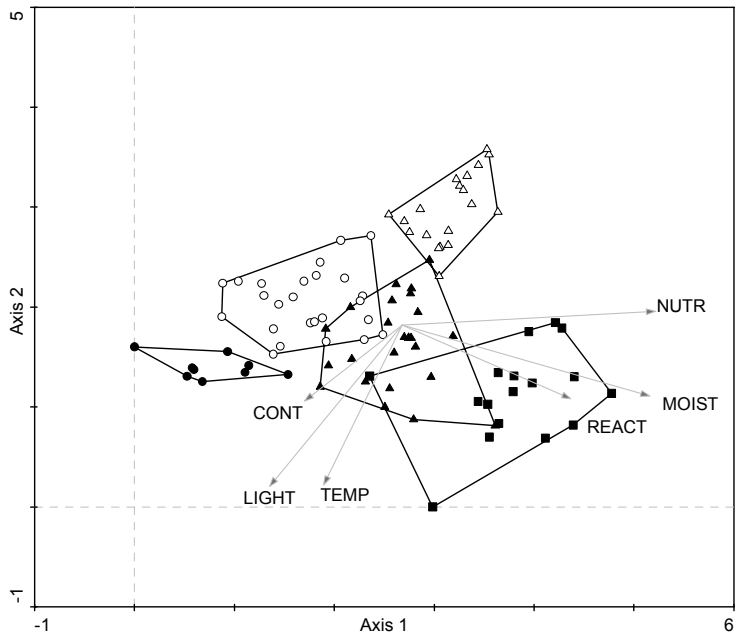
Why yes NMDS?



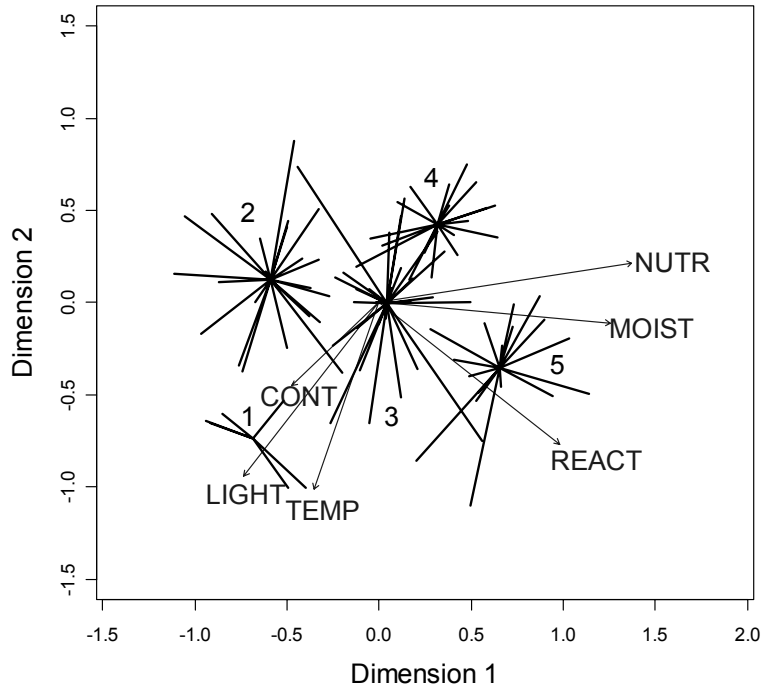
NMDS analysis of 2000 forest releves randomly selected from Czech National Vegetation Database.

Comparison of DCA and NMDS

DCA

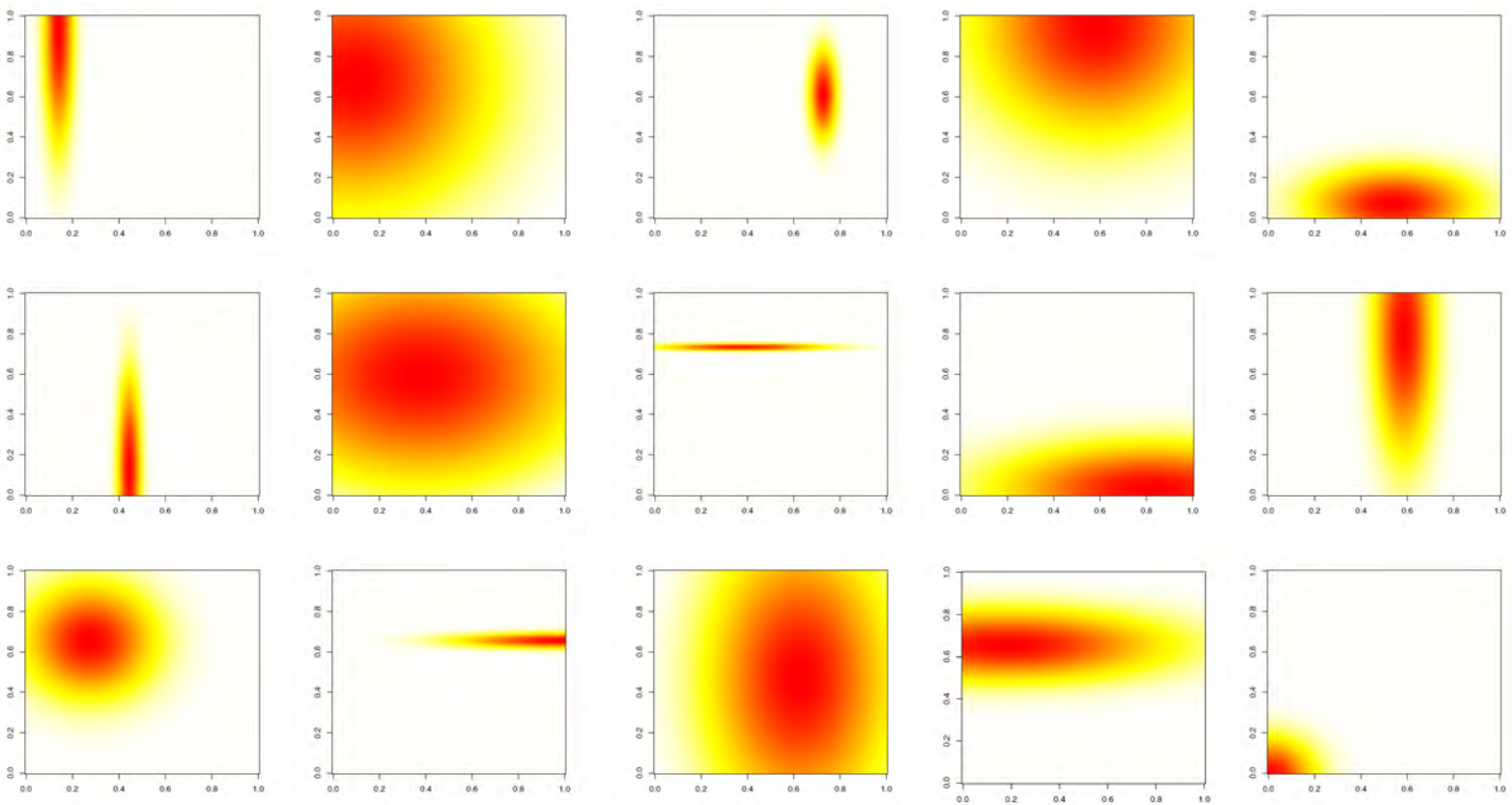


NMDS



Zelený D. & Chytrý M. (2007): Environmental control of vegetation pattern in deep river valleys of the Bohemian Massif. *Preslia*, 79: 205-222.

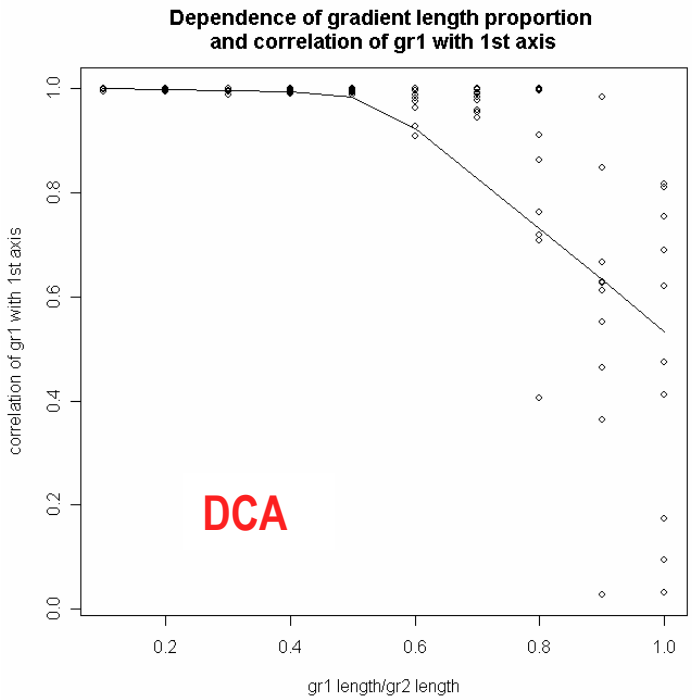
Simulation of species distribution in 2D ecospace

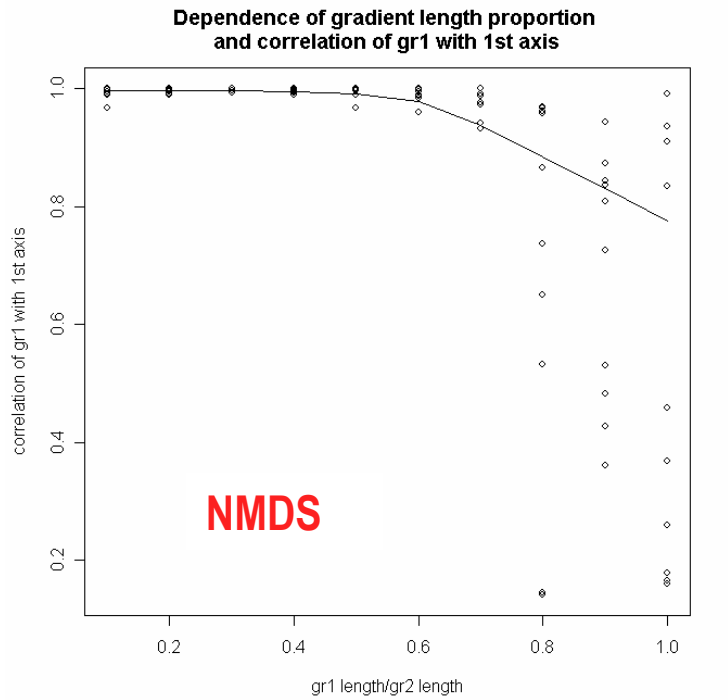


Which method can better recover simulated gradients in data?

- Simulated data contains two not-correlated gradients.
- **Question**: how successfully can DCA and NMDS recover these gradients from species-plot data?
- **Results:**
  - if one of the gradients is stronger (longer) than the other, both methods work well;
  - if the gradients are of similar length, both methods fails to recover the simulated pattern.

Which method can better recover simulated gradients in data?

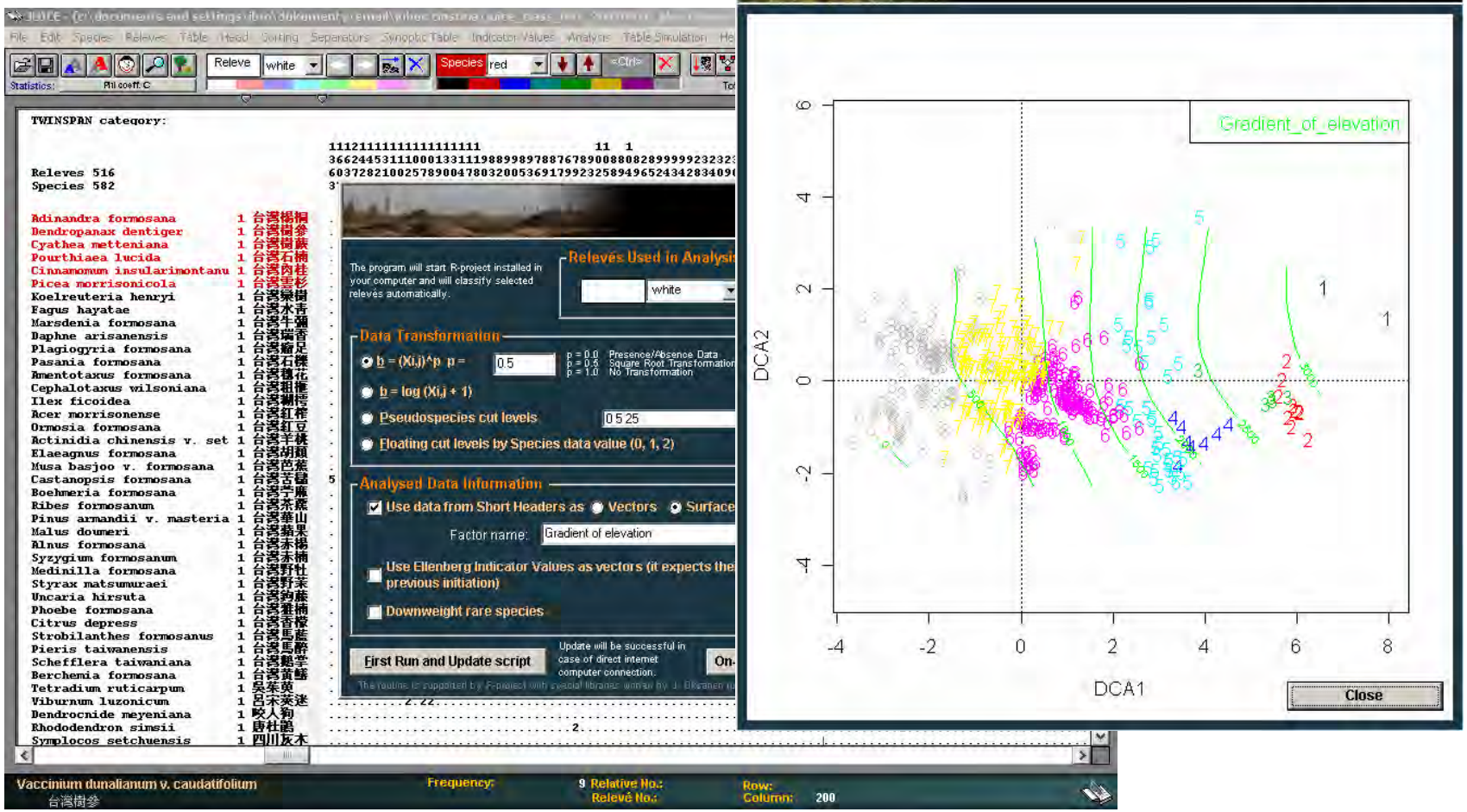




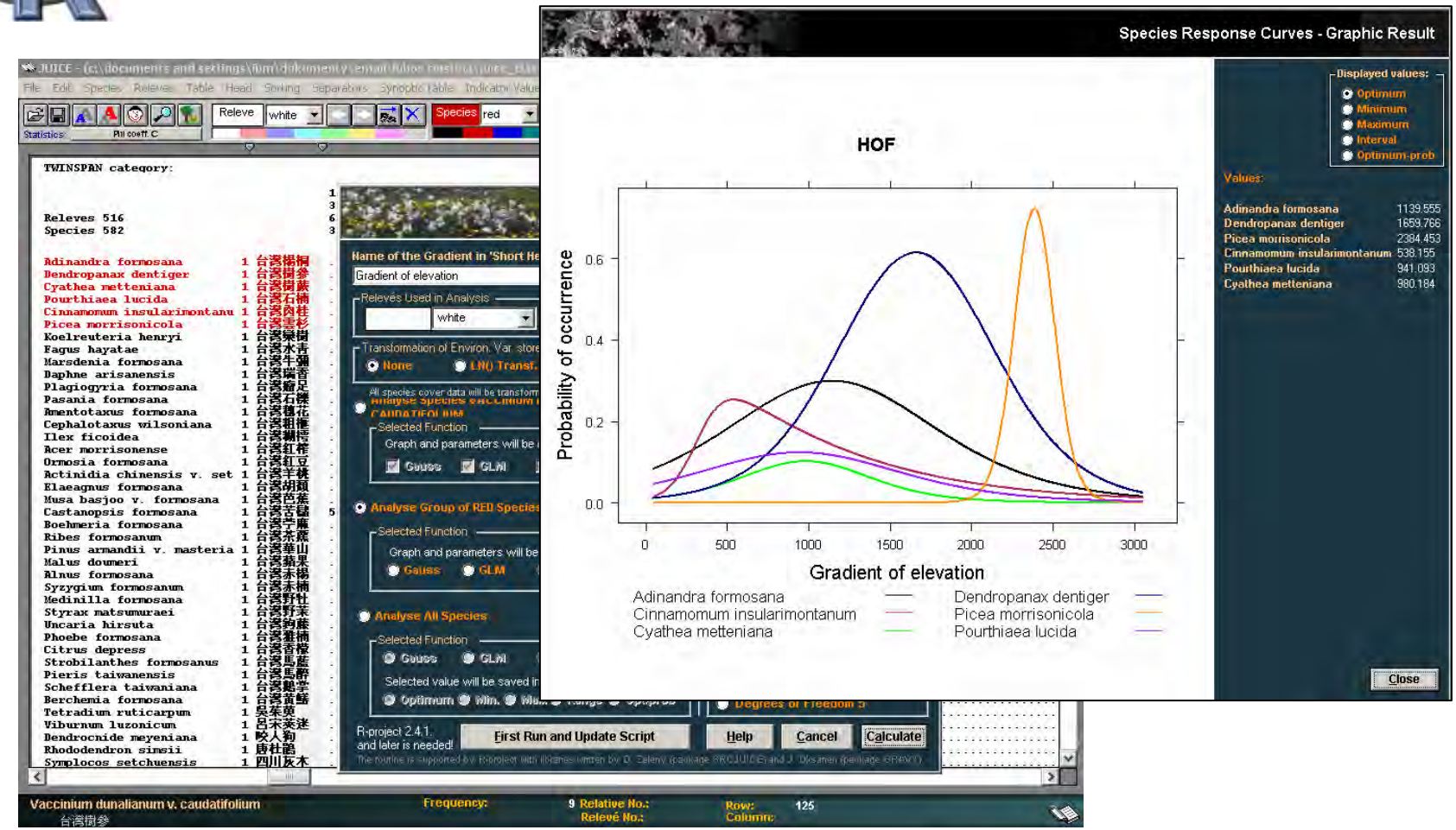
Employment of R by other software

&

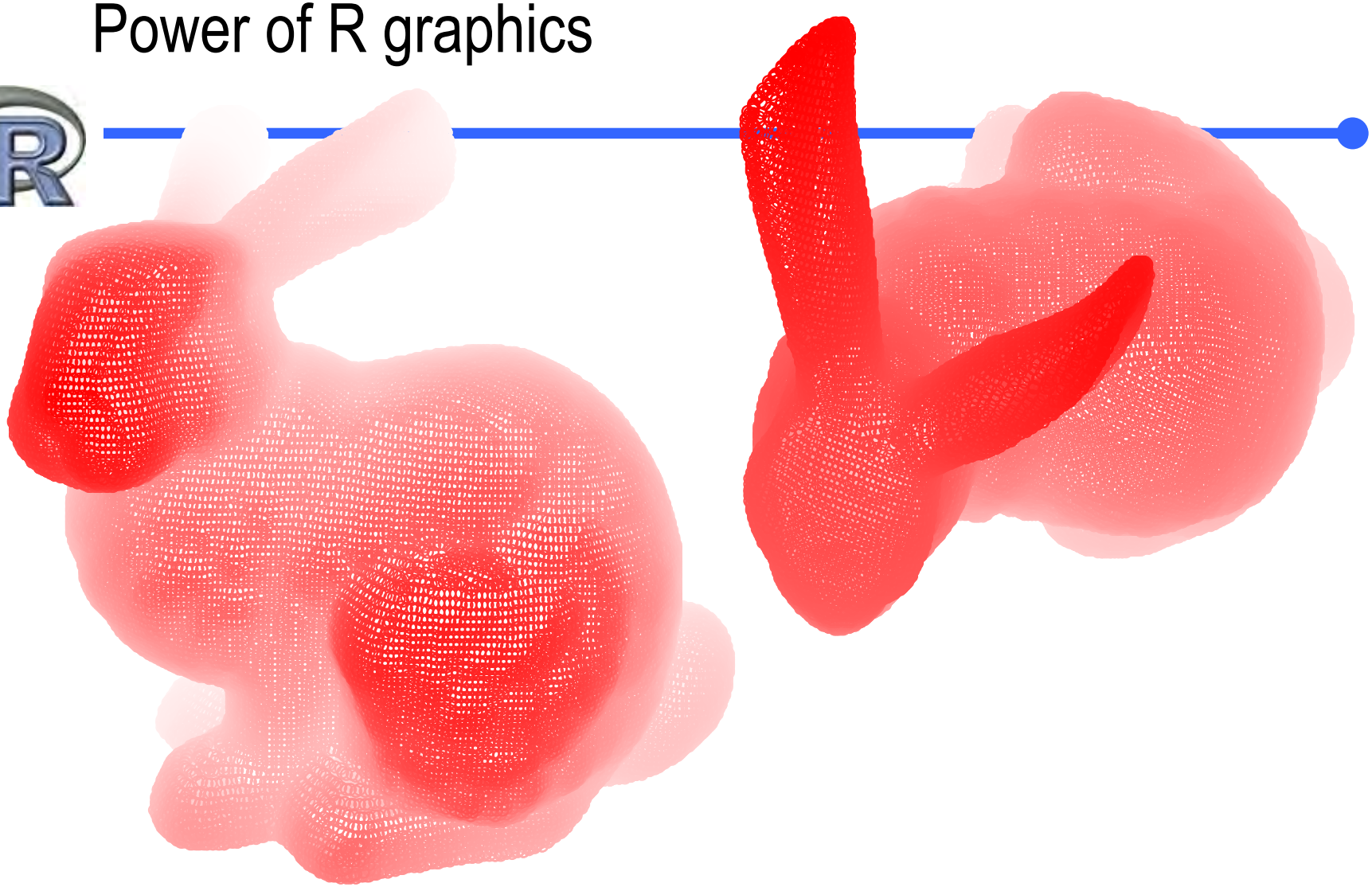
JUICE & DCA, NMDS and PCA



JUICE & Species Response Curves

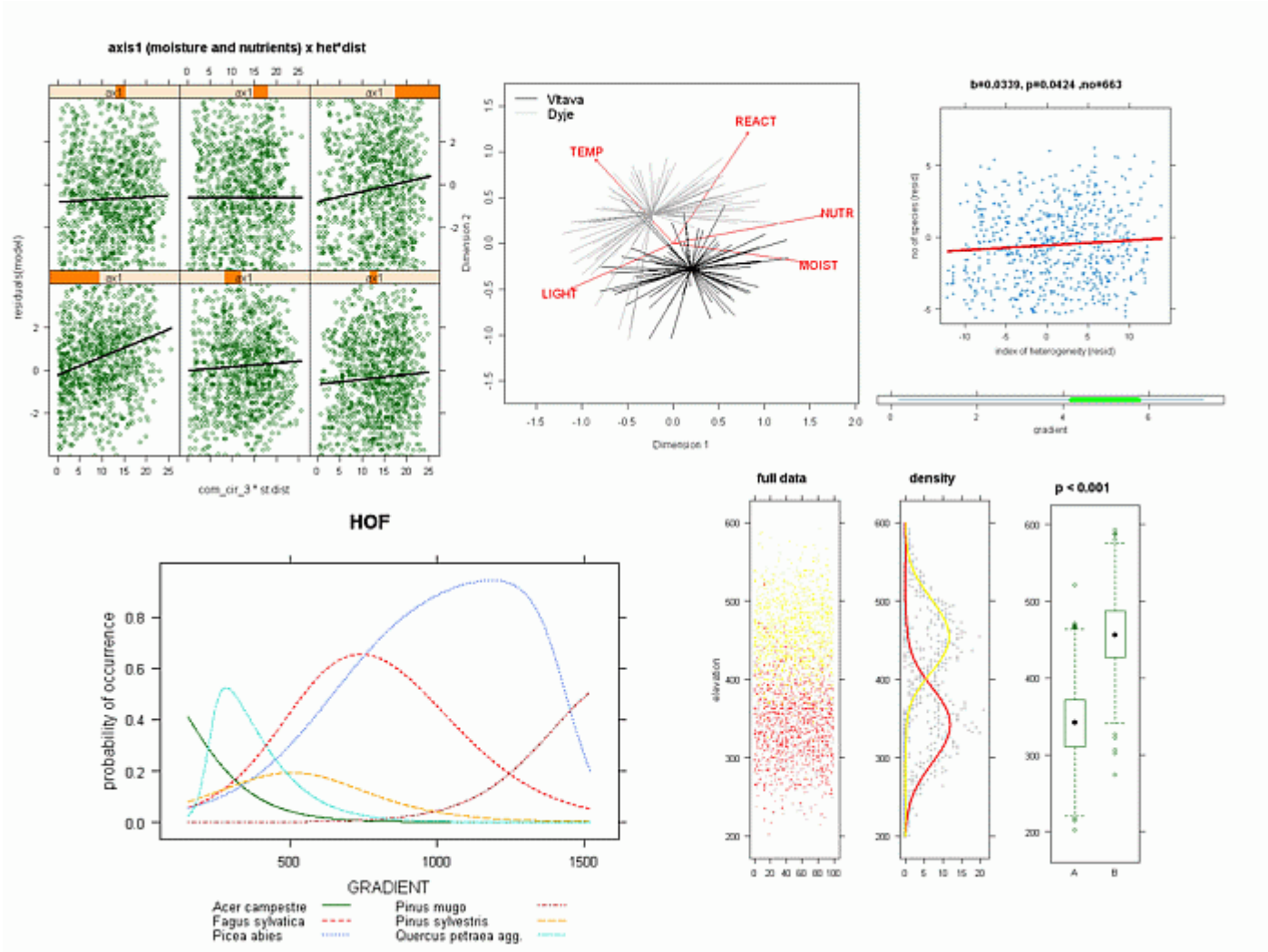


Power of R graphics

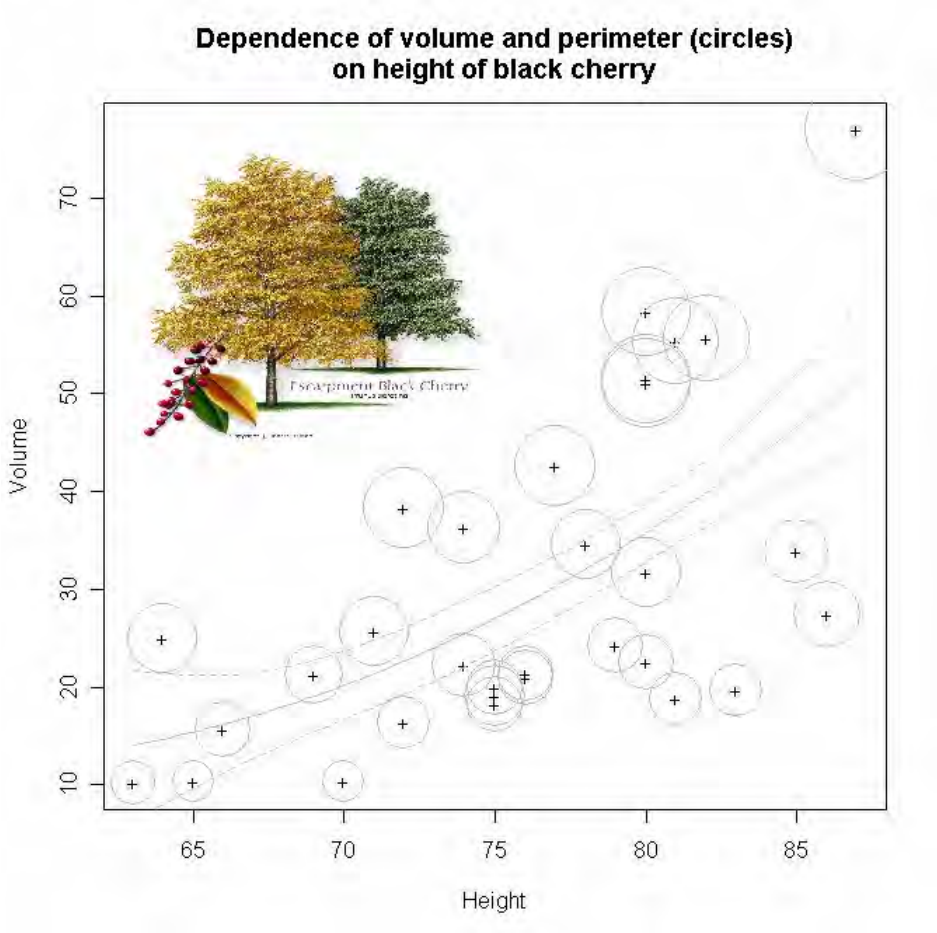


3D bunny, package *onion*, **author Robin K. S. Hankin** .

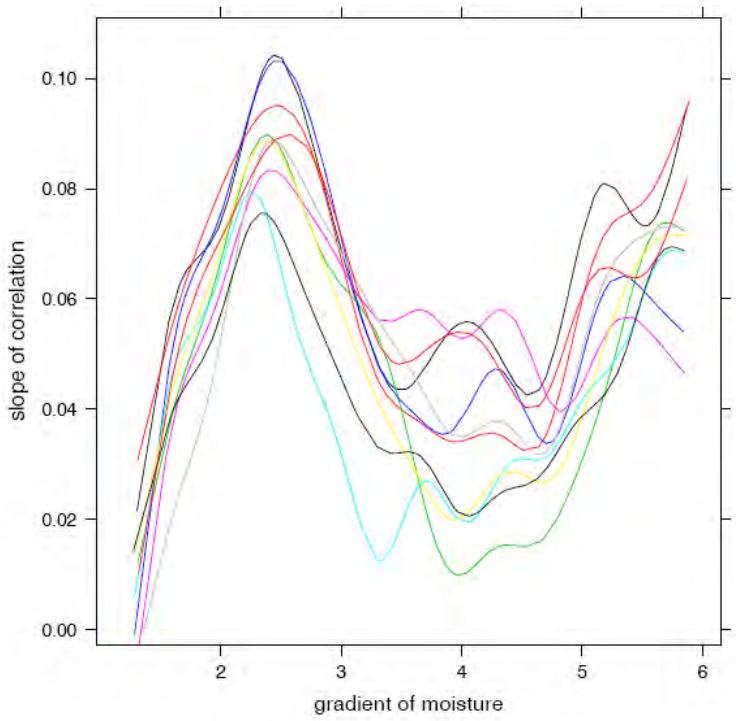
Some examples of graphical outputs from R

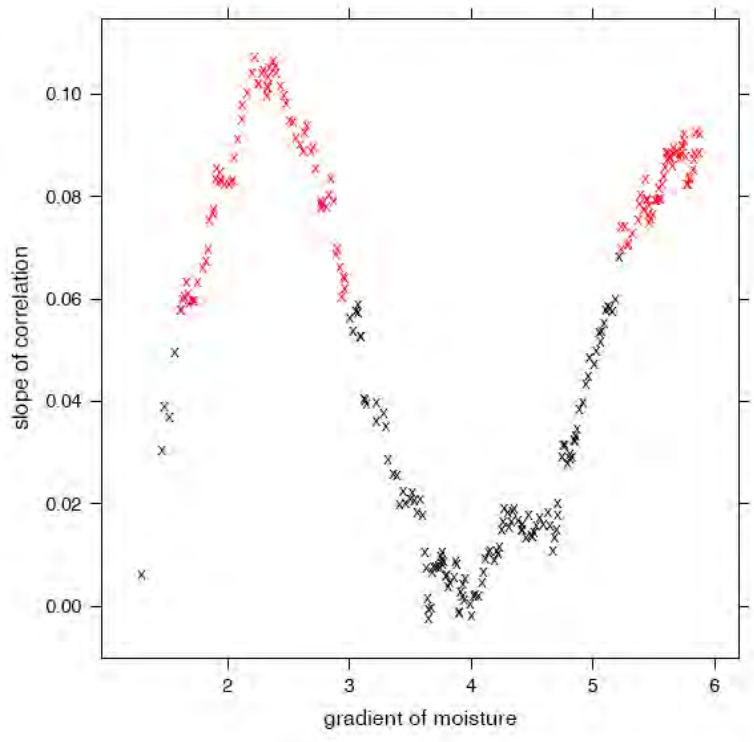


Cherry tree – relationship of volume, perimeter and tree height

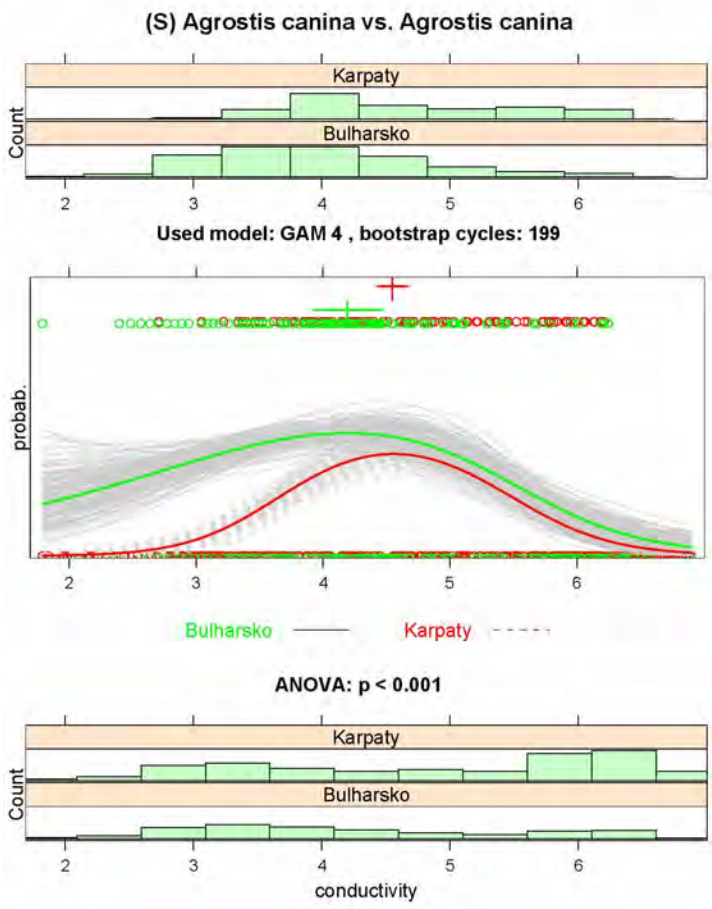


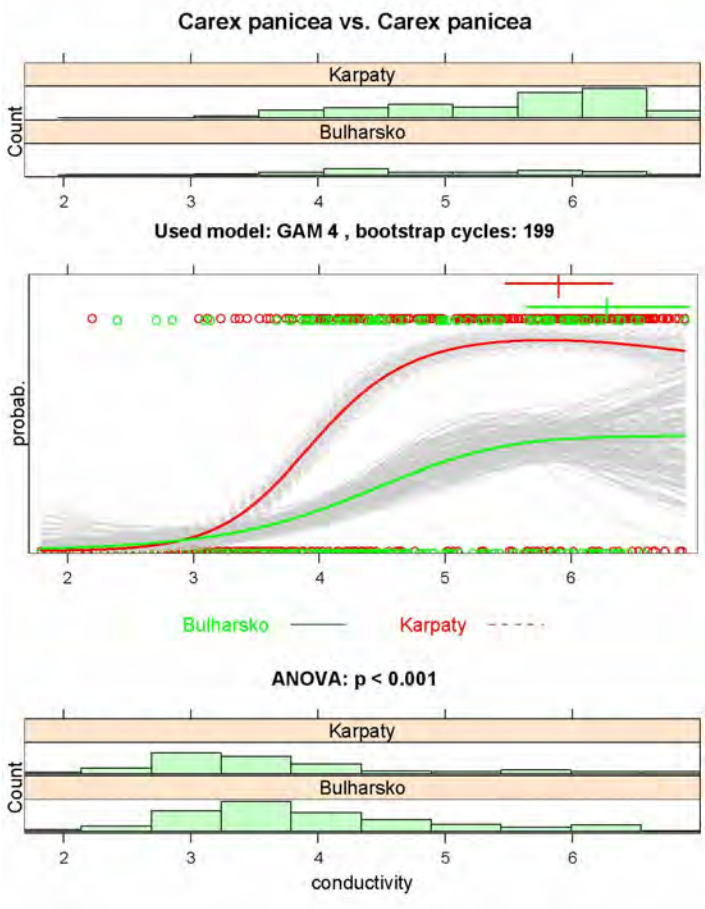
Randomized Results of Moving Window regression



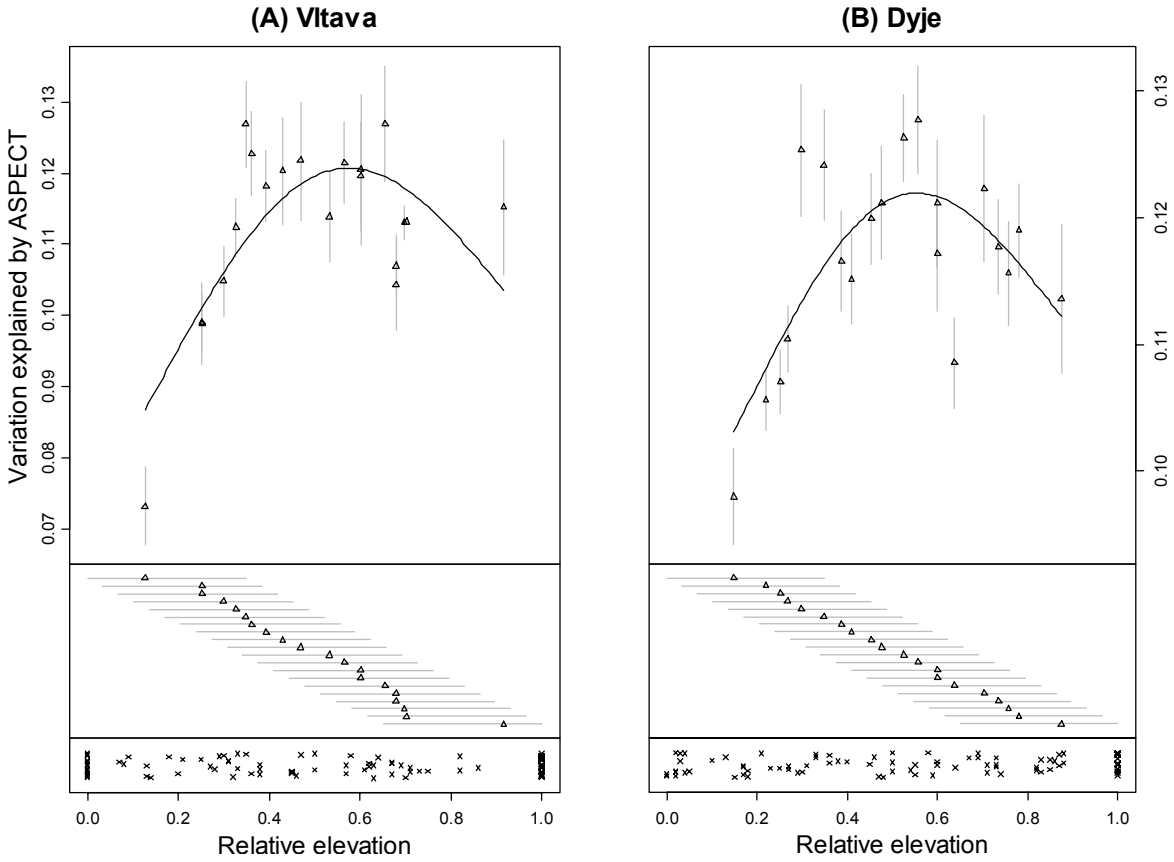


Bootstrapped species response curves



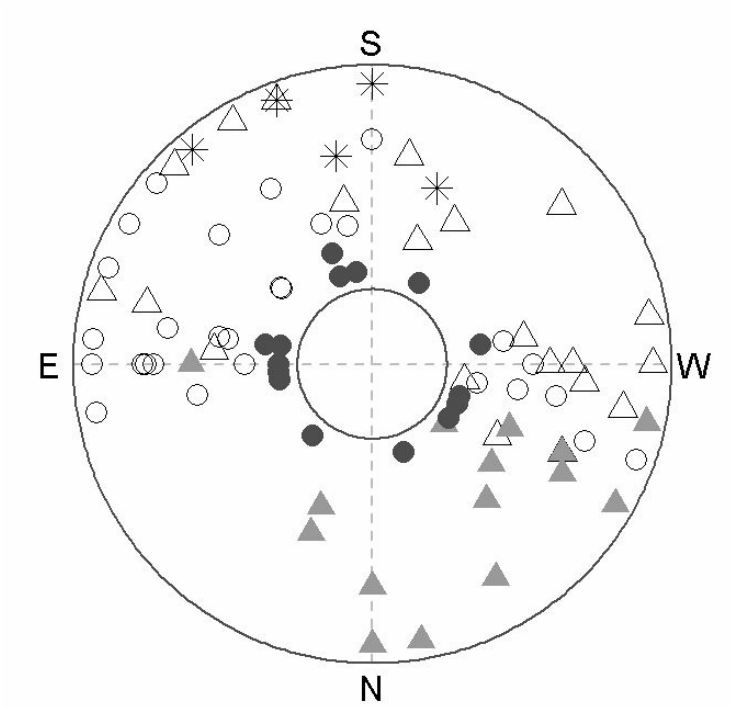


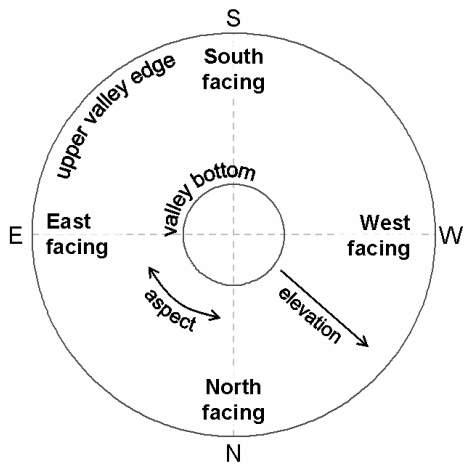
Moving frame CCA



Zelený D. & Chytrý M. (2007): Environmental control of vegetation pattern in deep river valleys of the Bohemian Massif. *Preslia*, 79: 205-222.

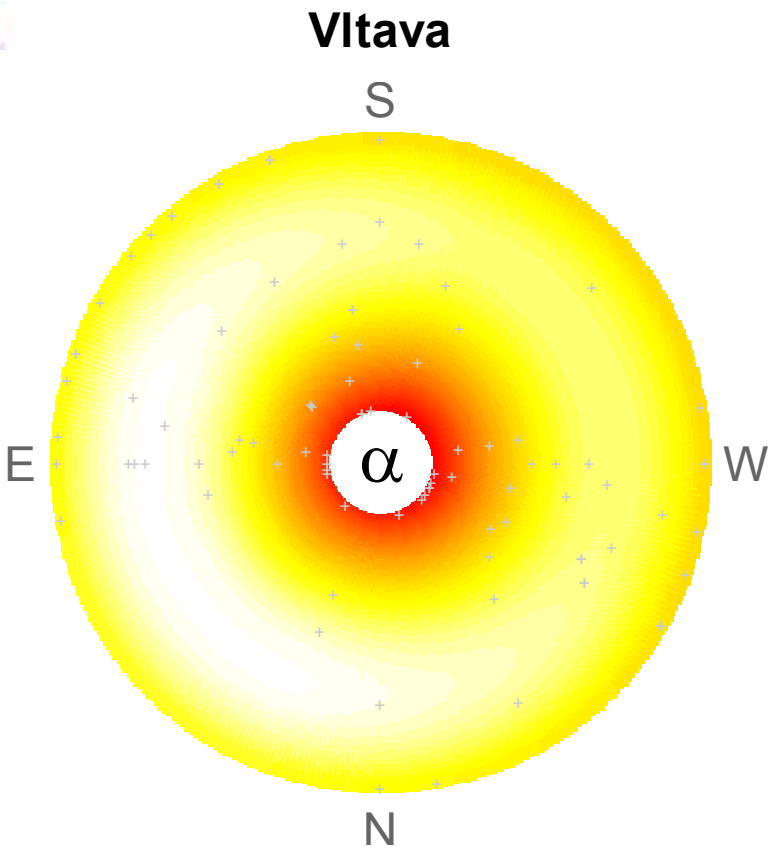
Iris diagram – distribution of vegetation types in river valley

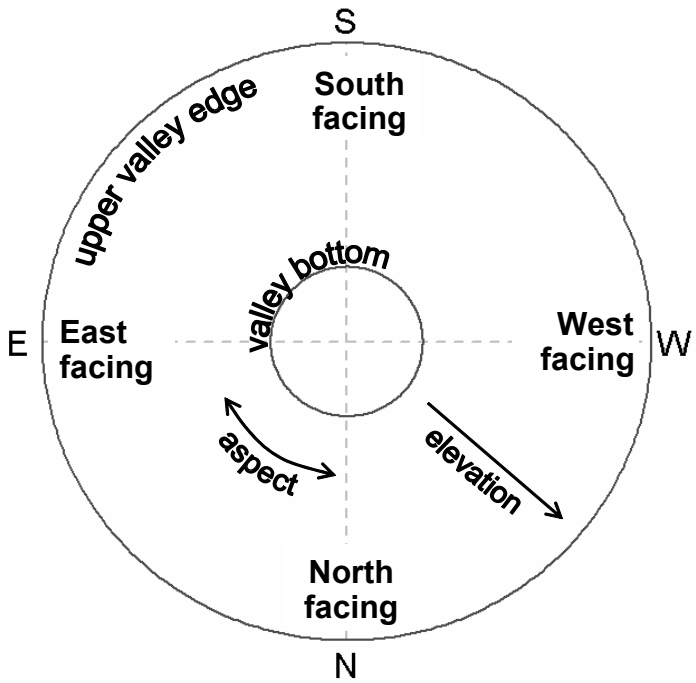




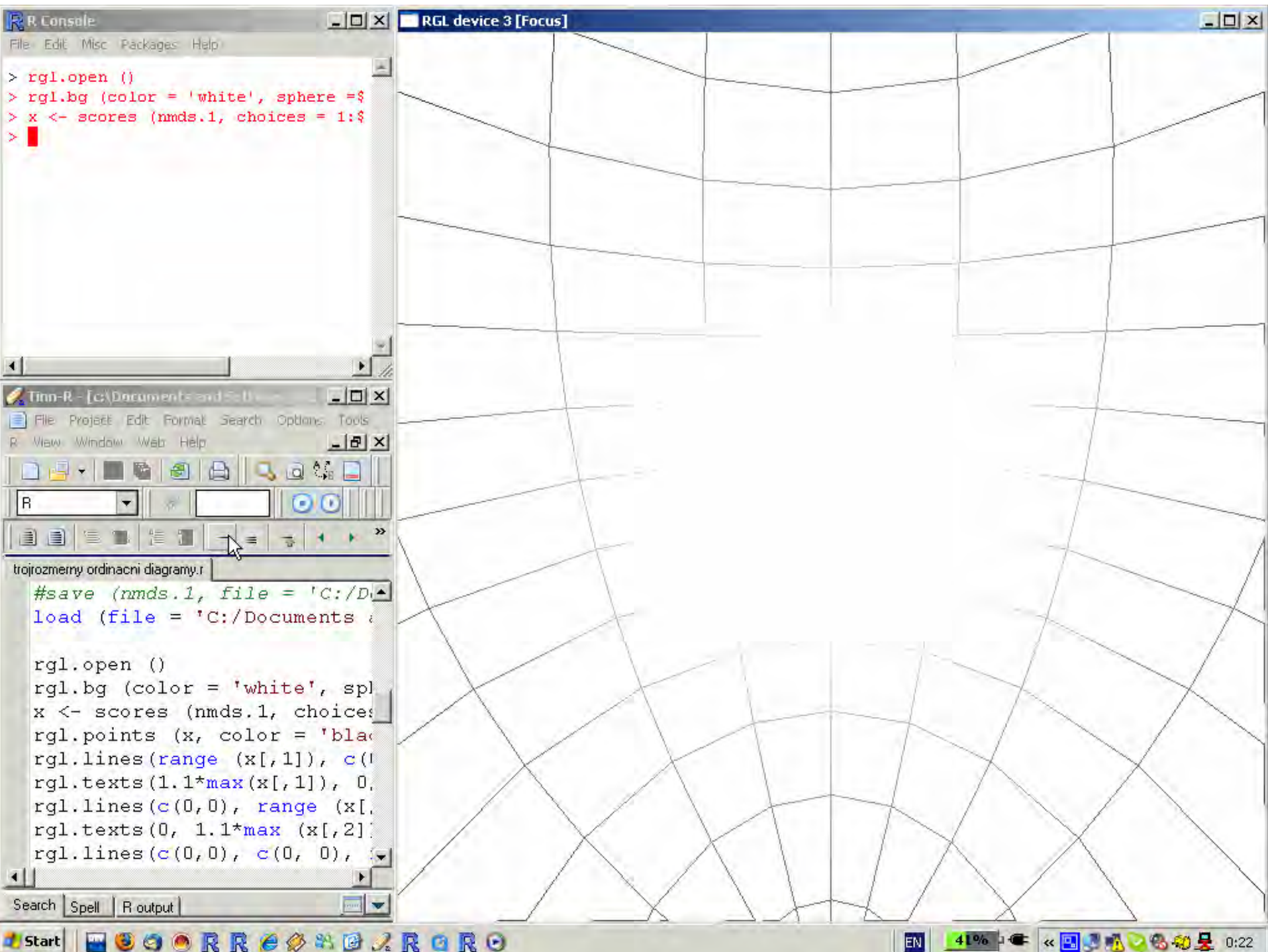
Zelený D. & Chytrý M. (2007): Environmental control of vegetation pattern in deep river valleys of the Bohemian Massif. *Preslia*, 79: 205-222.

Donut plot – pattern of alpha and beta diversity in deep river valleys





Zelený & Chytrý (2007): Pattern of α- and β-diversity of vegetation in deep river valleys of Bohemia Massif (European Vegetation Survey, 16th Workshop, Rome (Italy), 22.-26.03.2007, poster presentation)


R Console
File Edit Misc Packages Help
> rgl.open ()
> rgl.bg (color = 'white', sphere =$
> x <- scores (nmds.1, choices = 1:$
>
Tinn-R - [c:\Documents and Se
File Project Edit Format Search Options Tools
R View Window Web Help
R
trojrozmerny ordinacni diagramy.r
#save (nmds.1, file = 'C:/D
load (file = 'C:/Documents
rgl.open ()
rgl.bg (color = 'white', spl
x <- scores (nmds.1, choices
rgl.points (x, color = 'bla
rgl.lines(range (x[,1]), c(
rgl.texts(1.1*max(x[,1]), 0,
rgl.lines(c(0,0), range (x[,
rgl.texts(0, 1.1*max (x[,2]
rgl.lines(c(0,0), c(0, 0),
Search Spell R output
RGL device 3 [Focus]
Start
EN
41%
0:22

So, why to use R?

- Freedom – you are not bounded by some software engineer imagination about how the software should look like, you just simply create everything by your self!
- Openness – it's an open system, and it's growing every day. Do you need some new analytical tool and suitable software? Try to search in R libraries – this function is most probably already available there, or soon will be...
- Did you publish some new analytical method in your paper? Attach R script as appendix and make it available for others!
- It's free – and it's fun! (sometimes...)

This is the end... Thank you for your attention!