

R project: powerful tool for vegetation data analysis and visualization

David Zelený

Masaryk University Brno, Czech Republic



Overview of my talk



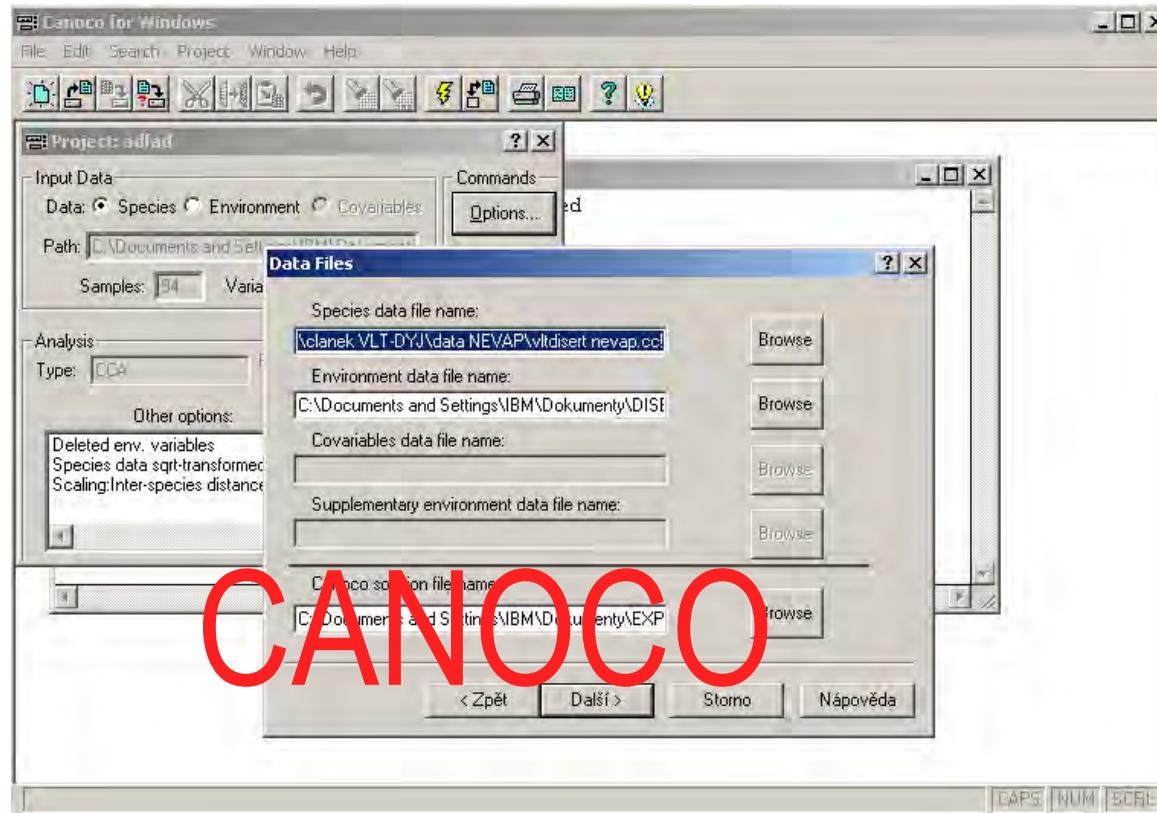
- Short introduction: what is R and what is R not?
- What says Web of Science about R?
- What R offers to vegetation ecologists?
- Examples of using R

What is R project?

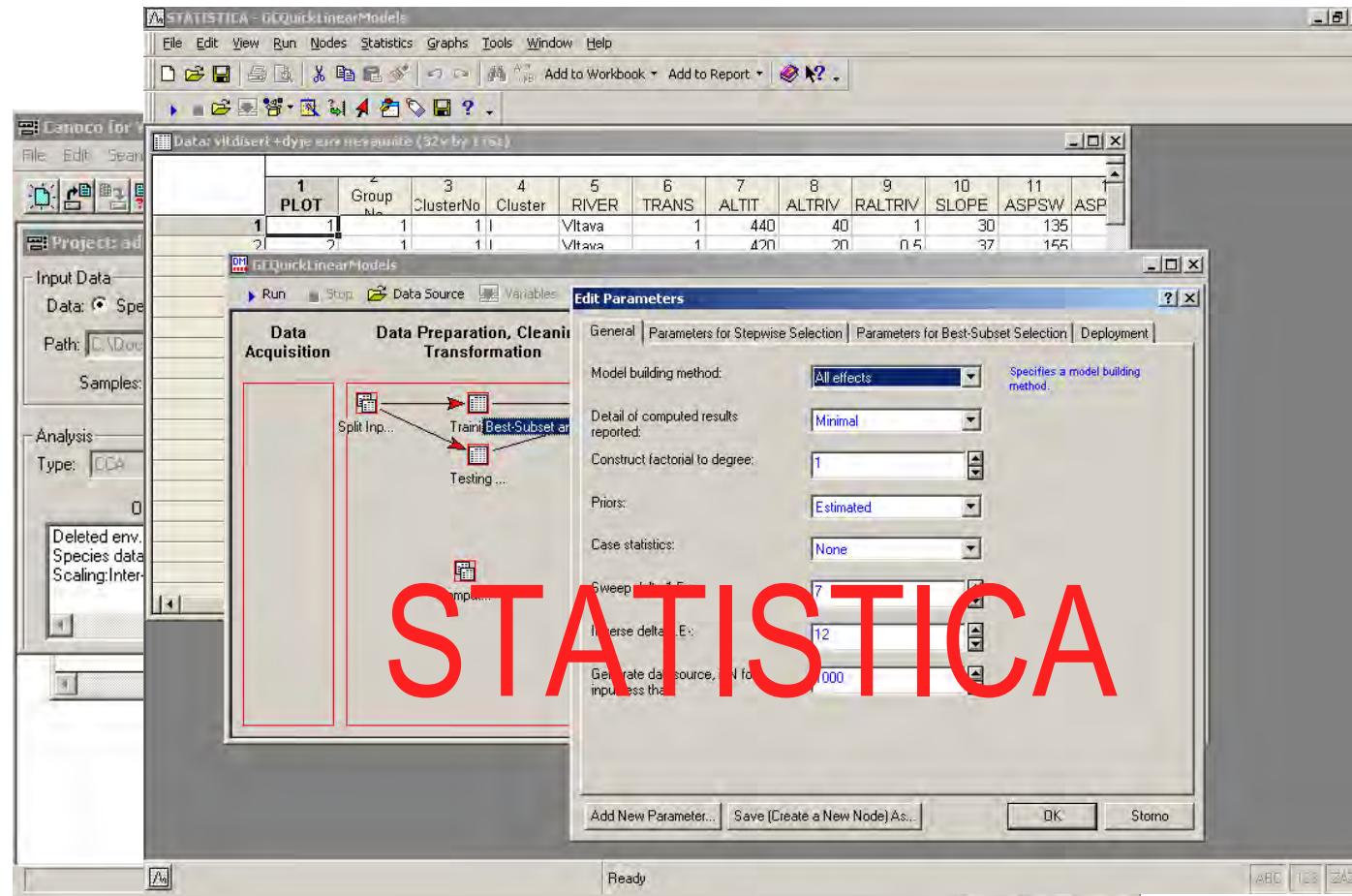


- Programming language based on S syntax
- Free software package, frequently updated
- Open community – everybody is welcome to contribute changes, improvement or new functions to this software
- People, who like to meet and discuss (yearly conferences useR! held in different parts of the world)

What R is not?



What R is not?



What R is not?



TURBOVEG

The screenshot shows the TurboVeg software interface. The main window displays a database table with columns: Relevé number, Cover abundance scale, Country code, Date (year/month/day), Syntaxon code, Relevé area (m²), Altitude (m), and Remarks. A red box highlights the 'Data Acquisition' section on the left. The central panel shows a 'Species list' with a search bar for '490 - Alopecurus pratensis'. To the right, there's a 'Selected species' list and a detailed table of species names, layers, and cover percentages. The bottom status bar shows 'Press F1 for Help', 'Add New Parameter...', 'Save (Create a New Node) As...', 'OK', 'Cancel', 'Remarks', 'Header data', '10/351', '18/07/2007', 'Europe', 'Europe', 'ReadWrite', 'Form edit', 'Species data', 'Rec: 1/32', 'Num', 'Caps', 'Ins', '24/10/2007', and 'Ready'.

What R is not?



The screenshot displays two overlapping software windows. The background window is JUICE (version 6.0), a program for cluster analysis. It shows a 'Cluster Analysis - PC-ORD 4.1' dialog box with various parameters set for a 'white' relevé. The main JUICE interface lists 'Relevés 690' and 'Species 626'. A large orange watermark 'JUICE' is overlaid across the center of the screen. In the foreground, another window titled 'STATISTICA - GROUP' is visible, showing a 'Data: vtdisert +dyt' table with numerous entries. A red rectangular box highlights the first few rows of this table.

JUICE - (c) JUICE and settings\um\diskoviny\april\juice\juice.exe - 2007-10-24 21:45:26

File Edit Species Relevé Table Head Sizing Separators Cryptic table Indicator Values Analyze Table Simulation Help

Relevé white Species red Separator hierarchy 6 Total time: 21 days 14 h 45 min 26 sec

Running number:

Relevés 690 Species 626

Cluster Analysis - PC-ORD 4.1

Cluster Analysis Setup

The program will start PC-ORD installed in your computer and will classify selected relevés automatically. Please, fill path of PC-ORD in the JUICE form Options.

Relevés Used in Analysis - white

Data Transformation

$b = (Xij)^p$ $p = 0.5$ $p = 0.0$ Presence/Absence Data

$b = \log(Xij + 1)$ $p = 0.4$ Square Root Transformation

Pseudospecies cut levels $p = 1.0$ No Transformation

Floating cut level by Species data value 0.525

Species data value

Clusters

No. of clusters: 690 Minimum = 2 Maximum = 600

Distance Measure

Sorenson (Bray-Curtis)

Rel. Sorenson (Rel. Manh.)

Jaccard

Euclidean (Pythagorean)

Group Average (UPGMA)

Correlation

Ward's Method

Flexible Beta -0.25

McQuitty's Method

Group Linkage Method

Nearest Neighbor

Farthest Neighbor

Median

Average Linkage

Centroid

Ward's Method

Flexible Beta -0.25

McQuitty's Method

Path of the output file: C:\Program Files\JUICE 6.0\pcord.csv Change path

Older version Cancel Continue >>

133. *Sinopanax formosana* [1] Frequency: 48 Relevé No.: Row: 16

Header data

Press F1 for Help

Add New Parameter... Save (Create a New Node) As... OK Storno

Ready

111111122121212
899898005190
868937280227

1. 華參
1. 山香圓
1. 尖葉楓
1. 小花屋樹
1. 厚皮東木董子
1. 檫木
1. 小葉樹杞
1. 杜虹花
1. 紅毛柃木
1. 台灣楠果
1. 台灣石楠
1. 山蒼
1. 水麻
1. 毛果柃木
1. 中原氏屋李
1. 牛樟
1. 石朴
1. 檉樹
1. 香葉樹
1. 山黃皮
1. 山茱萸
1. 瑞珊瑚
1. 俄氏柿
1. 野核桃
1. 小葉蘭
1. 食茱萸
1. 狗仔花
1. 片枝蘭
1. 台灣蘭
1. 阿里山
1. 金毛杜
1. 金毛杜
1. 金毛杜
1. 台灣杉
1. 阿里山十大功勞
1. 白雞油
1. 山芙蓉
1. 佛手
1. 錦大紫珠

444453533
811609099
346445524

555401 02
555402 02
555403 02
555404 02
555405 02
555406 02
555407 02
555408 02
555409 02
555410 02
555411 02
555412 02
555413 02
555414 02
555415 02
555416 02
555417 02
555418 02
555419 02
555420 02
555421 02
555422 02
555423 02
555424 02
555425 02
555426 02
555427 02
555428 02
555429 02

R has very simple interface...



The screenshot shows the R Console window with a blue header bar containing the title 'R Console' and menu options: File, Edit, Misc, Packages, Help. Below the header is a message box displaying the R startup sequence. At the bottom of the window, there is a text input field and a command prompt. A blue horizontal bar runs across the top of the slide, ending in a circular arrow icon on the right side.

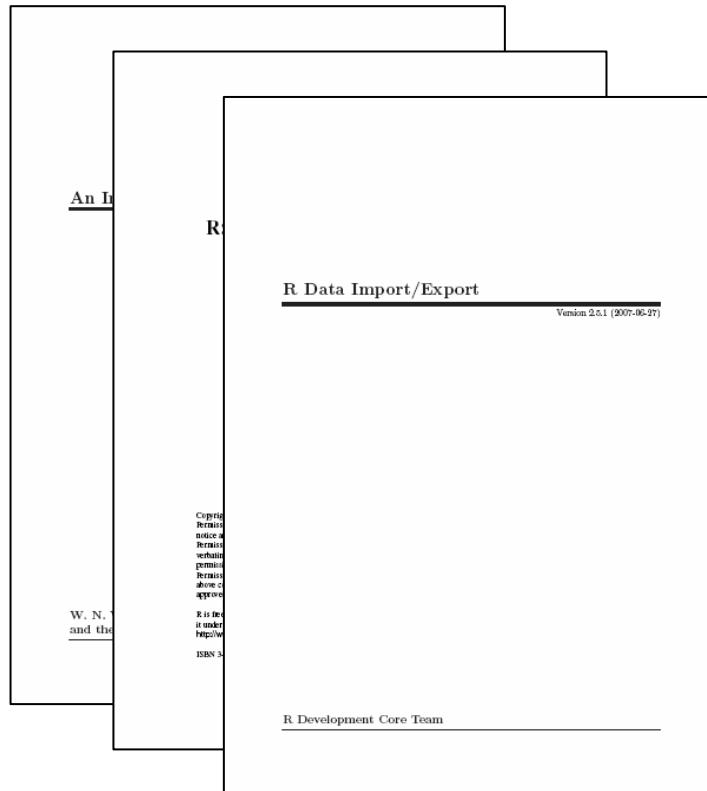
```
R version 2.5.1 (2007-06-27)
Copyright (C) 2007 The R Foundation for Statistical
Computing

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.

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.

> Hi, how are you?
Not bad, and you?
> Well, today is a bit busy, but otherwise fine!
So?
> So, may I ask you to calculate something for me?
Forget it! ... By the way, how is your mother?
```

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 Analysis



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

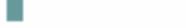
Field: Subject Category	Record Count	% of 2893	Bar Chart
ECOLOGY	549	18.9768 %	<input type="button" value="—"/>
STATISTICS & PROBABILITY	306	10.5773 %	<input type="button" value="—"/>
GENETICS & HEREDITY	216	7.4663 %	<input type="button" value="—"/>
ENVIRONMENTAL SCIENCES	178	6.1528 %	<input type="button"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



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 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 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 %	



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

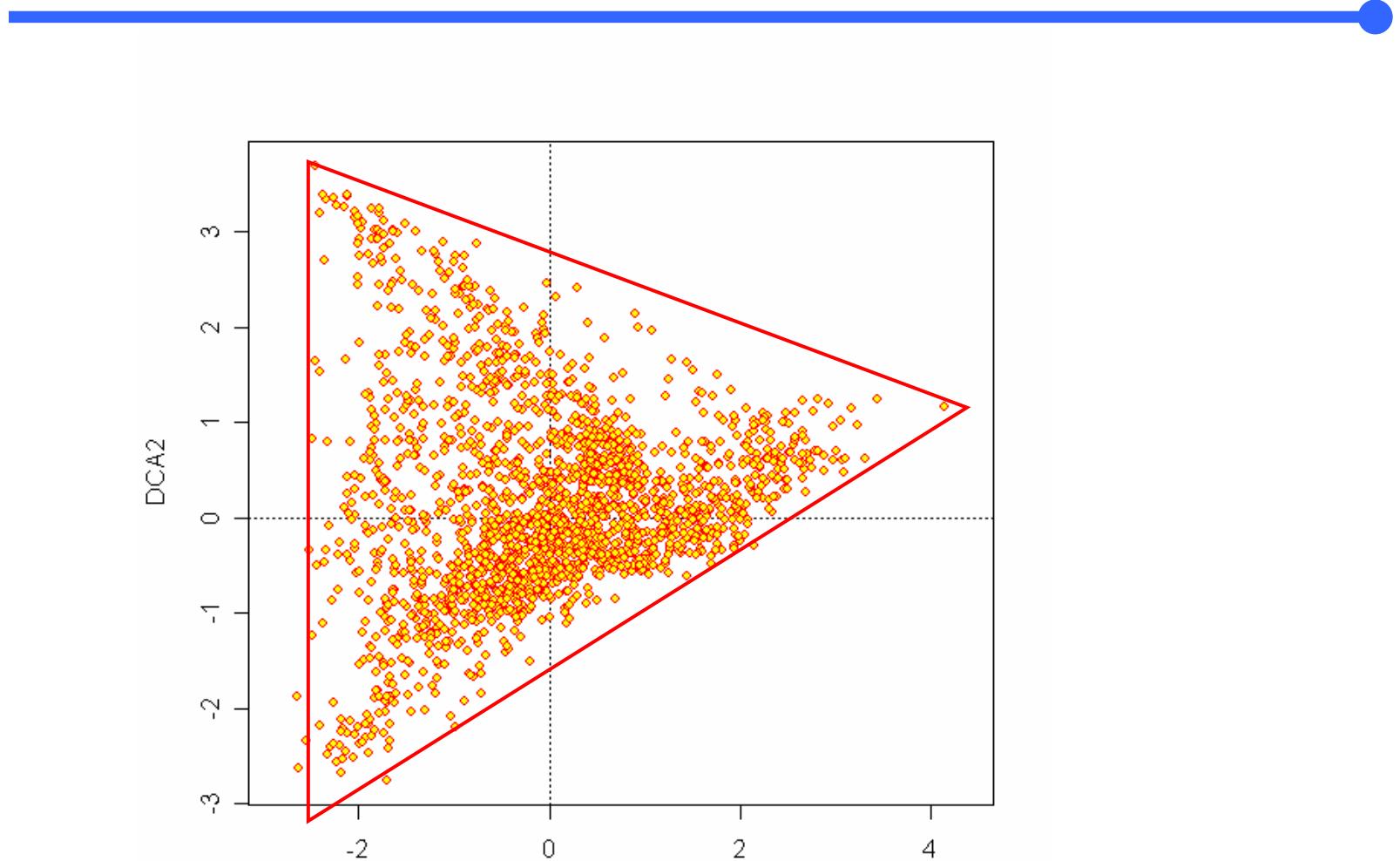
- non-metric alternative of unconstrained ordinations (PCA, DCA)
- the method tries to project samples into 2D figure, so as distances between these samples maximally correspond to the sample dissimilarities (measured by selected dissimilarity measure – Bray-Curtis etc.)

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?



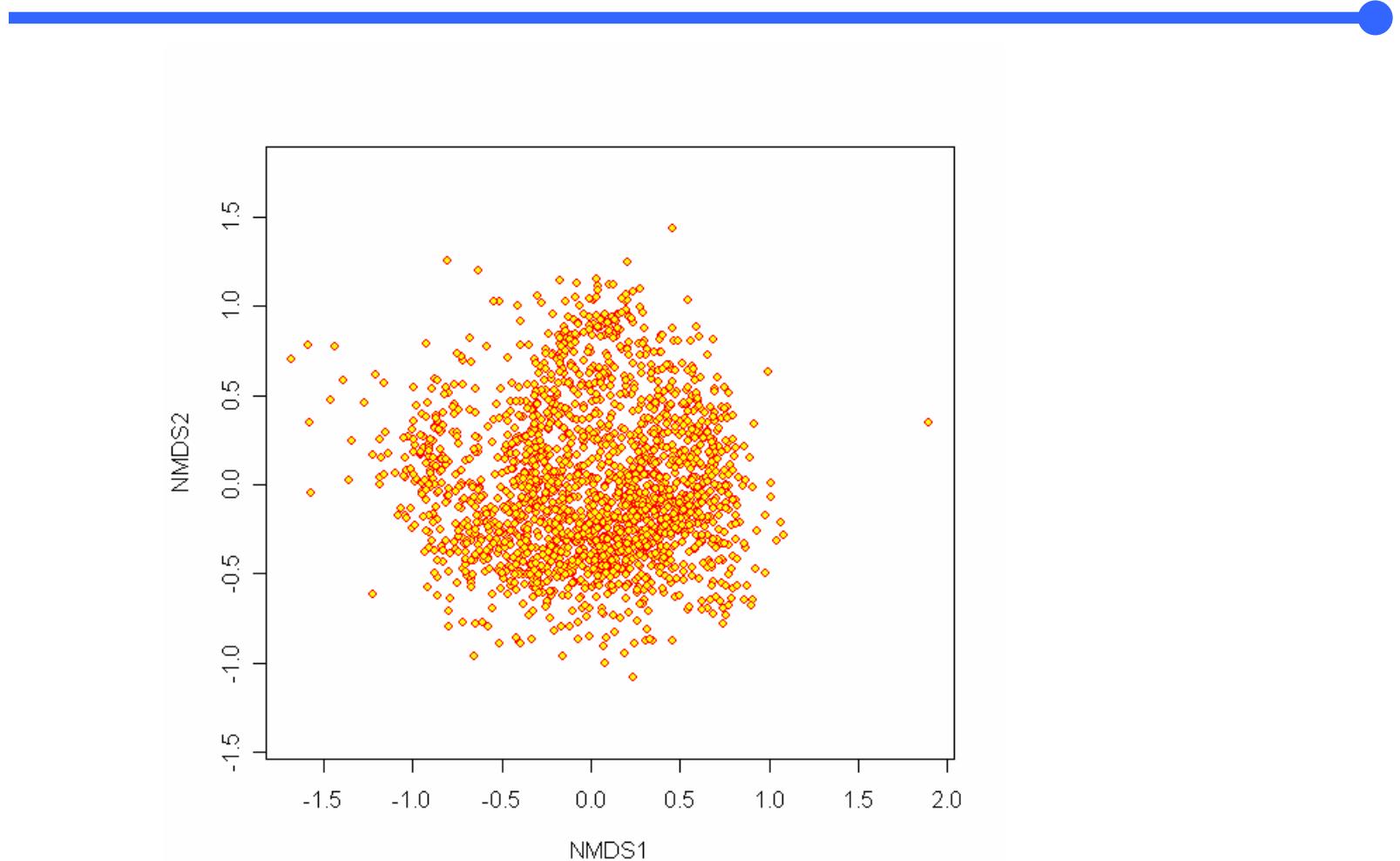
DCA analysis of 2000 forest relevés randomly selected from Czech National Vegetation Database.

Why yes NMDS?



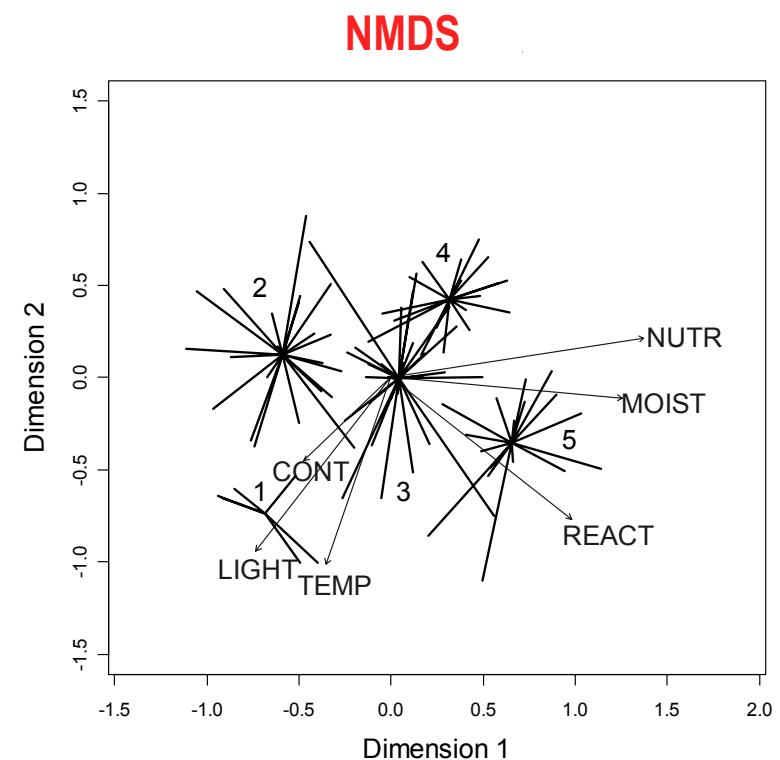
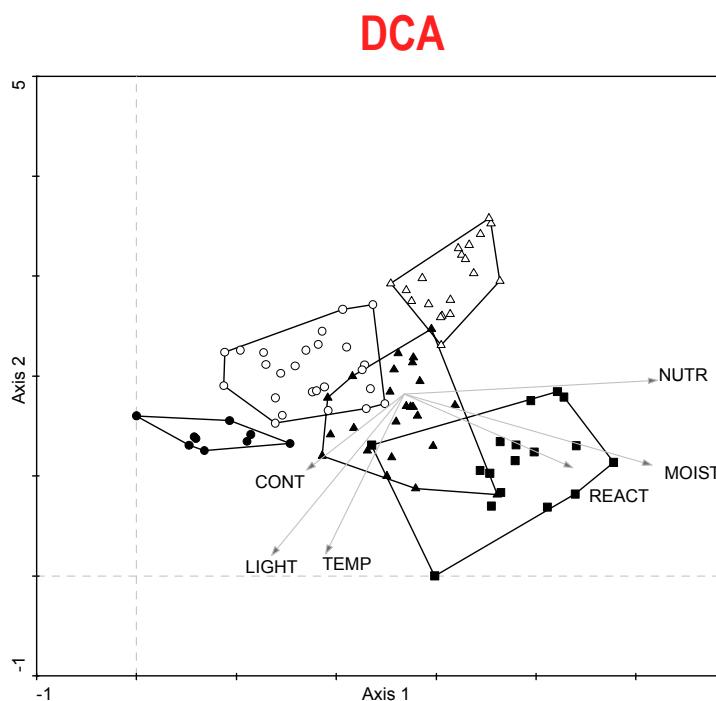
-
- It doesn't suffer from the artifacts of DCA.
 - According to Minchin (1987) it's the most robust unconstraint ordination method in vegetation ecology.
 - It's non-metric method, and doesn't assume the unimodal shape of species response curves.

Why yes NMDS?



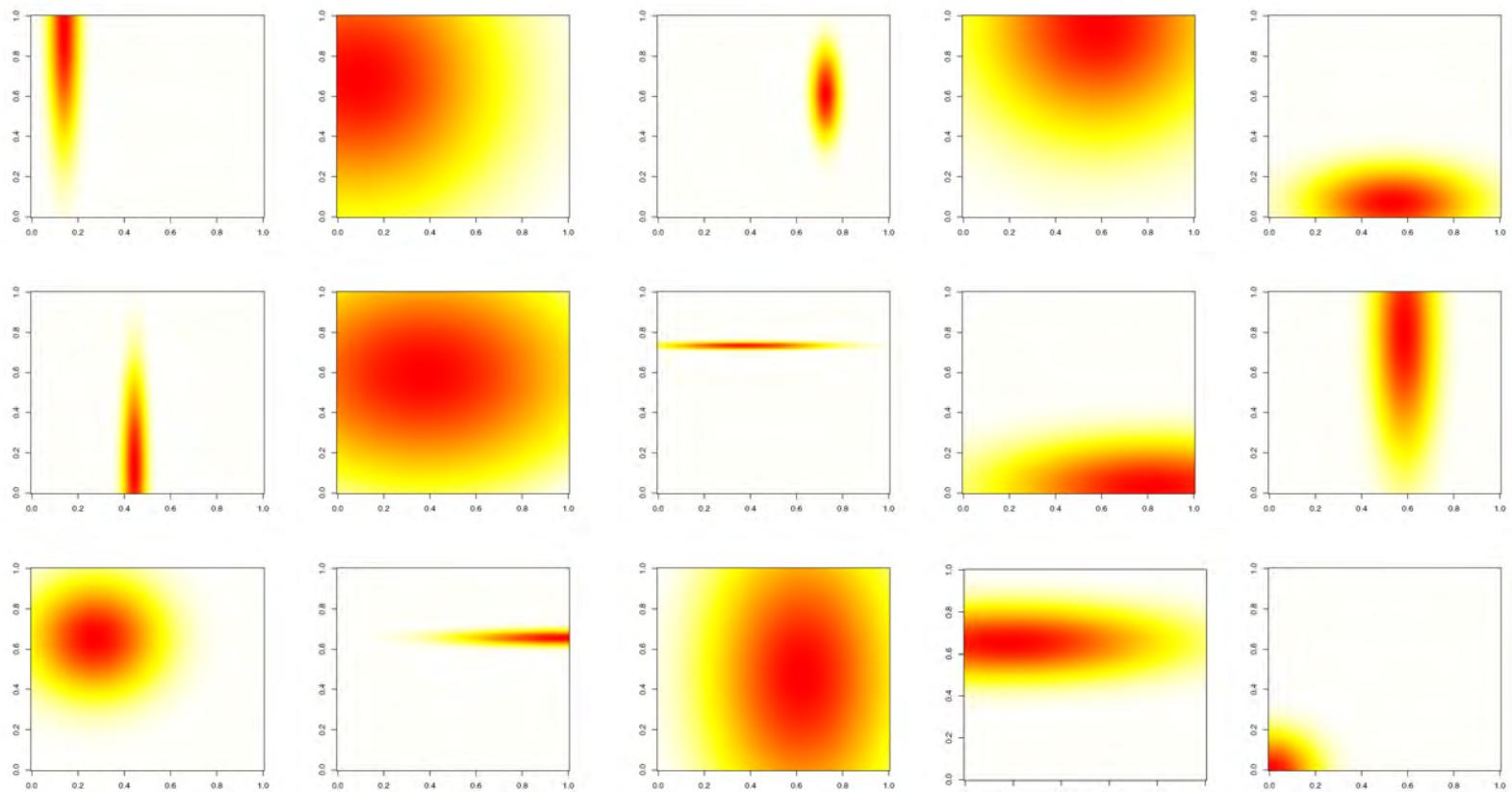
NMDS analysis of 2000 forest relevés randomly selected from Czech National Vegetation Database.

Comparison of DCA and 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

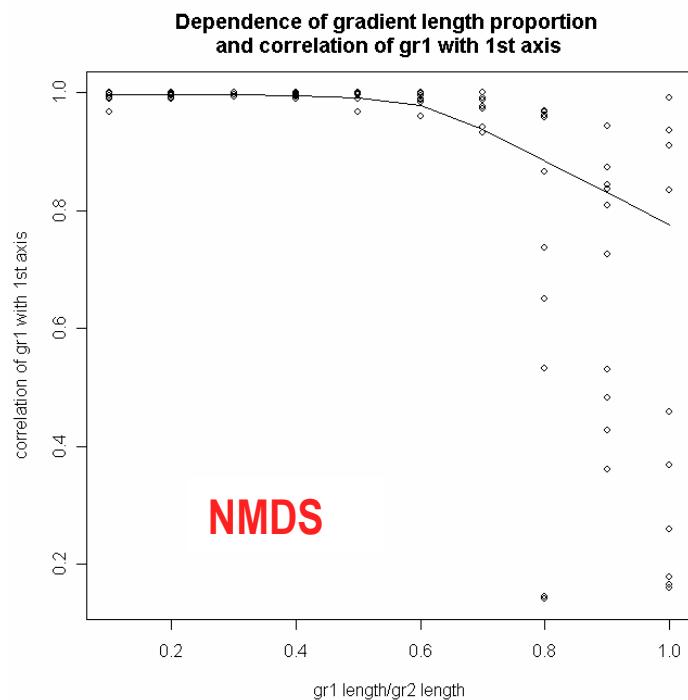
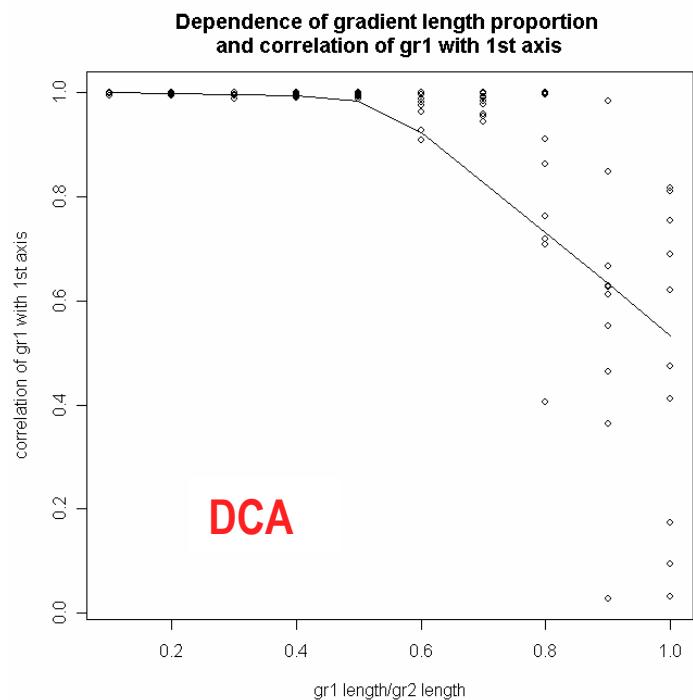


Which method can better recover simulated gradients in data?



- 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



File Edit Species Relieves Table Head Setting SynopticTable IndicatorValues Analysis Table Simulation Help

Statistics: Rho coeff. C

Relevé white Species red

TWINSPAN category:

Releves 516
Species 582

Data Transformation:

- $b = (X_{ij})^p$, $p = 0.5$
- $b = \log(X_{ij} + 1)$
- Pseudospecies cut levels 0.525
- Floating cut levels by Species data value (0, 1, 2)

Analysed Data Information:

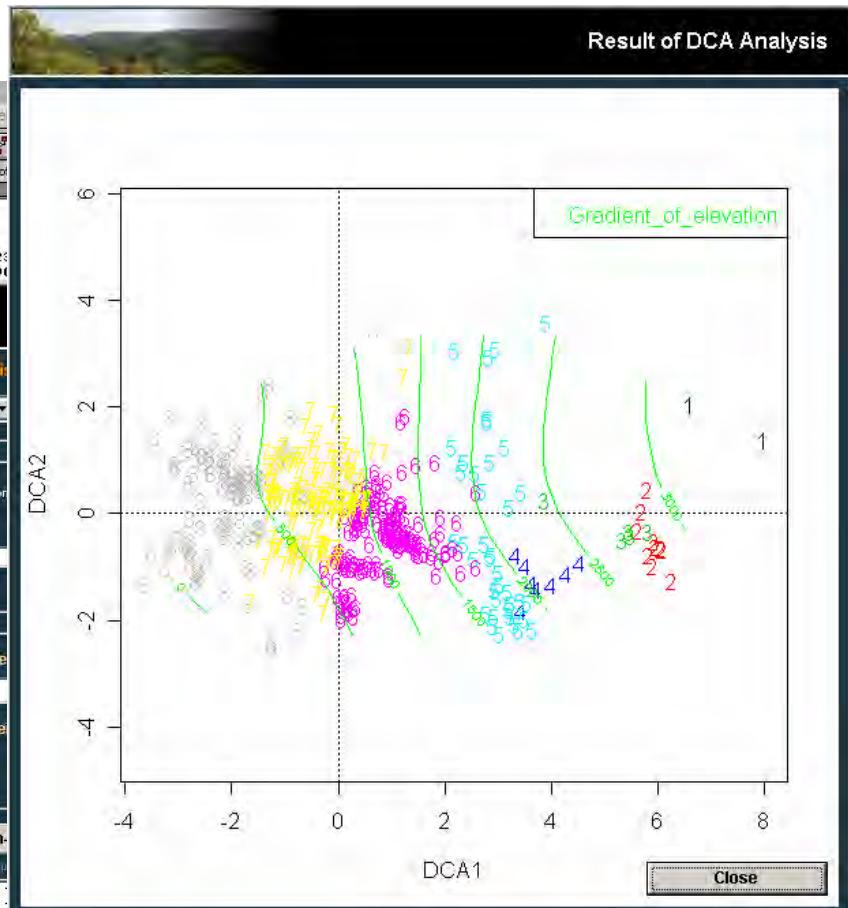
- Use data from Short Headers as Vectors Surface
- Factor name: Gradient of elevation
- Use Ellenberg Indicator Values as vectors (it expects the previous initiation)
- Downweight rare species

First Run and Update script:

The routine is supported by R-project with special libraries written by J. Oksanen et al.

Vaccinium dumalianum v. caudatifolium Frequency: 9 Relative No.: Relevé No.: Row: Column: 200

台灣樹參



JUICE & Species Response Curves



JUICE - (ct\documents and settings\um\dokumenty\email\juice\juice_main)

File Edit Species References Table Head Smiling Separators Synoptic Table Indication Values

Statistics Relievs white Species red

TWINSPAN category:

Releves 516 Species 582

Adinandra formosana

Dendropanax dentiger

Cyathia metteniana

Pourthiae lucida

Cinnamomum insularimontanum

Picea morrisonicola

Koelreuteria henryi

Fagus hayatae

Marsdenia formosana

Daphne arisanensis

Plagiogyria formosana

Pasania formosana

Amentotaxus formosana

Cephalotaxus wilsoniana

Ilex ficoidea

Acer morrisonense

Ormosia formosana

Actinidia chinensis v. set

Elaeagnus formosana

Musa basjoo v. formosana

Castanopsis formosana

Boehmeria formosana

Ribes formosanum

Pinus armandii v. masteiria

Malus doumeri

Alnus formosana

Syzygium formosanum

Medinilla formosana

Styrax matsumuraei

Uncaria hirsuta

Phoebe formosana

Citrus depress

Strobilanthes formosanus

Pieris taiwanensis

Schefflera taiwaniana

Berchemia formosana

Tetradium ruticarpum

Viburnum luzonicum

Bendrocnide meyeniana

Rhododendron simsei

Symplocos setchuensis

Vaccinium dunalianum v. caudatifolium

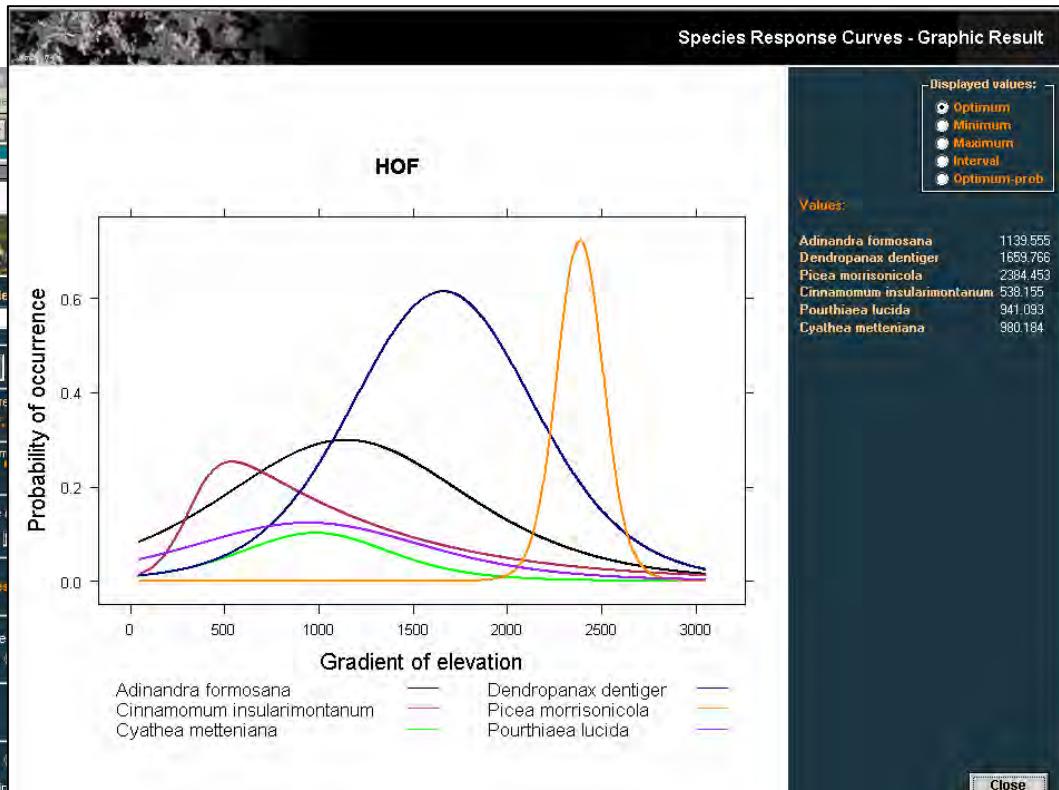
台湾樹參

Frequency: 9 Relative No.: 9 Relevé No.: Row: 125 Column:

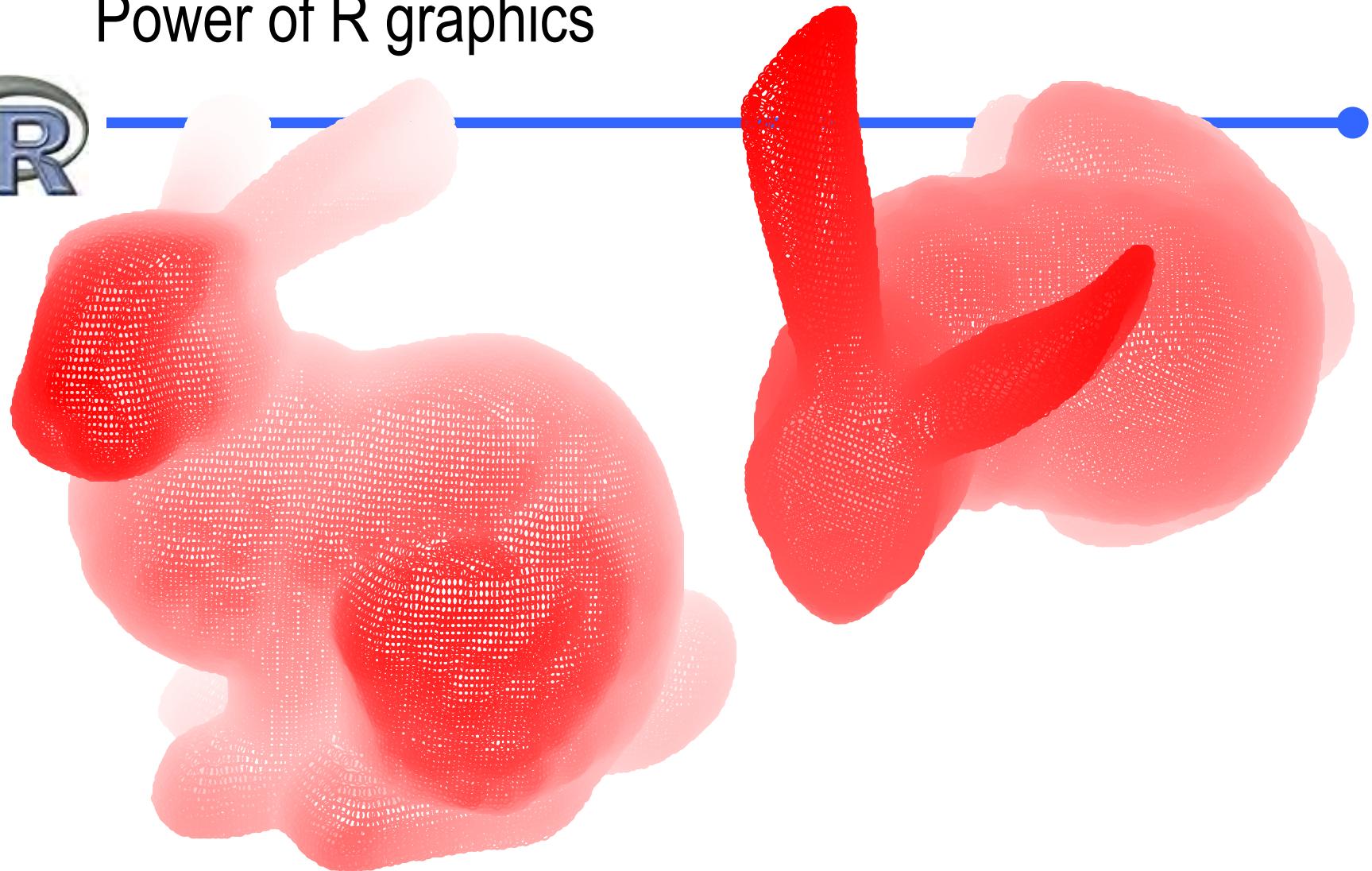
Project 2.4.1
and later is needed!

First Run and Update Script Help Cancel Calculate

The routine is supported by R-project with functions written by D. Zelený (package PROJUDICE) and J. Šíšma (package JUICE).

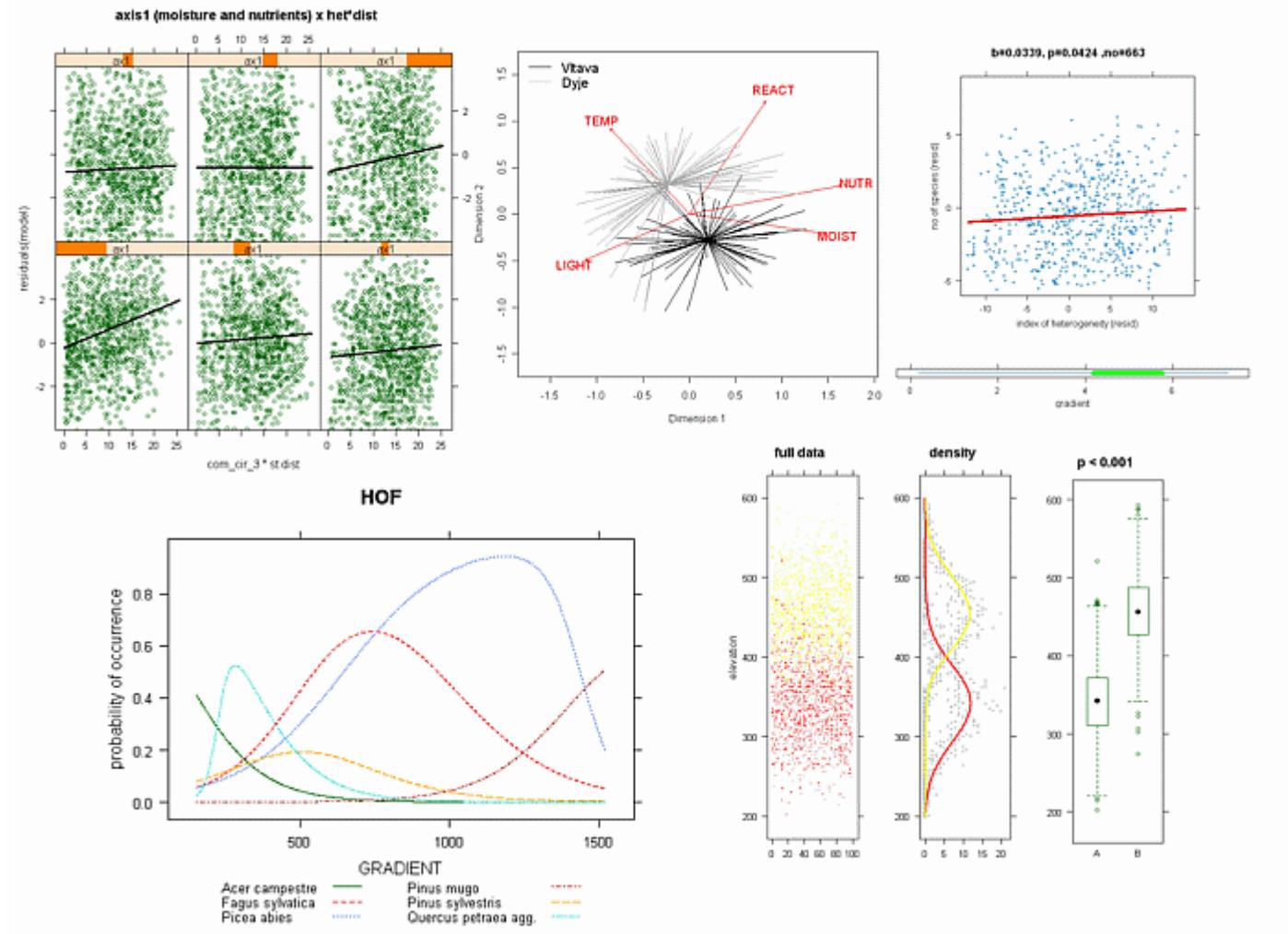


Power of R graphics

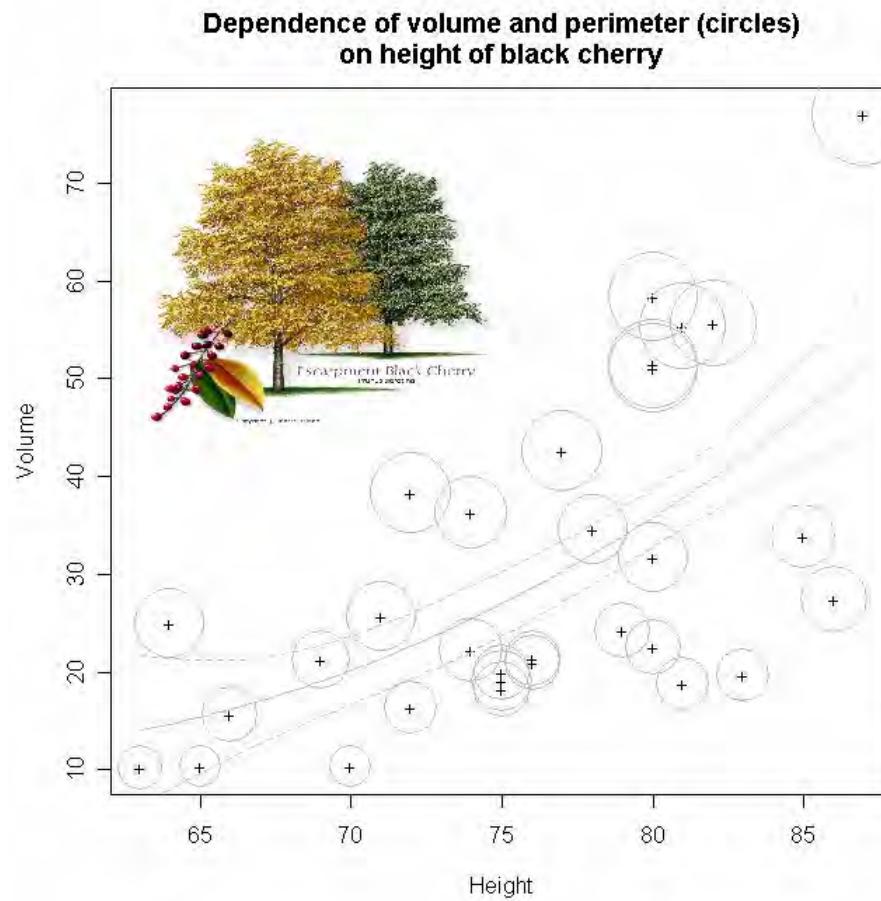


3D bunny, package *onion*, author Robin K. S. Hankin .

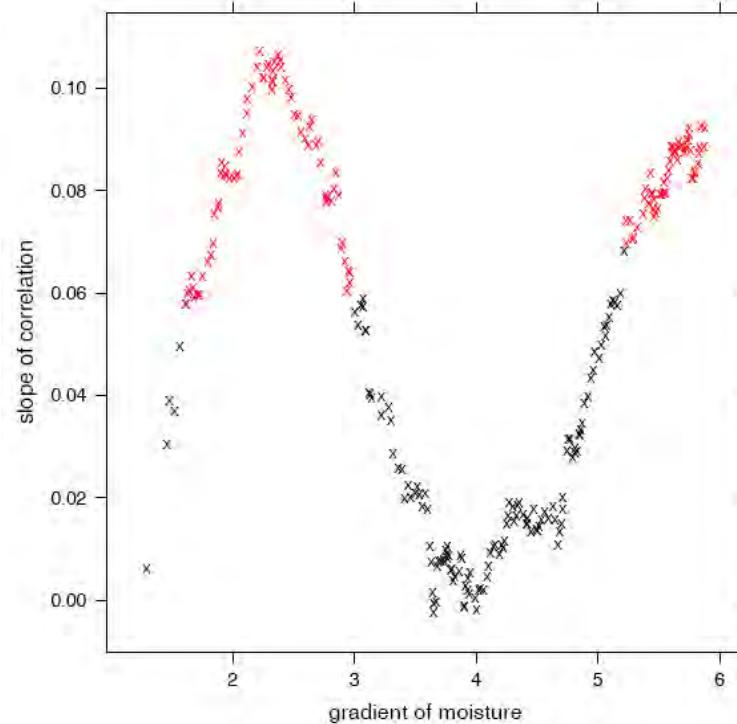
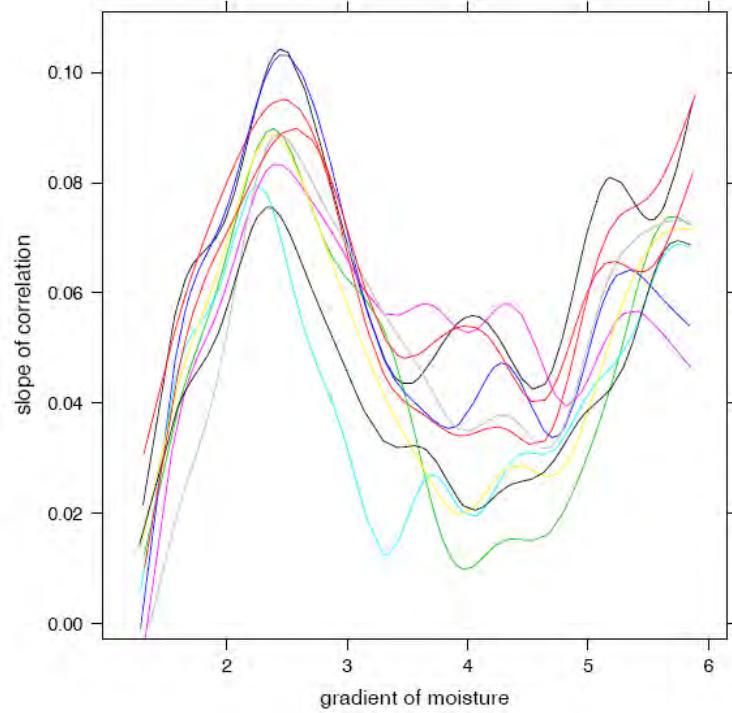
Some examples of graphical outputs from R



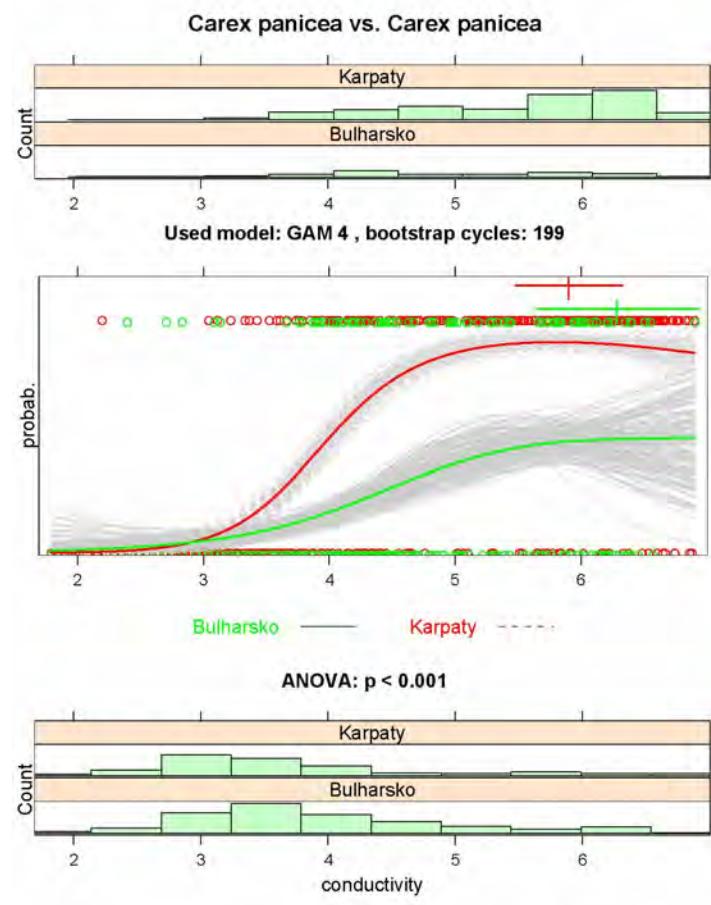
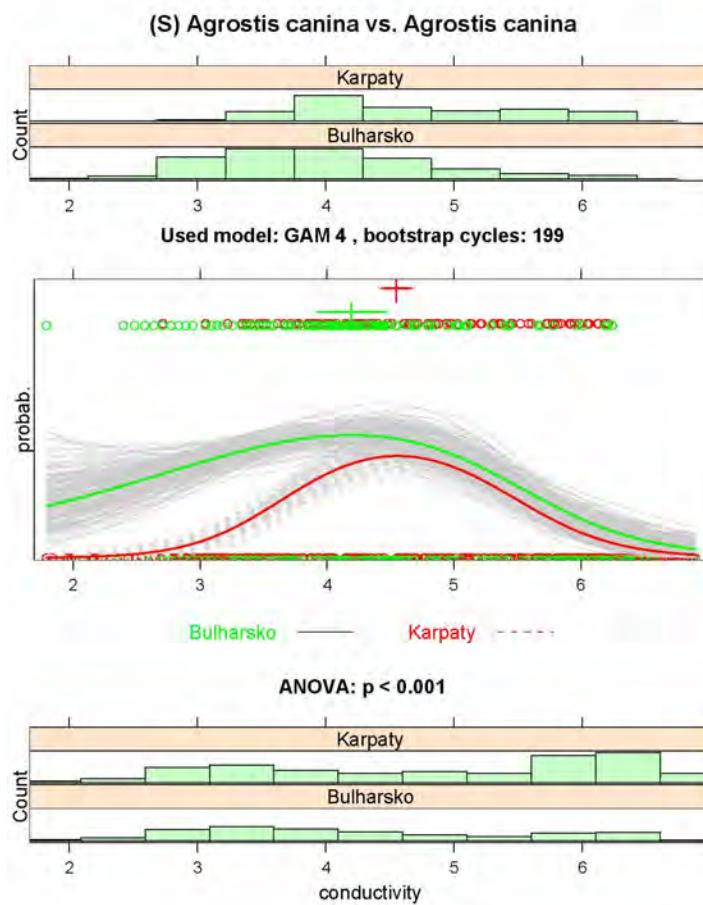
Cherry tree – relationship of volume, perimeter and tree height



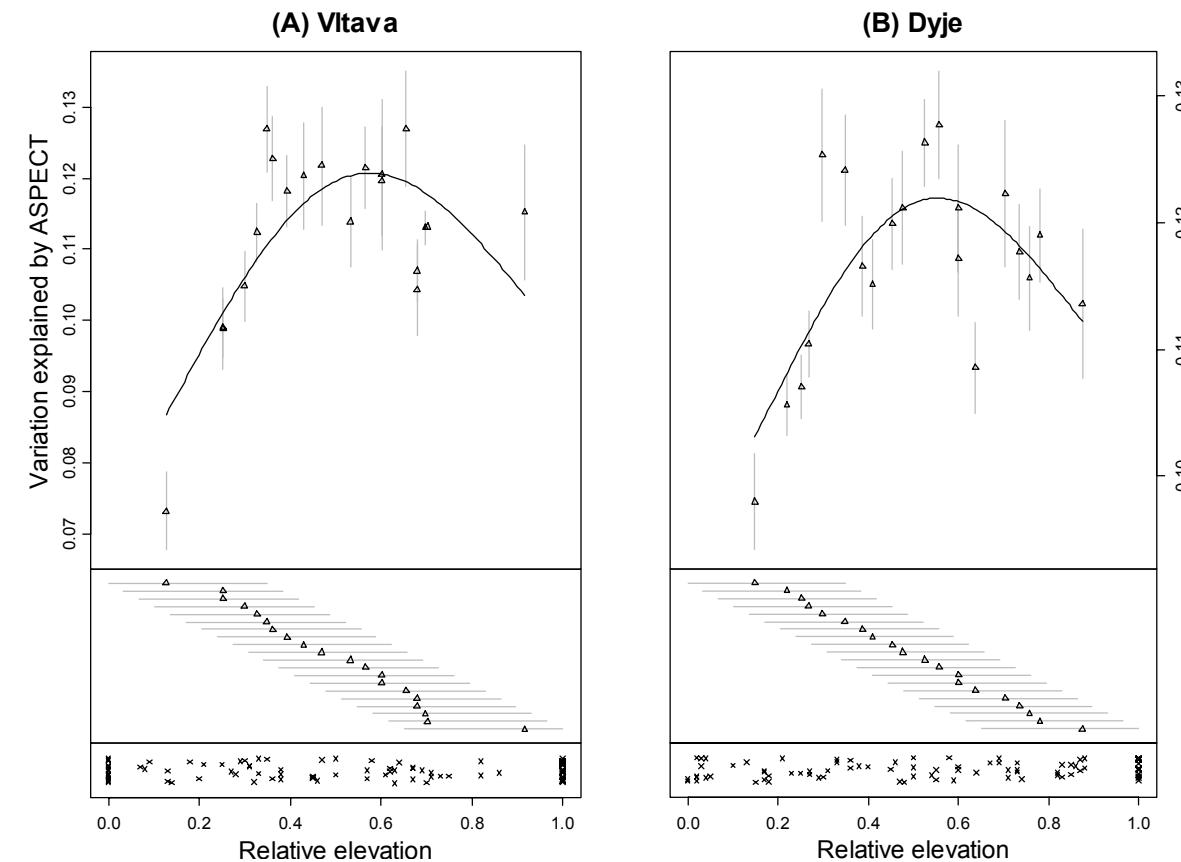
Randomized Results of Moving Window regression



Bootstrapped species response curves

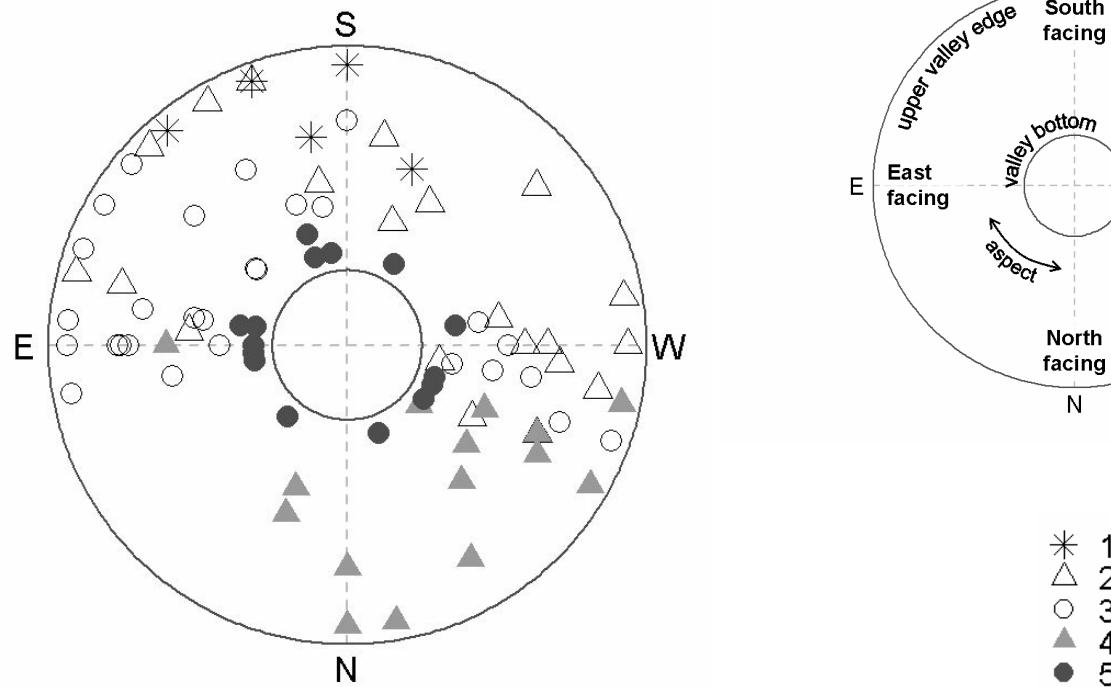


Moving frame CCA



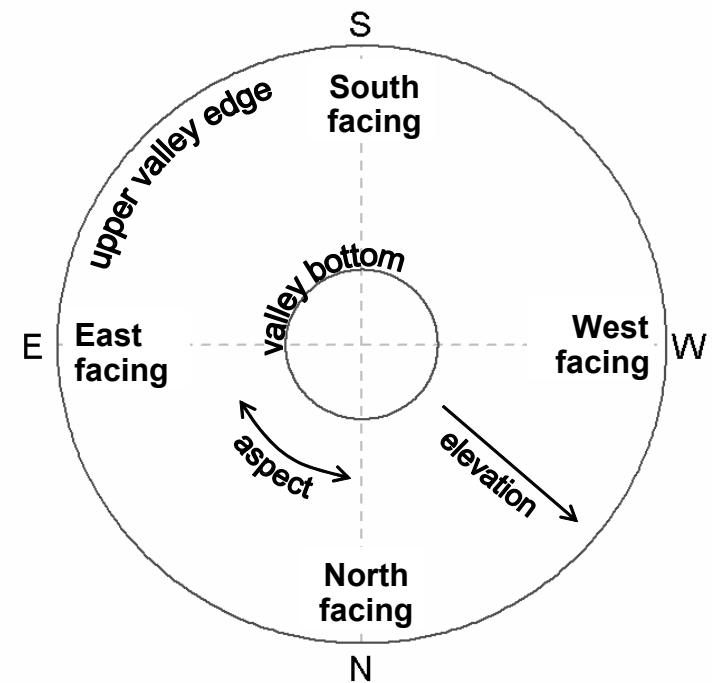
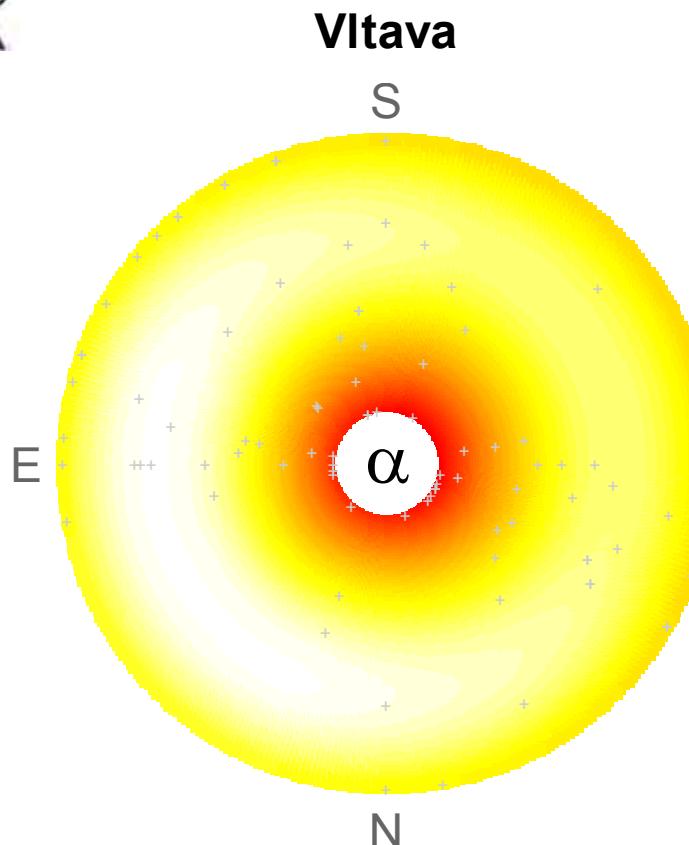
Zelený D. & Chytrý M. (2007): Environmental control of vegetation pattern in deep river valleys of the Bohemian Massif.
Preslia, 79: 205-222.

Iris diagram – distribution of vegetation types in river valley

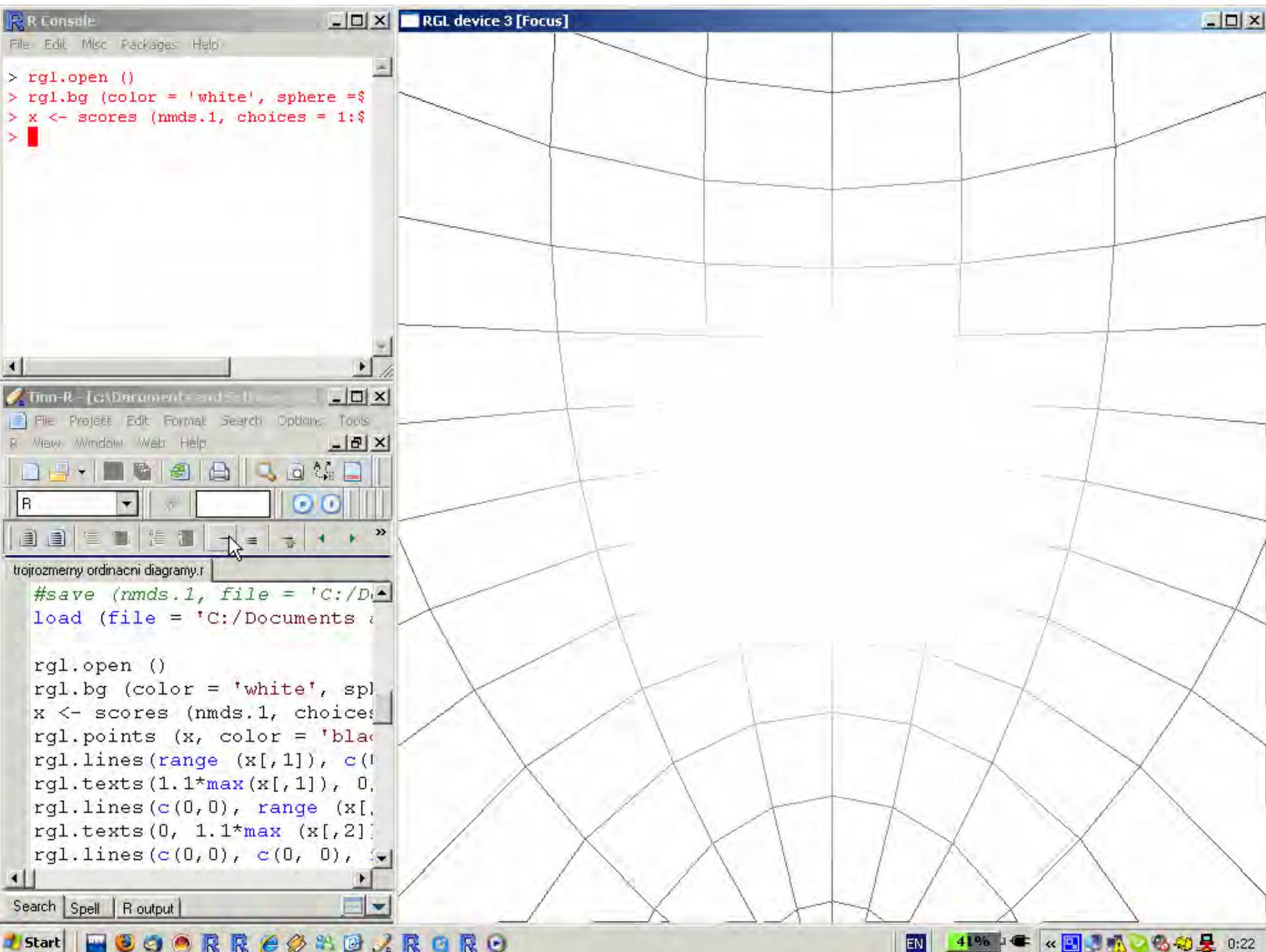


Zelený D. & Chytrý M. (2007): Environmental control of vegetation pattern in deep river valleys of the Bohemian Massif.
Preslia, 79: 205-222.

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)



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!

