LIFE WATCH GREECE

RvLAB

R virtual laboratory (Rvlab)

Anastasis Oulas

LIFE WATCH GREECE

FORTH-ICS

ESFRI

LifeWatch

NSRF 2007-2013 programme for development

IMBBC Institute of Marine Biology, Biotechnology and Aquaculture

hcmr ΕΛΚΕΘΕ

Rvlab scope and target audience
"R for LifeWatchers"

- Ecological/biological interest – users with little or no R programming or statisctical experience.
- Offer statistical and visualization tools for the LifeWatch Project, using R statistical language.

Main Objective of Rvlab:

- Support certain VEGAN package functions (Community Ecology Package), which include ordination methods, hypothesis testing and other *multivariate* diversity analysis functions for community and molecular ecologists.

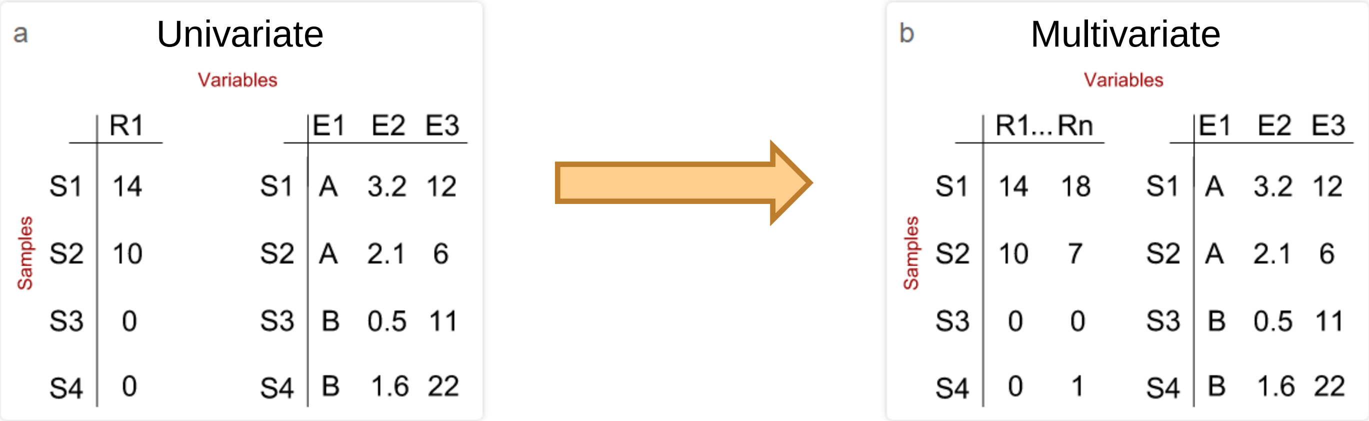
More specific issues addressed deal with *Optimization*:

1. Big data manipulation (overcome memory barriers)
2. Computational time speed-up (task segmentation multi-cores)

- Cluster computing environment at HCMR – recently upgraded by LlifeWatch)
- Develop an efficient and friendly user interface for analysis of ecological community data.

Multivariate analysis?

- Ecological phenomena are inherently complex.
- **multiple response variables** - such as the abundances of multiple species - are often required/measured to gain ecological insight.
- **multiple explanatory variables** - such as environmental parameters - are added to an analysis to explain the variation in the response data.
- Multivariate analyses address the complexity of ***simultaneously analysing multiple response variables***.



	R1
S1	14
S2	10
S3	0
S4	0

	E1	E2	E3
S1	A	3.2	12
S2	A	2.1	6
S3	B	0.5	11
S4	B	1.6	22

	R1...Rn	
S1	14	18
S2	10	7
S3	0	0
S4	0	1

	E1	E2	E3
S1	A	3.2	12
S2	A	2.1	6
S3	B	0.5	11
S4	B	1.6	22

https://sites.google.com/site/mb3gustame/home/

Multivariate analysis properties methods in a "*nutshell*"

Data reduction or structural simplification.

- For example principal components analysis, allows the summary of multiple variables through a projection of smaller 'synthetic' variables generated by the analysis.
- Thus, high-dimensional patterns are presented in a lower-dimensional space, aiding interpretation.

Sorting and grouping.

- Using similarity or dissimilarity for collecting objects and assignment to groups.
- Methods, such as cluster analysis and non-metric dimensional scaling are examples of this, allowing for detection of potential groups in the data.

Investigation of the dependency among variables.

- Dependency amongst variables is of key interest. Methods that detect dependency, such as canonical correspondence analysis (CCA) , are valuable in detecting influence or covariation in your data.

Prediction. Once the dependence among variables has been detected, multivariate models can be constructed to allow prediction. i.e. Regression analysis

Hypothesis construction and testing. Methods, such as ANOVA or ANOSIM allow the testing of statistical hypotheses on multivariate data..

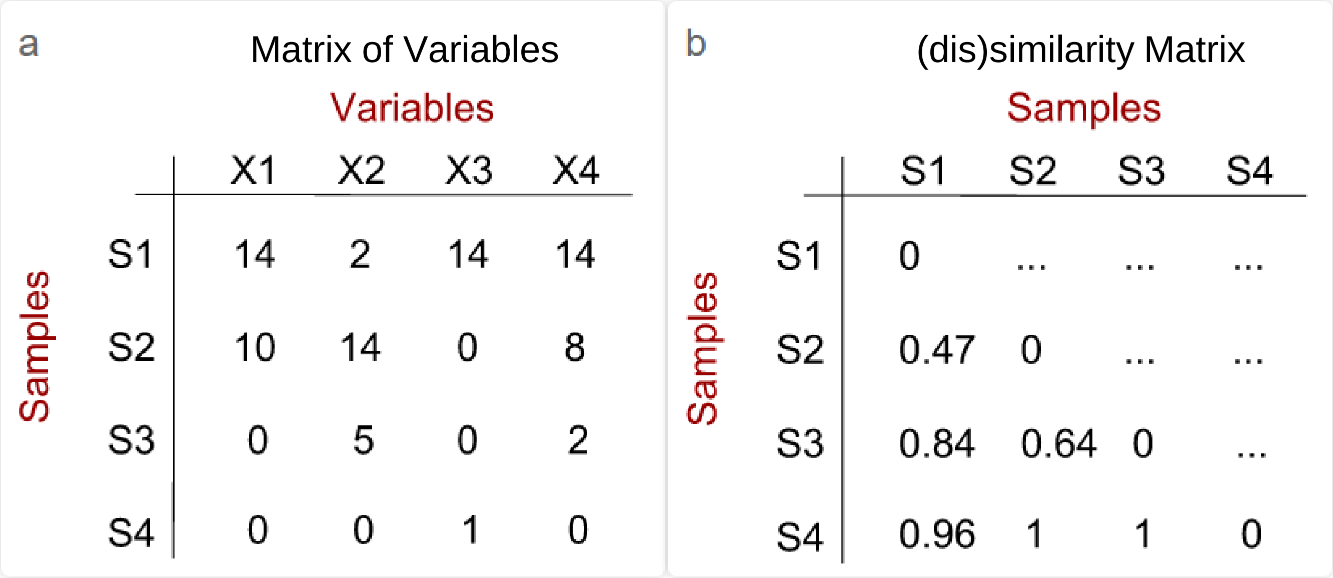
https://sites.google.com/site/mb3gustame/home/

(Dis)similarity-based methods – General Idea…

- Many ecological questions are concerned with finding how (dis)similar objects (i.e. station vs. species) are relative to one another.

- This (dis)similarity is established by comparing variable values through a (dis)similarity measure.

- The validity of (dis)similarity-based methods depends on the use of the correct (dis)similarity coefficient.



a

Matrix of Variables

Samples		Variables			
		X1	X2	X3	X4
	S1	14	2	14	14
	S2	10	14	0	8
	S3	0	5	0	2
	S4	0	0	1	0

b

(dis)similarity Matrix

Samples		Samples			
		S1	S2	S3	S4
	S1	0	...	...	...
	S2	0.47	0	...	...
	S3	0.84	0.64	0	...
	S4	0.96	1	1	0

https://sites.google.com/site/mb3gustame/home/

(Dis)similarity-based methods

Appropriate coefficients? Depends on the data…

Presence/absence and ordinal data

Jaccard coefficient

The Jaccard similarity coefficient assess the degree of overlap between two objects, ignoring double zeros (e.g. double absences). It is the quotient of the number of double presences ("1,1"s) and the sum of double presences and differences ("1,0"s and "0,1"s). When dealing with OTUs or species, its one complement may be used to assess turnover.

Abundance data – where high abundance and few zeros are treated equally to those with low abundance and many zeros

Bray-Curtis dissimilarity

This is an asymmetrical measure often used for raw count data. This is the one-complement of the Steinhaus similarity coefficient and a **popular measure of dissimilarity in ecology**

R: vegdist() in the vegan package

(Dis)similarity-based methods

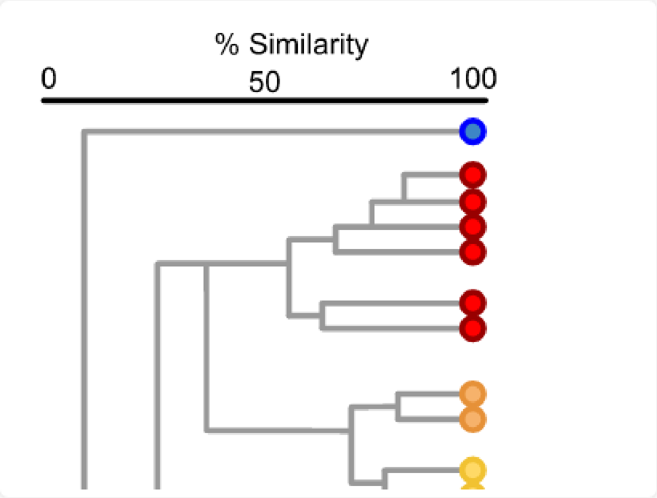
Cluster analysis - The main idea...

- Cluster analysis describes techniques suited for placing objects in groups, called clusters.
- The concept is that ***dissimilarities between objects within groups is smaller than those between groups***.
- The definition of a cluster depends on clustering algorithm (average linkage, complete linkage) and distance metric/coefficient.

Hierarchical cluster analysis

- First objects with the lowest dissimilarities are grouped together, before proceeding to group objects of increasing dissimilarity in a hierarchical manner

R – hclust()



https://sites.google.com/site/mb3gustame/home/

(Dis)similarity-based methods

The **BIOENV procedure (Clarke and Ainsworth, 1993)**

- A dissimilarity-based and exploratory method suited for identifying the subset of a set of explanatory variables (i.e Environmental data) that ***correlate maximally*** with response data (e.g. abundance data).

For Example:

Correlations:spearman
Dissimilarities: bray
Metric:Euclidean

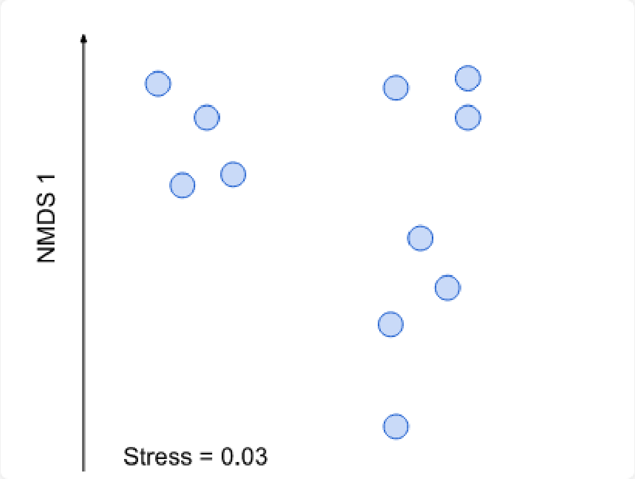
Best model has 1 parameters (max. 2 allowed):
maximumDepthInMeters
with correlation 0.9072356
R: bioenv() - vegan package

(Dis)similarity-based methods

Non-metric multidimensional scaling

- Non-metric multidimensional scaling (NMDS) is an indirect gradient analysis approach which produces an **ordination** based on a distance or dissimilarity matrix.
- Unlike methods which attempt to maximise the variance or correspondence between objects in an ordination (i.e. PCA), ***NMDS attempts to represent, as closely as possible, the pairwise dissimilarity between objects in a low-dimensional space***.
- Uses dissimilarity coefficient or distance measure to build the distance matrix to perform the analysis.

R: metaMDS() vegan package.



https://sites.google.com/site/mb3gustame/home/

(Dis)similarity-based methods

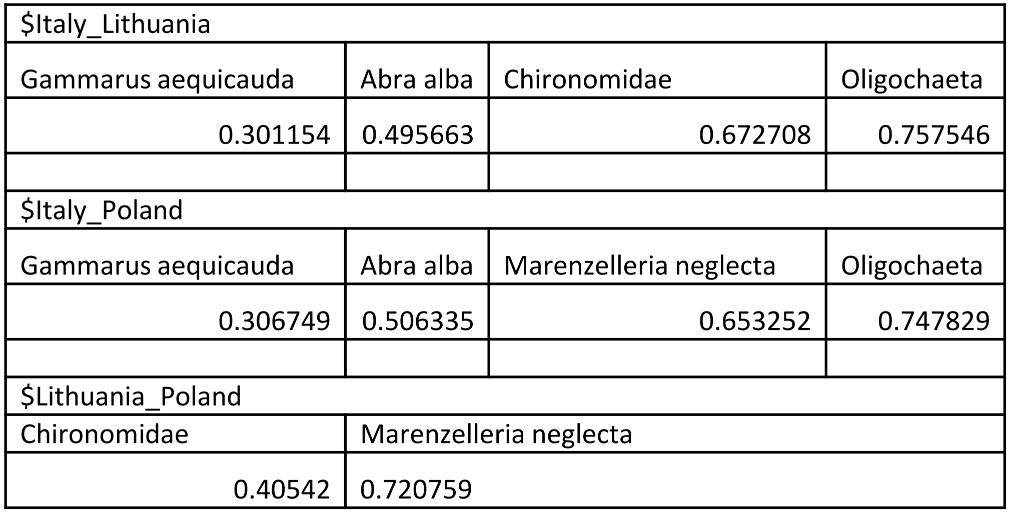
SIMPER (The similarity percentages breakdown) procedure (Clarke, 1993)

- Allows for assessment of the ***average percent contribution*** of individual variables to the dissimilarity between objects.

- For example: allows you to ***identify variables that are likely to be the major contributors (i.e. abundant species)*** to pairwise comparison of groups.

R: simper() – vegan package

cumulative contributions of most influential species:



$Italy_Lithuania			
Gammarus aequicauda	Abra alba	Chironomidae	Oligochaeta
0.301154	0.495663	0.672708	0.757546
$Italy_Poland			
Gammarus aequicauda	Abra alba	Marenzelleria neglecta	Oligochaeta
0.306749	0.506335	0.653252	0.747829
$Lithuania_Poland			
Chironomidae	Marenzelleria neglecta		
0.40542	0.720759		

Constrained analyses

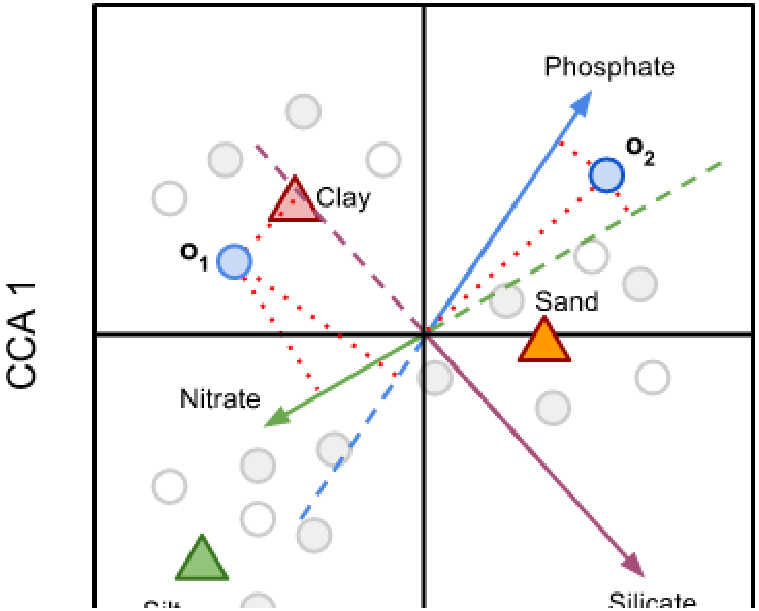
General idea –

- Constrained analysis is a form of **direct gradient analysis**, which attempts to explain variation in data *directly* through the variation in a set of explanatory variables (e.g. environmental factors).

- When constrained analysis are used with ordination techniques the result is an bi- or triplot, the axes built to represent high-dimensional data in a low-dimensional space are *constrained* to be functions of the explanatory factors.

Example: Canonical Correspondence Analysis (CCA)

R: cca() – vegan package



https://sites.google.com/site/mb3gustame/home/

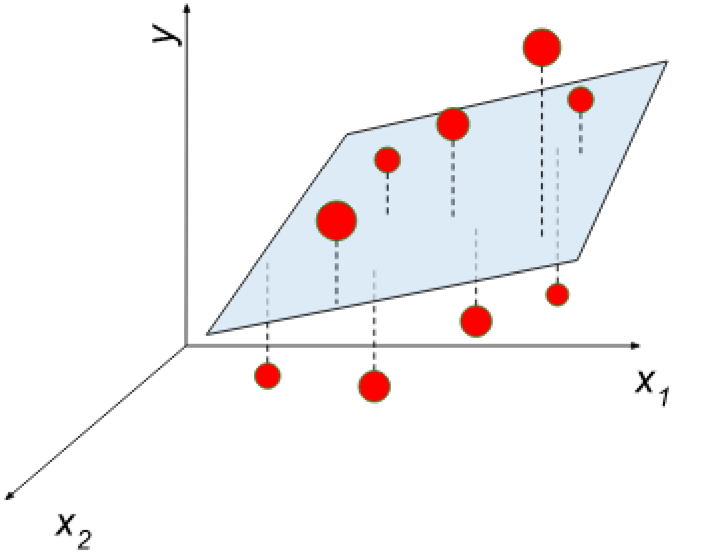
Constrained analyses

Linear regression

- Linear regression (LR) aims to quantify the degree of *linear* association between **several** explanatory variables (i.e. environmental data).
- LR may be used to answer general questions of the kind:

"Is there a significant, linear relationship between my explanatory variables?“

R - The lm() function builds a linear model and can be used for LR when a matrix of explanatory variables is used as input.



https://sites.google.com/site/mb3gustame/home/

Hypothesis tests

General idea….

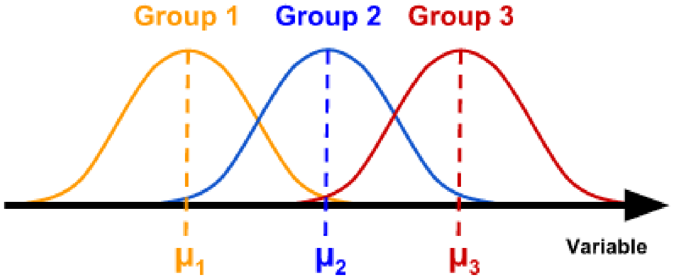
- Hypothesis testing describes a range of methods which attempt to help us make a decision concerning the truth or falsity of a given hypothesis using data from an appropriate experimental design.
- *For example, the hypotheses that two groups of samples have equal mean abundances of a given set of species/OTUs*
- Null hypothesis – accept or reject

Hypothesis tests

ANOVA (Analysis of Variance) - tests whether the assignment of objects groups of one or more explanatory variables (i.e. grouping variables) is statistically significant.

Null hypothesis - The means of two or more groups of objects are equal

R: aov() - The aov() function may also be used to test for differences between models generated by the lm() (also used in regression analysis). ***For this reason ANOVA can be thought of as the designated hypothesis test to complement regression analysis.***



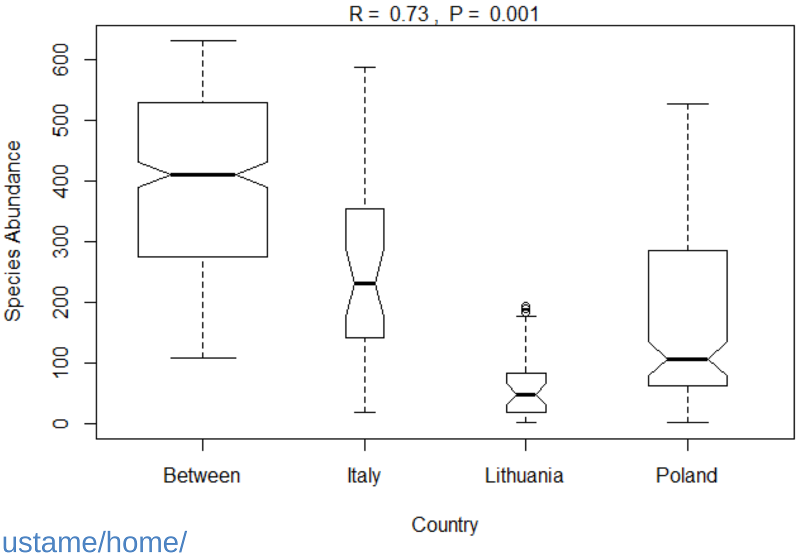
https://sites.google.com/site/mb3gustame/home/

Hypothesis tests

ANOSIM (The ANalysis Of SIMilarity) test - (similar to an ANOVA), however, it is used to evaluate a dissimilarity matrix rather than raw data (Clarke, 1993).

- ANOSIM is the designated hypothesis test for non-metric multidimensional scaling (NMDS) procedure.
- ***Together, the dimension reduction and visualization capacities of NMDS and the hypothesis testing offered by ANOSIM are complementary approaches in evaluating nonparametric multivariate data.***

General Idea: If two groups of samples are different in their species composition, then dissimilarities ***between the groups should be greater than those within the groups***.
The ANOSIM R statistic (0-1) gives an indication of this.



Null hypothesis - There is no difference between the means of two or more groups of (ranked) dissimilarities.

R: anosim() – vegan package

https://sites.google.com/site/mb3gustame/home/

Hypothesis tests

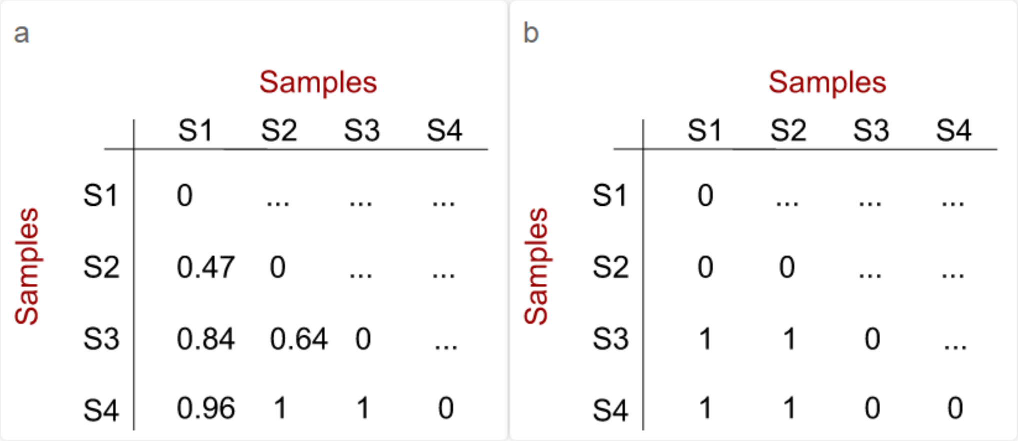
Mantel test (Mantel, 1967) - may be used to ***calculate correlations between corresponding positions of two (dis)similarity or distance matrices***.

- The matrices being tested must be calculated from data sets with the ***same objects, but with different number of variables,*** that should be independent of one another.

Null hypothesis: The distances among objects in a matrix of response variables are not linearly correlated with another matrix of explanatory variables.

Example do sample abundances correlate with certain sample environmental parameters.

R: mantel() – vegan package.



a

Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

b

Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0	0	...	...
S3	1	1	0	...
S4	1	1	0	0

https://sites.google.com/site/mb3gustame/home/

Indirect gradient analysis

Principal Components Analysis

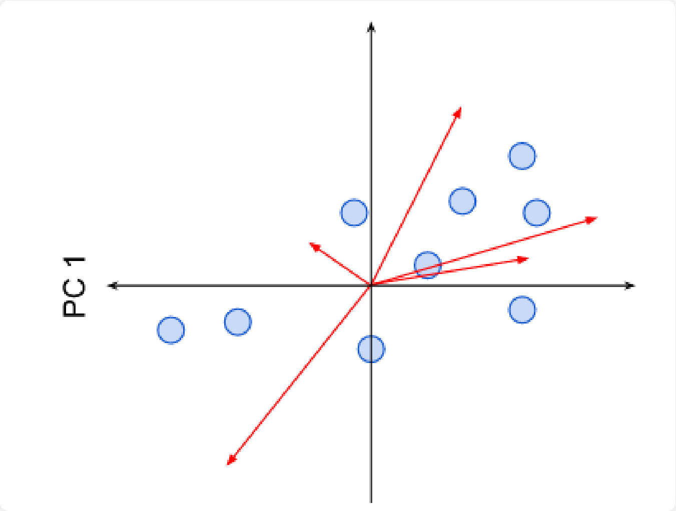
Principal components analysis (PCA) is a method to project, in a low-dimensional space, the variance in a multivariate scatter of points.

- In doing so, it provides an overview of linear relationships between your objects and variables.
- Good starting or end point in multivariate data analysis by allowing you to note trends, groupings, key variables, and potential outliers.
- Good for high dimensional data i.e. variables and relatively few objects and multiple variables (i.e. few samples vs. multiple species table).

R: rda() - vegan package
(a redundancy analysis [RDA] performed

without a matrix of explanatory variables

is equivalent to a PCA)



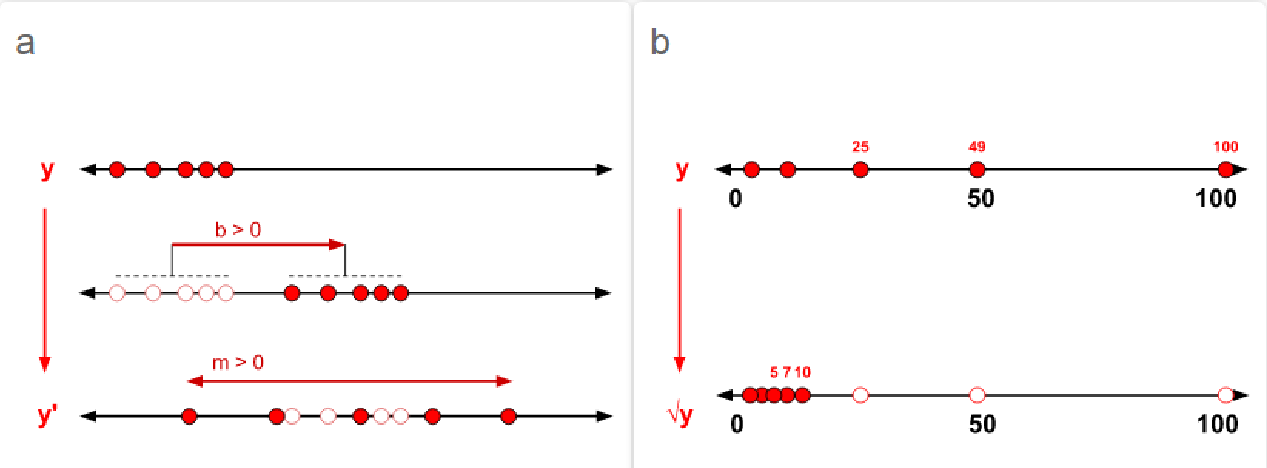
https://sites.google.com/site/mb3gustame/home/

Data Transformations

The main idea...

Occasionally, the variables in a "raw" data set have properties that violate an assumption of a statistical procedure (e.g. normally distributed dataset or when data cannot be compared to other variables due to differences in scale or variability). ***This is were transformation is applicable!***

- For example, principal components analysis (PCA) requires that variables be linearly related to one another and on roughly the same scale or will perform poorly.
- Rather than abandoning an analysis due to inappropriate data structure, it may be possible to transform the variables so they satisfy the conditions in question.



https://sites.google.com/site/mb3gustame/home/

Data Transformations

Scaling: x to presence/absence scale (0/1) .

Chi squared transformation: divide by row sums and square root of column sums (default MARGIN = 1).

Logarithmic transformation: particularly for data with uneven dimensions.

Standardize: scale x to zero mean and unit variance (default MARGIN = 2).

Normalize: make margin sum of squares equal to one (default MARGIN = 1).

Hellinger: ***Particularly suited to species abundance data***, this transformation gives low weights to variables with low counts and many zeros. The transformation itself comprises dividing each value in a data matrix by its row sum, and taking the square root of the quotient.

R: decostand() – vegan package

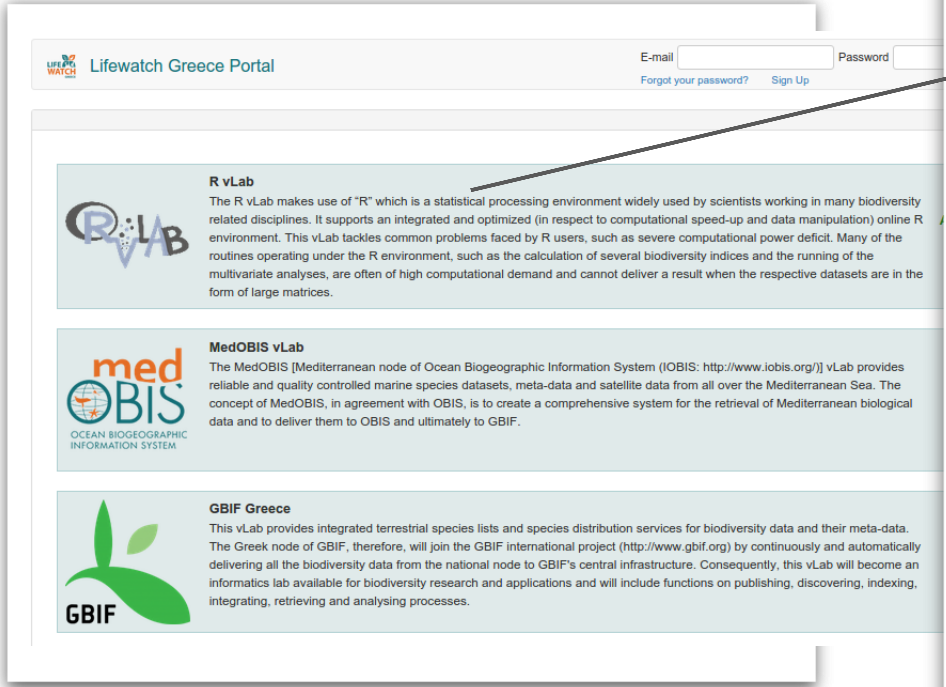
Things to consider!

- Missing Data! - Remove, interpolation?
- Pseudoreplication! – design your experiment carefully, consult statistician.
- Outliers! – Often basic plotting allow you to spot outliers or incorrectly filed data.
- See here for overview:

https://sites.google.com/site/mb3gustame/wizards/screening

Accessed via the LifeWatchGreece Portal

portal.lifewatchgreece.eu



Lifewatch Greece Portal

E-mail
Password
Forgot your password? Sign Up

R vLab

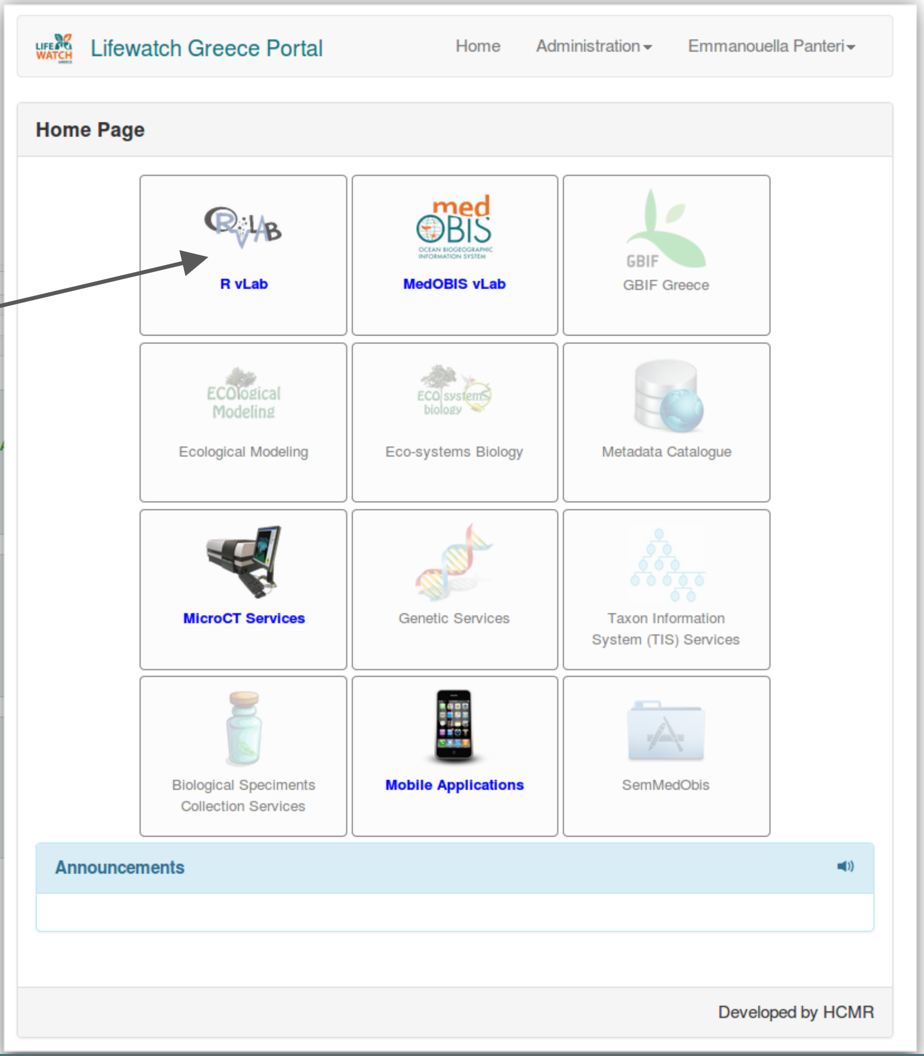
The R vLab makes use of "R" which is a statistical processing environment widely used by scientists working in many biodiversity related disciplines. It supports an integrated and optimized (in respect to computational speed-up and data manipulation) online R environment. This vLab tackles common problems faced by R users, such as severe computational power deficit. Many of the routines operating under the R environment, such as the calculation of several biodiversity indices and the running of the multivariate analyses, are often of high computational demand and cannot deliver a result when the respective datasets are in the form of large matrices.

MedOBIS vLab

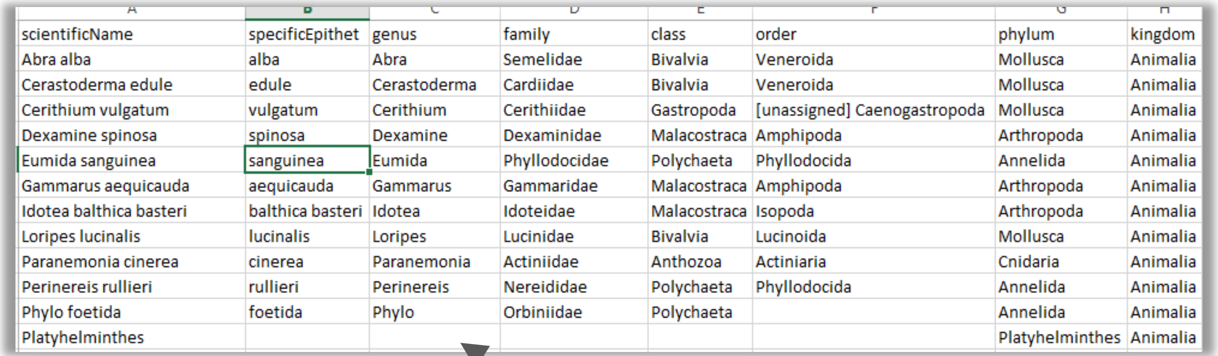
The MedOBIS [Mediterranean node of Ocean Biogeographic Information System (IOBIS: http://www.iobis.org/)] vLab provides reliable and quality controlled marine species datasets, meta-data and satellite data from all over the Mediterranean Sea. The concept of MedOBIS, in agreement with OBIS, is to create a comprehensive system for the retrieval of Mediterranean biological data and to deliver them to OBIS and ultimately to GBIF.

GBIF Greece

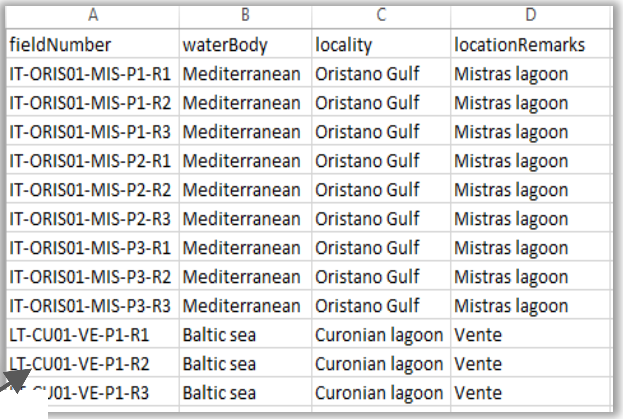
This vLab provides integrated terrestrial species lists and species distribution services for biodiversity data and their meta-data. The Greek node of GBIF, therefore, will join the GBIF international project (http://www.gbif.org) by continuously and automatically delivering all the biodiversity data from the national node to GBIF's central infrastructure. Consequently, this vLab will become an informatics lab available for biodiversity research and applications and will include functions on publishing, discovering, indexing, integrating, retrieving and analysing processes.



Main file formats required for R file.csv



A	B	C	D	E	F	G	H
scientificName	specificEpithet	genus	family	class	order	phylum	kingdom
Abra alba	alba	Abra	Semelidae	Bivalvia	Veneroida	Mollusca	Animalia
Cerastoderma edule	edule	Cerastoderma	Cardiidae	Bivalvia	Veneroida	Mollusca	Animalia
Cerithium vulgatum	vulgatum	Cerithium	Cerithiidae	Gastropoda	[unassigned] Caenogastropoda	Mollusca	Animalia
Dexamine spinosa	spinosa	Dexamine	Dexaminidae	Malacostraca	Amphipoda	Arthropoda	Animalia
Eumida sanguinea	sanguinea	Eumida	Phyllodocidae	Polychaeta	Phyllodocida	Annelida	Animalia
Gammarus aequicauda	aequicauda	Gammarus	Gammaridae	Malacostraca	Amphipoda	Arthropoda	Animalia
Idotea balthica basteri	balthica basteri	Idotea	Idoteidae	Malacostraca	Isopoda	Arthropoda	Animalia
Loripes lucinalis	lucinalis	Loripes	Lucinidae	Bivalvia	Lucinoida	Mollusca	Animalia
Paranemonia cinerea	cinerea	Paranemonia	Actiniidae	Anthozoa	Actiniaria	Cnidaria	Animalia
Perinereis rullieri	rullieri	Perinereis	Nereididae	Polychaeta	Phyllodocida	Annelida	Animalia
Phylo foetida	foetida	Phylo	Orbiniidae	Polychaeta		Annelida	Animalia
Platyhelminthes						Platyhelminthes	Animalia



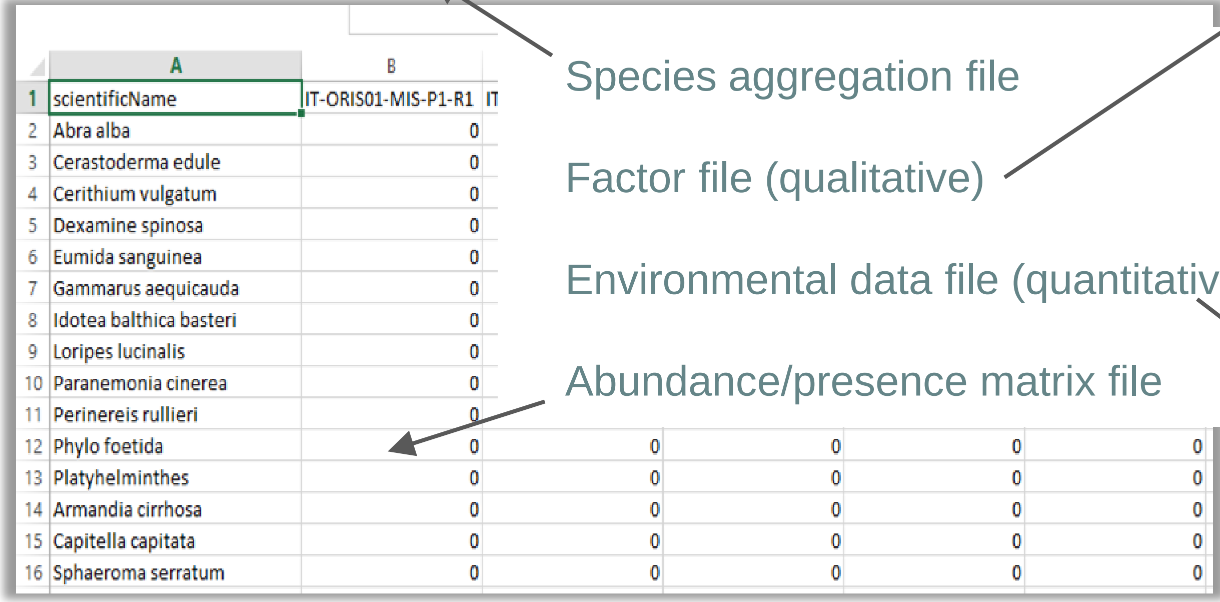
A	B	C	D
fieldNumber	waterBody	locality	locationRemarks
IT-ORIS01-MIS-P1-R1	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P1-R2	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P1-R3	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P2-R1	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P2-R2	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P2-R3	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P3-R1	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P3-R2	Mediterranean	Oristano Gulf	Mistras lagoon
IT-ORIS01-MIS-P3-R3	Mediterranean	Oristano Gulf	Mistras lagoon
LT-CU01-VE-P1-R1	Baltic sea	Curonian lagoon	Vente
LT-CU01-VE-P1-R2	Baltic sea	Curonian lagoon	Vente
LT-CU01-VE-P1-R3	Baltic sea	Curonian lagoon	Vente

Species aggregation file

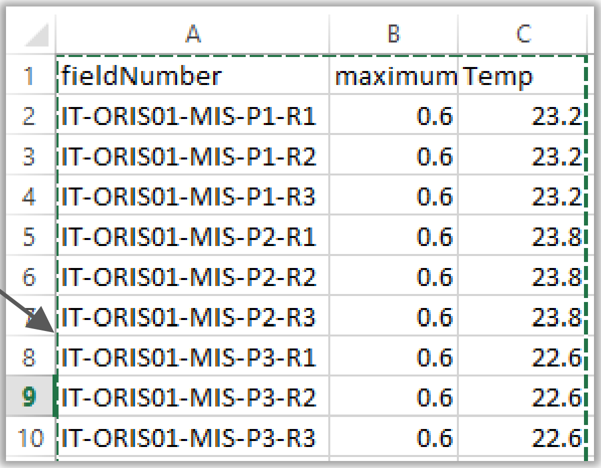
Factor file (qualitative)

Environmental data file (quantitative)

Abundance/presence matrix file



	A	B				
1	scientificName	IT-ORIS01-MIS-P1-R1	IT			
2	Abra alba	0				
3	Cerastoderma edule	0				
4	Cerithium vulgatum	0				
5	Dexamine spinosa	0				
6	Eumida sanguinea	0				
7	Gammarus aequicauda	0				
8	Idotea balthica basteri	0				
9	Loripes lucinalis	0				
10	Paranemonia cinerea	0				
11	Perinereis rullieri	0				
12	Phylo foetida	0	0	0	0	0
13	Platyhelminthes	0	0	0	0	0
14	Armandia cirrhosa	0	0	0	0	0
15	Capitella capitata	0	0	0	0	0
16	Sphaeroma serratum	0	0	0	0	0



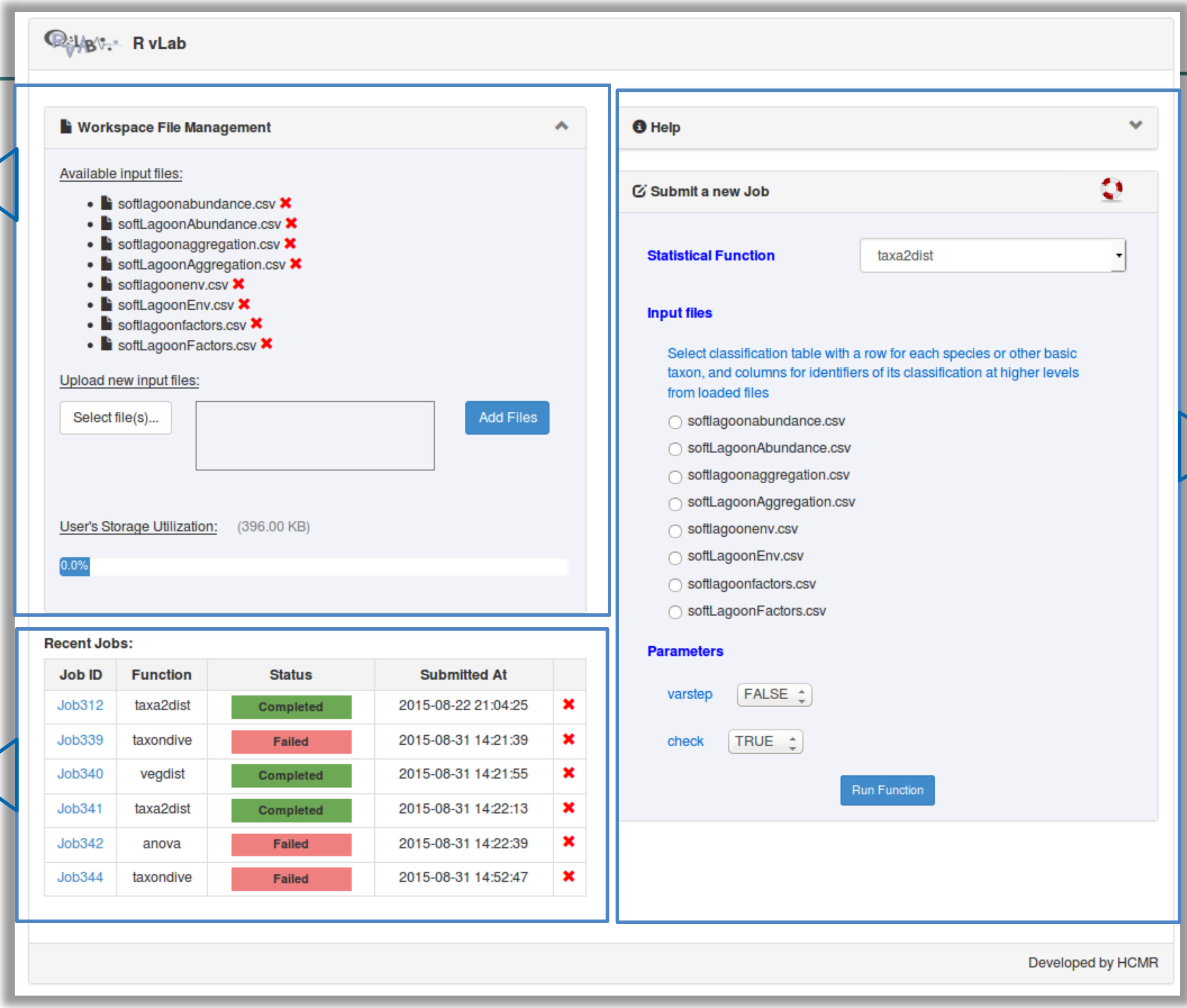
	A	B	C
1	fieldNumber	maximum	Temp
2	IT-ORIS01-MIS-P1-R1	0.6	23.2
3	IT-ORIS01-MIS-P1-R2	0.6	23.2
4	IT-ORIS01-MIS-P1-R3	0.6	23.2
5	IT-ORIS01-MIS-P2-R1	0.6	23.8
6	IT-ORIS01-MIS-P2-R2	0.6	23.8
7	IT-ORIS01-MIS-P2-R3	0.6	23.8
8	IT-ORIS01-MIS-P3-R1	0.6	22.6
9	IT-ORIS01-MIS-P3-R2	0.6	22.6
10	IT-ORIS01-MIS-P3-R3	0.6	22.6

Rvlab Main Page

Workspace (Files)

Jobs Submitted

Functions area



R vLab

Workspace File Management

Available input files:

- softlagoonabundance.csv ✖
- softLagoonAbundance.csv ✖
- softlagoonaggregation.csv ✖
- softLagoonAggregation.csv ✖
- softlagoonenv.csv ✖
- softLagoonEnv.csv ✖
- softlagoonfactors.csv ✖
- softLagoonFactors.csv ✖

Upload new input files:

User's Storage Utilization: (396.00 KB)

0.0%

Recent Jobs:

Job ID	Function	Status	Submitted At	
Job312	taxa2dist	Completed	2015-08-22 21:04:25	✖
Job339	taxondive	Failed	2015-08-31 14:21:39	✖
Job340	vegdist	Completed	2015-08-31 14:21:55	✖
Job341	taxa2dist	Completed	2015-08-31 14:22:13	✖
Job342	anova	Failed	2015-08-31 14:22:39	✖
Job344	taxondive	Failed	2015-08-31 14:52:47	✖

Help

Submit a new Job

Statistical Function: taxa2dist

Input files

Select classification table with a row for each species or other basic taxon, and columns for identifiers of its classification at higher levels from loaded files

○ softlagoonabundance.csv
○ softLagoonAbundance.csv
○ softlagoonaggregation.csv
○ softLagoonAggregation.csv
○ softlagoonenv.csv
○ softLagoonEnv.csv
○ softlagoonfactors.csv
○ softLagoonFactors.csv

Parameters

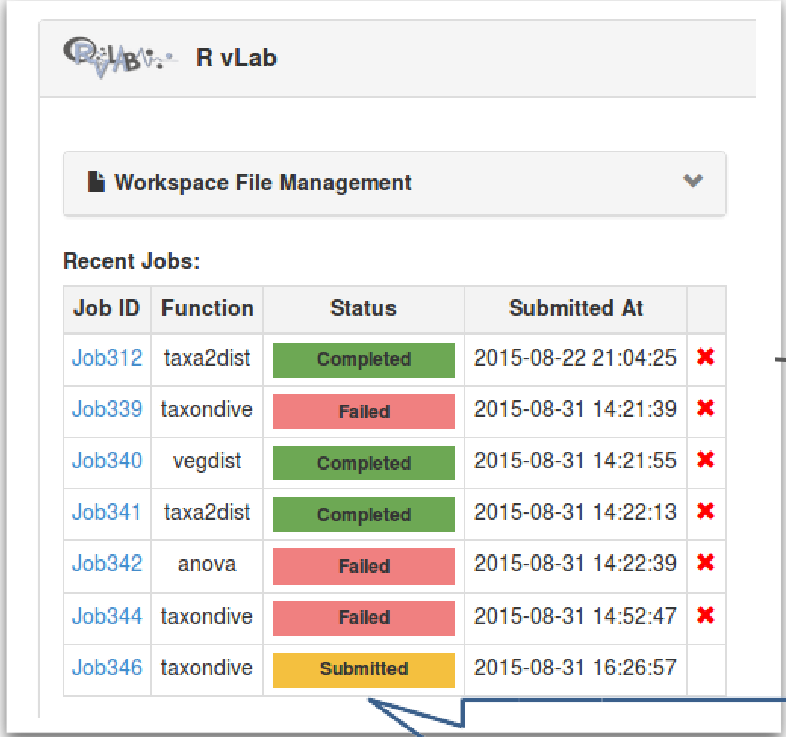
varstep: FALSE

check: TRUE

Developed by HCMR

Rvlab Jobs and Results

Jobs



R vLab

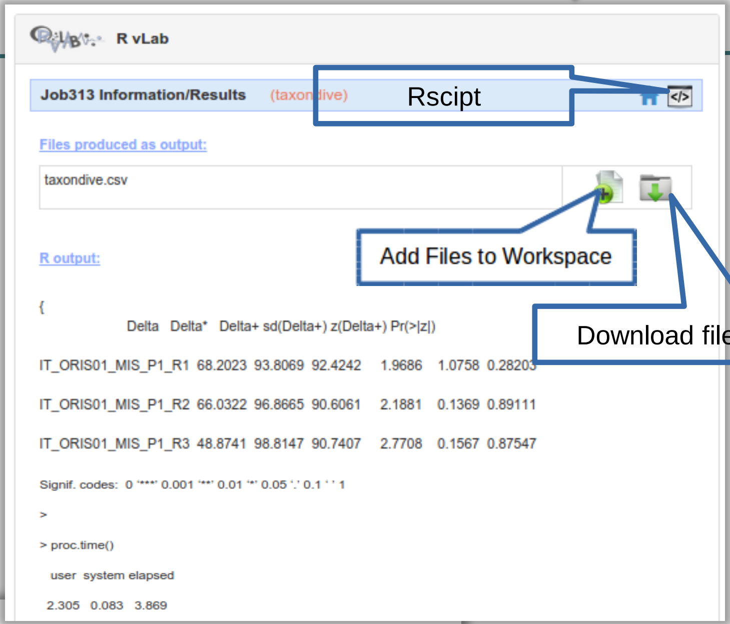
Workspace File Management

Recent Jobs:

Job ID	Function	Status	Submitted At	
Job312	taxa2dist	Completed	2015-08-22 21:04:25	✖
Job339	taxondive	Failed	2015-08-31 14:21:39	✖
Job340	vegdist	Completed	2015-08-31 14:21:55	✖
Job341	taxa2dist	Completed	2015-08-31 14:22:13	✖
Job342	anova	Failed	2015-08-31 14:22:39	✖
Job344	taxondive	Failed	2015-08-31 14:52:47	✖
Job346	taxondive	Submitted	2015-08-31 16:26:57	

Wait for results to be generated

Jobs Results



R vLab

Job313 Information/Results (taxondive)

Files produced as output:

taxondive.csv

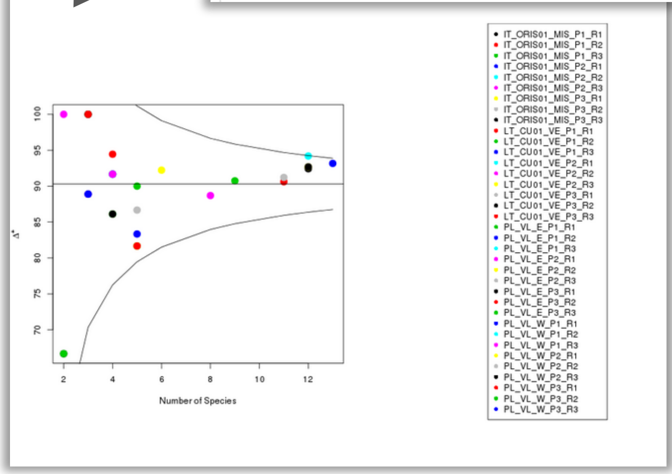
R output:

```
{
                 Delta   Delta*   Delta+ sd(Delta+) z(Delta+) Pr(>|z|)
IT_ORIS01_MIS_P1_R1  68.2023  93.8069  92.4242    1.9686    1.0758  0.28203
IT_ORIS01_MIS_P1_R2  66.0322  96.8665  90.6061    2.1881    0.1369  0.89111
IT_ORIS01_MIS_P1_R3  48.8741  98.8147  90.7407    2.7708    0.1567  0.87547
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> proc.time()
   user  system elapsed
  2.305   0.083   3.869
```

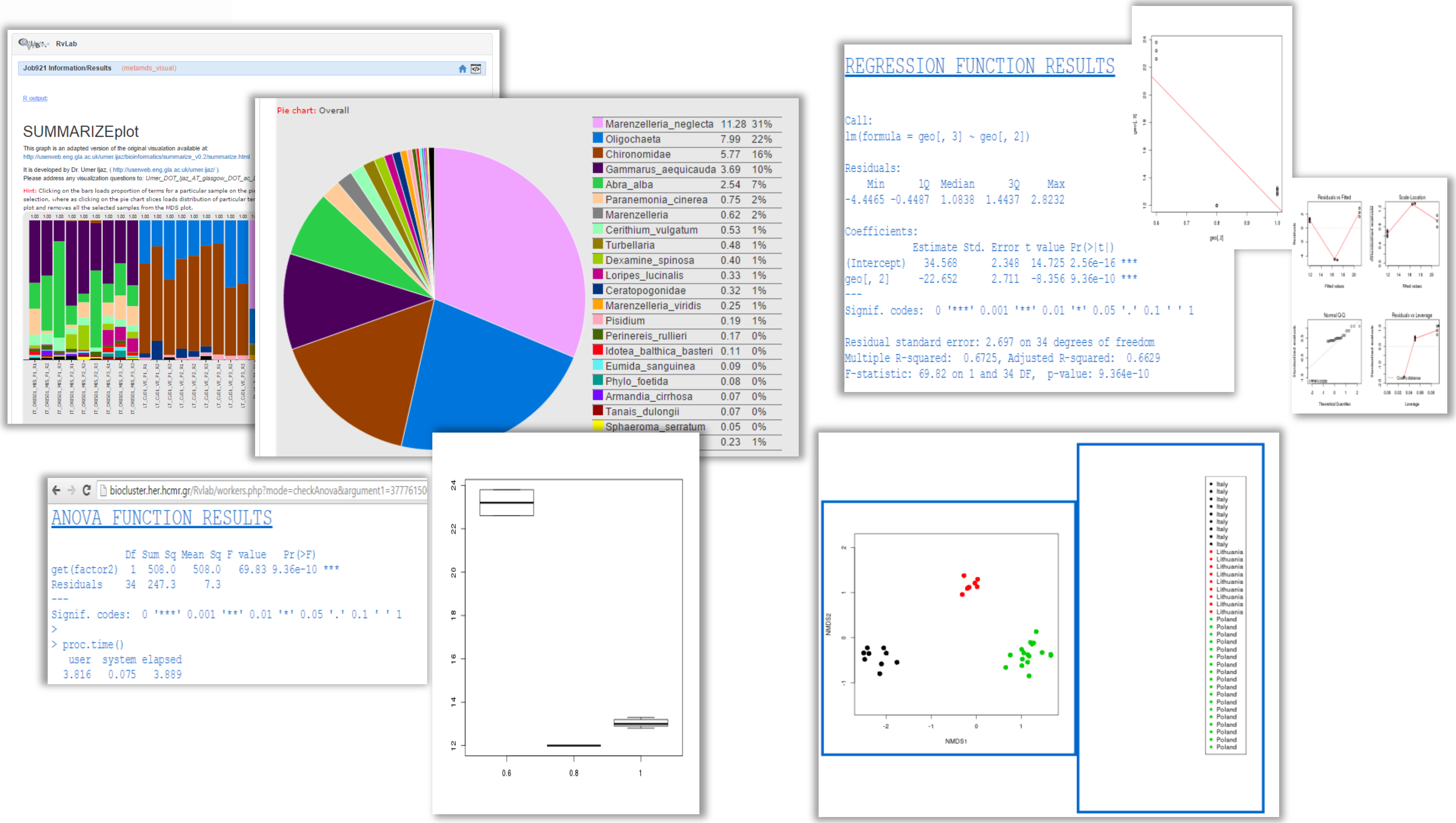
Rscipt

Add Files to Workspace

Download files



Rvlab Jobs and Results



Job921 Information/Results (metamds_visual)

R output:

SUMMARIZEplot

This graph is an adapted version of the original visualation available at: http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/summarize_v0.2/summarize.html

It is developed by Dr. Umer Ijaz, (http://userweb.eng.gla.ac.uk/umer.ijaz/).
Please address any visualization questions to: Umer_DOT_Ijaz_AT_glasgow_DOT_ac_

Hint: Clicking on the bars loads proportion of terms for a particular sample on the pie
selection, where as clicking on the pie chart slices loads distribution of particular ter
plot and removes all the selected samples from the MDS plot.

🟪	Marenzelleria_neglecta	11.28	31%
🟦	Oligochaeta	7.99	22%
🟫	Chironomidae	5.77	16%
🟪	Gammarus_aequicauda	3.69	10%
🟩	Abra_alba	2.54	7%
🟧	Paranemonia_cinerea	0.75	2%
⬛	Marenzelleria	0.62	2%
🟩	Cerithium_vulgatum	0.53	1%
🟫	Turbellaria	0.48	1%
🟩	Dexamine_spinosa	0.40	1%
🟪	Loripes_lucinalis	0.33	1%
🟦	Ceratopogonidae	0.32	1%
🟧	Marenzelleria_viridis	0.25	1%
🟥	Pisidium	0.19	1%
🟩	Perinereis_rullieri	0.17	0%
🟥	Idotea_balthica_basteri	0.11	0%
🟦	Eumida_sanguinea	0.09	0%
🟩	Phylo_foetida	0.08	0%
🟪	Armandia_cirrhosa	0.07	0%
🟥	Tanais_dulongii	0.07	0%
🟨	Sphaeroma_serratum	0.05	0%
		0.23	1%

REGRESSION FUNCTION RESULTS

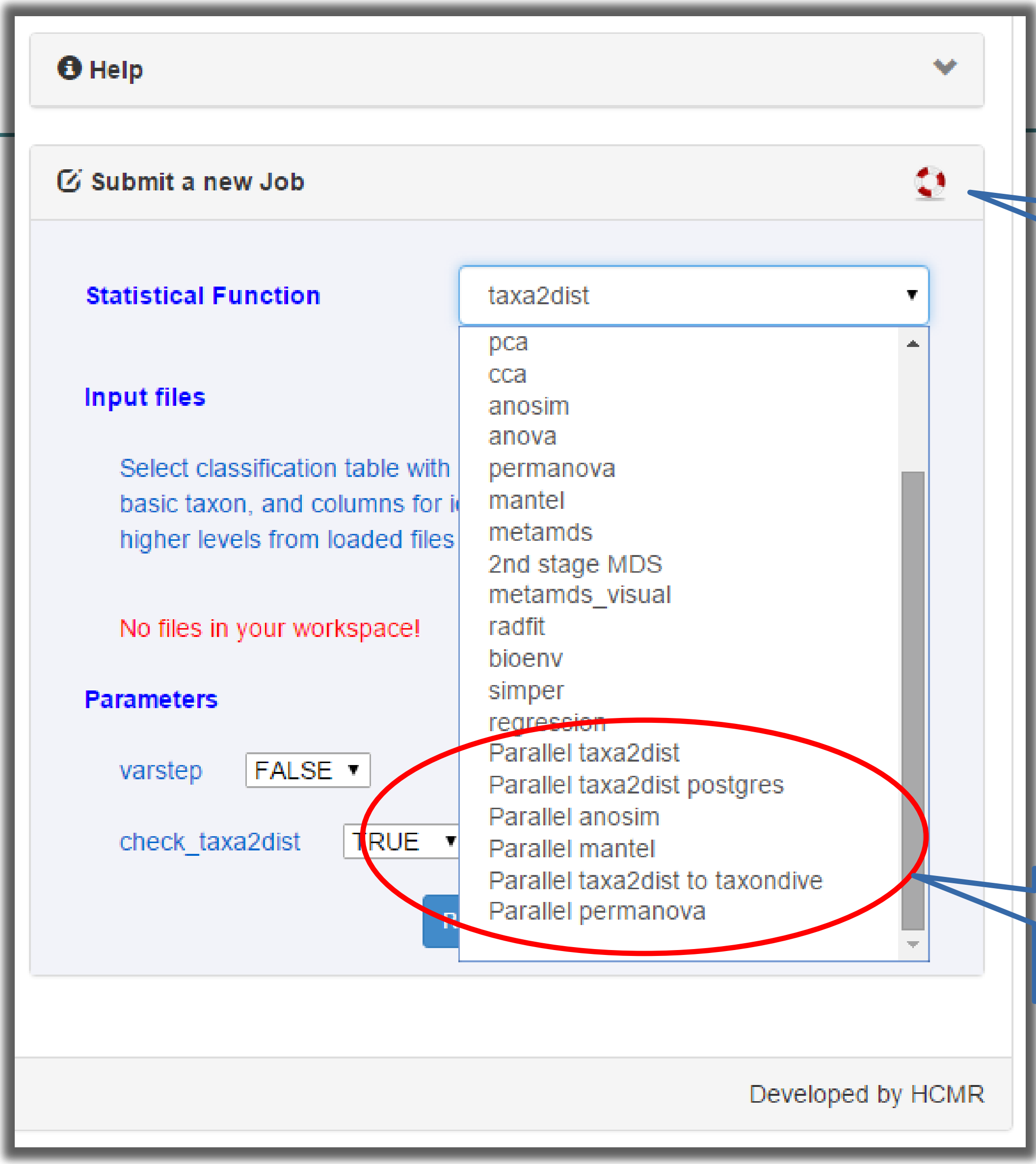
```
Call:
lm(formula = geo[, 3] ~ geo[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4465 -0.4487  1.0838  1.4437  2.8232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   34.568      2.348  14.725 2.56e-16 ***
geo[, 2]     -22.652      2.711  -8.356 9.36e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.697 on 34 degrees of freedom
Multiple R-squared:  0.6725, Adjusted R-squared:  0.6629 
F-statistic: 69.82 on 1 and 34 DF,  p-value: 9.364e-10
```

biocluster.her.hcmr.gr/Rvlab/workers.php?mode=checkAnova&argument1=37776150

ANOVA FUNCTION RESULTS

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
get(factor2)  1  508.0   508.0   69.83 9.36e-10 ***
Residuals    34  247.3     7.3                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
> proc.time()
   user  system elapsed 
  3.816   0.075   3.889 
```

Functions supported by Rvlab



Help

Submit a new Job

Statistical Function: taxa2dist

pca
cca
anosim
anova
permanova
mantel
metamds
2nd stage MDS
metamds_visual
radfit
bioenv
simper
regression
Parallel taxa2dist
Parallel taxa2dist postgres
Parallel anosim
Parallel mantel
Parallel taxa2dist to taxondive
Parallel permanova

Input files

Select classification table with basic taxon, and columns for i higher levels from loaded files

No files in your workspace!

Parameters

varstep: FALSE

check_taxa2dist: TRUE

Developed by HCMR

Function documentation

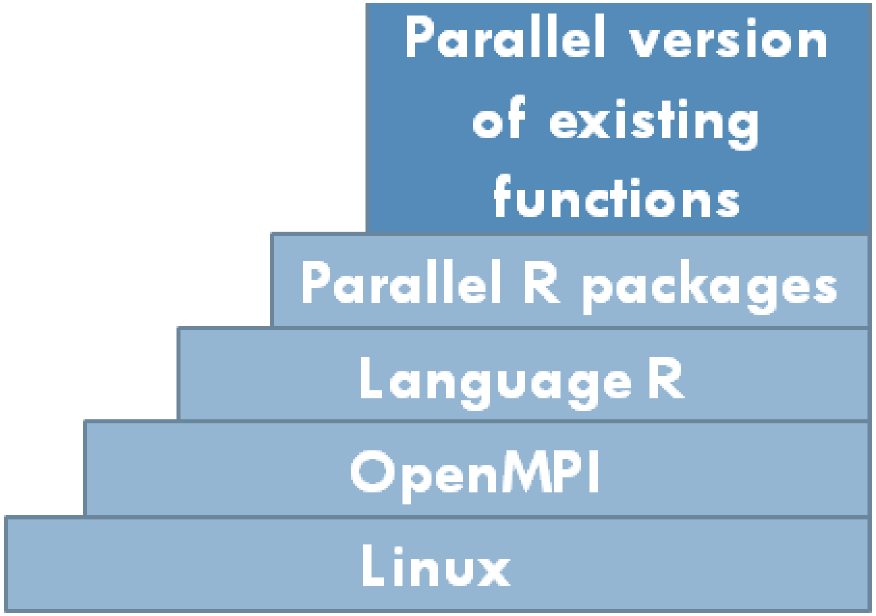
Parallel implementations

Parallel R packages utilized by Rvlab

R packages, such as:

- Vegan CRAN
- pbdR, bigmemory
- RMPI, pbdMPI
- parallel, multicore, snow
- RPostgreSQL, dplyr
- profiling packages (profr, proftools)
- graphical packages (grid, Rgraphviz)

Linux environment.



Parallel version of existing functions

Parallel R packages

Language R

OpenMPI

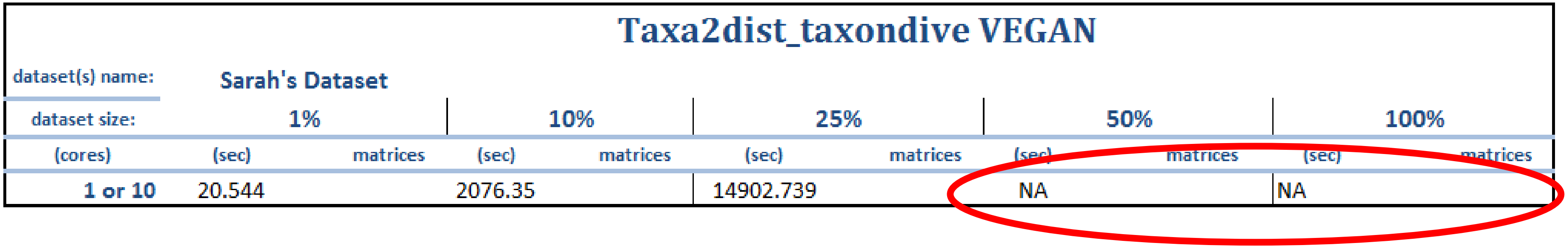
Linux

Some examples of parallel implementations

1. Big data manipulation (overcome memory barriers)
2. Computational time speed-up (task segmentation, multi-cores)

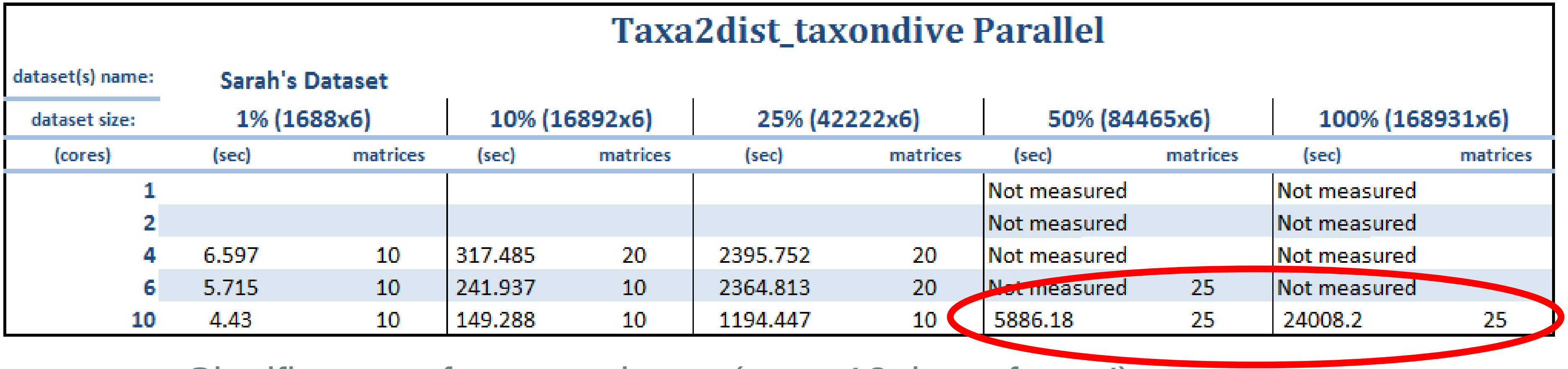
Cluster computing environment at HCMR – recent upgrade from LlifeWatch)

Taxa2Dist > Taxondive



Taxa2dist_taxondive VEGAN

dataset(s) name:	Sarah's Dataset									
dataset size:	1%		10%		25%		50%		100%	
(cores)	(sec)	matrices	(sec)	matrices	(sec)	matrices	(sec)	matrices	(sec)	matrices
1 or 10	20.544		2076.35		14902.739		NA		NA	



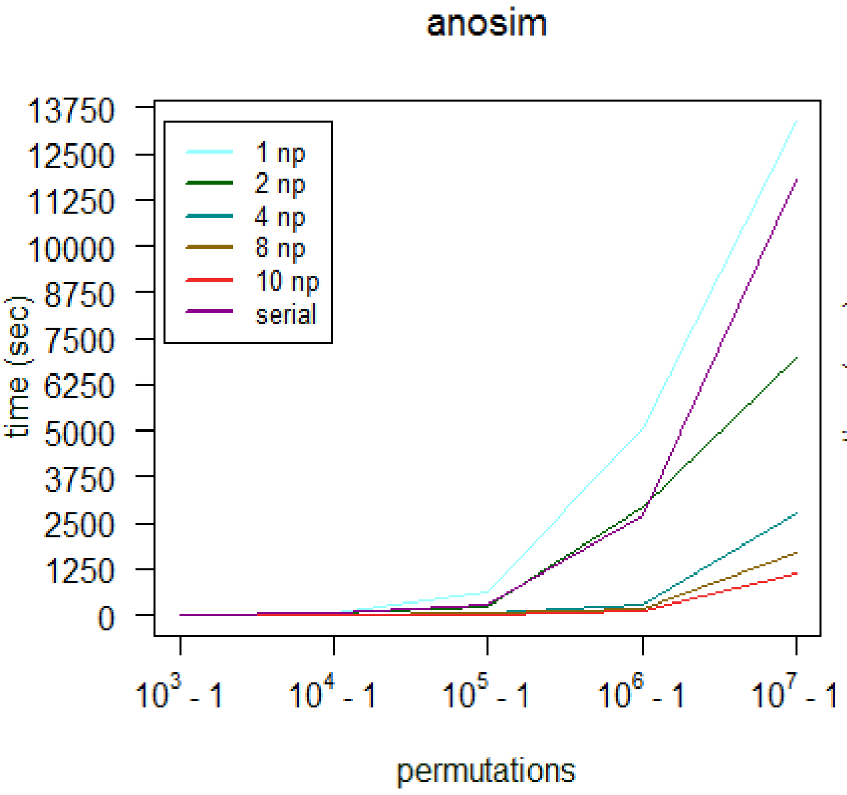
Taxa2dist_taxondive Parallel

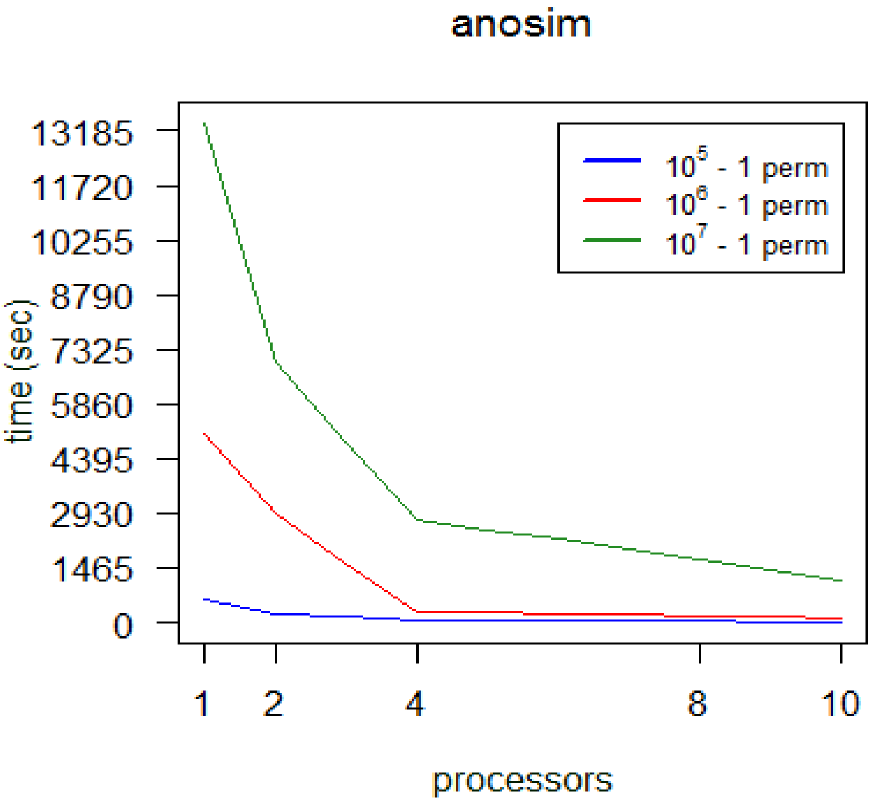
dataset(s) name:	Sarah's Dataset									
dataset size:	1% (1688x6)		10% (16892x6)		25% (42222x6)		50% (84465x6)		100% (168931x6)	
(cores)	(sec)	matrices	(sec)	matrices	(sec)	matrices	(sec)	matrices	(sec)	matrices
1							Not measured		Not measured	
2							Not measured		Not measured	
4	6.597	10	317.485	20	2395.752	20	Not measured		Not measured	
6	5.715	10	241.937	10	2364.813	20	Not measured	25	Not measured	
10	4.43	10	149.288	10	1194.447	10	5886.18	25	24008.2	25

Significant performance boost (up to 12 times faster!)

No memory barriers apply

Anosim speed-up





~2 fold speed-up

Parallel function guidelines

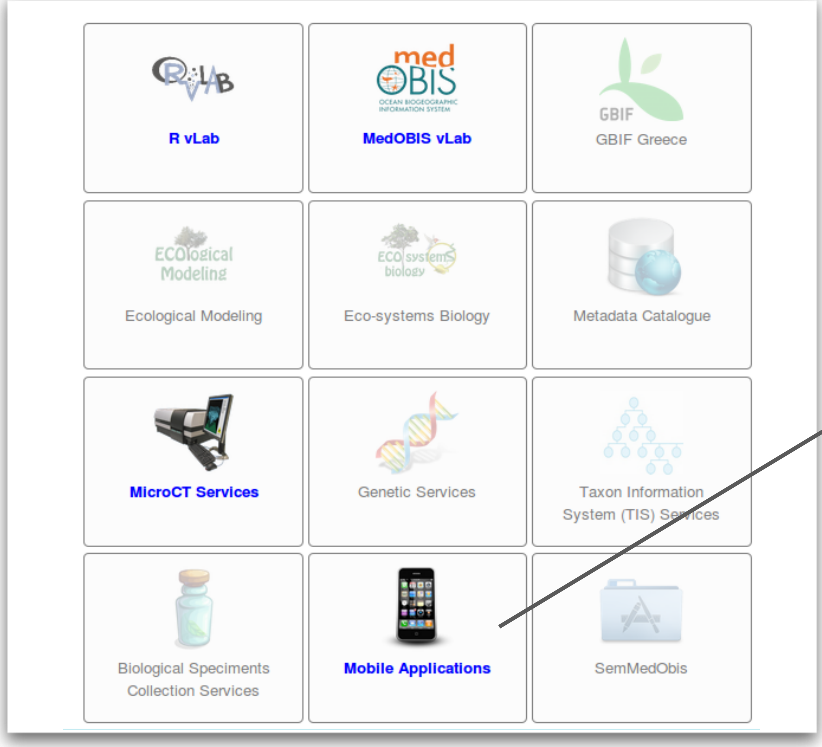
- Parallel Anosim – Only more than 10^6 permutations parallel versions are faster than the serial one.
- Parallel Taxa2Dist -> TaxonDive - Only for datasets exceeding memory resources otherwise performance is slower.

Things to consider! –

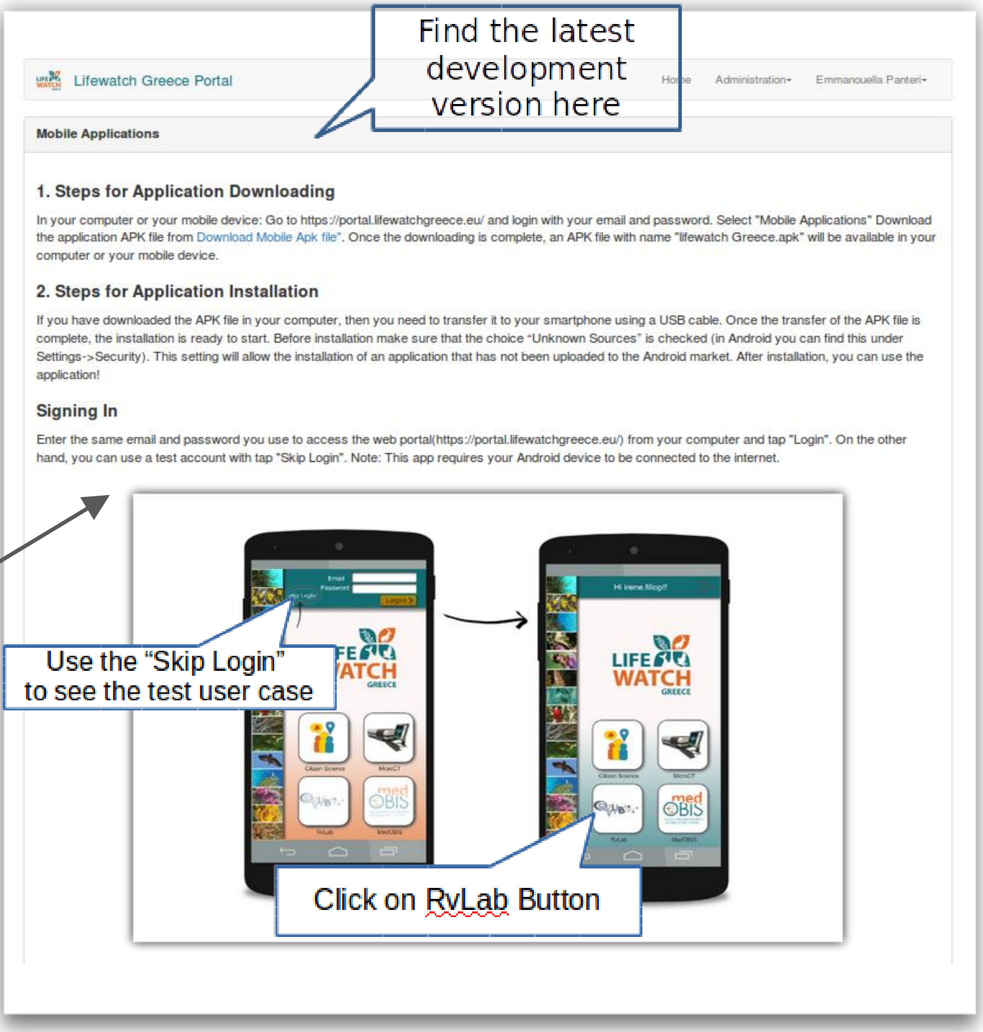
- **Assign jobs as a function of available resources (i.e. available cores & submitted jobs in queue)**
- **Size of data.**
- **Parameters selected.**
- **Decide on optimal function to use**

Rvlab Mobile Application – links to your account in portal!

portal.lifewatchgreece.eu



Filiopoulou Irene, Panteri Emmanouela, Gougousis Alexandros



Find the latest development version here

Mobile Applications

1. Steps for Application Downloading

In your computer or your mobile device: Go to https://portal.lifewatchgreece.eu/ and login with your email and password. Select "Mobile Applications" Download the application APK file from Download Mobile Apk file". Once the downloading is complete, an APK file with name "lifewatch Greece.apk" will be available in your computer or your mobile device.

2. Steps for Application Installation

If you have downloaded the APK file in your computer, then you need to transfer it to your smartphone using a USB cable. Once the transfer of the APK file is complete, the installation is ready to start. Before installation make sure that the choice "Unknown Sources" is checked (in Android you can find this under Settings->Security). This setting will allow the installation of an application that has not been uploaded to the Android market. After installation, you can use the application!

Signing In

Enter the same email and password you use to access the web portal(https://portal.lifewatchgreece.eu/) from your computer and tap "Login". On the other hand, you can use a test account with tap "Skip Login". Note: This app requires your Android device to be connected to the internet.

Use the "Skip Login" to see the test user case

Click on RvLab Button

Acknowledgments

LifeWatchGreece Research Infrastructure , funded by the GSRT (Greek government: structural funds), is the national effort to address the above requirement and to support relevant studies.

Coordinator: Christos Arvanitidis

Rvlab Contributors: Patkos Theodore, Oulas Anastasis, Gougousis Alexandros, Varsos Constantinos, Pattakos Nikos, Vanden Berghe Edward, Fernandez-Guerra Antonio, Pavloudi Christina, Bekiari Chrysoula, Doerr Martin, Panteri Emmanouela, Arvanitidis Christos

Hernandez Francisco

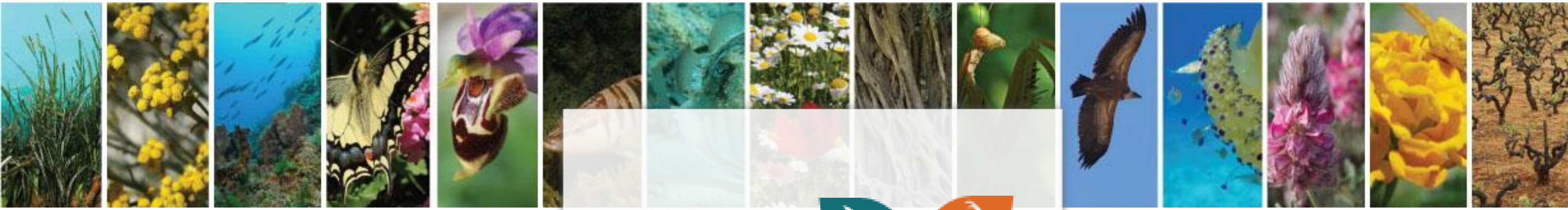
Klaas Deneudt

Stefanie Dekeyze

Useful links and material

https://sites.google.com/site/mb3gustame/home/

Citation: Buttigieg PL, Ramette A (2014) A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol*. **90**: 543–550



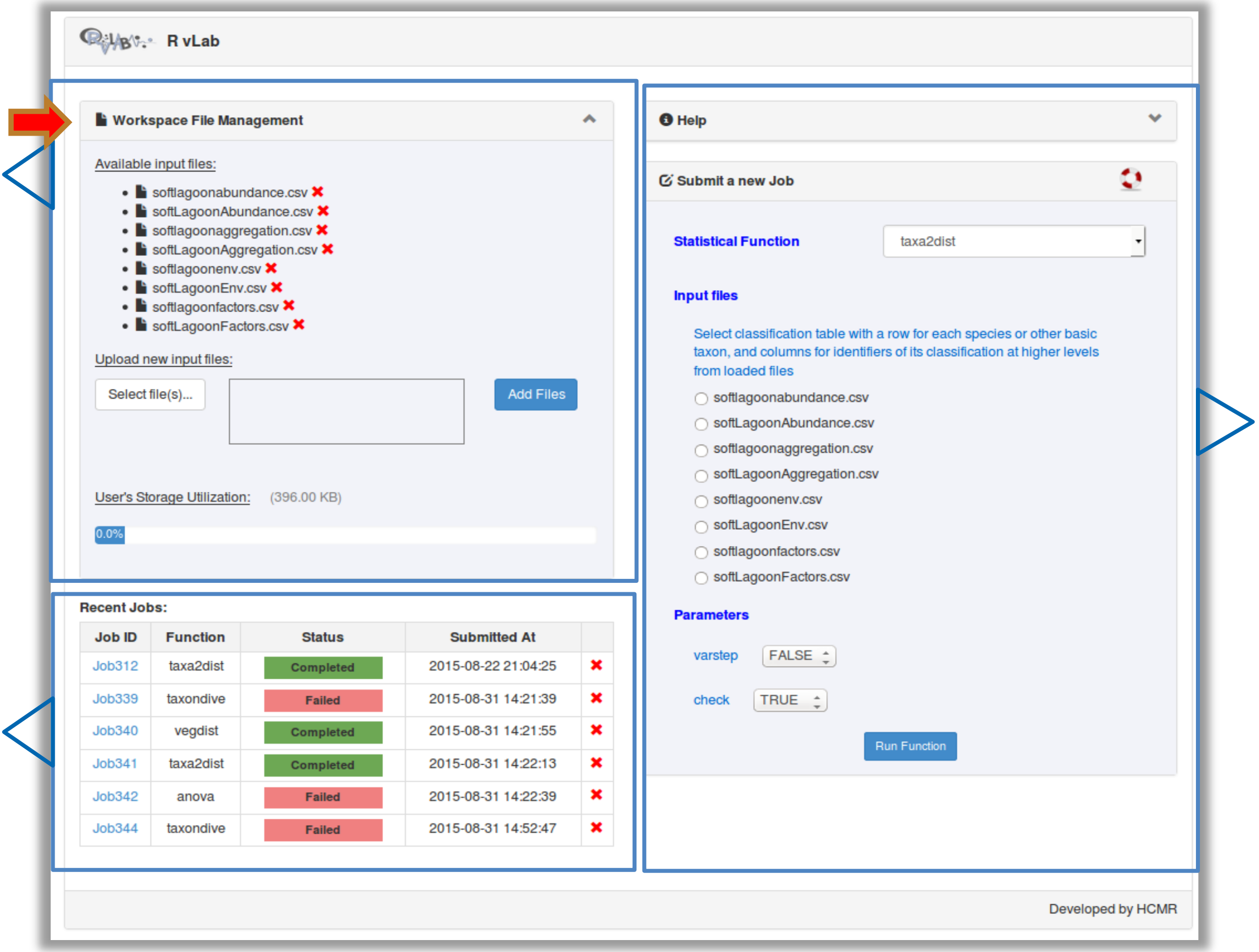
LIFE WATCH GREECE

Thank you

Hands-on
https://rvlab.portal.lifewatchgreece.eu/registration

Guideline for hands-on

Load file NSBS_All_Stations_bath_env_data.csv - use the *workspace management* tab of Rvlab to get started!



Start here

Workspace (Files)

Jobs Submitted

Functions area

R vLab

Workspace File Management

Available input files:

- softlagoonabundance.csv ✖
- softLagoonAbundance.csv ✖
- softlagoonaggregation.csv ✖
- softLagoonAggregation.csv ✖
- softlagoonenv.csv ✖
- softLagoonEnv.csv ✖
- softlagoonfactors.csv ✖
- softLagoonFactors.csv ✖

Upload new input files:

Select file(s)... Add Files

User's Storage Utilization: (396.00 KB)

0.0%

Recent Jobs:

Job ID	Function	Status	Submitted At	
Job312	taxa2dist	Completed	2015-08-22 21:04:25	✖
Job339	taxondive	Failed	2015-08-31 14:21:39	✖
Job340	vegdist	Completed	2015-08-31 14:21:55	✖
Job341	taxa2dist	Completed	2015-08-31 14:22:13	✖
Job342	anova	Failed	2015-08-31 14:22:39	✖
Job344	taxondive	Failed	2015-08-31 14:52:47	✖

Help

Submit a new Job

Statistical Function: taxa2dist

Input files

Select classification table with a row for each species or other basic taxon, and columns for identifiers of its classification at higher levels from loaded files

- ○ softlagoonabundance.csv
- ○ softLagoonAbundance.csv
- ○ softlagoonaggregation.csv
- ○ softLagoonAggregation.csv
- ○ softlagoonenv.csv
- ○ softLagoonEnv.csv
- ○ softlagoonfactors.csv
- ○ softLagoonFactors.csv

Parameters

varstep: FALSE

check: TRUE

Run Function

Developed by HCMR

Guideline for hands-on

Format data

- ➢ ***Execute Covert to R*** function of Rvlab - use “NSBS_All_Stations_bath_env_data.csv”
- You should get an abundance file of Species vs. Stations and an environmental parameter file for all Stations and selected environmental parameters.
- Care with missing data. You may decide to remove or substitute NA values.
- Care with spaces and non-alphanumeric characters in header names.
- ➢ ***Execute Rscript* -** Help is given using available R script (Data_format.R), Read-in files “transformed_dataAbu_job1350.csv” and “ENV.csv”. They should be the same as the files you generated using the “convert 2 R” function executed previously.
- You should generate 2 files *Formatted_Abu.csv* and *Formatted_ENV.csv*
- You should also generate a pairs plot for the Environmental parameters.
- *Are there any outliers in your data?*
- You will also get an additional file *Desc_Depth_labels.csv* of discretized maximum depth labels for your stations with the following depth intervals: 0-20, 21-50, 51-80, 81-110, 111-150. You can use the available R script or a spreadsheet to do this.
- ***Load all newly generated files into Rvlab.***

Guideline for hands-on

Get a first impression of your data

- ***Execute metaMDS_visual*** function of Rvlab using “Formatted_Abu.csv” file. This should give you a first indication of the most abundant species in your data both at the station level as well as complete picture for all stations.
- You can play with the parameters of ***metaMDS_visual*** and choose different ***number of top ranked species*** to display and
- You may also add a new matrix with the most abundant species into your Rvlab workspace for additional analyses.

Guideline for hands-on

(Dis)similarity and Ordination analysis of data

- **Execute BIOENV** – use Formatted_Abu.csv and Formatted_ENV.csv files - This analysis allows you to get an indication of the environmental parameters that best correlate with your data. Select different parameters like number of variables to (*upto parameter*) include in the analysis and asses the models using the correlation values. *Hint: use Bray Curtis coefficient*

- **Execute Regression** – use Formatted_ENV.csv - You may want to run a regression analysis for environmental parameters selected from BIOENV.

- **Execute SIMPER** – use Formatted_Abu.csv and Desc_Depth_labels.csv - this analysis gives you an indication of the most influential species for all pairwise comparisons for the discretized depth groups. *How do the results compare with the metaMDS_visual analysis performed earlier?*

- **Execute metaMDS** – use Formatted_Abu.csv and Desc_Depth_labels.csv - Use this analysis to visualize your data by colour coding using the discretized depth column labels you created using the Rscript.

Guideline for hands-on

Hypothesis testing

- **Execute ANOSIM** - use Formatted_Abu.csv and Desc_Depth_labels.csv – This will allow you to get a significance value for between and within level of dissimilarity for the discretized depth groups. *Hint: Try a Hellinger transformation to improve the R statistic*.
- **Execute ANOVA** – use Formatted_ENV.csv – this analysis can be used to complement regression analysis performed earlier on selected environmental parameters.

Constrained analysis

- **Execute CCA** – use Formatted_Abu.csv and Formatted_ENV.csv - perform cca using the environmental parameters that were found to best correlate with your data according to the BIOENV analysis. *You can transfer the script to Rstudio and use the identify() function to observe specific points (species or stations) in your data plot.*

Guideline for hands-on

Transformation and Indirect gradient analysis ordination.

- **Execute PCA** – use Formatted_Abu.csv - Try different transformation methods (*i.e. Hellinger transformation*) and run **PCA analysis** using the discretized depth labels to colour code your samples. *Do you observe any differences in the ordination plot?*

End of Hands-on

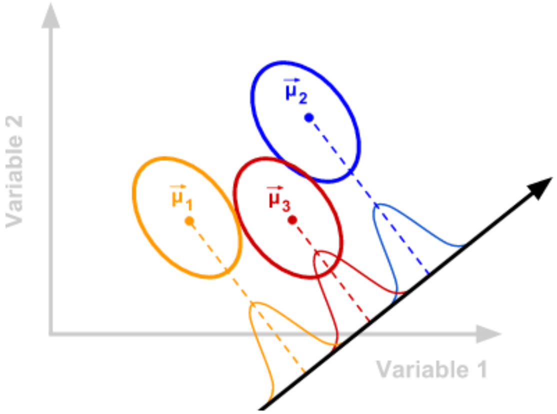
Hypothesis tests

MANOVA

Multivariate analysis of variance (MANOVA) is the multivariate analogues of univariate ANOVA test. It thus offers a very powerful method to examine the influence of factors and their interactions across groups. Similar to ANOVA, MANOVA tests whether the assignment of objects to levels of one or more nominal explanatory variables (i.e. grouping variables) is statistically supported by response data. However, this response data is contained in multiple continuous response variables rather than a single response variable (ANOVA). MANOVA is, therefore, suitable for testing the effect of different factors (e.g. experimental treatments or sampling site properties) on multiple response variables (e.g. OTU abundances).

Null hypothesis - The (multivariate) vectors of means of two or more groups of objects are equal.

R: manova() - stats package



https://sites.google.com/site/mb3gustame/home/why-multivariate-analysis