Short Communication:

R AND ITS APPLICATIONS IN ECOLOGICAL RESEARCH

Hendra F. Sihaloho

Marine Molecular Genetics Laboratory, LIPI Research Centre for Oceanography
E-mail: hend017@lipi.go.id or hfsihaloho@gmail.com

Received: August 2015 Accepted: December 2015

ABSTRACT

The increase in research activities in recent years has generated a lot of data to be analysed. Research-related communities need powerful software to perform their analyses, which can be a problem, particularly for those who live in developing countries, where the proprietary programs are often unaffordable. R may provide a solution, since it is open source software which can be installed on major operating systems. In addition, it is well maintained by the R Core Team, which ensures that the program and its packages work across platforms. The increasing usage of R, especially in universities, is not only proof that the program can be relied on, but it is also a guarantee that the software will continue developing. R and its capability, particularly for ecological research activities, will be described in this short note.

Keywords: RStudio, statistical analysis, advanced graphics, ecological applications, big data

Research is increasing and improving nowadays but this produces more data to be evaluated. Analysis, however, almost always relies on software, which enhances researchers' ability to compute and evaluate the acquired data. Unfortunately, most of the computer programs are not only proprietary, but also very expensive, which means that some students or early level research scientists cannot afford them. Thus, R (R Core Team, 2015) was introduced for those who want free powerful analytic software. This paper will discuss R and its applications, mainly for ecological research.

R and CRAN

R is a high-level programming language for data analysis and graphics (Crawley, 2007). It is open source, and can be used with mainstream computer operating systems: Windows, Macintosh and Linux. It offers applications to a wide range of users for statistical analysis, ecological studies, genetic and molecular science, prediction and modelling, economics as well social sciences. Furthermore, R also offers the latest statistical analyses, which may not be available on proprietary programs for several years. Those latest statistical analyses are usually available on several packages, which are strictly selected and maintained.

Many hours of practice is required to master R but the expended hours are an investment for future work. Once someone has a basic knowledge of R coding, s/he will understand the more complicated scripts and be able to apply them to her/his research. R also provides powerful graphical analysis, which extends from the basic, such as simple pie-charts or histograms to advanced graphic visualisations, like dendograms or waterfall graphs.

DOI : 10.14203/mri.v40i1.75

R was originally written by Ross Ihaka and Robert Gentleman at the Statistics Department of the University of Auckland - New Zealand (Crawley, 2007; Logan, 2010). It is a non-commercial version of S-plus software. Later on, the R core team was founded, which maintains the software as well as screens its packages on CRAN (Comprehensive R Archive Network), a complete and official reference website of R. CRAN provides the latest version of R and its packages and documentation. It can be accessed on http://cran.r-project.org. There have been several versions of this software since it was launched in 1996, and the latest version, at the time of writing, was R version 3.2.1, released on June 18th, 2015.

The uniqueness of CRAN is that it is placed on many mirrors in several countries. The mirrors are usually placed in universities or research institutions from both the government and private sectors. Recently, there were 101 mirrors in 49 different countries (Supplement 1). The number of servers is increasing as R communities are developing too. For example, in 2010 Indonesia did not have any CRAN repository servers, but it has one now, and perhaps the number will increase in the future. Currently, there are 6786 available packages on this website. The high number of packages increases the possibility of the users finding the package they need. CRAN also introduced Task Views that contain information about applying R to certain research fields (Supplement 2). Task Views lists available packages for a particular research field, and also suggests several books or related publications.

In addition to CRAN, there are also several other websites for particular purposes, such as *Rgeo* (https://geodacenter.asu.edu/projects/rsp) which offers news and guides for spatial analyses data; *biocunductor* (www.biocunductor.org) provides R packages related to genetics analyses; *Robust Statistics* (www.statistik.tuwien.ac.at/rsr) which organizes the R tools for many widely used models in Robust Statistics methods; *Rmetrics* (www.rmetrics.org) which is open source software for quantitative finance. The internet also helps the R users to interact with each other. Moreover, several blogs also facilitate discussion of certain R applications.

R can be run either on its native console version or on its GUI ones (Table 1). The preference for R GUI is normally based on the R users' experiences, and certainly on the GUI's stability. Mainly depending on how they got taught R during their previous education or training, people become familiar with a specific GUI. However, learning the native R console gives people the advantage of understanding the philosophy behind R itself and it will enhance their ability when they employ the scripts on one of GUIs. Besides, in order to implement more advanced analyses, one has to code the scripts or at least, modify others' scripts. In this article, the author used RStudio to run the codes for several R examples, which are compiled in Supplement 4 and hfs appli.R file.

Furthermore, reading and understanding the documentation are important in order to use the package properly, because complete information about the package can only be obtained through the original documentation, which is always included in the R package itself, and sometimes in certain journal articles or books.

R on ecological analyses

As mentioned above, we can easily find the necessary packages by using CRAN Task Views. It can connect R users to others Task Views, which provide related analysis packages. *Environmetrics*, *Multivariate*, *SpatioTemporal* and *Cluster* are very useful Task Views that give a lot of information about ecological analyses. Several routine ecological analyses, which are frequently carried out on ecological data, will be discussed below.

Table 1. Non-extensive list of R GUIs

GUI	Version	Links	Last update
RStudio	0.99.447	http://www.rstudio.com	2015-06-21
Rcmdr	2.1-7	http://socserv.mcmaster.ca/jfox/Misc/Rcmdr	2015-02-22
Tinn-R	4.0.3.4	http://nbcgib.uesc.br/lec/software/editores/tinn-r/en	2015-06-11
RKWard	0.6.3	https://rkward.kde.org	2015-03-07
StatET	3.4.1	http://www.walware.de/goto/statet	2015-03-01
JGR	1.7-16	https://www.rforge.net/JGR	2013-12-21

Simple application

R can be used as scientific calculator by directly typing on the console panel (Supplement 3). However, using a script panel is better and very helpful in order to save codes. Univariate analysis, such as mean, range, standard deviation and others (Sample 1) is easily performed in R. These basic statistical measurements are always performed on biodiversity data, for instance to calculate the number of species or the number of individuals of a species. The diversity indices, such as Shannon-Wiener, Pielou or Simpson may also be calculated on R using *vegan* package (Oksanen et al., 2015), for instance with the dune data set (Jongman et al., 2005) as in Sample 2.

```
# Sample 1
data("co2") # call co2 data set
mean(co2) # calculate the average of CO2
median(co2) # calculate the median
range(co2) # calculate the minimum and
maximum concentration of CO2
sd(co2) # calculate standard deviation
var(co2) # calculate variance
summary(co2)
boxplot(co2) # to create box-whisker plot

# Sample 2
library (vegan)
data(dune) # call dune data set
H <- diversity (dune) # calculate Shannon
Index
Lambda <- diversity(dune, "simpson") #
calculate Simpson Index
J <- H/log(specnumber(dune)) # calculate
Pielou's evenness index
dune.diversity<- data.frame (H,Lambda,J) #
create matrix of indices from previous
objects
```

Multivariate analysis

Ecologists are interested in interaction between biological and environmental data. The increase of environmental parameters attracts them to conduct multivariate analysis. It is also due to the fact that biodiversity phenomenon cannot be explained by a single environmental factor. R packages, such as the *Picante* package (Kembel et al., 2010), which applies *vegan* (Oksanen et al., 2015), *ape* (Paradis, 2012; Paradis et al., 2004), *permute* (Simpson, 2015), *lattice* (Sarkar, 2008) and *nlme* packages (Pinheiro et al., 2007) can perform multivariate analysis, such as Cluster Analysis (Sample 3), Multidimensional Scaling (MDS) (Sample 4) and Principal Component Analysis (PCA) (Sample 5) are examples of multivariate analysis which can be carried out in R (Borcard et al., 2011; Everitt & Hothorn, 2011; Kembel & Cahill Jr, 2011).

```
# Sample 3
data(dune)
dune.matrix<- vegdist (dune, "bray") #
matrix resemblance with
braycurtisdisimilarity
dune.matrix.average<- hclust
(dune.matrix, method = "average") #
cluster with average method
plot (dune.matrix.average) # create
cluster dendogram

# Sample 4
data(dune)
data(dune.env)
dune.bc.mds<- metaMDS(dune, dist =
"bray")
# Assess goodness of ordination fit
(stress plot)
stressplot(dune.bc.mds)
## plot site scores as text
ordiplot(dune.bc.mds, display = "sites",
type = "text")
dune.mds.fig<- ordiplot(dune.bc.mds, type
= "none", xlim=c(-1.1,1.9), ylim=c(-
0.8,1.2)) # no image
points(dune.mds.fig, "sites", pch = 15,
cex=1.2, col = "green", select =
dune.env$Management == "SF")
points(dune.mds.fig, "sites", pch = 16,
cex=1.2, col = "red", select =
dune.env$Management == "BF")
points(dune.mds.fig, "sites", pch = 17,
cex=1.2, col = "black", select =
dune.env$Management == "HF")
points(dune.mds.fig, "sites", pch = 18,
cex=1.2, col = "blue", select =
dune.env$Management == "NM")
legend("topright", legend=c("Management",
"SF","BF","HF","NM"),
pch=c(25,15,16,17,18),cex=1.3,
col=c("white","green","red", "black",
"blue"), bty="n")
```

Ecological modelling

"*A model is a simplified representation of a complex phenomenon*" (Soetaert & Herman, 2009). R is capable of conducting modelling and simulation in spite the fact that R was not originally established for modelling. Several kinds of modelling analyses for fitting ecological data to standard statistical model, such as simple and multiple regression (Sample 7), Generalised Linear Models (GLM), Generalised Additive Models (GAM), mixed effects models and others can be implemented in R (see James et al., 2014; Logan, 2010; Zuur et al., 2009 for details).

```
># Sample 5
data(dune)
data(dune.env)
dune.pca<- rda(dune, scale = TRUE)
summary(dune.pca)
# plotpca
biplot(dune.pca, scaling=3)
## plotpca, site as factor of land-use
biplot(dune.pca, type = c("text",
"points"), scaling=3)
points(dune.pca, "sites", pch = 15, col =
"green", select = dune.env$Use ==
"Hayfield", scaling =3)
points(dune.pca, "sites", pch = 16, col =
"black", select = dune.env$Use ==
"Haypastu", scaling =3)
points(dune.pca, "sites", pch = 17, col =
"blue", select = dune.env$Use ==
"Pasture", scaling =3)
legend("bottomright", legend=c("Land-
Use","Hayfield","Haypastu","Pasture"),pch
= c(25,15,16,17,18), cex = 1.1,
col=c("white","green","black", "blue"),
bty="n")

># Sample 6
# Regression
data(BOD)
attach(BOD)
shapiro.test(Time) # normal
shapiro.test(demand) # normal
cor.test (Time, demand,
method="spearman") # no correlation
reg1 <- lm (demand ~ Time, data = BOD)
summaryreg1)
# checking assumption
e.reg1 <- resid(reg1) # define residuals
for log.Dens~log.Sec
e.reg1[abs(e.reg1) > 3* sd(e.reg1)] #
define the outliers
# plotting
plot(Time, demand, pch=16, col="red")
abline(reg.1, col="blue")
legend("bottomright", legend=c("p-value =
0.05435", "R^2 = 0.5562"))
```

Non-parametric analysis

In the case of data sets that do not fulfil the parametric analysis requirements, R offers non-parametric analysis, for instance the Wilcoxon Signed-Rank test, the Kruskal-Wallis test and multiple comparison (Sample 7). Packages such as *Rfit* and *npsm* (Kloke & McKean, 2012; 2014; 2015) can perform more advanced non-parametric analysis.

A glance of big data with R

The advances in certain fields of research have produced a lot data to be analysed, which also requires high-performance computing (HPC). The *pbdR* (programming with big data in R) (Ostrouchov et al., 2012) is a project that enables analysis of big data on HPC platforms with multi cores, but it still uses the same R programming language.

CONCLUSION

R is very powerful for analysing ecological data, and it will develop in the future. Investment for training including self-training and practice are necessary in order to enhance the speed of analysis and the knowledge of R users. The R GUIs, such as RStudio, are very helpful, however, the native R console offers stability while several R GUIs still are not fully stable.

ACKNOWLEDGEMENT

The author thanks Udhi E. Hernawan and Yaya I. Ulumudin for suggestions on an earlier version of the manuscript.

```
># Sample 7
# Non-Parametric
abalone<-
read.csv(file="http://archive.ics.uci.edu
/ml/machine-learning-
databases/abalone/abalone.data",header =
FALSE)
# "Sex" = S, "Length" = L, "Diameter" =
D, "Height" = H, "Whole weight" = WW,
# "Shucked weight" = SW, "Viscera weight"
= VW, "Shell weight" = SHW, "Rings" = R
names(abalone) <- c("S", "L", "D", "H",
"WW", "SW", "VW", "SHW", "R")
dim(abalone)
attach(abalone)
head (abalone)
plot(L ~ S, data=abalone, col=c("red",
"green", "blue"), xlab= "Sex",
ylab="Length of Shell (mm)")
tapply (L, S, shapiro.test)
# test normality of Length for each sex
resulted F, I & M were not normal
distributed

library(car) #(Fox & Weisberg, 2015)
leveneTest (L, S)
# testhomgenity variance of Length for
each sex showed F, I & M were not
homogeneous

# subsetting
male.l<- subset (abalone, S=="M",
select=c("S", "L"))
female.l<- subset (abalone, S=="F",
select=c("S", "L"))
infant.l<- subset (abalone, S=="I",
select=c("S", "L"))
wilcox.test(male.l$L, female.l$L)
wilcox.test(male.l$L, infant.l$L)
wilcox.test(female.l$L, infant.l$L)

kruskal.test(male.l$L, female.l$L)
# kruskal.test couldn't be performed due
to unequal sample size
dim(male.l); dim(female.l); dim(infant.l)
# check dimension of data
# for instance, select the first 1300
data of each sex for equal sample size
male.l<- male.l[1:1300,]
female.l<- female.l[1:1300,]
infant.l<- infant.l[1:1300,]
kruskal.test(male.l$L, female.l$L)
kruskal.test(male.l$L, infant.l$L)
kruskal.test(female.l$L, infant.l$L)

# UsingRfit&npsm packages
library(Rfit)
library (npsm)
np.length<-
with(abalone,oneway.rfit(L,S))
np.length
summary(np.length)
```

REFERENCES

Borcard, Daniel, Gillet, François, & Legendre, Pierre. (2011). *Numerical Ecology with R*. New York: Springer.

Crawley, Michael J. (2007). *The R Book*. West Susex: John Wiley & Sons Ltd.

Everitt, Brian, & Hothorn, Torsten. (2011). *An Introduction to Applied Multivariate Analysis with R*. New York: Springer.

Fox, J., & Weisberg, S. (2015). car: Companion to Applied Regression. R package version 2.0-25. Retrieved from http://CRAN.R-project.org/package=car

http://nbcgib.uesc.br/lec/software/editores/tinn-r/en accessed on June 23rd, 2015

http://socserv.mcmaster.ca/jfox/Misc/Rcmdr. The R Commander: A Basic-Statistics GUI for R, accessed on June 23rd, 2015.

http://www.rstudio.com. Welcome to Rstudio, accessed on June 23rd, 2015.

http://www.walware.de/goto/statet. StatET for R, accessed on June 23rd, 2015.

https://rkward.kde.org. Welcome to RKWard, accessed on June 23rd, 2015.

https://www.rforge.net/JGR. JGR - Java GUI for R, accessed on June 23rd, 2015.

James, Gareth, Witten, Daniela, Hastie, Trevor, & Tibshirani, Robert. (2014). *An introduction to statistical learning : with applications in R*. New York: Springer .

Jongman, R. H. G., Ter Braak, C. J. F., & Van Tongeren, O. F. R. (2005). *Data analysis in community and landscape ecology* (2 ed.). New York: Cambridge University Press.

Kembel, Steven W., & Cahill Jr, James F. (2011). Independent Evolution of Leaf and Root Traits within and among Temperate Grassland Plant Communities. *PLoS ONE, 6*(6), e19992. doi: 10.1371/journal.pone.0019992.

Kembel, Steven W., Cowan, Peter D., Helmus, Matthew R., Cornwell, William K., Morlon, Helene, Ackerly, David D., Blomberg, Simon P., & Webb, Campbell O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics, 26*(11), 1463-1464. doi: 10.1093/bioinformatics/btq166.

Kloke, John D, & McKean, Joseph W. (2012). Rfit: Rank-based estimation for linear models. *The R Journal, 4*(2), 57-64.

Kloke , John, & McKean, Joseph. (2014). npsm: Package for Nonparametric Statistical Methods using R. R package version 0.5. Retrieved from http://CRAN.R-project.org/package=npsm

Kloke, John, & McKean, Joseph W. (2015). *Nonparametric Statistical Methods Using R*. Boca Raton: CRC Press.

Lichman, M. (2013). *UCI Machine Learning Repository*. Retrieved from: http://archive.ics.uci.edu/ml/datasets/Abalone

Logan, Murray. (2010). *Biostatistical Design and Analysis using R A practical Guide*.

Oksanen, Jari, Blanchet, F. Guillaume, Kindt, Roeland, Legendre, Pierre, Minchin, Peter R., O'Hara, R. B., Simpson, Gavin L., Solymos, Peter, Stevens, M. Henry H., & Wagner, Helene. (2015). vegan: Community Ecology Package. R package version 2.3-0. Retrieved from http://CRAN.R-project.org/package=vegan

Ostrouchov, George , Chen, Wei-Chen, Schmidt, Drew, & Patele, Pragneshkumar. (2012). Programming with Big Data in R. from http://r-pbd.org/

Paradis, Emmanuel. (2012). *Analysis of Phylogenetics and Evolution with R* (2 ed.). New York: Springer.

Paradis, Emmanuel, Claude, Julien, & Strimmer, Korbinian. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics, 20*(2), 289-290. doi: 10.1093/bioinformatics/btg412.

Pinheiro, J, Bates, D, DebRoy, S, Sarkar, D, & R Core Team. (2007). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-121. Retrieved from http://CRAN.R-project.org/package=nlme

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sarkar, Deepayan. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Simpson, Gavin L. (2015). permute: Functions for generating restricted permutations of data. R package version 0.8–4. Retrieved from http://CRAN.R-project.org/package=permute

Soetaert, Karline , & Herman, Peter M. J. (2009). *A Practical Guide to Ecological Modelling*.

Zuur, Alain F., Ieno, Elena N., Walker, Neil J., Saveliev, Anatoly A., & Smith, Graham M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

R and its Applications in the Ecological Research

Hendra F. Sihaloho[1]

[1]Marine Molecular Genetics Laboratory, LIPI Research Centre for Oceanography
Correspondence author: hend017@lipi.go.id or hfsihaloho@gmail.com

Supplement 1. List of CRAN mirrors

Countries	Web addresses	Institutions
0-Cloud	http://cran.rstudio.com/	Rstudio, automatic redirection to servers worldwide
Algeria	http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	http://cran.csiro.au/	CSIRO
	http://cran.ms.unimelb.edu.au/	University of Melbourne
Austria	http://cran.at.r-project.org/	Wirtschaftsuniversitaet Wien
Belgium	http://www.freestatistics.org/cran/	K.U.Leuven Association
Brazil	http://nbcgib.uesc.br/mirrors/cran/	Center for Comp. Biol. at Universidade Estadual de Santa Cruz
	http://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
	http://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
	http://www.vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
	http://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Canada	http://cran.stat.sfu.ca/	Simon Fraser University, Burnaby
	http://mirror.its.dal.ca/cran/	Dalhousie University, Halifax
	http://cran.utstat.utoronto.ca/	University of Toronto
	http://cran.skazkaforyou.com/	iWeb, Montreal
	http://cran.parentingamerica.com/	iWeb, Montreal
Chile	http://dirichlet.mat.puc.cl/	Pontificia Universidad Catolica de Chile, Santiago
China	http://ftp.ctex.org/mirrors/CRAN/	CTEX.ORG
	http://mirror.bjtu.edu.cn/cran/	Beijing Jiaotong University, Beijing
	http://mirrors.opencas.cn/cran/	Chinese Academy of Sciences, Beijing
	http://mirrors.tuna.tsinghua.edu.cn/CRAN/	TUNA Team, Tsinghua University
	http://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
	http://mirrors.xmu.edu.cn/CRAN/	Xiamen University
Colombia	http://www.icesi.edu.co/CRAN/	Icesi University
Czech Republic	http://mirrors.nic.cz/R/	CZ.NIC, Prague
Denmark	http://mirrors.dotsrc.org/cran/	dotsrc.org, Aalborg
Ecuador	http://cran.espol.edu.ec/	Escuela Superior Politecnica del Litoral
El Salvador	http://cran.salud.gob.sv/	Ministry of Health (Ministerio de Salud)
Estonia	http://ftp.eenet.ee/pub/cran/	EENet
France	http://cran.univ-lyon1.fr/	Dept. of Biometry & Evol. Biology, University of Lyon
	http://mirror.ibcp.fr/pub/CRAN/	CNRS IBCP, Lyon
	http://ftp.igh.cnrs.fr/pub/CRAN/	Institut de Genetique Humaine, Montpellier
	http://cran.univ-paris1.fr/	Universite Paris 1 Pantheon-Sorbonne
Germany	http://mirrors.softliste.de/cran/	Softliste.de, Berlin
	http://ftp5.gwdg.de/pub/misc/cran/	GWDG Goettingen

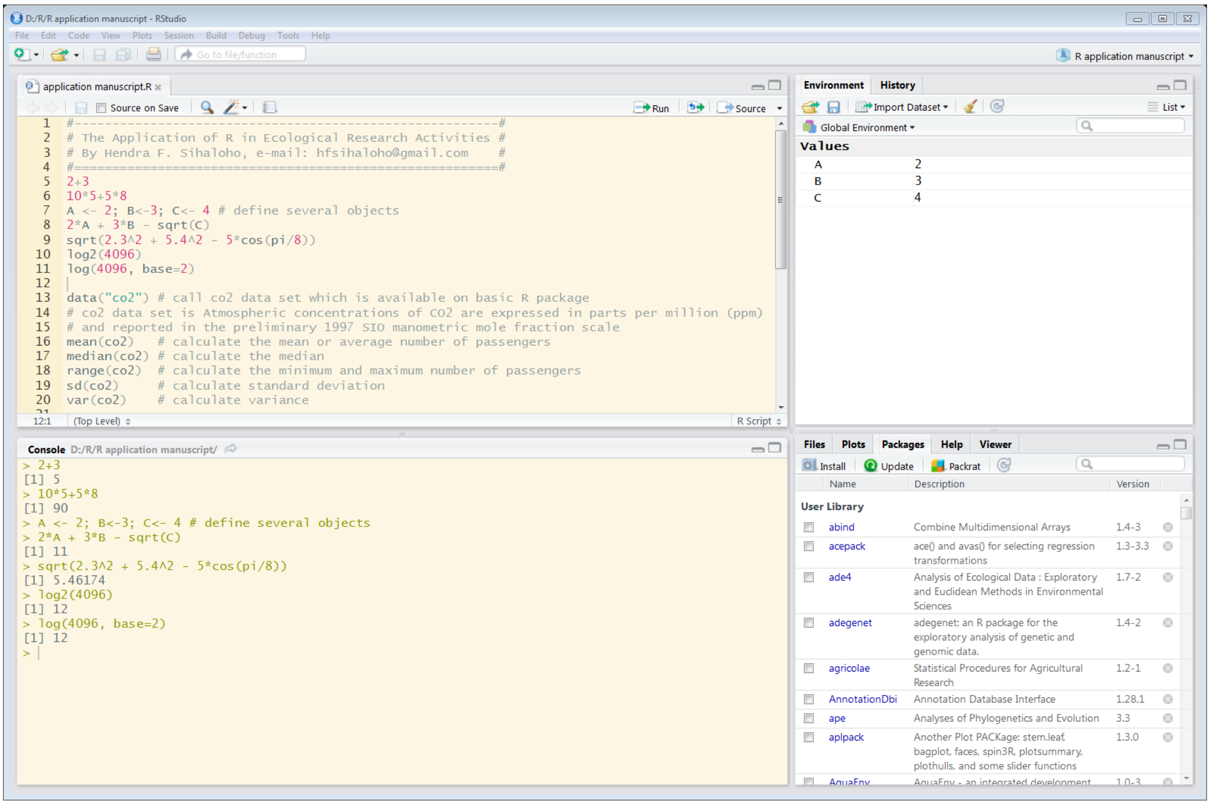
Countries	Web addresses	Institutions
	http://cran.sciserv.eu/	Michael Marz (private)
	http://cran.uni-muenster.de/	University of Münster, Germany
Greece	http://cran.cc.uoc.gr/mirrors/CRAN/	University of Crete
Hungary	http://cran.rapporter.net/	Rapporter.net, Budapest
Iceland	http://cran.hafro.is/	Marine Research Institute
India	http://ftp.iitm.ac.in/cran/	Indian Institute of Technology Madras
Indonesia	http://cran.repo.bppt.go.id/	Agency for The Application and Assessment of Technology
Iran	http://cran.um.ac.ir/	Ferdowsi University of Mashhad
Ireland	http://ftp.heanet.ie/mirrors/cran.r-project.org/	HEAnet, Dublin
Italy	http://cran.mirror.garr.it/mirrors/CRAN/	Garr Mirror, Milano
	http://cran.stat.unipd.it/	University of Padua
	http://dssm.unipa.it/CRAN/	Universita degli Studi di Palermo
Japan	http://cran.ism.ac.jp/	Institute of Statistical Mathematics, Tokyo
	http://ftp.yz.yamagata-u.ac.jp/pub/cran/	Yamagata University
Korea	http://cran.nexr.com/	NexR Corporation, Seoul
	http://healthstat.snu.ac.kr/CRAN/	Graduate School of Public Health, Seoul National University, Seoul
	http://cran.biodisk.org/	The Genome Institute of UNIST (Ulsan National Institute of Science and Technology)
Lebanon	http://rmirror.lau.edu.lb/	Lebanese American University, Byblos
Mexico	http://cran.itam.mx/	Instituto Tecnologico Autonomo de Mexico
	http://www.est.colpos.mx/R-mirror/	Colegio de Postgraduados, Texcoco
Netherlands	http://cran.xl-mirror.nl/	XL-Data, Amsterdam
	http://cran-mirror.cs.uu.nl/	Utrecht University
New Zealand	http://cran.stat.auckland.ac.nz/	University of Auckland
Norway	http://cran.uib.no/	University of Bergen
Philippines	http://cran.stat.upd.edu.ph/	University of the Philippines and PREGINET
Poland	http://r.meteo.uni.wroc.pl/	University of Wroclaw
Portugal	http://cran.dcc.fc.up.pt/	University of Porto
Russia	http://cran.gis-lab.info/	GIS-Lab.info
	http://www.go-parts.com/mirrors-ru/cran/	Go-Parts
Slovakia	http://cran.fyxm.net/	FYXM.net, Bratislava
South Africa	http://cran.mirror.ac.za/	TENET, Johannesburg
Spain	http://ftp.cixug.es/CRAN/	Oficina de software libre (CIXUG)
	http://cran.es.r-project.org/	Spanish National Research Network, Madrid
Sweden	http://ftp.acc.umu.se/mirror/CRAN/	Academic Computer Club, Umeå University
Switzerland	http://stat.ethz.ch/CRAN/	ETH Zuerich
Taiwan	http://ftp.yzu.edu.tw/CRAN/	Department of Computer Science and Engineering, Yuan Ze University
	http://cran.csie.ntu.edu.tw/	National Taiwan University, Taipei
Thailand	http://mirrors.psu.ac.th/pub/cran/	Prince of Songkla University, Hatyai
Turkey	http://cran.pau.edu.tr/	Pamukkale University, Denizli
UK	http://www.stats.bris.ac.uk/R/	University of Bristol

Countries	Web addresses	Institutions
	http://mirrors.ebi.ac.uk/CRAN/	EMBL-EBI (European Bioinformatics Institute)
	http://mirrors-uk2.go-parts.com/cran/	Go-Parts
	http://cran.ma.imperial.ac.uk/	Imperial College London
	http://mirror.mdx.ac.uk/R/	Middlesex University London
	http://star-www.st-andrews.ac.uk/cran/	St Andrews University
USA	http://cran.cnr.Berkeley.edu/	University of California, Berkeley, CA
	http://cran.stat.ucla.edu/	University of California, Los Angeles, CA
	http://mirror.las.iastate.edu/CRAN/	Iowa State University, Ames, IA
	http://ftp.ussg.iu.edu/CRAN/	Indiana University
	http://rweb.quant.ku.edu/cran/	University of Kansas, Lawrence, KS
	http://watson.nci.nih.gov/cran_mirror/	National Cancer Institute, Bethesda, MD
	http://cran.mtu.edu/	Michigan Technological University, Houghton, MI
	http://www.go-parts.com/mirrors-usa/cran/	Go-Parts
	http://cran.wustl.edu/	Washington University, St. Louis, MO
	http://cran.case.edu/	Case Western Reserve University, Cleveland, OH
	http://iis.stat.wright.edu/CRAN/	Wright State University, Dayton, OH
	http://ftp.osuosl.org/pub/cran/	Oregon State University
	http://lib.stat.cmu.edu/R/CRAN/	Statlib, Carnegie Mellon University, Pittsburgh, PA
	http://cran.mirrors.hoobly.com/	Hoobly Classifieds, Pittsburgh, PA
	http://mirrors.nics.utk.edu/cran/	National Institute for Computational Sciences, Oak Ridge, TN
	http://cran.revolutionanalytics.com/	Revolution Analytics, Dallas, TX
	http://cran.fhcrc.org/	Fred Hutchinson Cancer Research Center, Seattle, WA
	http://cran.cs.wwu.edu/	Western Washington University, Bellingham, WA
Venezuela	http://camoruco.ing.uc.edu.ve/cran/	Universidad de Carabobo Venezuela
Vietnam	http://cran.vinastat.com/	VinaStat.com

Supplement 2. List of CRAN Task Views and its related research field

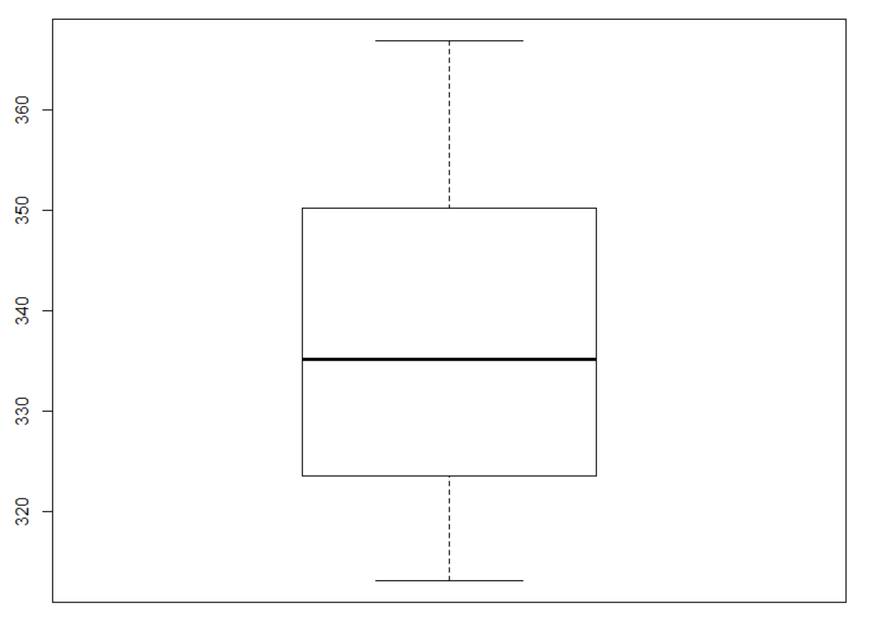
Task Views	Research Fields
Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

Supplement 3. The screenshot of RStudio window which has 4 main panels: Script on top left, Console on bottom left, Environment and History on the upper right, while the lower right panel shows files, plots, packages, help and viewer tabs



Supplement 4. Results of hfs_script.R

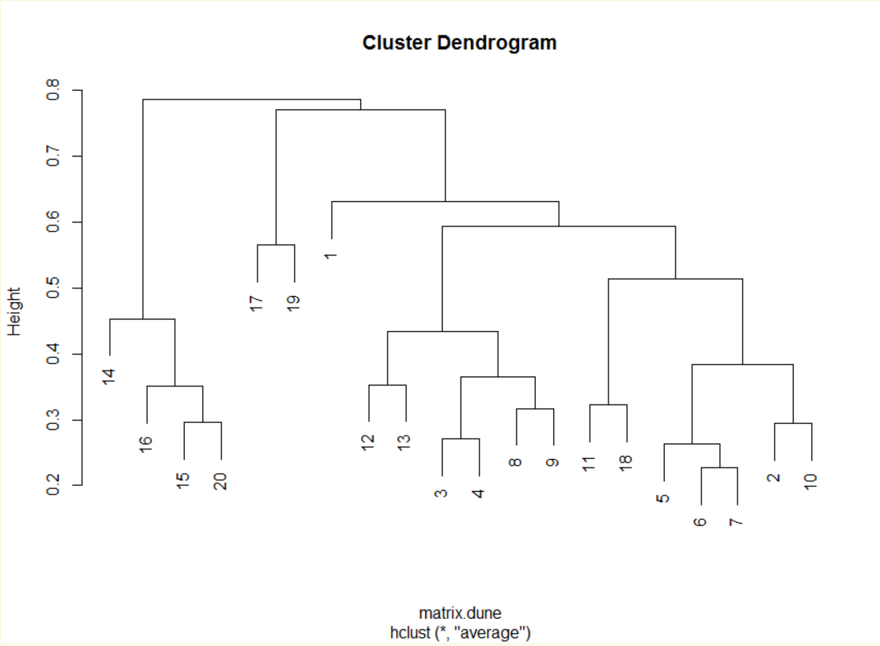
```
> # Sample 1
> # co2 data set is Atmospheric concentrations of CO2 are expressed in parts per mi
llion (ppm)
> # and reported in the preliminary 1997 SIO manometric mole fraction scale
> data("co2") # call co2 data set
> mean(co2) # calculate the average of CO2
[1] 337.0535
> median(co2) # calculate the median
[1] 335.17
> range(co2) # calculate the minimum and maximum concentration of CO2
[1] 313.18 366.84
> sd(co2) # calculate standard deviation
[1] 14.96622
> var(co2) # calculate variance
[1] 223.9877
> summary(co2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  313.2   323.5   335.2   337.1   350.3   366.8 
> boxplot(co2) # to draw the data on box-whisker plot
```



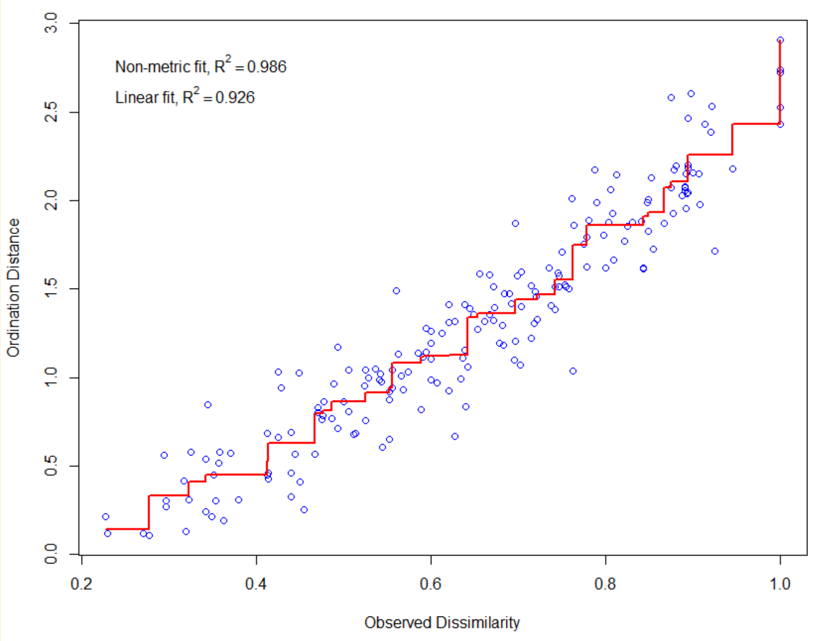

```
> # Sample 2
> library (vegan)
Loading required package: permute
Loading required package: lattice
This is vegan 2.3-0
> # The dune meadow vegetation data, dune, has cover class values of 30 species on
20 sites
> data(dune) # call dune data set
> H <- diversity (dune) # calculate Shannon Index
> Lambda <- diversity(dune, "simpson") # calculate Simpson Index
> J <- H/log(specnumber(dune)) # calculate Pielou's evenness index
> dune.diversity <- data.frame (H,Lambda,J) # create matrix of indices from previou
s objects
> dune.diversity
          H    Lambda         J
1  1.440482 0.7345679 0.8950216
2  2.252516 0.8900227 0.9782554
3  2.193749 0.8787500 0.9527332
4  2.426779 0.9007407 0.9461313
5  2.544421 0.9140076 0.9641403
6  2.345946 0.9001736 0.9783356
```

```
7  2.471733 0.9075000 0.9636577
8  2.434898 0.9087500 0.9798749
9  2.493568 0.9115646 0.9721705
10 2.398613 0.9031909 0.9652727
11 2.106065 0.8671875 0.9585114
12 2.114495 0.8685714 0.9623483
13 2.099638 0.8521579 0.9118610
14 1.863680 0.8333333 0.9577421
15 1.979309 0.8506616 0.9518463
16 1.959795 0.8429752 0.9424621
17 1.876274 0.8355556 0.9642139
18 2.079387 0.8614540 0.9463699
19 2.134024 0.8740895 0.9712362
20 2.048270 0.8678460 0.9850099
```

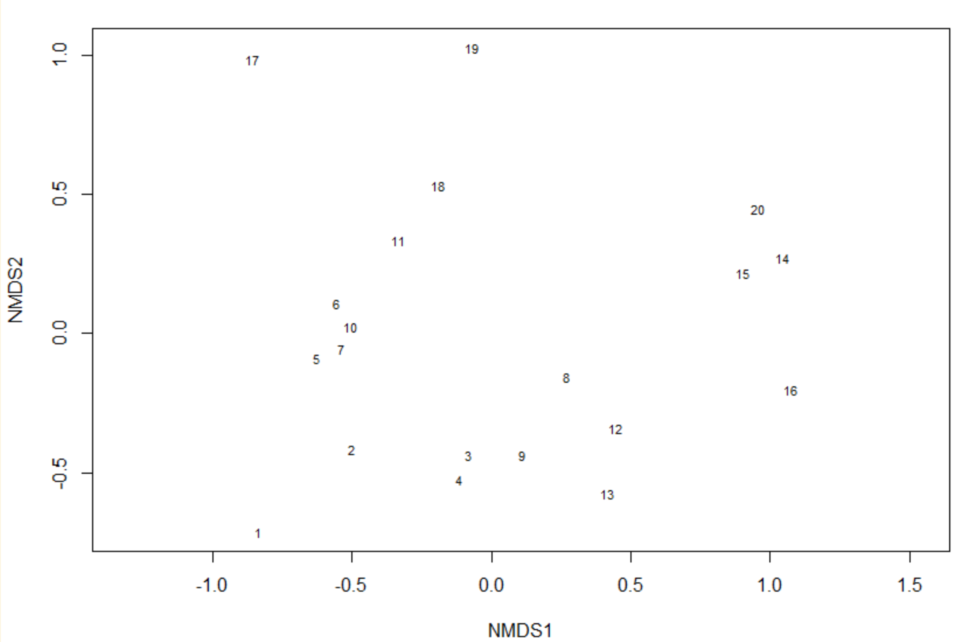
```
> # Sample 3
> # Dendogram
> data(dune)
> dune.matrix <- vegdist (dune, "bray") # matrix resemblance with braycurtis disimi
larity
> dune.matrix.average <- hclust (dune.matrix, method = "average") # cluster with av
erage method
> plot (dune.matrix.average) # create cluster dendogram
```



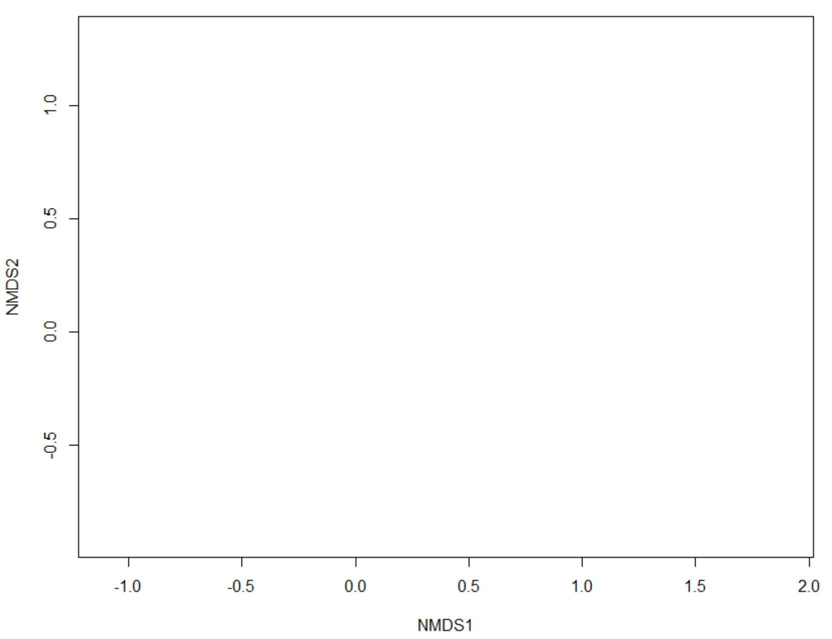

```
> # Sample 4
> # MDS
> # The metaMDS function automatically transforms data and checks solution robustne
s
> data(dune)
> data(dune.env)
> dune.bc.mds <- metaMDS(dune, dist = "bray")
Run 0 stress 0.1192678
Run 1 stress 0.1192682
... procrustes: rmse 0.0003200494  max resid 0.0009810563
*** Solution reached
> # Assess goodness of ordination fit (stress plot)
> stressplot(dune.bc.mds)
> # plot site scores as text
```



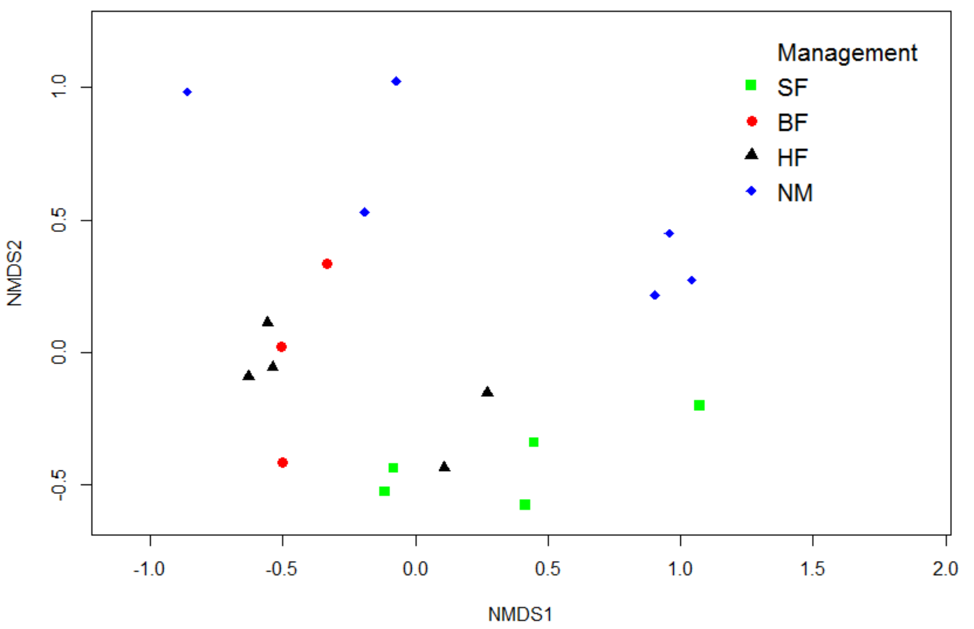

```
> ordiplot(dune.bc.mds, display = "sites", type = "text")
```




```
>
> ## plot site as management of the dune
> dune.mds.fig <- ordiplot(dune.bc.mds, type = "none", xlim=c(-1.1,1.9), ylim=c(-0.
8,1.2)) # no image
```



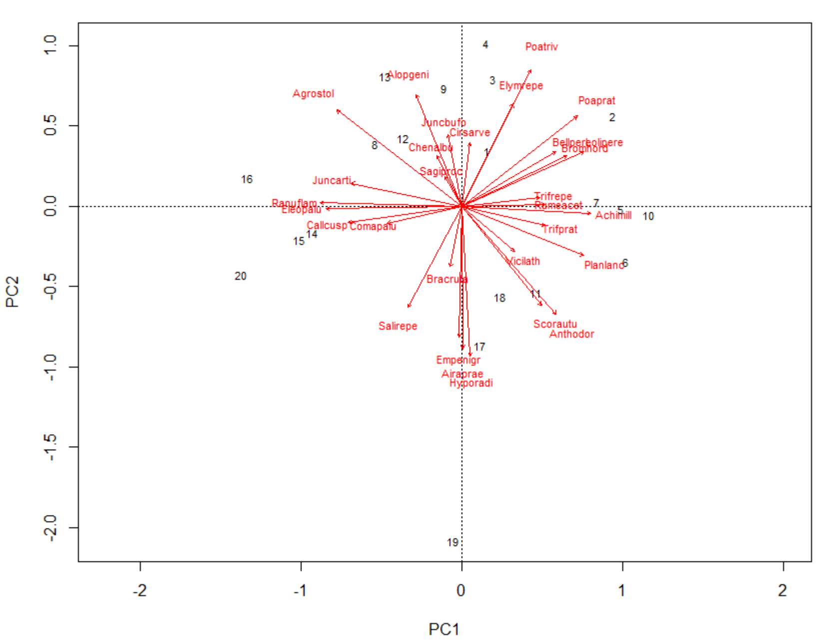

```
> points(dune.mds.fig, "sites", pch = 15, cex=1.2, col = "green", select = dune.env
$Management == "SF")
> points(dune.mds.fig, "sites", pch = 16, cex=1.2, col = "red", select = dune.env$M
anagement == "BF")
> points(dune.mds.fig, "sites", pch = 17, cex=1.2, col = "black", select = dune.env
$Management == "HF")
> points(dune.mds.fig, "sites", pch = 18, cex=1.2, col = "blue", select = dune.env$
Management == "NM")
> legend("topright", legend=c("Management","SF","BF","HF","NM"),
+        pch=c(25,15,16,17,18),cex=1.3, col=c("white","green","red", "black", "blue
"), bty="n")
```



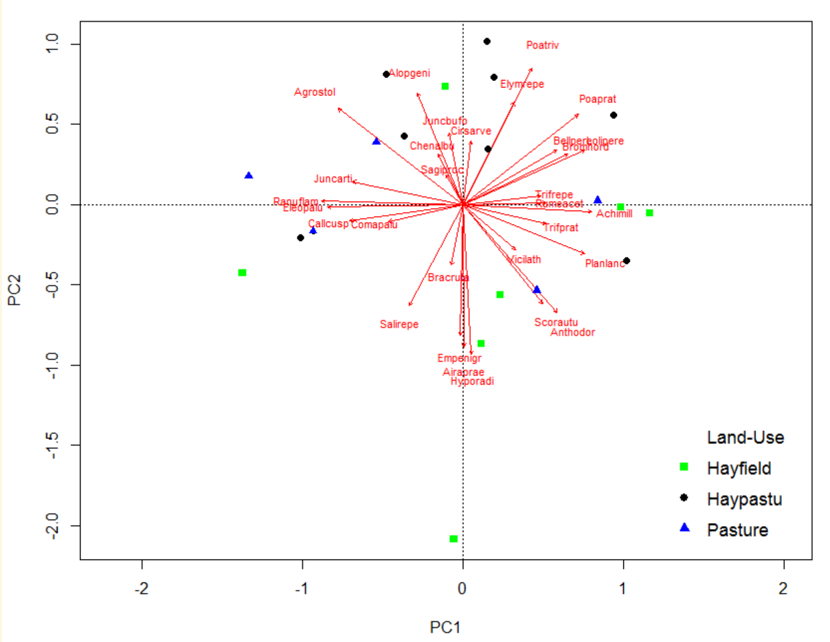

```
> # Sample 5
> # PCA
> data(dune)
> data(dune.env)
> dune.pca <- rda(dune, scale = TRUE)
> summary(dune.pca)
```

```
Call:
rda(X = dune, scale = TRUE)

Partitioning of correlations:
              Inertia Proportion
Total              30          1
Unconstrained      30          1

Eigenvalues, and their contribution to the correlations

Importance of components:
                         PC1    PC2    PC3     PC4    PC5     PC6     PC7     PC8
PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17   PC18    PC19
Eigenvalue            7.0324 4.9973 3.5548 2.64405 2.1389 1.75781 1.47834 1.31640 1
.10787 0.80898 0.74527 0.69660 0.57488 0.35796 0.22253 0.21974 0.15071 0.1319 0.063
50
Proportion Explained  0.2344 0.1666 0.1185 0.08813 0.0713 0.05859 0.04928 0.04388 0
.03693 0.02697 0.02484 0.02322 0.01916 0.01193 0.00742 0.00732 0.00502 0.0044 0.002
12
Cumulative Proportion 0.2344 0.4010 0.5195 0.60762 0.6789 0.73751 0.78679 0.83067 0
.86760 0.89456 0.91941 0.94263 0.96179 0.97372 0.98114 0.98846 0.99349 0.9979 1.000
00

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
* General scaling constant of scores:  4.886172

Species scores
              PC1       PC2       PC3       PC4      PC5       PC6
Achimill  0.65615 -0.031721 -0.288522 -0.071330 -0.38877 -0.021586
Agrostol -0.63980  0.449378  0.126398 -0.033358  0.03338  0.220887
Airaprae  0.00383 -0.666918  0.461324  0.081841 -0.27604  0.114044
Alopgeni -0.23508  0.523916  0.341214  0.266377  0.10359 -0.125309
Anthodor  0.47855 -0.504530 -0.042095  0.216161 -0.33567  0.184658
Bellpere  0.47686  0.254417  0.065363 -0.467748 -0.09334  0.101603
Bromhord  0.53293  0.236962 -0.005037 -0.364163 -0.29092  0.117422
Chenalbu -0.12921  0.237490  0.227366  0.337179 -0.15836 -0.419720
Cirsarve  0.03939  0.297900  0.325356 -0.274435  0.07615  0.494720
Comapalu -0.38201 -0.079690 -0.314566 -0.153780 -0.25028 -0.166589
Eleopalu -0.69457 -0.012911 -0.382956 -0.137600 -0.06047  0.101890
Elymrepe  0.25958  0.479231  0.224739 -0.135805 -0.10774  0.310247
Empenigr -0.01570 -0.611852  0.490145  0.060573 -0.11700  0.110506
Hyporadi  0.04145 -0.703508  0.484633  0.006194 -0.04594 -0.005173
Juncarti -0.56496  0.108411 -0.175409  0.007844  0.15665  0.182690
Juncbufo -0.07304  0.336593  0.270540  0.572440  0.07798 -0.237750
Lolipere  0.61693  0.253253 -0.089528 -0.251018  0.25974  0.022426
Planlanc  0.61659 -0.231282 -0.424325  0.220501  0.16319  0.052640
Poaprat   0.58572  0.421592  0.077315 -0.214655  0.24276 -0.040105
Poatriv   0.34924  0.639286  0.166348  0.283144 -0.18103 -0.033875
Ranuflam -0.72448  0.015720 -0.284143 -0.041979 -0.07417 -0.050091
Rumeacet  0.41931  0.008718 -0.345174  0.590333  0.07281  0.291534
Sagiproc -0.08780  0.139706  0.676217  0.120676  0.24558  0.137039
Salirepe -0.27415 -0.474527  0.058351 -0.146213  0.21067  0.140351
Scorautu  0.40433 -0.462826  0.228004 -0.156482  0.21035 -0.205059
Trifprat  0.42518 -0.091054 -0.438682  0.477095  0.08450  0.275019
Trifrepe  0.39602  0.039044 -0.190829 -0.032309 -0.10496 -0.381565
Vicilath  0.26587 -0.211444 -0.048313 -0.345497  0.50668 -0.416800
Bracruta -0.06077 -0.283080 -0.237426  0.131228  0.66972  0.189213
Callcusp -0.57940 -0.075740 -0.373983 -0.164042 -0.18122  0.021965


Site scores (weighted sums of species scores)
        PC1      PC2      PC3      PC4      PC5        PC6
1   0.22304  0.54162  0.04175 -0.53653 -0.39474  2.381e-01
2   1.34700  0.87580  0.23831 -1.58267 -1.42302 -6.496e-01
```

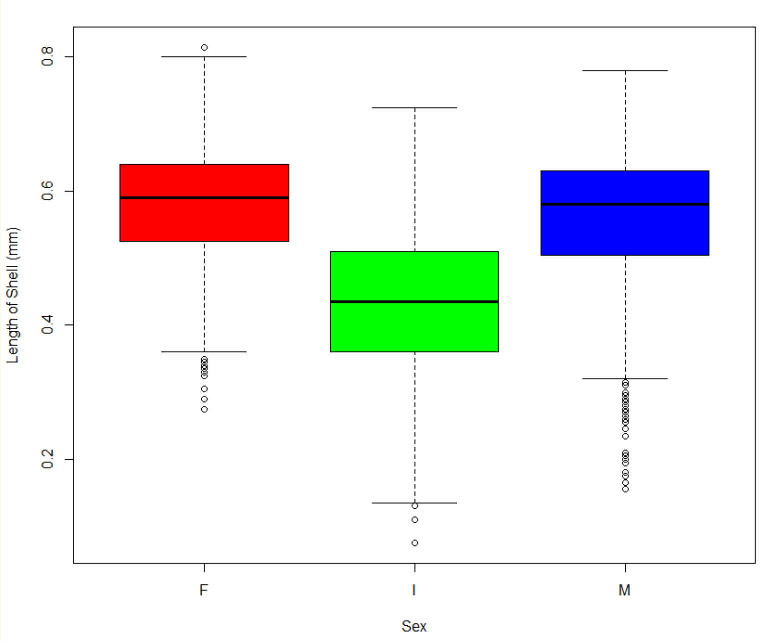
```
3    0.27213  1.24009  0.61014 -0.49098  0.25415 -1.796e-03
4    0.21028  1.59035  1.73693 -1.46508  0.40655  2.641e+00
5    1.41354 -0.02314 -0.90200  0.83760 -0.93543  1.329e+00
6    1.45857 -0.54292 -1.88699  2.06801  0.89204  1.037e+00
7    1.20519  0.03873 -0.84458  1.02246 -0.04956  1.315e-01
8   -0.78210  0.60749  0.09209 -0.10130  0.68298 -6.133e-02
9   -0.16230  1.14986  0.70997  1.03446  0.50332  1.967e-01
10   1.66958 -0.08276 -0.71009 -1.31231 -0.83698 -1.073e+00
11   0.65855 -0.83922  0.19382 -1.00039  2.53899 -1.737e+00
12  -0.52762  0.66633  1.04159  1.84314  0.80106 -5.501e-01
13  -0.68977  1.26785  1.21380  1.80005 -0.84543 -2.241e+00
14  -1.34682 -0.26055 -1.28433 -0.76651 -1.66391 -1.224e+00
15  -1.46036 -0.32505 -1.02725 -0.36354 -0.17529 -6.804e-05
16  -1.92513  0.27484 -0.95166 -0.02451 -0.08830  8.722e-01
17   0.16164 -1.36109  0.40568  0.28323 -1.65437  1.857e-01
18   0.33407 -0.87974 -0.28097 -1.00217  2.08747 -6.589e-01
19  -0.08379 -3.26639  2.61666  0.32337 -0.62462  5.899e-01
20  -1.97570 -0.67211 -1.01287 -0.56633  0.52511  9.759e-01

> biplot(dune.pca, scaling=3)
```




```
> ## plot pca, site as factor of land-use
> biplot(dune.pca, type = c("text", "points"), scaling=3)
> points(dune.pca, "sites", pch = 15, col = "green", select = dune.env$Use == "Hayf
ield", scaling =3)
> points(dune.pca, "sites", pch = 16, col = "black", select = dune.env$Use == "Hayp
astu", scaling =3)
> points(dune.pca, "sites", pch = 17, col = "blue", select = dune.env$Use == "Pastu
re", scaling =3)
> legend("bottomright", legend=c("Land-Use","Hayfield","Haypastu","Pasture"), pch =
c(25,15,16,17,18), cex=1.1,
+         col=c("white","green","black", "blue"), bty="n")
>
```




```
> # Sample 6
> # Non-Parametric
> abalone <- read.csv(file="http://archive.ics.uci.edu/ml/machine-learning-database
s/abalone/abalone.data",
+                      header = FALSE)
> # "Sex" = S, "Length" = L, "Diameter" = D, "Height" = H, "Whole weight" = WW,
> # "Shucked weight" = SW, "Viscera weight" = VW, "Shell weight" = SHW, "Rings" = R
> names(abalone) <- c("S", "L", "D", "H", "WW", "SW", "VW", "SHW", "R")
> write.csv(abalone, "abalone.csv", row.names=TRUE)
> dim(abalone)
[1] 4177    9
> head(abalone)
  S     L     D     H     WW     SW     VW   SHW  R
1 M 0.455 0.365 0.095 0.5140 0.2245 0.1010 0.150 15
2 M 0.350 0.265 0.090 0.2255 0.0995 0.0485 0.070  7
3 F 0.530 0.420 0.135 0.6770 0.2565 0.1415 0.210  9
4 M 0.440 0.365 0.125 0.5160 0.2155 0.1140 0.155 10
5 I 0.330 0.255 0.080 0.2050 0.0895 0.0395 0.055  7
6 I 0.425 0.300 0.095 0.3515 0.1410 0.0775 0.120  8
> attach(abalone)
The following object is masked _by_ .GlobalEnv:
    H
> plot(L ~ S, data=abalone, col=c("red", "green", "blue"), xlab= "Sex", ylab="Lengt
h of Shell (mm)")
```

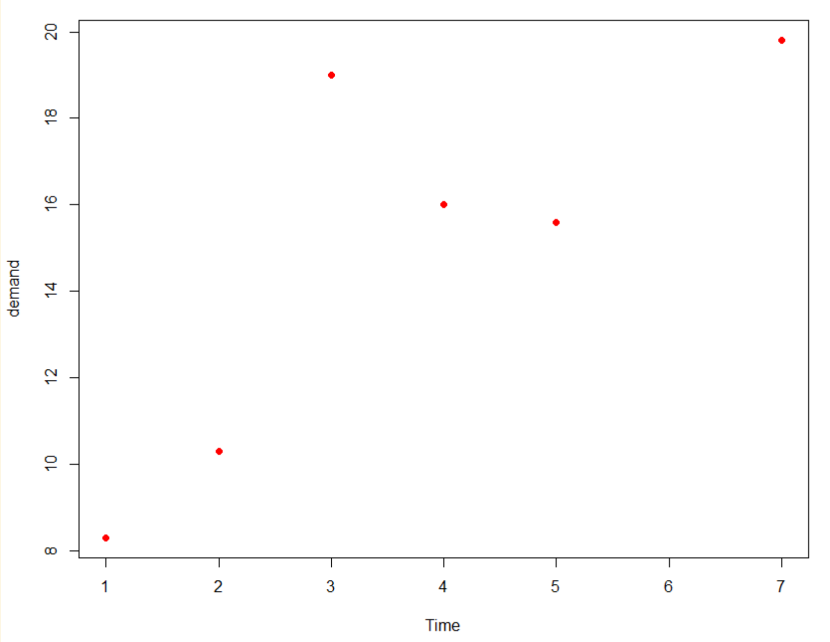



```
> tapply (L, S, shapiro.test)
$F
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.98054, p-value = 2.841e-12
$I
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.98821, p-value = 6.02e-09
$M
        Shapiro-Wilk normality test
data:  X[[i]]
W = 0.95274, p-value < 2.2e-16
> # test normality of Length for each sex resulted F, I & M were not normal distrib
uted

> library(car)
> leveneTest (L, S)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  32.078 1.496e-14 ***
      4174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # test the variance homogenity of Length for each sex showed F, I & M were not ho
mogeneous
> # subsetting
> male.l <- subset (abalone, S=="M", select=c("S", "L"))
> female.l <- subset (abalone, S=="F", select=c("S", "L"))
> infant.l <- subset (abalone, S=="I", select=c("S", "L"))

> wilcox.test(male.l$L, female.l$L)
        Wilcoxon rank sum test with continuity correction
data:  male.l$L and female.l$L
W = 915030, p-value = 0.0001205
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(male.l$L, infant.l$L)
        Wilcoxon rank sum test with continuity correction
data:  male.l$L and infant.l$L
W = 1688500, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

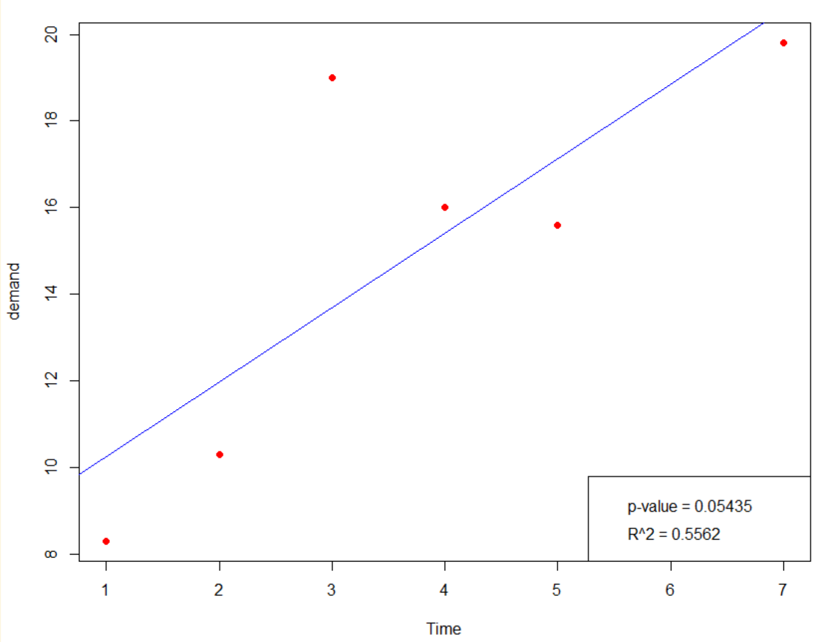
```
> wilcox.test(female.l$L, infant.l$L)
        Wilcoxon rank sum test with continuity correction
data:  female.l$L and infant.l$L
W = 1514400, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> kruskal.test(male.l$L, female.l$L)
Error in kruskal.test.default(male.l$L, female.l$L) :
  'x' and 'g' must have the same length
> # kruskal.test can be performed due to unequal sample size
> dim(male.l); dim(female.l); dim(infant.l) # check dimension of data
[1] 1528    2
[1] 1307    2
[1] 1342    2
 > # for instance, select the first 1300 data of each sex to create an equal sample
size
> male.l <- male.l[1:1300,]
> female.l <- female.l[1:1300,]
> infant.l <- infant.l[1:1300,]

> kruskal.test(male.l$L, female.l$L)
        Kruskal-Wallis rank sum test
data:  male.l$L and female.l$L
Kruskal-Wallis chi-squared = 127.8, df = 90, p-value = 0.005446

> kruskal.test(male.l$L, infant.l$L)
        Kruskal-Wallis rank sum test
data:  male.l$L and infant.l$L
Kruskal-Wallis chi-squared = 134.88, df = 110, p-value = 0.05372

> kruskal.test(female.l$L, infant.l$L)
        Kruskal-Wallis rank sum test
data:  female.l$L and infant.l$L
Kruskal-Wallis chi-squared = 187.72, df = 110, p-value = 5.575e-06

> library(Rfit)
Loading required package: quantreg
Loading required package: SparseM
Attaching package: ‘SparseM’
The following object is masked from ‘package:base’:
    backsolve
Attaching package: ‘Rfit’
The following object is masked from ‘package:car’:
    subsets
> library (npsm)
> np.length <- with(abalone,oneway.rfit(L,S))
> np.length
Call:
oneway.rfit(y = L, g = S)

Overall Test of All Locations Equal
Drop in Dispersion Test
F-Statistic     p-value
     778.71        0.00

        Pairwise comparisons using Rfit
data:  L and S
  F I
M - -
F 1 -
I 1 1

P value adjustment method: none
> summary(np.length)

Multiple Comparisons
```

```
Method Used  none
  I J Estimate  St Err Lower Bound CI Upper Bound CI
1 M F   -0.150 0.00418       -0.15819       -0.14181
2 M I   -0.015 0.00405       -0.02294       -0.00706
3 F I   -0.135 0.00402       -0.14289       -0.12711
>
```

```
> # Sample 7 Regression
> data(BOD)
> attach(BOD)
> shapiro.test(Time) # normal
        Shapiro-Wilk normality test
data:  Time
W = 0.98259, p-value = 0.9637

> shapiro.test(demand) # normal
        Shapiro-Wilk normality test
data:  demand
W = 0.90893, p-value = 0.4294

> cor.test (Time, demand , method="spearman") # no correlation
        Spearman's rank correlation rho
data:  Time and demand
S = 8, p-value = 0.1028
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7714286

> reg1 <- lm (demand ~ Time, data = BOD)
> summary(reg1)
Call:
lm(formula = demand ~ Time, data = BOD)
Residuals:
      1       2       3       4       5       6
-1.9429 -1.6643  5.3143  0.5929 -1.5286 -0.7714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5214     2.6589   3.205   0.0328 *
Time          1.7214     0.6387   2.695   0.0544 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.085 on 4 degrees of freedom
Multiple R-squared:  0.6449,     Adjusted R-squared:  0.5562
F-statistic: 7.265 on 1 and 4 DF,  p-value: 0.05435

> # checking assumption
> e.reg1 <- resid(reg1) # define residuals for log.Dens~log.Sec
> e.reg1[abs(e.reg1) > 3* sd(e.reg1)] # define the outliers
named numeric(0)
> # plotting
> plot(Time, demand, pch=16, col="red")
```




```
abline(reg1, col="blue")
legend("bottomright", legend=c("p-value = 0.05435", "R^2 = 0.5562"))
```




```
> #========================================= End ========================================================================#
```

References

1. co2 {datasets}: Mauna Loa Atmospheric CO2 Concentration
Description: Atmospheric concentrations of CO2 are expressed in parts per million (ppm) and reported in the preliminary 1997 SIO manometric mole fraction scale. Source: Keeling, C. D. and Whorf, T. P., Scripps Institution of Oceanography (SIO), University of California, La Jolla, California USA 92093-0220. ftp://cdiac.esd.ornl.gov/pub/maunaloa-co2/maunaloa.co2.

2. dune {vegan} : Vegetation and Environment in Dutch Dune Meadows.
Description: The dune meadow vegetation data, dune, has cover class values of 30 species on 20 sites.
The corresponding environmental data frame dune.env has a data frame of 20 observations on the following 5 variables: A1 (a numeric vector of thickness of soil A1 horizon.), Moisture (an ordered factor with levels: 1 < 2 < 4 < 5), Management (a factor with levels: BF (Biological farming), HF (Hobby farming), NM (Nature Conservation Management), and SF (Standard Farming)), Use (an ordered factor of land-use with levels: Hayfield < Haypastu < Pasture), Manure (an ordered factor with levels: 0 < 1 < 2 < 3 < 4). Source: Jongman, R.H.G, ter Braak, C.J.F & van Tongeren, O.F.R. (1987). Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen.

3. abalone:
Description: Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. Source: Lichman, M. (2013). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Abalone

4. BOD {datasets}: Biochemical Oxygen Demand
Description: The BOD data frame has 6 rows and 2 columns giving the biochemical oxygen demand versus time in an evaluation of water quality. Source: Bates, D.M. and Watts, D.G. (1988), Nonlinear Regression Analysis and Its Applications, Wiley, Appendix A1.4. Originally from Marske (1967), Biochemical Oxygen Demand Data Interpretation Using Sum of Squares Surface M.Sc. Thesis, University of Wisconsin – Madison.