Verso l'Open Science in ISMAR

Giovedì 23 ottobre 2014

Riproducibilità nella ricerca
e nelle applicazioni ecologiche

Marco Sigovini

CNR-ISMAR - Venezia

OPEN ACCESS

OPEN SOURCE

FREE SOFTWARE

OPEN DATA

OPEN SCIENCE

OPEN STANDARDS

REPRODUCIBLE RESEARCH

Open Knowledge Foundation

OPEN KNOWLEDGE

FREE SOFTWARE FOUNDATION

FOAS

foss

open source initiative

OPEN ACCESS

OPEN SOURCE

FREE SOFTWARE

OPEN DATA

OPEN STANDARDS

REPRODUCIBLE RESEARCH
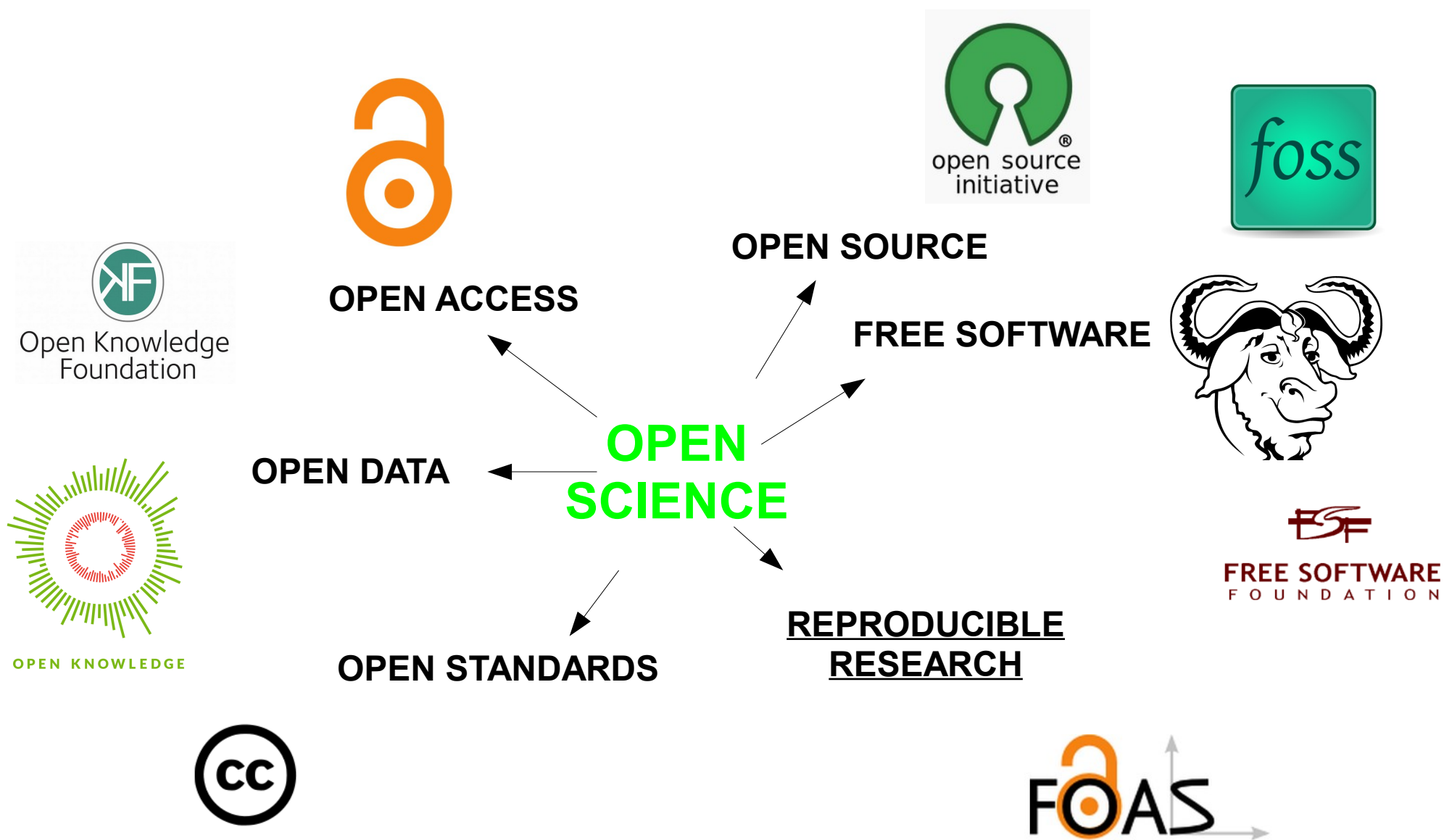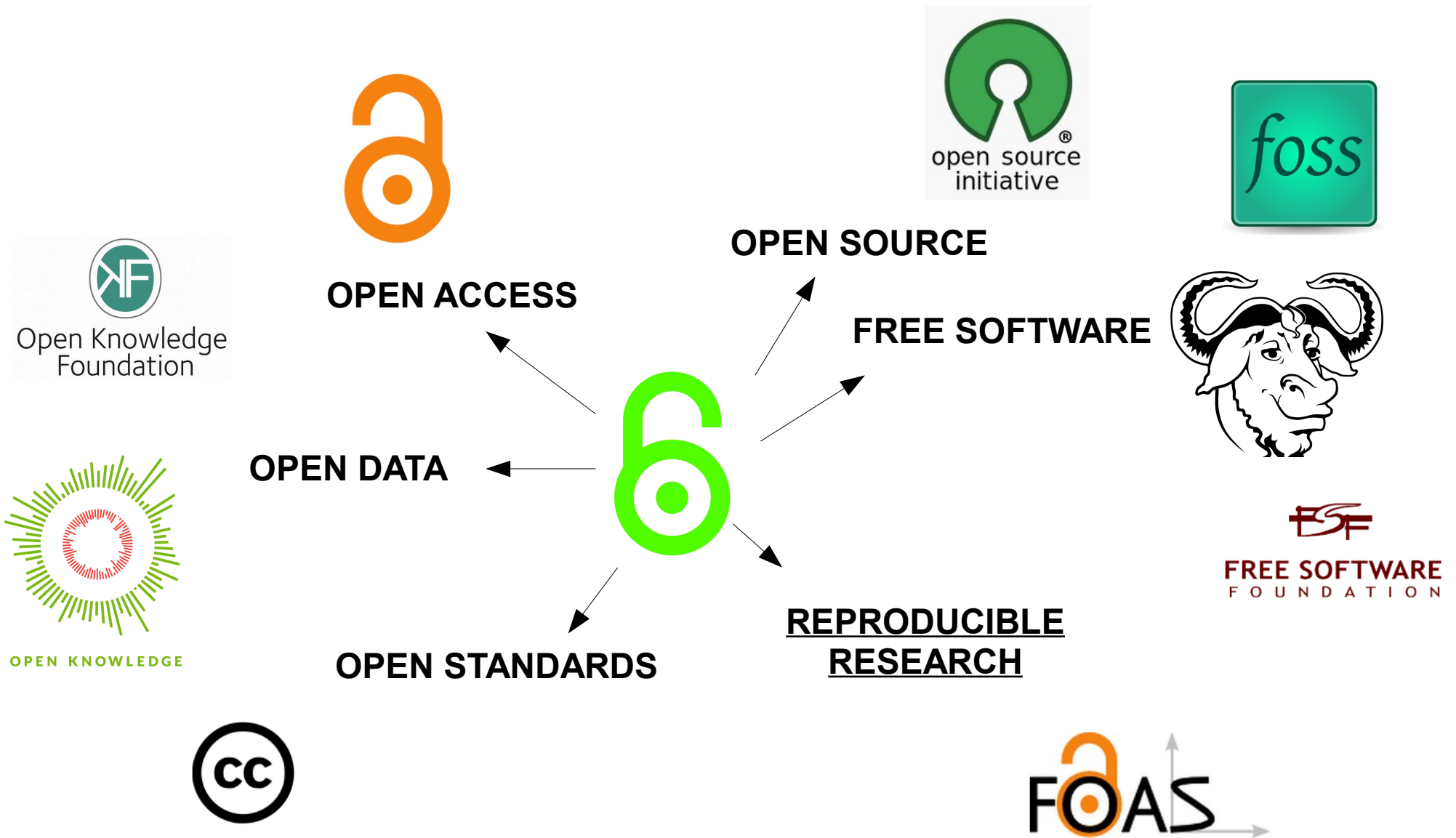
REPRODUCIBILITY

Reproducibility is a main tenet of the scientific method.

It can be defined as "*the extent to which consistent results are obtained when produced repeatedly*" (Oxford English Dictionary). It is the ability of an experiment or study to be reproduced, either by the researcher or by a third party working independently.

Reproducibility should involve all steps in experimental studies, including the ability to recompute analysis on known data sets.

Ecology is characterized by high complexity and context-specific contingency, and it is often based on field observations and studies. Therefore, it is inherently harder to reproduce. Moreover, by aiming at maximizing reproducibility of ecological studies, relevance and generalisability of results is reduced. However, clearly, full reproducibility is expected for the analysis of raw data.

Reproducibility is here focused to the phase of data analysis.

Marco Sigovini, 23/10/2014

REPRODUCIBLE RESEARCH

Reproducible research (computational science): *The ability to repeat the calculations for analyzing the data and obtaining the computational results* (Mesirov, 2010).

Big data sets and complex mathematical approaches (including inherently stochastic methods) reduce reproducibility even when data are provided and methods reported!

Gilbert et al. (2012) analyzed 23 papers with available data sets about genetic structure (software STRUCTURE): <u>results were not reproduced in 30% of them</u>.

In practice, RR is the <u>paper</u> along with the <u>full computational environment</u> used to perform analysis, including the <u>code</u>, <u>data</u>, etc. that can be used to reproduce the results. When a new algorithm or computational approach is proposed, the code should be fully available.

Both raw data and code need to be properly licensed (public copyright).

REPRODUCIBLE RESEARCH

In Ecology:

• methods and algorithms for ecological indices in <u>bioindication and biomonitoring</u>

• <u>multivariate data analysis</u> in community studies (many similar methods with confused nomenclature, different existing version of the algorithms, a number of parameters, complex plots, …)

• <u>time series analysis</u>

• <u>spatial analysis</u>

• ...

REPRODUCIBLE RESEARCH

Reproducible Research allows:

• <u>traceability</u> and full control of methods, approaches and results

• proper and independent <u>interpretation</u> of the results

• <u>linking provided data to results</u>

• <u>testing and validation</u> of methods and results

• <u>teaching</u> and disseminating methods

• <u>updating</u> the analysis or an entire manuscript or manuscript when changes occur in software, source data, analysis

• <u>coordinating</u> different researchers

• <u>reuse and improvement</u> of the code in the same or other contexts

• <u>promoting the debate</u>

RR approaches include eg. the **Science Code Manifesto** (Barnes, 2013) and **Linked Open Science** (Kauppinen & de Espindola, 2011).

Marco Sigovini, 23/10/2014

REPRODUCIBLE RESEARCH

Science Code Manifesto

Code — All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.

Copyright — The copyright ownership and license of any released source code must be clearly stated.

Citation — Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.

Credit — Software contributions must be included in systems of scientific assessment, credit, and recognition.

Curation — Source code must remain available, linked to related materials, for the useful lifetime of the publication.

Marco Sigovini, 23/10/2014

Linked Open Science

An approach in the context of Reproducible Research to solve challenges of an "executable paper":

1) publication of scientific data, metadata, results, and provenance information using **Linked Data** principles,

2) **open source and web-based environments** for executing, validating and exploring research,

3) **Cloud Computing** for efficient and distributed computing,

4) **Creative Commons** for the legal infrastructure.

(Kauppinen & de Espindola, 2011)

REPRODUCIBLE RESEARCH
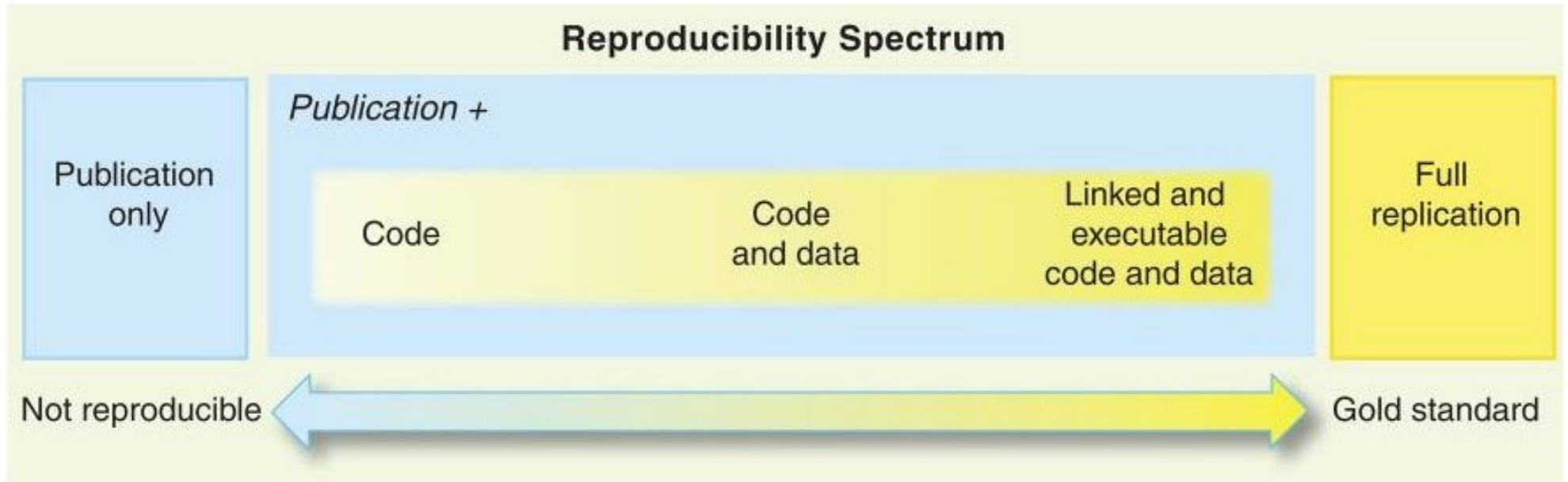
(Peng, 2011)

CASE STUDY

Reproducibility in bioindication:

the M-AMBI quality index
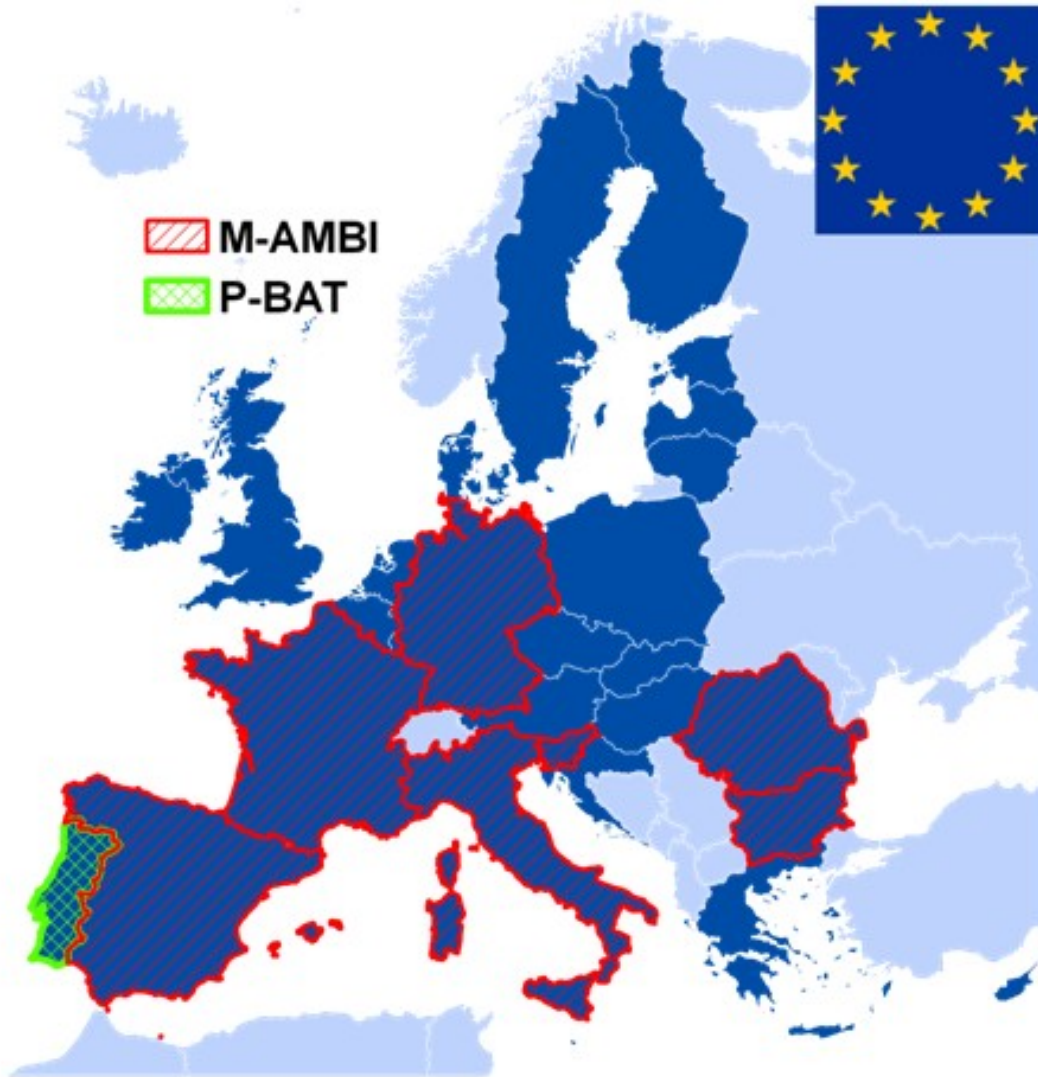
Marco Sigovini, 23/10/2014

M-AMBI

- **multivariate index of ecological quality** (Muxika et al., 2007)

- **integrates biotic index AMBI with richness (S) and Shannon diversity (H')**

- **macrozoobenthos** (soft bottom) **in marine and transitional waters**

- **WFD:** officially enforced in 7 EU countries
(+ Portuguese index P-BAT, Teixeira *et al.*, 2009)

biotic index AMBI (Borja *et al.*, 2000): taxa attributed to 5 ecological groups (EG, based on sensitivity); biotic coefficient BC [0; 6] is calculated based on relative abundance of EG:

$$= \{(0 \times \% \, GI) + (1.5 \times \% \, GII)$$
$$+ (3 \times \% \, GIII) + (4.5 \times \% \, GIV)$$
$$+ (6 \times \% \, GV)\}/100.$$

Water Framework Directive 2000/60/CE

SOFTWARE

CODE

FREEWARE?

USER-FRIENDLY?

OPEN?

Marco Sigovini, 23/10/2014

http://ambi.azti.es/

Marco Sigovini, 23/10/2014

Marco Sigovini, 23/10/2014

SOFTWARE · CODE · FREEWARE? · USER-FRIENDLY? · OPEN?

AMBI by azti tecnalia — MATLAB The Language of Technical Computing

R — free software environment for statistical computing and graphics

GPLv2

Marco Sigovini, 23/10/2014

http://ambi.azti.es/
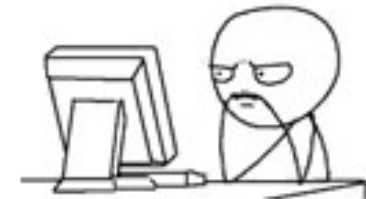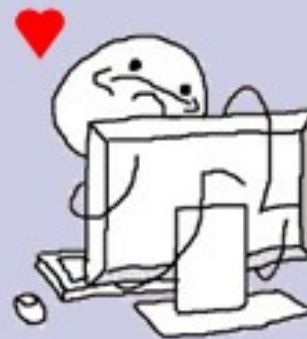
http://www.r-project.org/

The R Project for Statistical Computing

www.r-project.org
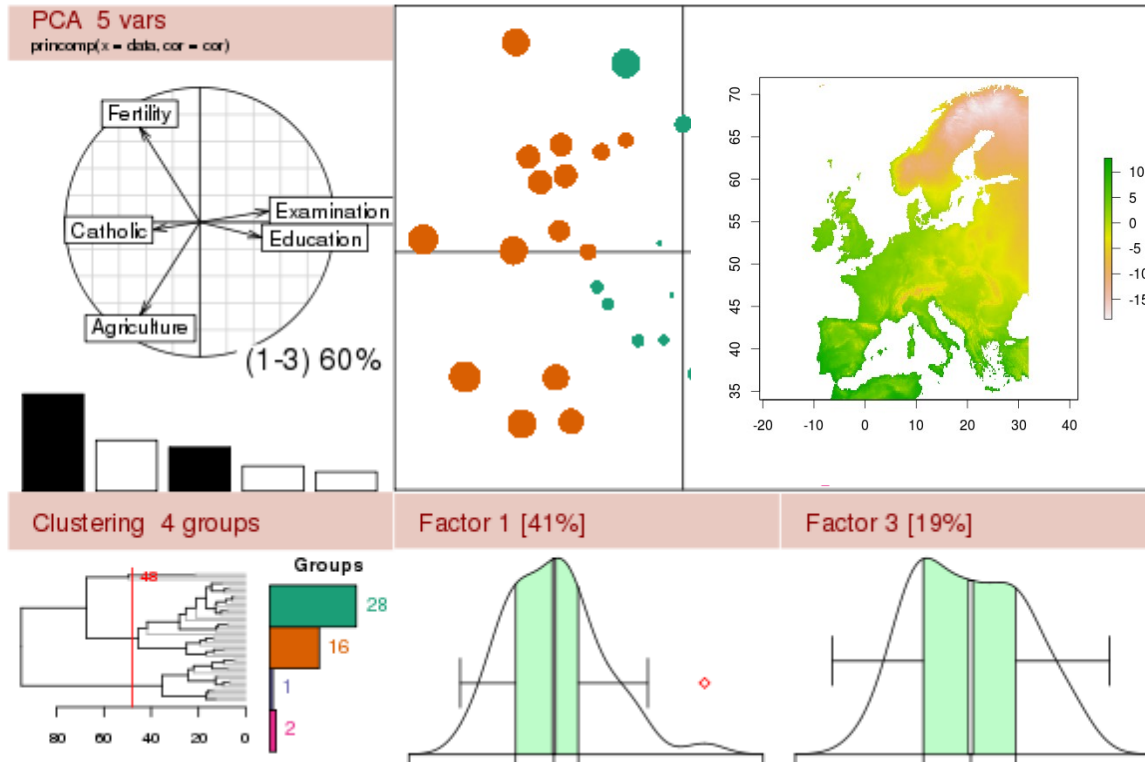
R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

Marco Sigovini, 23/10/2014

ORIGINAL ALGORITHM

Instructions for the use of the AMBI index software (Version 5.0)

Ángel Borja, Julien Mader, Iñigo Muxika[1]

… *When running the software, it uses **richness, diversity and AMBI as variables**. Then, the **variables are standardized**, by subtracting the mean and dividing by the standard deviation. The **Factorial Analysis uses principal components analysis as extraction method, using the correlation matrix and extracting 3 factors**. The software uses **the Varimax rotation, the regression as method for calculating the factor scores**, and missing values are excluded using listwise. The **factor scores (X, Y, Z)** are the new coordinates of each sampling station in the new factor space. These coordinates **are used in deriving the EQR or M-AMBI value**.*

If you are using other software to calculate the EQRs, you must be aware that each software may use different algorithms for the Factor Analysis, providing different results (see Borja et al.,2008a).
...

ORIGINAL ALGORITHM

Instructions for the use of the AMBI index software (Version 5.0)

Ángel Borja, Julien Mader, Iñigo Muxika[1]

… *When running the software, it uses* **richness, diversity and AMBI as variables**. *Then, the* **variables are standardized**, *by subtracting the mean and dividing by the standard deviation. The* **Factorial Analysis uses principal components analysis as extraction method, using the correlation matrix and extracting 3 factors**. *The software uses* **the Varimax rotation, the regression as method for calculating the factor scores**, *and missing values are excluded using listwise. The* **factor scores (X, Y, Z)** *are the new coordinates of each sampling station in the new factor space. These coordinates* **are used in deriving the EQR or M-AMBI value**.

If you are using other software to calculate the EQRs, you must be aware that each software may use different algorithms for the Factor Analysis, providing different results (see Borja et al.,2008a).
...

principal( )	princomp( )	prcomp( )	factanal( )
psych	stats	stats	?
eigendec. on cov	eigendec. on cov	svd on centered scaled data matrix	?
var: /(n-1)	var: /(n)	var: /(n-1)	?
eigenval	sdev=sqrt(eigenval)	sdev=sqrt(eigenval)	?
eigenvec rescaled by sdev	eigenvec	eigenvec	?

Main function in R for factor analysis

Marco Sigovini, 23/10/2014

```r
mambisimpl <- function(x, eg, m.values, metrics = c("S", "H1", "BC"), trasf = "f", high = "def", bad = "def",
            st = rownames(x)) {

    ## 1) METRICS

    if (missing(m.values)) {
            x <- as.matrix(x)[order(rownames(x)), order(colnames(x))]

            AMBI <- function(x, eg) {                                    ## function to calculate AMBI
                    eg <- eg[which(rownames(eg) %in% colnames(x)), , drop = F]

                    if(length(which(colnames(x) %in% rownames(eg))) < length(colnames(x))) {
                            eg.na <- colnames(x)[-which(colnames(x) %in% rownames(eg))]
                            eg <- rbind(eg, rep(NA, length(eg.na)))
                            rownames(eg)[(nrow(eg)-length(eg.na)+1):nrow(eg)] <-
                                    colnames(x)[-which(colnames(x) %in% rownames(eg))]
                    }

                    eg <- eg[order(rownames(eg)), , drop = F]

                    eg.matr <- as.data.frame(matrix(0, nrow = ncol(x), ncol = 6))
                    colnames(eg.matr) <- c("EG1", "EG2", "EG3", "EG4", "EG5", "NA")
                    rownames(eg.matr) <- rownames(eg)
                    options(warn = -1)
                    eg.matr[which(eg[, 1] == 1), 1] <- 1
                    eg.matr[which(eg[, 1] == 2), 2] <- 1
                    eg.matr[which(eg[, 1] == 3), 3] <- 1
                    eg.matr[which(eg[, 1] == 4), 4] <- 1
                    eg.matr[which(eg[, 1] == 5), 5] <- 1
                    eg.matr[which(eg[, 1] == 0), 6] <- 1
                    eg.matr[which(is.na(eg[, 1]) == T), 6] <- 1
                    options(warn = 0)

                    ambi <- as.data.frame(x %*% as.matrix(eg.matr))
                    ambi <- cbind("N" = apply(x, 1, sum), ambi, "BC" = rep(NA, nrow(x)))
                    for (i in 1:nrow(x)) {
                            for (j in 2:6) {
                                    ambi[i, j] <- ambi[i, j] / (ambi[i, "N"] - ambi[i, "NA"]) * 100
                            }
                    }
                    for (i in 1:nrow(x)) {
                            ambi[i, "BC"] <- round((1.5*ambi[i, "EG2"] + 3*ambi[i, "EG3"] + 4.5*ambi[i, "EG4"] +
                                    6*ambi[i, "EG5"]) / 100, 3)
                    }
                    colnames(ambi)[2:6] <- c("EG1(%)", "EG2(%)", "EG3(%)", "EG4(%)", "EG5(%)")
                    ambi[, 2:6] <- round(ambi[, 2:6], 2)

                    return(ambi)
            }

            ambi <- AMBI(x, eg)

            ## samples with N = 0
            if(any(is.nan(ambi$BC))) {ambi$BC[which(is.nan(ambi$BC))] <- 7}

            ## calculation of AMBI-BC on each replicate (other metrics are calculated on pooled replicates)
            st <- st[order(rownames(x))]
            st <- as.factor(st)
            ambi.mu <- round(aggregate(ambi$BC ~ st, FUN = mean)[2], 3)[, 1]

            x.st <- aggregate(x ~ st, FUN = sum)
            rownames(x.st) <- x.st[, 1]
            x.st <- x.st[-1]
            colnames(x.st) <- colnames(x)

            ## calculation of richness/diversity metrics
            require(vegan)
            m.values <- data.frame(
                    "S" = specnumber(x.st),                             ## richness
                    "d" = round((specnumber(x.st) - 1)/(log(apply(x.st, 1, sum))), 3),     ## Margalef index
                    "H1" = round(diversity(x.st, index = "shannon", base = 2), 3),         ## Shannon index
                    "BC" = ambi.mu)                                     ## AMBI-BC

            sample.names <- rownames(x.st)
            METRICS <- NULL
            for (i in 1:length(metrics)) METRICS <- cbind(METRICS, m.values[, metrics[i]])
            colnames(METRICS) <- metrics
            rownames(METRICS) <- sample.names
    }

    else {
            ambi <- NULL

            sample.names <- rownames(m.values)
            METRICS <- as.matrix(m.values)
            colnames(METRICS) <- metrics
            rownames(METRICS) <- sample.names
    }


    ## 2) "HIGH" AND "BAD" REFERENCES VALUES ADDED AS FICTITIOUS SAMPLES

    options(warn = -1)
    if (is.numeric(high) == T) METRICS.high <- high
    if (is.numeric(bad) == T) METRICS.bad <- bad
    if (high == "def") {
            METRICS.high <- rep(NA, length(metrics))
            for (i in 1:length(metrics)) {
                    if (metrics[i] == "BC") METRICS.high[i] <- min(METRICS[, i])
                    else METRICS.high[i] <- max(METRICS[, i])
            }
    }
    if (bad == "def") {
            METRICS.bad <- rep(NA, length(metrics))
            for (i in 1:length(metrics)) {
                    if (metrics[i] == "BC") METRICS.bad[i] <- 6
                    else METRICS.bad[i] <- 0
            }
    }
    options(warn = 0)
```

```r
METRICS.tot <- rbind(METRICS, METRICS.bad, METRICS.high)
rownames(METRICS.tot)[1:nrow(METRICS)] <- sample.names
rownames(METRICS.tot)[nrow(METRICS) + 1] <- "B"
rownames(METRICS.tot)[nrow(METRICS) + 2] <- "H"


## 3) STANDARDISATION OR NORMALISATION

if (trasf == "s") METRICS.tr <- scale(METRICS.tot)                    ## -> M-AMBI*, S-AMBI

NORM <- function(data) {
        norm <- NULL
        for (i in 1:ncol(data)) {
                norm <- cbind(norm, (data[, i] - data["B", i]) / (data["H", i] - data["B", i]))
        }
        rownames(norm) <- rownames(data)
        colnames(norm) <- colnames(data)
        return(norm)
}

if (trasf == "n") METRICS.tr <- NORM(METRICS.tot)                    ## -> M-AMBI*(n), S-AMBI(n)


## 4) FACTOR ANALYSIS

if (trasf == "f") {

        ## calculation with package 'psych'
        require(psych)
        METRICS.fa <- principal(scale(METRICS.tot), nfactors = 3, rotate = "varimax", scores = T)
        METRICS.scores <- factor.scores(METRICS.tot, f = METRICS.fa, method = c("components"))$scores
#       METRICS.scores <- scale(METRICS.tot) %*% METRICS.fa$loadings
        colnames(METRICS.scores) <- c("x", "y", "z")

        ## direct calculation, which produces the factor scores with the same signs of the scores produced by
        ## the AZTI-Tecnalia AMBI software
#       options(warn = -1)
#       METRICS.fa <- princomp(METRICS.tot, cor = T, covmat = cov(METRICS.tot))
#       options(warn = 0)
#       METRICS.fa.load <- loadings(METRICS.fa) %*% diag(METRICS.fa$sdev)
#       METRICS.fa.load <- eigen(cor(METRICS.tot))$vectors %*% diag(sqrt(eigen(cor(METRICS.tot))$values))
#       METRICS.fa.load.varimax <- loadings(varimax(METRICS.fa.load))
#       METRICS.scores <- scale(METRICS.tot) %*% METRICS.fa.load.varimax
#       colnames(METRICS.scores) <- c("x", "y", "z")

        METRICS.tr <- METRICS.scores                    ## -> M-AMBI
}


## 5) EQR CALCULATION

EQR <- function(data) {
        segm <- data[nrow(data),] - data[(nrow(data)-1),]
        vett <- matrix(NA, nrow = nrow(data), ncol = ncol(data))
        for (k in 1: ncol(data)) {vett[, k] <- data[(nrow(data)-1), k]}
        vett <- data - vett
        ris <- round((vett %*% segm / sqrt(sum(segm*segm))) / sqrt(sum(segm*segm)), 3)
        return(ris)
}

eqr <- EQR(METRICS.tr)

#       if (trasf == "n") eqr <- round(apply(NORM(METRICS.tot), 1, mean), 3)          ## direct calculation

if(ncol(METRICS.tot) == 3) name <- "M-AMBI*"
if(ncol(METRICS.tot) == 2) name <- "S-AMBI"
if(trasf == "n") name <- paste(name, "(n)", sep = "")
if(trasf == "f") name <- "M-AMBI"

if(ncol(METRICS.tot) == 2) {
indices <- data.frame(cbind(METRICS.tot, rep(NA, length(eqr)), eqr))
colnames(indices) <- c(metrics, "NA", name)
}

if(ncol(METRICS.tot) == 3) {
indices <- data.frame(cbind(METRICS.tot, eqr))
colnames(indices) <- c(metrics, name)
}

if(trasf == "f") {
        indices <- data.frame(cbind(METRICS.tot, eqr, round(METRICS.tr, 5)))
        colnames(indices) <- c(metrics, name, colnames(METRICS.tr))
}

return(list(ambi, indices))

}


#*****************************************************************************************************************
## EXAMPLE

## data set (provided with the original AZTI-Tecnalia AMBI software)
taxa.ex <- c("Abarenicola_claparedei", "Abra_alba", "Abra_prismatica", "Abra_sp", "Alkmaria_romijni", "Astarte_sp",
        "Bathyporeia_elegans", "Callianassa_tyrrhena", "Capitella_capitata", "Carcinus_maenas", "Caulleriella_alata",
        "Caulleriella_zetlandica", "Corbula_gibba", "Corophium_sp", "Corophium_volutator", "Cossura_sp",
        "Cumopsis_fagei", "Cyathura_carinata", "Desdemona_ornata", "Diogenes_pugilator", "Diptera",
        "Dispio_uncinata", "Eteone_spetsbergensis", "Eumida_sp", "Glycera_tridactyla", "Glycera_unicornis",
        "Hediste_diversicolor", "Heteromastus_filiformis", "Hydrobia_ulvae", "Lekanesphaera_hookeri",
        "Lekanesphaera_rugicauda", "Lepidochitona_cinerea", "Lunatia_alderi", "Malacoceros_fuliginosus",
        "Malacoceros_sp", "Mediomastus_fragilis", "Melita_palmata", "Mysidacea", "Nassarius_pygmaeus",
        "Nassarius_reticulatus", "Nematoda", "Nemertea", "Oligochaeta", "Ovatella_myosotis", "Paradoneis_armata",
        "Paraonathia_formica", "Parvicardium_scabrum", "Pectinaria_koreni", "Polydora_ligni", "Prionospio_fallax",
        "Prionospio_steenstrupi", "Pseudopolydora_paucibranchiata", "Pseudopythina_macandrewi", "Pygospio_elegans",
        "Scoloricia_sp", "Scrobicularia_plana", "Solen_marginatus", "Spio_martinensis", "Spiochaetopterus_costarum",
        "Streblospio_shrubsolii", "Tapes_decussatus", "Tellina_compressa", "Tellina_sp", "Tubulanus_polymorphus",
        "Urothoe_brevicornis")
abundances.ex <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 29, 0, 0, 0, 35, 23, 180, 0, 0, 0, 1,
        0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
```

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 10, 505, 518, 117, 80, 58, 11, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 1, 0,
0, 0, 0, 2, 0, 0, 1, 1, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 9, 14, 5, 21, 0, 32, 11, 7, 6, 5, 2, 4, 17, 1, 38, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 5, 7, 6, 0, 1, 2, 5, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 6, 13, 8,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 5, 0, 1, 1, 0, 0, 3, 2, 1, 2, 33, 0, 20, 54, 116, 64, 1, 1, 0, 0, 4, 5,
1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 12, 4, 19, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 16, 5, 19, 0, 0, 2, 1, 0, 1, 0, 0, 9, 0, 0, 0, 0, 0, 2, 1, 2, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 8, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 3, 5, 4, 2, 5, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 4, 0, 0, 0, 2, 7, 28, 0, 1, 38, 266, 62, 44, 15, 3,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 4, 0, 27, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6, 2, 2,
0, 0, 0, 0, 0, 3, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 4, 0, 0, 0, 0, 0, 0,
0, 0, 0, 2, 7, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 3, 0, 3,
2, 1, 8, 13, 0, 3, 24, 3, 29, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 3, 0, 0, 0, 5, 5, 28, 1, 4, 2, 0, 12, 5, 0, 0, 0, 0, 10, 4, 11, 0, 1, 0,
0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
st.ex <- paste("st.", as.character(sprintf("%02d", c(rep(1, 3), rep(2, 3), rep(3, 3), rep(4, 3), rep(5, 3),
     rep(6, 3), rep(7, 3), rep(8, 2), rep(9, 2), rep(10, 2), rep(11, 2)))), sep = "")
rep.ex <- paste(st.ex, c(rep(c("a", "b", "c"), 7), rep(c("a", "b"), 4)), sep = "")
dataset.ex <- matrix(abundances.ex, ncol = 65)
rownames(dataset.ex) <- rep.ex
colnames(dataset.ex) <- taxa.ex
dataset.ex <- dataset.ex[, -which(colnames(dataset.ex) == "Mysidacea")] ## taxON ignored (as in the original example)

dataset_st.ex <- aggregate(dataset.ex ~ st.ex, FUN = sum)
rownames(dataset_st.ex) <- dataset_st.ex[, 1]
dataset_st.ex <- dataset_st.ex[-1]
colnames(dataset_st.ex) <- colnames(dataset.ex)

## EG (source: list of taxa provided with the AZTI-Tecnalia AMBI software, March 2012 version)
eg.mar2012 <- matrix(c(1, 3, 3, 3, 3, 1, 1, 3, 5, 3, 4, 4, 4, 3, 3, 4, 2, 3, 2, 2, 4, 3, 3, 2, 2, 2, 3, 4, 3, 3, 3,
     2, 2, 5, 3, 3, 1, NA, 2, 2, 3, 3, 5, NA, 3, 3, 1, 4, 4, 4, 4, 4, NA, 3, 1, 3, 1, 3, 1, 3, 1, 1, 1, 2, 1),
     ncol = 1)
rownames(eg.mar2012) <- taxa.ex
colnames(eg.mar2012) <- "EG_v.Mar2012"

## direct calculation of the metrics
mambisimpl(dataset_st.ex, eg.mar2012)                                    ## M-AMBI (AMBI on stations)
mambisimpl(dataset.ex, eg.mar2012, st = st.ex)                           ## M-AMBI (AMBI on replicates)

mambisimpl(dataset.ex, eg.mar2012, trasf = "n", st = st.ex)              ## M-AMBI*(n)
mambisimpl(dataset.ex, eg.mar2012, metrics = c("S", "d", "BC"), st = st.ex)   ## BAT index

mambisimpl(dataset.ex, eg.mar2012, metrics = c("S", "BC"), trasf = "n", st = st.ex)   ## S-AMBI(n) with richness

## Directive 2000/60/EC monitoring: italian ref. values for non-tidal coastal lagoons (D.M. 260, 8/11/2010)
IT_MAT1 <- c(25, 3.3, 1.85)
mambisimpl(dataset.ex, eg.mar2012, high = IT_MAT1, st = st.ex)

## metrics provided to mambisimpl by the user
metrics.ex <- mambisimpl(dataset.ex, eg.mar2012, trasf = "n", st = st.ex)[[2]][1:11, 1:3]
mambisimpl(m.values = metrics.ex, trasf = "n")                           ## M-AMBI*(n)

###############################################################################################################
###############################################################################################################MS#v1.0##
```

M-AMBI

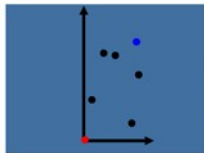(Muxika *et al.*, 2007; Borja *et al.*, 2012)
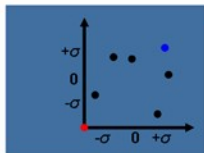

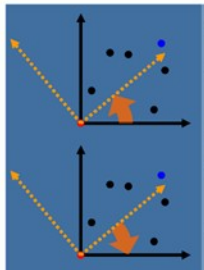1. species / abundance matrix

2. three metrics (S, H', BC)

(plot 3D → 2D for displaying purposes)

3. High/Bad reference values added

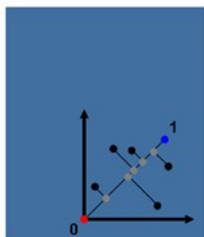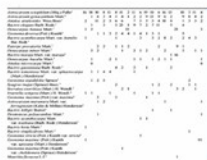4. standardisation

5. factor analysis:

i. eigendecomposition (→ PCA)

ii. ~~factor selection~~

iii. eigenvectors rescaling by √(eigenvalues)

iv. Varimax rotation

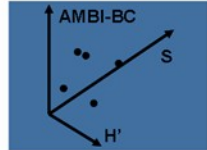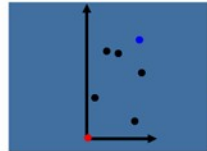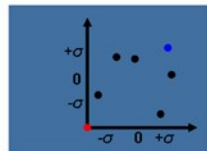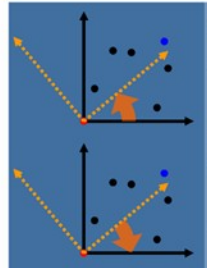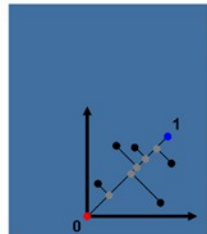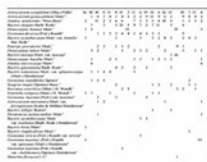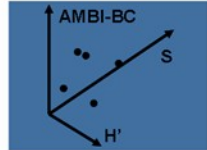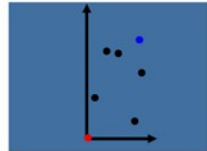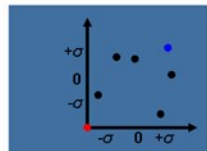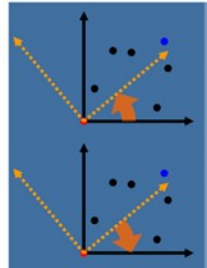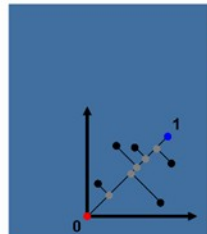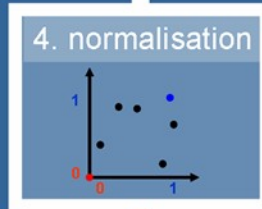6. scores (new coordinates) projected on the line identified by ref. values and normalised 0-1

Marco Sigovini, 23/10/2014

M-AMBI
(Muxika *et al.*, 2007; Borja *et al.*, 2012)

M-AMBI*(n)

1. species / abundance matrix

2. three metrics (S, H', BC)

AMBI-BC
S
H'

(plot 3D → 2D for displaying purposes)

3. High/Bad reference values added

4. standardisation

5. factor analysis:
   i. eigendecomposition (→ PCA)
   ii. ~~factor selection~~
   iii. eigenvectors rescaling by √(eigenvalues)
   iv. Varimax rotation

6. scores (new coordinates) projected on the line identified by ref. values and normalised 0-1

4. normalisation

(mathematically identical to the simple mean)

Marco Sigovini, 23/10/2014

M-AMBI
(Muxika et al., 2007; Borja et al., 2012)

1. species / abundance matrix

2. three metrics (S, H', BC)

AMBI-BC
S
H'

(plot 3D → 2D for displaying purposes)

3. High/Bad reference values added

4. standardisation

5. factor analysis:
  i. eigendecomposition (→ PCA)
  ii. factor selection
  iii. eigenvectors rescaling by √(eigenvalues)
  iv. Varimax rotation

6. scores (new coordinates) projected on the line identified by ref. values and normalised 0-1

M-AMBI*$_{(n)}$

4. normalisation

(mathematically identical to the simple mean)

S-AMBI$_{(n)}$

2. two metrics (S, BC)

AMBI-BC
S

4. normalisation

Marco Sigovini, 23/10/2014
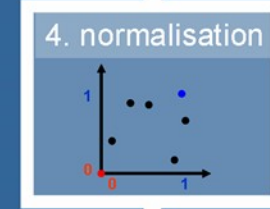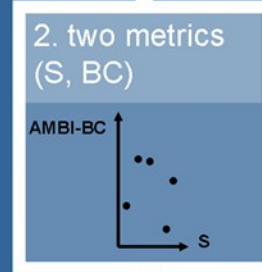
PROPOSED ALGORITHMS

$$\text{M-AMBI*}_{(n)} = \frac{S_{(n)} + H'_{(n)} + \text{AMBI BC}_{(n)}}{3}$$

$$\text{S-AMBI*}_{(n)} = \frac{S_{(n)} + \text{AMBI BC}_{(n)}}{2}$$

The less mutually correlated the metrics are, the more the multimetric index is interpretable.
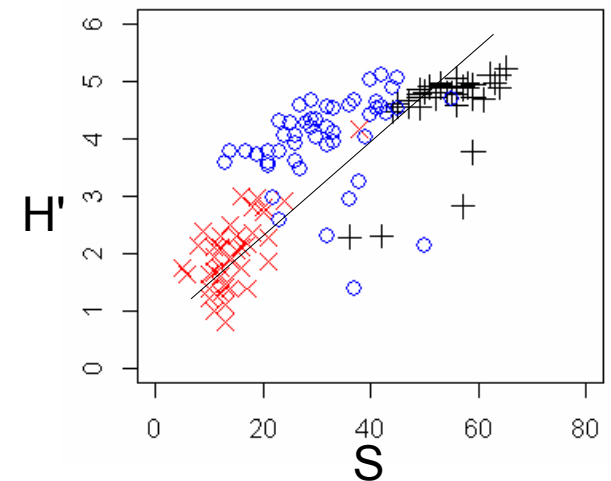
DATASETS

1. Ekofisk

Survey of the Ekofisk oil field (Norway), 1987

Gray *et al.*,1990; in PRIMER software for community analysis (Clarke & Gorley, 2006)

39 samples (0.1 m², 3 rep., 1 mm mesh), 13,883 individuals, 173 taxa.
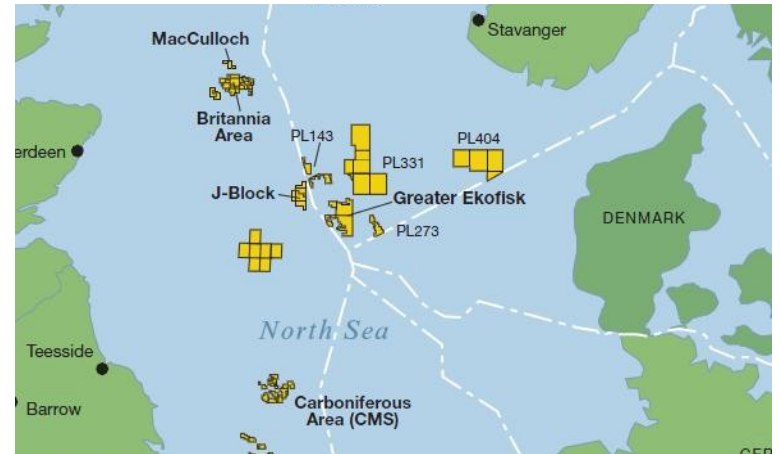
2. Venice

Palude della Rosa, Lagoon of Venice (Italy), 1991

Tagliapietra *et al.*, 1998

42 stations (0.1 m², 1 mm mesh), 34,732 individuals, 62 taxa.
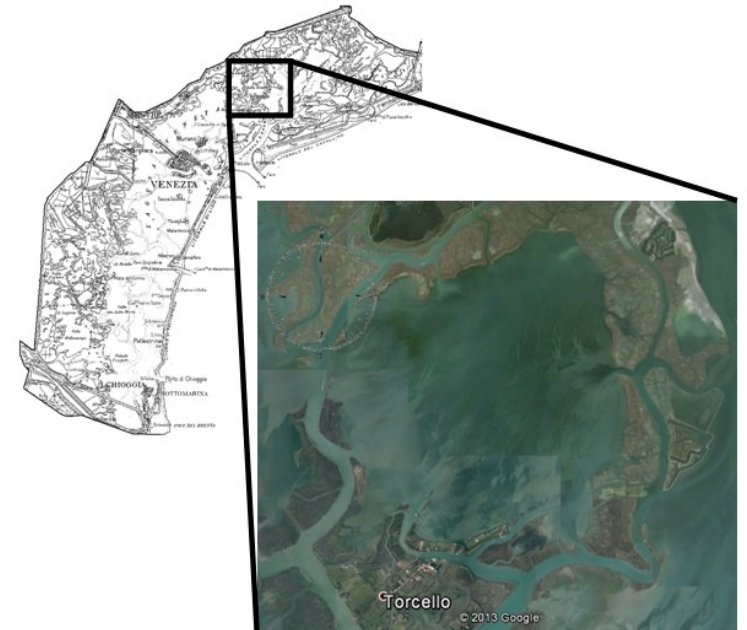
3. Simulated

Poisson lognormal distribution, with parameters μ and σ randomly generated (from a normal distribution) for each sample. EG randomly assigned. 50 samples, 120 fictitious species.

-> **independence among AMBI and S, H'**

RESULTS: M-AMBI*$_{(n)}$
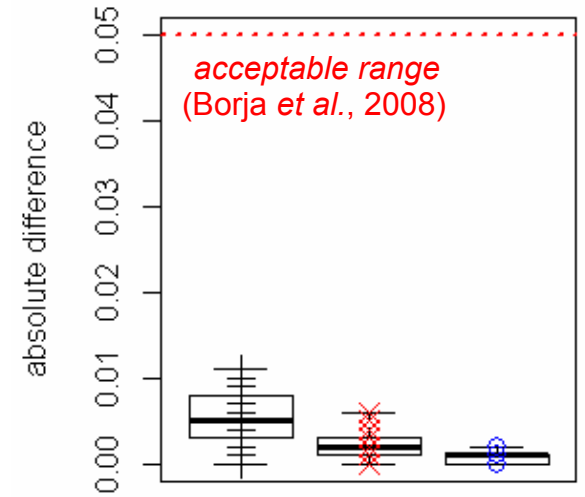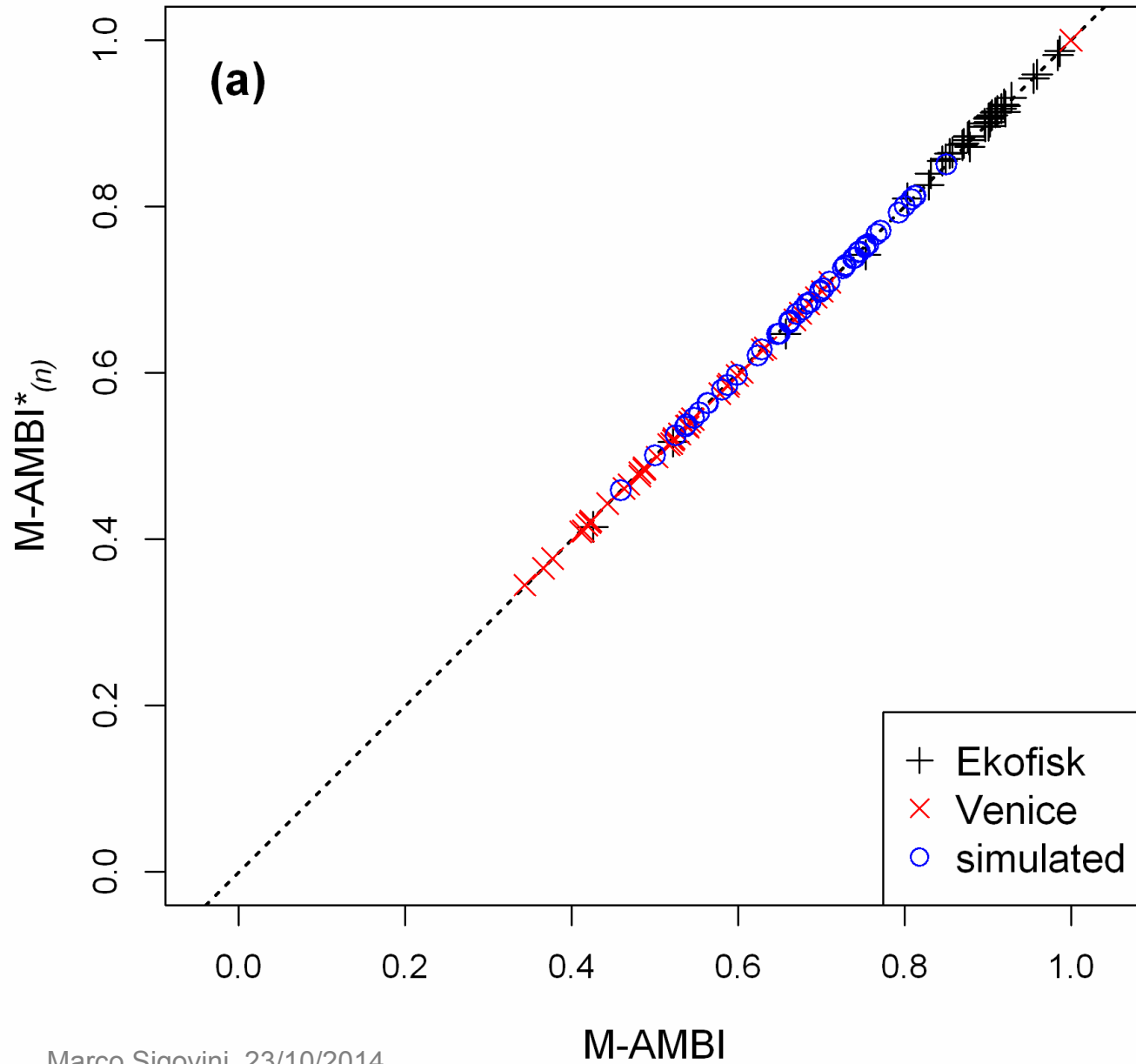
RESULTS: S-AMBI$_{(n)}$
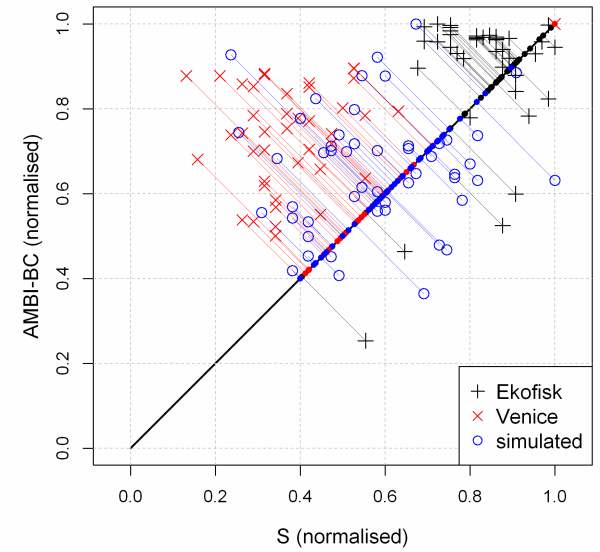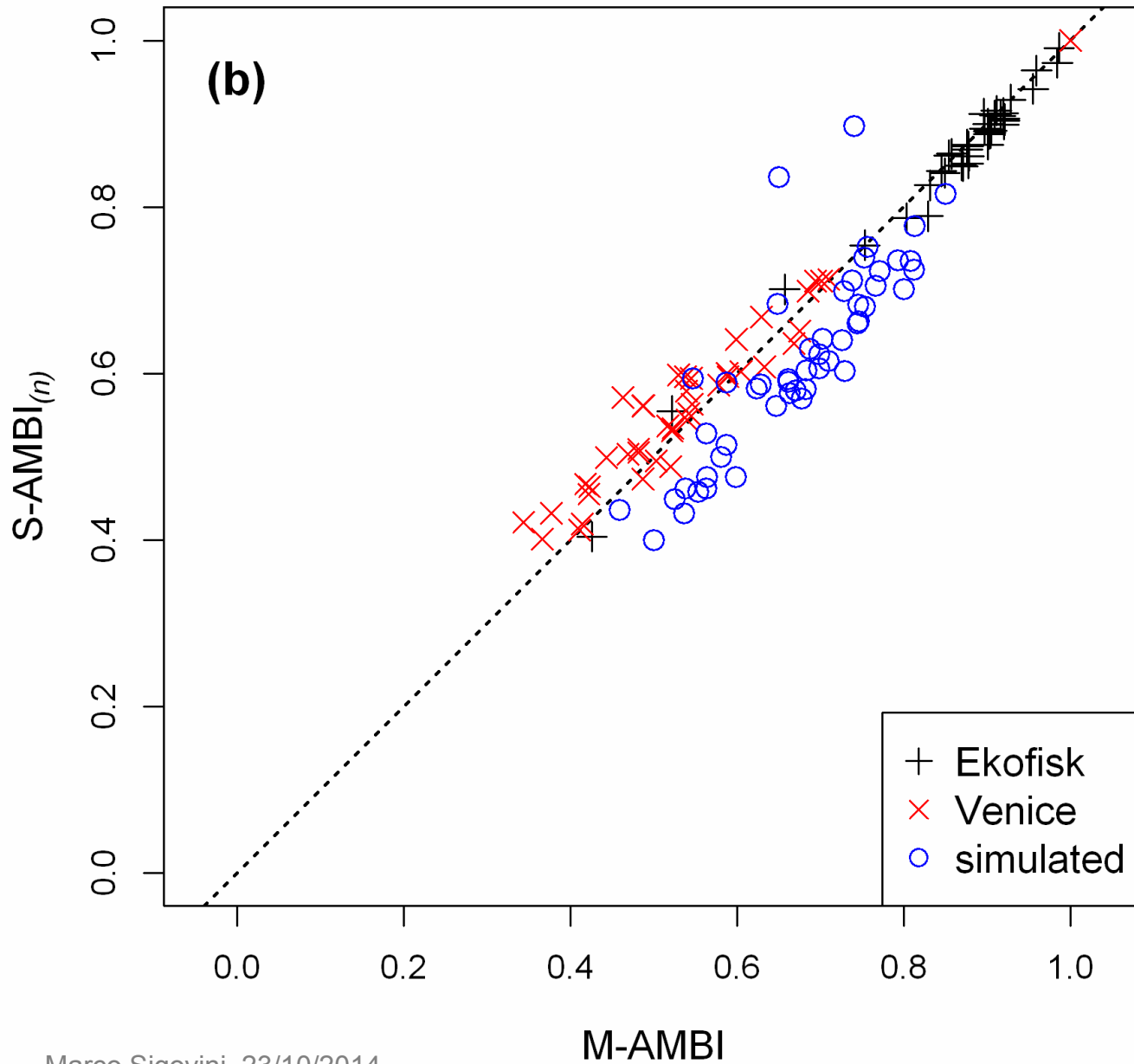
**Sensitivity + Diversity components**

CONCLUSIONS

1) Factor Analysis is demonstrated not functional to M-AMBI

FA results are affected by the criteria used to choose the model, decide how many factors to retain, select the rotation method.; addition of new data gives different results (in fact, very small deviations).

2) M-AMBI is closely approximated by the simple mean of the normalised metrics independent of the number of sample

3) Bivariate version of M-AMBI (S-AMBI: diversity + species sensitivity index) still highly correlated.

4) → free the species (list)!

Marco Sigovini, 23/10/2014
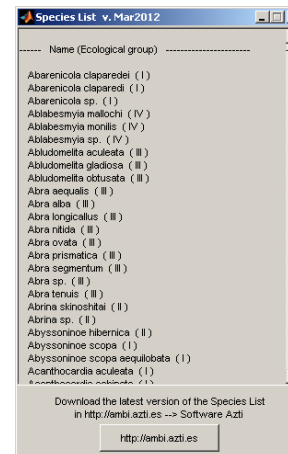
CONCLUSIONS

***lex parsimoniae* (Ockham's Razor)**

"The model with the fewest assumptions should be preferred"

Desirable proprieties (proposed algorithms):

• Transparency, Reproducibility, Openness

• Testability, Falsifiability

• Simplicity

• Generalisability

• Robustness

• Ecological meaningfulness (and significance)

Reproducing the algorithm on open source software was instrumental to understand and test its features, identify flaws and propose improvements.

Ecologists should consider to provide source code (and data) when proposing mathematical tools or publishing results (eg. as suppl. material)
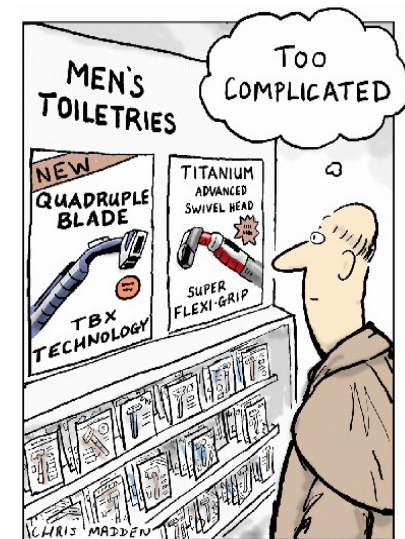
M-AMBI revisited: looking inside a widely-used benthic index

**Marco Sigovini · Erica Keppel ·
Davide Tagliapietra**

ONLINE SOURCES about Reproducible Research with R:

www.r-project.org

http://cran.r-project.org/web/views/ReproducibleResearch.html

http://www.r-bloggers.com/reproducible-research-is-still-a-challenge/

http://www.r-bloggers.com/making-reproducible-research-enjoyable/

http://www.r-bloggers.com/brief-introduction-on-sweave-and-knitr-for-reproducible-research/

https://osf.io/s9tya/ (Stodden et al., 2013, Implementing Reproducible Research, 448 pp.)

http://ropensci.org/blog/2014/06/09/reproducibility/

REFERENCES

Borja et al., 2008. Marine Pollution Bulletin, 56: 1377–1379

Borja et al., 2012. Revista de Investigación Marina, AZTI-Tecnalia 19: 71–82

Gilbert et al., 2012. Molecular Ecology, 21: 4925–4930

Gray et al., 1990. Marine Ecology Progress Series, 66: 285–299

Kauppinen & de Espindola, 2011. Procedia Computer Science, 4: 726–731

Mesirov, 2010. Science, 327(5964)

Muxika et al., 2007. Marine Pollution Bulletin, 55: 16–29

Peng, 2011. Science, 334(6060): 1226–1227

Tagliapietra et al., 1998. Estuarine Coast and Shelf Science, 47: 217–226

Teixeira et al., 2009. Marine Pollution Bulletin, 58: 1477–1486

CONTACTS: marco.sigovini@ve.ismar.cnr.it