Package 'octopucs'

August 17, 2024

Title Statistical Support for Hierarchical Clusters

Version 0.1.1

Description
Generates n hierarchical clustering hypotheses on subsets of classifiers (usually species in community ecology studies). The n clustering hypotheses are combined to generate a generalized cluster, and computes three metrics of support. 1) The average proportion of elements conforming the group in each of the n clusters (integrity). And 2) the contamination, i.e., the average proportion of elements from other groups that enter a focal group. 3) The probability of existence of the group gives the integrity and contamination in a Bayesian approach.

Date 2024-08-01

Depends vegan, stringr, progress

License GPL-3

Encoding UTF-8

RoxygenNote 7.3.1

NeedsCompilation no

Author Roger Guevara [aut, cre],
Ian MacGregor-Fors [aut]

Maintainer Roger Guevara <roger.guevara@inecol.mx>

Repository CRAN

Date/Publication 2024-08-16 23:10:03 UTC

Contents

`compsim2` *Complementarity and disimilarity*

Description

Calculates the complementaritry and dissimilarity between two or more communities base on .

Usage

```
compsim2(M)
```

Arguments

`M`	A species times site matrix.

Value

Generates a disimilarity matrix.

Author(s)

Roger Guevara (roger.guevara@inecol.mx)

References

Aguirre, A., Guevara, R., García, M., & López, J. C. (2010). Fate of epiphytes on phorophytes with different architectural characteristics along the perturbation gradient of Sabal mexicana forests in Veracruz, Mexico. Journal of Vegetation Science, 21(1), 6-15. https://doi.org/10.1111/j.1654-1103.2009.01131.x

Examples

```
data(dune)
M2 <- compsim2(t(dune[1:2,]))
M2
```

`draw.octopucs` *Draw the Generalized Clustering Hypothesis*

Description

Draws the resulting generalized cluster with tips facing either down, left or right.

Usage

```
draw.octopucs(M2, cex=1, lab.pos=0.09, n=NULL,
distance=NULL, method=NULL, prop=NULL, Name=NULL, ylim=NULL, xlim=NULL,
verbose=NULL, orientation=NULL, support=NULL, sep.lab=NULL, disp = NULL)
```

Arguments

`M2`	A lis object generated by glzcluster or glzcluster2
`cex`	Size of text in the figure
`lab.pos`	This argument modifies the position of the labels relative to the tips of the dendrogram (sites/conditions).
`n`	This is defined in the glzcluster or glzcluster2 functions. Number of bootstrap samples performed.
`distance`	Distance method used (default="bray"). This is defined in the glzcluster or glzcluster2 functions. Options include those of the function "vegdist" ("vegan"): "manhattan", "euclidean", "canberra", "bray", "kulczynski", "gower", "morisita", "horn", "mountford", "jaccard", "raup", "binomial", "chao", "altGower", "cao" and "mahalanobis", as well as those referred in Koleff et al. (2003), available either by name or number: 1 or "w", 2 or "-1", 3 or "c", 4 or "wb", 5 or "r", 6 or "I", 7 or "e", 8 or "t", 9 or "me", 10 or "j", 11 or "sor", 12 or "m", 13 or "-2", 14 or "co", 15 or "cc", 16 or "g", 17 or "-3", 18 or "1", 19 or "", 20 or "hk", 21 or "rlb", 22 or "sim", 23 or "gl", 24 or "z". For open distance metrics (not restricted to zero-one ranges), the axis is re-scaled to cover a zero-one range. If so, the axis will be labeled a "Similarity (rescaled)"; if no re-scaling is necessary, the label will appear as "Similarity".
`method`	This is defined in the glzcluster or glzcluster2 functions. Agglomerative procedure for hierarchical clustering, including: "average" (=UPGMA), "ward.D", "ward.D2", "single", "complete", "mcquitty" (=WPGMA), "median" (=WPGMC) or "centroid" (=UPGMC).
`prop`	This is defined in the glzcluster or glzcluster2 functions. Proportion of information used in the bootstrapping procedures. When matrices contain more than 18 columns, it refers to: (i) the proportion of columns used in every run to produce a new grouping array when "resample=TRUE" and (ii) the amount of cells in the body of the matrix (1–prop; e.g., 1–0.8=0.2) that will be replaced with column averages when "resample=FALSE". When matrices contain less than 18 columns, the procedure with "resample=FALSE" will be forced. In such scenario, "prop" corresponds to the amount of cells in the body of the matrix (1–prop; e.g., 1–0.8=0.2) that will be replaced in the bootstrapping procedures.

Name	Name of the database run on octopucs
xlim	Limits of the x-axis
ylim	Limits of the y-axis
verbose	TRUE = display all relevant setup values used in the analysis. FALSE = display only the dendrogram with marked significantly grouped clusters.
orientation	Any of d, l, or r denoting the orientation of the tips (down, left, and right, respectively)
support	A matrix object generated by the function glzcs
sep.lab	This argument modifies the position of supporting values from the nodes and lines and the dendrogram.
disp	What to display in relation to supporting values for the groups and subgroups. By default is set to "sig" and then support values are shown only in those groups statistically supported.

Value

This function draws and if requested, populate with support metrics the consesnsus hypothesis of clustering

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence–absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
rownames(dune)<- paste("S",1:20, sep="")
OCT <- glzcluster(dune, n=100)
GRP <- IDscluster(OCT)
SUP <- glzcs(GRP, OCT)
draw.octopucs(OCT, support=SUP)
```

F1 *Labeling of memberships of two groups*

Description

Internal function to label the membership of two clusters

Usage

```
F1(A, B)
```

Arguments

A	Memberships in the first cluster
B	Membership in the second cluster

Value

Produced a two-column matrix populated with zeros and ones to indicate what IDs (row names) are included in each of two vectors in which some element may be shared

Author(s)

Roger Guevara (roger.guevara@inecol.mx)

---

glzcluster *Generalized Cluster*

Description

Generates a cluster hypothesis based on n subsets of the available columns.

Usage

```
glzcluster(data, n = NULL, prop=NULL, distance =NULL, method=NULL, Name=NULL)
```

Arguments

data	A matrix object of r rows that will be grouped by c columns.
n	The number of subsets to be generated is based on which the generalized cluster will be generated.
prop	Either the proportion of the c columns used in every run to produce a new grouping array or the 1-prop number of cells in the body of the matrix whose values are replaced by the corresponding column average. This will be automatically decided based on the total number of columns in the matrix. When there are fewer than 19 columns in the matrix, the procedure will replace cell values with column averages even if the resample is TRUE.
distance	Distance method to be used (default="bray"). Options include those of the function "vegdist" ("vegan"): "manhattan", "euclidean", "canberra", "bray", "kulczynski", "gower", "morisita", "horn", "mountford", "jaccard", "raup", "binomial", "chao", "altGower", "cao" and "mahalanobis", as well as those referred in Koleff et al. (2003), available either by name or number: 1 or "w", 2 or "-1", 3 or "c", 4 or "wb", 5 or "r", 6 or "I", 7 or "e", 8 or "t", 9 or "me", 10 or "j", 11 or "sor", 12 or "m", 13 or "-2", 14 or "co", 15 or "cc", 16 or "g", 17 or "-3", 18 or "1", 19 or "", 20 or "hk", 21 or "rlb", 22 or "sim", 23 or "gl", 24 or "z". The axis is re-scaled to cover a zero-one range for open distance metrics (not restricted to zero-one ranges). If so, the axis will be labeled a "Similarity (rescaled)"; if no re-scaling is necessary, the label will appear as "Similarity".

method	Agglomerative procedure for hierarchical clustering, including: "average" (=UPGMA), "ward.D", "ward.D2", "single", "complete", "mcquitty" (=WPGMA), "median" (=WPGMC) or "centroid" (=UPGMC).
Name	Name of the object matrix analyzed

Value

Generates a list object that contains the n cluster assableged with random subsets of the data and in the n+1 possition the generalized cluster, an hclust output

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
M2 <- glzcluster(dune, n=100,
prop=0.8, distance="bray",
method="average")
length(M2)
ls(M2[[length(M2)]])
```

---

glzcluster2 *Generalized Cluster*

Description

Generates a cluster hypothesis based on the transformation of the 1-prop ratio of the cells in the matrix's body into their corresponding column averages.

Usage

```
glzcluster2(data=data, n = NULL, prop=NULL, distance =NULL, method=NULL,
Name=NULL)
```

Arguments

data	A matrix object of r rows that will be gropued by c columns.
n	The number of subsets to be generated is based on which the generalized cluster will be generated.

prop	Either the proportion of the c columns used in every run to produce a new grouping array or the 1-prop number of cells in the body of the matrix whose values are replaced by the corresponding column average. This will be automatically decided based on the total number of columns in the matrix. When there are less than 19 columns in the matrix, the procedure will replace cell values with column averages even if the resample is TRUE.
distance	Distance method to be used (default="bray"). Options include those of the function "vegdist" ("vegan"): "manhattan", "euclidean", "canberra", "bray", "kulczynski", "gower", "morisita", "horn", "mountford", "jaccard", "raup", "binomial", "chao", "altGower", "cao" and "mahalanobis", as well as those referred in Koleff et al. (2003), available either by name or number: 1 or "w", 2 or "-1", 3 or "c", 4 or "wb", 5 or "r", 6 or "I", 7 or "e", 8 or "t", 9 or "me", 10 or "j", 11 or "sor", 12 or "m", 13 or "-2", 14 or "co", 15 or "cc", 16 or "g", 17 or "-3", 18 or "1", 19 or "", 20 or "hk", 21 or "rlb", 22 or "sim", 23 or "gl", 24 or "z". For open distance metrics (not restricted to zero-one ranges), the axis is re-scaled to cover a zero-one range. If so, the axis will be labeled a "Similarity (rescaled)"; if no re-scaling is necessary, the label will appear as "Similarity".
method	Agglomerative procedure for hierarchical clustering, including: "average" (=UPGMA), "ward.D", "ward.D2", "single", "complete", "mcquitty" (=WPGMA), "median" (=WPGMC) or "centroid" (=UPGMC).
Name	Name of the matrix object to be analyzed

Value

Generates a list object that contains the n cluster assembled with random subsets of the data, and in the n+1 position the generalized cluster, an hclust output

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
M2 <- glzcluster(dune, n=100,
prop=0.8, distance="bray",
method="average")
length(M2)
ls(M2[[length(M2)]])
```

glzcs *Support for Groups and Subgroups in the Generalized Dendrogram*

Description

Generates three metrics of support for the observed clusters at all hierarchical levels

Usage

```
glzcs(GRPs=GRPs, M2=M2)
```

Arguments

GRPs	List object generated by the functions IDscluster & IDscluster2
M2	Class list object generated by the functions glzcluster or glzcluster2

Value

A table with supporting values and size of every group: integrity, contamination, probability and group size.

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
OCT <- glzcluster(dune, n=100)
GRP <- IDscluster(OCT)
SUP <- glzcs(GRP, OCT)
SUP
```

IDscluster *Memberships of Groups and Subgroups*

Description

Identifies the memberships at every level of the topology of the generalized cluster.

Usage

```
IDscluster(M2=M2)
```

Arguments

M2	List object generated by glzcluster or glzcluster2

Value

A list object with the memberships for the groups and subgroups all through the topology of the generalized cluster

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence–absence data. Journal of Animal Ecology 71:367

Examples

```
data(dune)
M2 <- glzcluster(dune, n=5, prop=0.8, distance="bray", method="average")
GRPs <- IDscluster(M2)
GRPs
```

IDscluster2 *Memberships of Groups and Subgroups*

Description

Identify the memberships at every level of the topology in the bootstrapped clusters.

Usage

```
IDscluster2(M2=M2)
```

Arguments

M2	List object generated by glzcluster or glzcluster2

Value

A list object with the memberships for the groups and subgroups all through the topology of the generalized cluster

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
M2 <- glzcluster(dune, n=5, prop=0.8, distance="bray", method="average")
GRPs <- IDscluster2(M2[[length(M2)]])
GRPs
```

---

octopucs *Support for Hierarchical Clusters*

Description

The global function of the "octopucs" package is to estimate support for hierarchical clusters throughout the dendrogram's topology, considering their repeatability, integrity, and contamination based on bootstrapping procedures.

Usage

```
octopucs(data, n=NULL, prop=NULL, resample = TRUE, distance =NULL,
method=NULL, orientation=NULL, disp = NULL, lab.pos=NULL,
sep.lab=NULL, xlim=NULL, ylim=NULL, verbose=NULL)
```

Arguments

data	A matrix whose rows (sites/conditions) will be grouped based on the columns (species/elements).
n	The number of bootstrap repetitions.

prop	Proportion of information used in the bootstrapping procedures. When matrices contain more than 18 columns, it refers to (i) the proportion of columns used in every run to produce a new grouping array when "resample=TRUE" and (ii) the number of cells in the body of the matrix (1–prop; e.g., 1–0.8=0.2) that will be replaced with column averages when "resample=FALSE". When matrices contain less than 18 columns, the procedure with "resample=FALSE" will be forced. In such a scenario, "prop" corresponds to the number of cells in the body of the matrix (1–prop; e.g., 1–0.8=0.2) that will be replaced in the bootstrapping procedures.
resample	When TRUE, "octopucs" will select random subsets of columns in every bootstrap run to produce a new grouping array. When FALSE, "octopucs" will replace cells in the body of the matrix with column averages instead of entire columns (see description for "prop" for further details).
distance	Distance method to be used (default="bray"). Options include those of the function "vegdist" ("vegan"): "manhattan", "euclidean", "canberra", "bray", "kulczynski", "gower", "morisita", "horn", "mountford", "jaccard", "raup", "binomial", "chao", "altGower", "cao" and "mahalanobis", as well as those referred in Koleff et al. (2003), available either by name or number: 1 or "w", 2 or "-1", 3 or "c", 4 or "wb", 5 or "r", 6 or "I", 7 or "e", 8 or "t", 9 or "me", 10 or "j", 11 or "sor", 12 or "m", 13 or "-2", 14 or "co", 15 or "cc", 16 or "g", 17 or "-3", 18 or "1", 19 or "", 20 or "hk", 21 or "rlb", 22 or "sim", 23 or "gl", 24 or "z". For open distance metrics (not restricted to zero-one ranges), the axis is re-scaled to cover a zero-one range. If so, the axis will be labeled a "Similarity (rescaled)"; if no re-scaling is necessary, the label will appear as "Similarity".
method	Agglomerative procedure for hierarchical clustering, including: "average" (=UPGMA), "ward.D", "ward.D2", "single", "complete", "mcquitty" (=WPGMA), "median" (=WPGMC) or "centroid" (=UPGMC).
orientation	Orientation of the plot (default="r": right; metric in the x-axis is similarity). "l" = left-wise orientation (metric in the x-axis is dissimilarity). "d" = down-wise orientation (metric in the y-axis is dissimilarity).
disp	Result display in the dendrogram (default="sig"; only supported nodes will populated with given results). "all" = all nodes will be populated with support metrics, with significantly supported nodes in blue/red and non-significantly supported nodes in black.
lab.pos	This argument modifies the position of the labels relative to the tips of the dendrogram (sites/conditions).
sep.lab	This argument modifies the position of supporting values from the nodes and lines in the dendrogram.
xlim	Limits of the x-axis.
ylim	Limits of the y-axis.
verbose	TRUE = display all relevant setup values used in the analysis. FALSE = display only the dendrogram with marked significantly grouped clusters.

Value

"octopucs" generates a list object that contains:

`$gcluster`	hclust output of the generalized clustered dendrogram.
`$support`	Table with the support value for every group and subgroup across the topology of the generalized clustered dendrogram.
`$grps`	List with the memberships of every group and subgroup in the topology.

Author(s)

Roger Guevara (roger.guevara@inecol.mx) & Ian MacGregor-Fors (ian.macgregor@helsinki.fi)

References

Koleff, P., Gaston, K.J., Lennon, J. J. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 71:367. https://doi.org/10.1046/j.1365-2656.2003.00710.x

Examples

```
data(dune)
octopucs(dune, n=100, prop=0.8,
resample=TRUE,distance="bray",
method="average", disp = "sig",
verbose=TRUE)
```

Index