Package 'ClustGeo'

September 30, 2021

Type Package

Title Hierarchical Clustering with Spatial Constraints

Version 2.1

Author Marie Chavent [aut, cre],
Vanessa Kuentz [aut],
Amaury Labenne [aut],
Jerome Saracco [aut]

Maintainer Marie Chavent <Marie.Chavent@u-bordeaux.fr>

Description Implements a Ward-like hierarchical clustering algorithm including soft spatial/geographical constraints.

Depends R (>= 3.0.0)

Imports graphics, stats, sp, spdep

License GPL (>= 2.0)

LazyData true

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 7.1.2

NeedsCompilation no

Repository CRAN

Date/Publication 2021-09-30 14:20:13 UTC

R topics documented:

`choicealpha` *Choice of the mixing parameter*

Description

This function calculates the proportion of inertia explained by the partitions in `K` clusters for a range of mixing parameters `alpha`. When the proportion of explained inertia calculated with `D0` decreases, the proportion of explained inertia calculated with `D1` increases. The plot of the two curves of explained inertia (one for `D0` and one for `D1`) helps the user to choose the mixing parameter `alpha`.

Usage

```
choicealpha(D0, D1, range.alpha, K, wt = NULL, scale = TRUE, graph = TRUE)
```

Arguments

`D0`	a dissimilarity matrix of class `dist`. The function `as.dist` can be used to transform an object of class `matrix` to object of class `dist`.
`D1`	an other dissimilarity matrix of class `dist`.
`range.alpha`	a vector of real values between 0 and 1.
`K`	the number of clusters.
`wt`	vector with the weights of the observations. By default, wt=NULL corresponds to the case where all observations are weighted by 1/n.
`scale`	if TRUE the two dissimilarity matrices are scaled i.e. divided by their max.
`graph`	if TRUE, two graphics (proportion and normalized proportion of explained inertia) are drawn.

Value

An object with S3 class "choicealpha" and the following components:

`Q`	a matrix of dimension `length(range.alpha)` times `2` with the proportion of explained inertia calculated with `D0` (first column) and calculated with `D1` (second column)
`Qnorm`	a matrix of dimension `length(range.alpha)` times `2` with the proportion of normalized explained inertia calculated with `D0` (first column) and calculated with `D1` (second column)

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

See Also

`plot.choicealpha`, `hclustgeo`

Examples

```
data(estuary)
D0 <- dist(estuary$dat) # the socio-demographic distances
D1 <- as.dist(estuary$D.geo) # the geographic distances between the cities
range.alpha <- seq(0,1,0.1)
K <- 5
cr <- choicealpha(D0,D1,range.alpha,K,graph=TRUE)
cr$Q # proportion of explained pseudo inertia
cr$Qnorm # normalized proportion of explained pseudo inertia
```

`estuary` *estuary data*

Description

Data refering to n=303 French municipalities of gironde estuary (a south-ouest French county). The data are issued from the French population census conducted by the National Institute of Statistics and Economic Studies. The dataset is an extraction of four quantitative socio-economic variables for a subsample of 303 French municipalities located on the atlantic coast between Royan and Mimizan. `employ.rate.city` is the employment rate of the municipality, that is the ratio of the number of individuals who have a job to the population of working age (generally defined, for the purposes of international comparison, as persons of between 15 and 64 years of age). `graduate.rate` refers to the level of education of the population that is the highest degree declared by the individual. It is defined here as the ratio for the whole population having completed a diploma equivalent or of upper level to two years of higher education (DUT, BTS, DEUG, nursing and social training courses, license, maitrise, master, DEA, DESS, doctorate, or Grande Ecole diploma). `housing.appart` is the ratio of apartment housing. `agri.land` is the part of agricultural area of the municipality.

Format

The R dataset estuary is a list of three objects:

- dat: a data frame with the description of the n=303 municipalities on p=4 socio-demographic variables.
- D.geo: a matrix with the geographical distances between the town hall of the n=303 municipalities.
- map: an object of class `SpatialPolygonsDataFrame` with the map of the gironde estuary.

Source

Original data are issued from the French population census of National Institute of Statistics and Economic Studies for year 2009. The agricultural surface has been calculated on data coming from the French National Institute of Geographical and Forestry Information. The calculation of the ratio and recoding of categories have been made by Irstea Bordeaux.

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

Examples

```
data(estuary)
names(estuary)
head(estuary$dat)
```

`hclustgeo` *Ward clustering with soft contiguity contraints*

Description

Implements a Ward-like hierarchical clustering algorithm including soft contiguity constraints. The algorithm takes as input two dissimilarity matrices `D0` and `D1` and a mixing parameter alpha between 0 an 1. The dissimilarities can be non euclidean and the weights of the observations can be non uniform. The first matrix gives the dissimilarities in the "feature space". The second matrix gives the dissimilarities in the "constraint" space. For instance, `D1` can be a matrix of geographical distances or a matrix build from a contiguity matrix. The mixing parameter `alpha` sets the importance of the constraint in the clustering process.

Usage

```
hclustgeo(D0, D1 = NULL, alpha = 0, scale = TRUE, wt = NULL)
```

Arguments

`D0`	an object of class `dist` with the dissimilarities between the n observations. The function `as.dist` can be used to transform an object of class `matrix` to object of class `dist`.
`D1`	an object of class "dist" with other dissimilarities between the same n observations.
`alpha`	a real value between 0 and 1. This mixing parameter gives the relative importance of `D0` compared to `D1`. By default, this parameter is equal to 0 and `D0` is used alone in the clustering process.
`scale`	if TRUE the two dissimilarity matric `D0` and `D1` are scaled i.e. divided by their max. If `D1`=NULL, this parameter is no used and D0 is not scaled.
`wt`	vector with the weights of the observations. By default, wt=NULL corresponds to the case where all observations are weighted by 1/n.

Details

The criterion minimized at each stage is a convex combination of the homogeneity criterion calculated with `D0` and the homogeneity criterion calculated with `D1`. The parameter `alpha` (the weight of this convex combination) controls the importance of the constraint in the quality of the solutions. When `alpha` increases, the homogeneity calculated with `D0` decreases whereas the homogeneity calculated with `D1` increases.

Value

Returns an object of class `hclust`.

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

See Also

`choicealpha`

Examples

```
data(estuary)
# with one dissimilarity matrix
w <- estuary$map@data$POPULATION # non uniform weights
D <- dist(estuary$dat)
tree <- hclustgeo(D,wt=w)
sum(tree$height)
inertdiss(D,wt=w)
inert(estuary$dat,w=w)
plot(tree,labels=FALSE)
part <- cutree(tree,k=5)
sp::plot(estuary$map, border = "grey", col = part)

# with two dissimilarity matrix
D0 <- dist(estuary$dat) # the socio-demographic distances
D1 <- as.dist(estuary$D.geo) # the geographical distances
alpha <- 0.2 # the mixing parameter
tree <- hclustgeo(D0,D1,alpha=alpha,wt=w)
plot(tree,labels=FALSE)
part <- cutree(tree,k=5)
sp::plot(estuary$map, border = "grey", col = part)
```

inert *Inertia of a cluster*

Description

Computes the inertia of a cluster i.e. on a subset of rows of a data matrix.

Usage

```
inert(
  Z,
  indices = 1:nrow(Z),
  wt = rep(1/nrow(Z), nrow(Z)),
  M = rep(1, ncol(Z))
)
```

Arguments

Z	matrix data
indices	vectors representing the subset of rows
wt	weight vector
M	diagonal distance matrix

Examples

```
data(estuary)
n <- nrow(estuary$dat)
Z <- scale(estuary$dat)*sqrt(n/(n-1))
inert(Z) # number of variables

w <- estuary$map@data$POPULATION # non uniform weights
inert(Z,wt=w)
```

inertdiss *Pseudo inertia of a cluster*

Description

The pseudo inertia of a cluster is calculated from a dissimilarity matrix and not from a data matrix.

Usage

```
inertdiss(D, indices = NULL, wt = NULL)
```

Arguments

D	an object of class "dist" with the dissimilarities between the n observations. The function as.dist can be used to transform an object of class matrix to object of class "dist".
indices	a vector with the indices of the subset of observations.
wt	vector with the weights of the n observations

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

Examples

```
data(estuary)
n <- nrow(estuary$dat)
Z <- scale(estuary$dat)*sqrt(n/(n-1))
inertdiss(dist(Z)) # pseudo inertia
inert(Z) #equals for euclidean distance

w <- estuary$map@data$POPULATION # non uniform weights
inertdiss(dist(Z),wt=w)
```

---

plot.choicealpha *Plot to choose the mixing parameter*

Description

Plot two curves of explained inertia (one for D0 and one for D1) calculated with choicealpha.

Usage

```
## S3 method for class 'choicealpha'
plot(
  x,
  norm = FALSE,
  lty = 1:2,
  pch = c(8, 16),
  type = c("b", "b"),
  col = 1:2,
  xlab = "alpha",
  ylab = NULL,
  legend = NULL,
  cex = 1,
  ...
)
```

Arguments

x	an object of class choicealpha.
norm	if TRUE, the normalized explained inertia are plotted. Otherwise, the explained inertia are plotted.
lty	a vector of size 2 with the line types of the two curves. See par
pch	a vector of size 2 specifying the symbol for the points of the two curves. See par
type	a vector of size 2 specifying the type of lines of the two curves. See par
col	a vector of size 2 specifying the colors the two curves. See par
xlab	the title fot the x axis.
ylab	the title fot the y axis.
legend	a vector of size two the the text for the legend of the two curves.
cex	text size in the legend.
...	further arguments passed to or from other methods.

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

See Also

choicealpha

Examples

```
data(estuary)
D0 <- dist(estuary$dat)
D1 <- as.dist(estuary$D.geo) # the geographic distances between the cities
range.alpha <- seq(0,1,0.1)
K <- 5
cr <- choicealpha(D0,D1,range.alpha,K,graph=FALSE)
plot(cr,cex=0.8,norm=FALSE,cex.lab=0.8,ylab="pev",
         col=3:4,legend=c("socio-demo","geo"), xlab="mixing parameter")
plot(cr,cex=0.8,norm=TRUE,cex.lab=0.8,ylab="pev",
         col=5:6,pch=5:6,legend=c("socio-demo","geo"), xlab="mixing parameter")
```

---

wardinit *Ward aggregation measures between singletons*

Description

This function calculates the Ward aggregation measures between pairs of singletons.

Usage

```
wardinit(D, wt = NULL)
```

Arguments

D	a object of class "dist" with the dissimilarities between the n obsevations. The function as.dist can be used to transform an object of class matrix to object of class "dist".
wt	vector with the weights of the observations. By default, wt=NULL corresponds to the case where all observations are weighted by 1/n.

Details

The Ward agreggation measure between to singletons i and j weighted by wi and wj is : (wiwj)/(wi+wj)dij^2 where dij is the dissimilarity between i and j.

Value

Returns an object of class dist with the Ward aggregation measures between the n singletons.

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

---

withindiss *Dissimilarity based pseudo within-cluster inertia of a partition*

Description

This function performs the pseudo within-cluster inertia of a partition from a dissimilarity matrix.

Usage

```
withindiss(D, part, wt = NULL)
```

Arguments

D	an object of class "dist" with the dissimilarities between the n observations. The function as.dist can be used to transform an object of class matrix to object of class "dist".
part	a vector with group membership.
wt	vector with the weights of the observations

References

M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. ClustGeo: an R package for hierarchical clustering with spatial constraints. Comput Stat (2018) 33: 1799-1822.

Index