Package ‘ClustGeo’

July 14, 2017

Type Package
Title Hierarchical Clustering with Spatial Constraints
Version 2.0
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 spatial/geographical constraints.
Depends R (>= 3.0.0)
Imports sp, spdep
License GPL (>= 2.0)
LazyData true
Suggests knitr, rmarkdown
VignetteBuilder knitr
RoxygenNote 5.0.1
NeedsCompilation no
Repository CRAN
Date/Publication 2017-07-14 13:40:16 UTC

R topics documented:

  choicealpha .................................................. 2
  estuary ....................................................... 3
  hclustgeo .................................................... 4
  inert .......................................................... 5
  inertdiss ...................................................... 6
  plot.choicealpha ............................................. 7
  wardinit ....................................................... 8
  withinindiss ................................................ 8
Empirical choice of the mixing parameter

Description

This function calculates the proportion (resp. normalized proportion) of explained inertia of the partitions in K clusters obtained with the Ward-like hclustgeo procedure for a range of mixing parameters alpha. When the proportion (resp. normalized proportion) of explained inertia based on D0 decreases, the proportion (resp. normalized proportion) of explained inertia based on D1 increases. The plot of these criteria can help the user in the choice of the mixing parameter alpha.

Usage

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

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.
- **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 matrix are scaled i.e. divided by their max.
- **graph**: if TRUE the two graphics (proportion and normalized proportion of explained inertia) are drawn.

References


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
Description

Data referring to n=303 French municipalities of Gironde estuary (a south-west 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


Examples

```r
data(estuary)
names(estuary)
head(estuary$dat)
sp::plot(estuary$map)
```
**Description**

This function implements a Ward-like hierarchical clustering algorithm including soft contiguity constraints. This algorithm takes as input two dissimilarity matrices $D_0$ and $D_1$ and a mixing parameter $\alpha$ between 0 and 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" (socio-demographic variables or grey levels for instance). The second matrix gives the dissimilarities in the "constraint" space. For instance, $D_1$ can be a matrix of geographical distances or a matrix build from the contiguity matrix $C$. The mixing parameter $\alpha$ sets the importance of the constraint in the clustering procedure.

**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 $D_0$ compared to $D_1$. By default, this parameter is equal to 0 and $D_0$ is used alone in the clustering process.
- **scale**: if TRUE the two dissimilarity matrix $D_0$ and $D_1$ are scaled i.e. divided by their max. If $D_1$=NULL, this parameter is no used and $D_0$ 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 $D_0$ and the homogeneity criterion calculated with $D_1$. The parameter $\alpha$ (the weight of this convex combination) controls the weight of the constraint in the quality of the solutions. When $\alpha$ increases, the homogeneity calculated with $D_0$ decreases whereas the homogeneity calculated with $D_1$ increases.

**Value**

Returns an object of class `hclust`.
inert

References


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 vector representing the subset of rows
wt weight vector
M diagonal distance matrix
Examples

```r
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

```r
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


Examples

```r
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 of the mixing parameter*

**Description**

Plot of the criterion Q or Qnorm obtained in output of the function `choicealpha`.

**Usage**

```r
## 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 criterion Qnorm is plotted. Otherwise, it is Q.
- `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 for the x axis.
- `ylab` the title for 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.

**See Also**

`choicealpha`

**Examples**

```r
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")
```
\texttt{wardinit} \hspace{1cm} \textit{Ward aggregation measures between singletons}

\textbf{Description}

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

\textbf{Usage}

\texttt{wardinit(D, wt = NULL)}

\textbf{Arguments}

- \texttt{D} a object of class "dist" with the dissimilarities between the \textit{n} observations. The function \texttt{as.dist} can be used to transform an object of class "matrix" to object of class "dist".

- \texttt{wt} vector with the weights of the observations. By default, \texttt{wt=NULL} corresponds to the case where all observations are weighted by 1/n.

\textbf{Details}

The Ward aggregation measure between two singletons \textit{i} and \textit{j} weighted by \textit{w\textsubscript{i}} and \textit{w\textsubscript{j}} is: \( \frac{w\textsubscript{i} w\textsubscript{j}}{w\textsubscript{i} + w\textsubscript{j}} d\textsubscript{ij}^2 \)

where \( d\textsubscript{ij} \) is the dissimilarity between \textit{i} and \textit{j}.

\textbf{Value}

Returns an object of class \texttt{dist} with the Ward aggregation measures between the \textit{n} singletons.

\textbf{References}


\texttt{withindiss} \hspace{1cm} \textit{Dissimilarity based pseudo within-cluster inertia of a partition}

\textbf{Description}

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

\textbf{Usage}

\texttt{withindiss(D, part, wt = NULL)}
Arguments

D    an object of class "dist" with the dissimilarities between the n observations. The function \texttt{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

Index

*Topic data
  estuary, 3

as.dist, 2, 4, 6, 8, 9
choicealpha, 2, 7
estuary, 3
hclust, 4
hclustgeo, 4
inert, 5
inertdiss, 6
par, 7
plot.choicealpha, 7
wardinit, 8
withindiss, 8