Package 'KnowBR'

July 21, 2025

Version 2.2

Title Discriminating Well Surveyed Spatial Units from Exhaustive Biodiversity Databases

Author Castor Guisande Gonzalez and Jorge M. Lobo

Maintainer Castor Guisande Gonzalez <castor@uvigo.es>

Description It uses species accumulation curves and diverse estimators to assess, at the same time, the levels of survey coverage in multiple geographic cells of a size defined by the user or polygons. It also enables the geographical depiction of observed species richness, survey effort and completeness values including a background with administrative areas.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.0), fossil, mgcv, plotrix, sp, vegan

Suggests raster, rgbif, usdm, car, IDPmisc, psych, candisc

NeedsCompilation no

Repository CRAN

Date/Publication 2023-10-07 06:30:28 UTC

Contents

adworld *Geographical coordinates*

Description

Latitude and longitude of all administrative areas.

Usage

```
data(adworld)
```

Format

A matrix of many rows and 3 columns (Latitude, Longitude and name of the administrative area)

Source

Latitude and longitude coordinates of the administrative areas were obtained from the web page https://www.openstreetmap.org.

Beetles *Individual counts of species of beetles*

Description

This database includes 15,142 records belonging to 54 Iberian species of the Scarabaeidae (Coleoptera) previously compiled in the so called BANDASCA database (Lobo & Martín-Piera, 2002) also freely available in GBIF (https://www.gbif.org/). Individual counts, longitude and latitude of species occurrences of the family Scarabaeidae in the Iberian Peninsula are provided.

Usage

```
data(Beetles)
```

Format

A matrix with four columns: species, longitude, latitude and individual counts.

References

Lobo, J.M. & Martín-Piera, F. 2002. Searching for a predictive model for Iberian dung beetle species richness based on spatial and environmental variables. *Conservation Biology* 16: 158-173.

Estimators *Estimators obtained with the function KnowBPolygon*

Description

Estimators obtained with the function KnowBPolygon using species of freshwater fish species in all the countries of the world (Guisande et al., 2010).

Usage

```
data(Estimators)
```

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, L.F., Duque, S. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

FishIrelandUK *Estimators obtained with the function KnowB*

Description

Estimators obtained with the function KnowB using a database that includes 121,709 records of freshwater fishes obtained from Pelayo-Villamil et al. (2015), with the Clench estimator and a cell resolution of 5'.

Usage

```
data(FishIrelandUK)
```

References

Pelayo-Villamil, P., Guisande, C., Vari, R.P., Manjarrés-Hernández, A., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Patti, B., Quinci, E.M., Jiménez, L.F., Granado-Lorencio, C., Tedesco, P.A., Lobo, J.M. (2015) Global diversity patterns of freshwater fishes-Potential victims of their own success. *Diversity and Distributions*, 21: 345-356.

KnowB *Discriminating well surveyed cell units from exhaustive biodiversity databases*

Description

Advances during the last decades in information technology allow us to store, retrieve, transmit and manipulate an unprecedented magnitude of massive information about species distributions (Guralnick *et al*., 2007). Unfortunately, this compilation process suffers from three main shortcomings:

i) *Unknown survey effort*. A lack of knowledge of the effort devoted to survey each territorial unit that is due to most occurrence records lacking any associated measure of the effort carried out to obtain them.

ii) *Unknown absences*. As almost all the available information involves only species occurrences (i.e., the localities in which a species has been collected), without any indication of the likelihood that a species is actually absent from the localities where it was not collected (whether these have been surveyed or not).

iii) *Unknown recurrence*. Which results from the incomplete compilation of species occurrences in many biodiversity databases, as multiple records of the same species in the same site or territorial unit are considered redundant and not reported (Hortal *et al*. 2007); this prevents teasing apart occasional records from the continued presence of the species in an area.

These three limitations are mutually interrelated, so when all known occurrences are compiled exhaustively it is possible to estimate survey effort with some reliability. Therefore, a biodiversity database that compiles exhaustively all available information on the identity and distribution of a group of species would enable both identifying well-surveyed areas (e.g. Hortal and Lobo 2005) and obtaining estimates of the repeated occurrence and/or the probability of absence of particular species (e.g. Guillera-Arroita *et al*. 2010).

Employing statistical shortcuts on data with unknown levels of error and bias can generate unreliable results. Consequently, good practice in biodiversity informatics requires knowledge about the number, location and degree of completeness of surveys for those territorial units that have been, at least relatively, well inventoried. Such knowledge would facilitate identifying localities where the lack of records for a target species can be reliably assumed to correspond to its actual absence. Nonetheless, it can be used to guide the location of future surveys and/or determine uncertain or ignorance areas in which biodiversity data are insufficiently consistent (Hortal and Lobo 2005, Ladle and Hortal 2013, Hortal *et al*. 2015, Ruete 2015, Meyer *et al*. 2015, 2016).

Despite the widely recognized importance of evaluating data quality as a preliminary step in any biodiversity study, this process is often neglected. Arguably, this is in part because such evaluation process is highly time-consuming, for it requires using analyses spread over several software applications and/or R packages, and repeating the same process for each one of the territorial units or sites considered (or, in general, for any type of spatial unit). Here we present KnowBR, a freely available R package to estimate the survey coverage of species inventories across an unlimited number of territorial units or sites simultaneously. Starting with any biodiversity database, KnowBR calculates the survey coverage per spatial unit as the final slope of the relationship between the number of collected species and the number of database records, which is used as a surrogate of the survey effort. To do this, KnowBR estimates the accumulation curve (the accumulated increase in the number of

species with the addition of database records) for each one of the spatial units according to the *exact* estimator of Ugland *et al*. (2003), as well as performing 200 permutations of the observed data (*random* estimator) to obtain a smoothed accumulation curve. This curve is subsequently adjusted to four different functions with three or less parameters, and the obtained extrapolated asymptotic value used to obtain a completeness percentage (the percentage representing the observed number of species against the predicted one) that also may be used to estimate the territorial units with probable reliable inventories.

These territorial units can be regular cells of any resolution (*cell* option) but also irregular polygons (*polygon* option) according to user preferences. RWizard includes in the "Area" argument the possibility of select the administrative spatial units (countries, regions, departments and/or provinces) or the rivers basins of different levels in which to perform the calculations. Instead of using the polygons available in RWizard, the user may also include any shapefile containing the desired irregular polygons (e.g. protected areas, countries, etc) by means of the "shape" argument.

Usage

```
KnowB(data, format="A", cell=60, curve= "Rational", estimator=1, cutoff=1,
cutoffCompleteness= 0, cutoffSlope= 1, largematrix=FALSE, Area="World",
extent=TRUE, minLon, maxLon, minLat, maxLat, colbg="transparent",
colcon="transparent", colf="black", pro=TRUE, inc=0.005, exclude=NULL,
colexc=NULL, colfexc="black", colscale=c("#C8FFFFFF","#64FFFFFF","#00FFFFFF",
"#64FF64FF","#C8FF00FF","#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"),
legend.pos="y", breaks=9, xl=0, xr=0, yb=0, yt=0, asp, lab=NULL, xlab="Longitude",
ylab="Latitude", main1="Observed richness", main2="Records", main3="Completeness",
main4="Slope", cex.main=1.6, cex.lab=1.4, cex.axis=1.2, cex.legend=1.2,
family="sans", font.main=2, font.lab=1, font.axis=1, lwdP=0.6, lwdC=0.1,
trans=c(1,1), log=c(0,0), ndigits=0, save="CSV", file1="Observed richness",
file2="List of species", file3="Species per site", file4="Estimators",
file5="Species per record", file6="Records", file7="Completeness", file8="Slope",
file9="Standard error of the estimators", na="NA", dec=",", row.names=FALSE,
jpg=TRUE, jpg1="Observed richness.jpg", jpg2="Records.jpg", jpg3="Completeness.jpg",
jpg4="Slope.jpg",cex=1.5, pch=15, cex.labels=1.5, pchcol="red", ask=FALSE)
```

Arguments

`data` The data is introduced as a CSV, TXT or RData file following two simple formats: one in which only four columns are included (see format A; species name, longitude, latitude and a number reflecting the incidence of the species) and another one including the longitude and latitude of each spatial unit and as many columns as species (see format B in the following table). The CSV file with the format A may be obtained using ModestR (see details).

Format A

Species	Longitude	Latitude	Counts
Sp1	-79.33	22.00	1
Sp2	-85.91	16.46	1
Sp2	-85.90	16.48	2
Sp2	-64.74	18.33	1
Sp3	-84.21	22.46	4

Format B

Longitude	Latitude	Sp1	Sp2	Sp3
-79.33	22.00	1	0	0
-85.91	16.46	0	1	0
-85.90	16.48	0	2	0
-64.74	18.33	0	1	0
-84.21	22.46	0	0	4

The primary matrix used in *KnowBR* has a special characteristic - it must be derived from an exhaustive database including all the available georeferenced information including even those apparently redundant records of a species from the same locality provided that is a difference in some of the collection conditions for a species at a locality (i.e. date of capture, food source, collector, type of microhabitat, etc.). Thus, any difference in any database field value yields a new database record regardless of the number of individuals (see for example Lobo & Martín-Piera, 2002). As biodiversity data can derive from heterogeneous sources with different collector methodologies, no universal sampling effort measure capable of offering reliable comparisons exists and the number of database records is used as a surrogate (see Soberón *et al*., 2007; Lobo, 2008). This approach is particularly appropriate for poorly surveyed groups and/or regions lacking sufficient information to correct unequal sampling efforts arising from standardized survey protocols.

`format` If it is "A" (default), the format of the data frame is species, longitude, latitude and a count value (format A of the table showed above). If it is "B" the format of the data frame is longitude, latitude and the rest of columns are the presence of the species in each site (format B of the table showed above). If numeric values higher than 1 are included in these data a (Count or Sp columns), a database record is considered for each unit. This in the example of format A, four different records are included for the Sp3 with same geographical coordinates.

`cell` Resolution of the cells (spatial units) in minutes on which calculations were carried out. In the present version the user can select any resolution between 1 and 60 minutes.

`curve` The smoothed accumulation curve generated by the accumulation curve can be adjusted to a "Clench", "Exponential", "Saturation" or "Rational" function (see equations in details section), calculating the asymptotic extrapolated values to further derive a completeness percentage (the percentage representing the observed number of species against the predicted one).

`estimator` Vector that defines the used estimator:

0 The data for the estimation of the accumulation curve and the final slope are obtained with both the "exact" and "random" procedures. When the predicted richness is estimated with the type of curve selected by the user ("Clench", "Nexponential", "Saturation" or "Rational") using the data generated by the methods "exact" and "random" at the same time, the mean of both richness values is used to calculate completeness (the percentage representing the observed number of species against the predicted one).

1 It is the "exact" estimator of Ugland *et al.* (2003) (default option) to obtain a smoothed accumulation curve.

2 If the chosen option is "random". It adds records at random performing 200 permutations in the order of records entry to generate the accumulation curve.

`cutoff` This number reflects the ratio between the number of database records and the number of species. If this ratio is lower than the selected threshold value in each considered spatial unit, any one of the estimators will be calculated and these spatial units are considered as lacking information.

`cutoffCompleteness`

If the value of completeness is lower than this threshold, the completeness is not

	calculated.
cutoffSlope	If the slope is higher than this threshold, the completeness is not calculated.
largematrix	When there many species and/or many records resulting in a species per record matrix with more than 2^{31} cells, it is impossible to create the CSV or RData file with the species per record due to memory limits in R. If this argument is TRUE, the function creates a TXT file with the species per record, but the process is computationally intensive, so it may takes several hours and it may create a large TXT file. The default value is FALSE.
Area	A character with the name of the administrative area or a vector with several administrative areas. If a vector with several administrative areas are used, it is necessary to use RWizard (see details).
extent	If TRUE the minimum and maximum longitudes and latitudes are delimited by the minimum and maximum of the data (default). If FALSE the minimum and maximum longitudes and latitudes are delimited by the arguments Area and, minLat, maxLat, minLon and maxLon.
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude (see details).
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude (see details).
colbg	Background color of the map (in some cases this is the sea).
colcon	Background color of the administrative areas.
colf	Color of administrative areas border.
pro	If it is TRUE an automatic calculation is made in order to correct the aspect ratio y/x along latitude.
inc	Adds some room along the map margins with the limits x and y thus not exactly the limits of the selected areas.
exclude	A character with the name of the administrative area or a vector with several administrative areas that can be plotted with a different color on the map.
colexc	Background color of areas selected in the argument exclude.
colfexc	Color of borders of the areas selected in the argument exclude.
colscale	Palette color.
legend.pos	Whether to have a horizontal (x) or vertical (y) color gradient.
breaks	Number of breakpoints of the color legend.
xl, xr, yb, yt	The lower left and upper right coordinates of the color legend in user coordinates.
asp	The y/x aspect ratio.
lab	A numerical vector of the form c(x. Y) which modified the default method by which axes are annotated. The values of x and y give the (approximate) number of tick marks on the x and y axes.
xlab	A title for the x axis.
ylab	A title for the y axis.

main1	An overall title for the plot of the observed species richness.
main2	An overall title for the plot of the records.
main3	An overall title for the plot of the completeness.
main4	An overall title for the plot of the slope between the last species richness value and the previous value for each one of the accumulation methods.
cex.main	The magnification to be used for main titles relative to the current setting of cex.
cex.lab	The magnification to be used for x and y labels relative to the current setting of cex.
cex.axis	The magnification to be used for axis annotation relative to the current setting of cex.
cex.legend	The magnification to be used in the numbers of the color legend relative to the current setting of cex.
family	The name of a font family for drawing text.
font.main	The font to be used for plot main titles.
font.lab	The font to be used for x and y labels.
font.axis	The font to be used for axis annotation.
lwdP	Line width of the plot.
lwdC	Line width of the borders.
trans	It is possible to multiply or divide the dataset by a value. For a vector with two values, the first may be 0 (divide) or 1 (multiply), and the second number is the value of the division or multiplication.
log	It is possible to apply a logarithmic transformation to the dataset. For a vector with two values, the first may be 0 (do not log transform) or 1 (log transformation), and the second number is the value to be added in case of log transformation.
ndigits	Number of decimals in legend of the color scale.
save	If "CSV" the files are save as CSV and if "RData" the files are save as RData.
file1	RData or CSV file. A character string naming the file with the observed richness.
file2	RData or CSV file. A character string naming the file with the list of species.
file3	RData or CSV file. A character string naming the file with the species incidences per site.
file4	RData or CSV file. A character string naming the file with the estimators per site.
file5	RData or CSV file. A character string naming the file with the species per records.
file6	RData or CSV file. A character string naming the file with the records.
file7	RData or CSV file. A character string naming the file with the completeness.
file8	RData or CSV file. A character string naming the file with the slopes of the accumulation analyses.
file9	RData or CSV file. A character string naming the file with the standard error of the estimators.

na	CSV FILE. Text that is used in the cells without data.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
row.names	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.
jpg	If TRUE the plots are exported to jpg files instead of using the windows device.
jpg1	Name of the jpg file with the values of the observed richness.
jpg2	Name of the jpg file with the records.
jpg3	Name of the jpg file with the completeness.
jpg4	Name of the jpg file with the slopes of the accumulation analyses.
cex	A numerical value giving the amount by which plotting symbols should be magnified relative to the default in the correlation matrix plot.
pch	Either an integer specifying a symbol or a single character to be used as the default in plotting points in the correlation matrix plot.
cex.labels	Size of labels in the correlation matrix plot.
pchcol	Color of the symbols in the correlation matrix plot.
ask	If TRUE (and the R session is interactive) the user is asked for input before a new figure is drawn.

Details

The CSV file required in the argument *data* with the format A (species, longitude, latitude and count) may be obtained using ModestR (available at the web site www.ipez.es/ModestR) juast selecting Export/Export maps of the select branch/To RWizard Applications/To KnowBR.

In ModestR is possible to export the valid samples or pseudosamples. The pseudosamples are grid cells for instance of 5' x 5', 30' x 30', 1º x 1º, etc. Therefore, the output of ModestR is a list of species within each of the grid cells with the cell size defined by the user. It is therefore possible to obtain the number of records for each species within the grid cell or just the records available for all the species, with the format described above.

Area = "World" to plot the entire world. If the coordinates minLon, maxLon, minLat and maxLat are not specified, they are calculated automatically based on the selected administrative areas. If some administrative areas are selected, e.g. some countries, so the argument is not "World", it only works with RWizard.

It is important to emphasize that the quality of the geographical records of the administrative areas is lower if it is used the entire world(*Area="World"*, because the file adworld is used), than if it is selected some countries, departments, etc., because the geographical records of the administrative areas available in RWizard are used. It means that the records inside the polygons may vary depending on the selection specified in the argument *Area*.

The type of curves are:

1) The curve of Clench (Clench, 1979), which is a modification of the function of Monod (Monod, 1950), and was proposed to butterflies.

2) The exponential (Miller & Wiegert, 1989) that was proposed for rare plant species.

3) The saturation curve that was used to show the relationship between growth of phytoplankton, a toxic algae of the genus *Alexandrium* and the concentration of phosphate (Frangópulos et al., 2004), which is similar to von Bertalanffy growth curve but adapting the coefficients to better explain the pattern of accumulation function.

4) The rational function (Ratkowski, 1990) that can be used when there is no clear criterion which model to use (Falther, 1996).

Name	**Function**	**Reference**
Clench	$y = \frac{ax}{1+bx}$	(Clench, 1979)
Negative exponential	$y = a\left(1 - e^{-bx}\right)$	(Miller & Wiegert, 1989)
Saturation	$y = a\left(1 - e^{-b(x-c)}\right)$	(Frangópulos et al., 2004)
Rational	$y = \frac{(a+bx)}{(1+cx)}$	(Ratkowski, 1990)

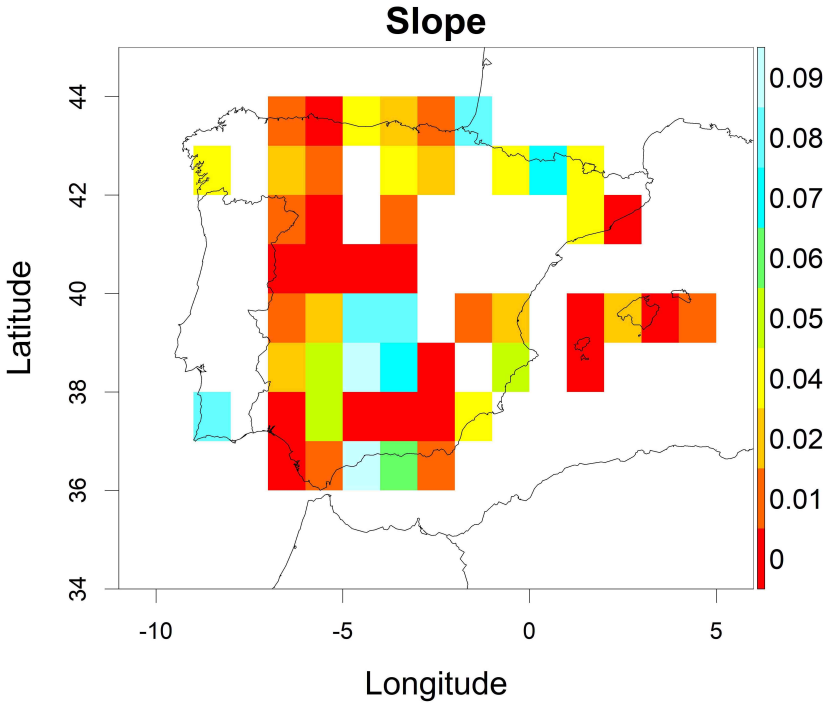
FUNCTIONS

The estimators exact and random were estimated with the function specaccum of the package vegan (Oksanen *et al*., 2014).

The color legend of the maps is depicted with the function color.legend of the package plotrix (Lemon et al., 2014).

EXAMPLE

The database of the example includes 15,142 records for the 54 Iberian species of the Scarabaeidae (Coleoptera) previously compiled in the so called BANDASCA database (Lobo & Martín-Piera, 2002). The following map show the slopes obtained in cells of 60ºx 60º using the estimator exact and the Rational's curve.



The maps may be easily modified using the function MapCell using the exported CSV or RData files detailing the observed species richness (with alias ObservedRichness), the records (with alias Records), the completeness (with alias Completeness) and the slope (with alias Slope).

Value

RData or CSV files: 1) Observed richness, 2) List of species, 3) Species per site, 4) Estimators, 5) Species per record, 6) Records, 7) Completeness, Slope and 9) Standard error of the estimators.

JPG files with maps: 1) Observed richness, 2) Records, 3) Completeness and 4) Slope.

Source

Spatial database of the location of the world's administrative areas (or administrative boundaries) was obtained from the Web Site http://www.openstreet.org/.

References

Clench, H.K. (1979) How to make regional lists of butterflies: some invoking empirically based criteria in selecting among thoughts. *The Journal of the Lepidopterists' Society*, 33: 216-231.

Flather, C.H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, 23: 155-168.

Frangópulos, M., Guisande, C., deBlas, E. y Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae*, 3: 131-139.

Guillera-Arroita, G., Ridout, M.S & Morgan, B.J.T. (2010)Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, 1: 131-139.

Guralnick, R.P., Hill, A.W. & Lane, M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10: 663-672.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review Ecology and Systematics*, 46: 523-549.

Hortal, J. & Lobo, J.M. 2005. An ED-based protocol for the optimal sampling of biodiversity. *Biodiversity and Conservation*, 14: 2913-2947.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A., 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). *Conservation Biology* 21, 853-863.

Ladle, R. & Hortal, J. (2013) Mapping species distributions: living with uncertainty. *Frontiers of Biogeography*, 5: 8-9.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duursma, R., Biancotto, E., Levy, O., Dutang, C., Solymos, P., Engelmann, R., Hecker, M., Steinbeck, F., Borchers, H., Singmann, H., Toal, T. & Ogle, D. (2017). Various plotting functions. R package version 3.6-5. Available at: http://CRAN.R-project.org/package=plotrix.

Lobo, J.M., Baselga, A., Hortal, J., Jiménez-Valverde, A. & Gómez, J.F. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions* 13:772-780.

Lobo, J.M. 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodiversity and Conservation* 17: 873-881.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*6: 8221.

Meyer, C., Weigelt, P. & Kreft, H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecoly Letters*, 19: 992-1006.

Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, 70: 16-22.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H. & Wagner, H. 2014. Community Ecology Package. R package version 2.0-10. Available at: https://CRAN.R-project.org/package=vegan.

Ratkowski, D.A. (1990) *Handbook of nonlinear regression models*. Marcel Dekker, New York, 241 pp.

Ruete, A. (2015) Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal*3: e5361.

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P., 2007. Assessing completeness of biodiversity databases at different spatial scales. *Ecography* 30, 152-160.

Ugland, K.I., Gray J. S. & Ellingsen, K.E. 2003. The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology* 72: 888-897.

Examples

```
## Not run:

#Example 1. Default conditions using estimator 1 (method exact)
#but only slopes lower than 0.1 are selected for depicting
#and, therefore, only the completeness is depicted for those
#cells with the slope lower than 0.1.
#If using RWizard, for a better quality of the geographic
#coordinates, replace data(adworld) by @_Build_AdWorld_

data(adworld)
data(Beetles)
KnowB(data=Beetles, save="RData", jpg=FALSE, cutoffSlope=0.1, xl=6.1, xr=6.3)

#Only to be used with RWizard.
#Example 2. Using @_Build_AdWorld_

data(Beetles)
@_Build_AdWorld_
KnowB(Beetles, cell=15, save="RData")


## End(Not run)
```

KnowBPolygon *Discriminating well surveyed polygon units from exhaustive biodiversity databases*

Description

It is the same function than KnowB but the estimation of the well surveyed units is on polygons instead of on cells.

Usage

```
KnowBPolygon(data, format="A", shape=NULL, shapenames=NULL, admAreas=FALSE,
Area="World", curve="Rational", estimator=1, cutoff=1, cutoffCompleteness=0,
cutoffSlope=1, extent=TRUE, minLon, maxLon, minLat, maxLat, int=30, colbg="#FFFFFF",
colcon="#C8C8C8", colf="black", pro = TRUE, inc=0.005, exclude=NULL, colexc=NULL,
colfexc="black", colscale=c("#C8FFFFFF","#64FFFFFF",
"#00FFFFFF","#64FF64FF","#C8FF00FF","#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"),
legend.pos="y", breaks=9, xl=0, xr=0, yb=0, yt=0, asp, lab=NULL,
xlab="Longitude", ylab="Latitude", main1="Records", main2="Observed richness",
main3="Completeness", main4="Slope", cex.main=1.6, cex.lab=1.4, cex.axis=1.2,
cex.legend=0.9, family="sans", font.main=2, font.lab=1, font.axis=1,
lwdP=0.6, lwdC=0.1, trans=c(1,1), ndigits=0, save="CSV", file1="Species per site",
file2="Estimators", file3="Standard error of the estimators", na="NA",
dec=",", row.names=FALSE, Maps=TRUE, jpg=TRUE, jpg1="Records.jpg",
```

```
jpg2="Observed richness.jpg",  jpg3="Completeness.jpg", jpg4="Slope.jpg",
cex=1.5, pch=15,cex.labels=1.5, pchcol="red", ask=FALSE)
```

Arguments

data	The data is introduced as a CSV, TXT or RData file following two simple formats (for further details see the description of the same argument in the function KnowB.): one in which only four columns are included (format A; species name, longitude, latitude and a number reflecting the incidence of the species) and another one including the longitude and latitude of each spatial unit and as many columns as species (format B). The CSV file with the format A may be obtained using ModestR (see details).
format	If it is "A" (default), the format of the data frame is species, longitude, latitude and a count value. If it is "B" the format of the data frame is longitude, latitude and the rest of columns are the presence of the species in each site (for further details see the description of the same argument in the function KnowB).
shape	Optionally it may be used a shape file with the information of the polygons.
shapenames	Variable in the shapefile with the names of the polygons.
admAreas	If it is TRUE the border lines of the countries are depicted in the map.
Area	A character with the name of the administrative area or a vector with several administrative areas (see details). If a vector with several administrative areas are used, it is necessary to use RWizard (see details section in the function KnowB).
curve	The smoothed accumulation curve generated by the accumulation curve can be adjusted to a "Clench", "Exponential", "Saturation" or "Rational" function (see equations in details section of function KnowB), calculating the asymptotic extrapolated values to further derive a completeness percentage (the percentage representing the observed number of species against the predicted one).
estimator	Vector that defines the used estimator (see the same argument in the function KnowB).
cutoff	This number reflects the ratio between the number of database records and the number of species. If this ratio is lower than the selected threshold value in each considered spatial unit, any one of the estimators will be calculated and these spatial units are considered as lacking information.
cutoffCompleteness	If the value of completeness is lower than this threshold, the completeness is not calculated.
cutoffSlope	If the slope is higher than this threshold, the completeness is not calculated.
extent	If TRUE the minimum and maximum longitudes and latitudes are delimited by the minimum and maximum of the data (default). If FALSE the minimum and maximum longitudes and latitudes are delimited by the arguments Area and, minLat, maxLat, minLon and maxLon.
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude (see details).
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude (see details).

int	Number of intervals of the color ramp.
colbg	Background color of the map (in some cases this is the sea).
colcon	Background color of the administrative areas.
colf	Color of administrative areas border.
pro	If it is TRUE an automatic calculation is made in order to correct the spect ratio y/x along latitude.
inc	Adds some room along the map margins with the limits x and y thus not exactly the limits of the selected areas.
exclude	A character with the name of the administrative area or a vector with several administrative areas that can be plotted with a different color on the map.
colexc	Background color of areas selected in the argument exclude.
colfexc	Color of borders of the areas selected in the argument exclude.
colscale	Palette color.
legend.pos	Whether to have a horizontal (x) or vertical (y) color gradient.
breaks	Number of breakpoints of the color legend.
xl, xr, yb, yt	The lower left and upper right coordinates of the color legend in user coordinates.
asp	The y/x aspect ratio.
lab	A numerical vector of the form c(x. Y) which modified the default method by which axes are annotated. The values of x and y give the (approximate) number of tick marks on the x and y axes.
xlab	A title for the x axis.
ylab	A title for the y axis.
main1	An overall title for the plot of the observed species richness.
main2	An overall title for the plot of the records.
main3	An overall title for the plot of the completeness.
main4	An overall title for the plot of the slope between the last species richness value and the previous value for each one of the accumulation methods.
cex.main	The magnification to be used for main titles relative to the current setting of cex.
cex.lab	The magnification to be used for x and y labels relative to the current setting of cex.
cex.axis	The magnification to be used for axis annotation relative to the current setting of cex.
cex.legend	The magnification to be used in the numbers of the color legend relative to the current setting of cex.
family	The name of a font family for drawing text.
font.main	The font to be used for plot main titles.
font.lab	The font to be used for x and y labels.
font.axis	The font to be used for axis annotation.
lwdP	Line width of the plot.

lwdC	Line width of the borders.
trans	It is possible to multiply or divide the dataset by a value. For a vector with two values, the first may be 0 (divide) or 1 (multiply), and the second number is the value of the division or multiplication.
ndigits	Number of decimals in legend of the color scale.
save	If "CSV" the files are save as CSV and if "RData" the files are save as RData.
file1	RData or CSV file. A character string naming the file with the species or CSV per site.
file2	RData file or CSV file. A character string naming the file with the estimators per shape.
file3	RData or CSV file. A character string naming the file with the standard error of the estimators.
na	CSV FILE. Text that is used in the cells without data.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
row.names	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.
Maps	If it is TRUE the maps are depicted.
jpg	If TRUE the plots are exported to jpg files instead of using the windows device.
jpg1	Name of the jpg file with the records.
jpg2	Name of the jpg file with of the values of observed richness.
jpg3	Name of the jpg file with the completeness.
jpg4	Name of the jpg file with the slopes of the accumulation analyses.
cex	A numerical value giving the amount by which plotting symbols should be magnified relative to the default in the correlation matrix plot.
pch	Either an integer specifying a symbol or a single character to be used as the default in plotting points in the correlation matrix plot.
cex.labels	Size of labels in the correlation matrix plot.
pchcol	Color of the symbols in the correlation matrix plot.
ask	If TRUE (and the R session is interactive) the user is asked for input before a new figure is drawn.

Details

FUNCTIONS

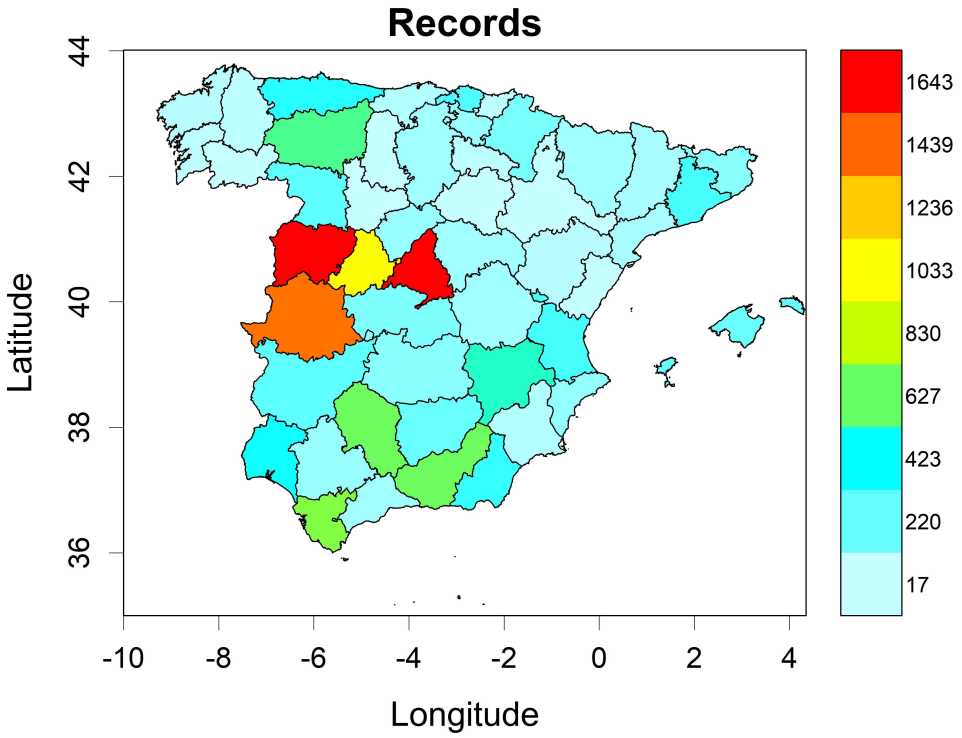
The estimators exact and random were estimated with the function specaccum of the package vegan (Oksanen *et al*., 2014).

The color legend of the maps is depicted with the function color.legend of the package plotrix (Lemon et al., 2014).

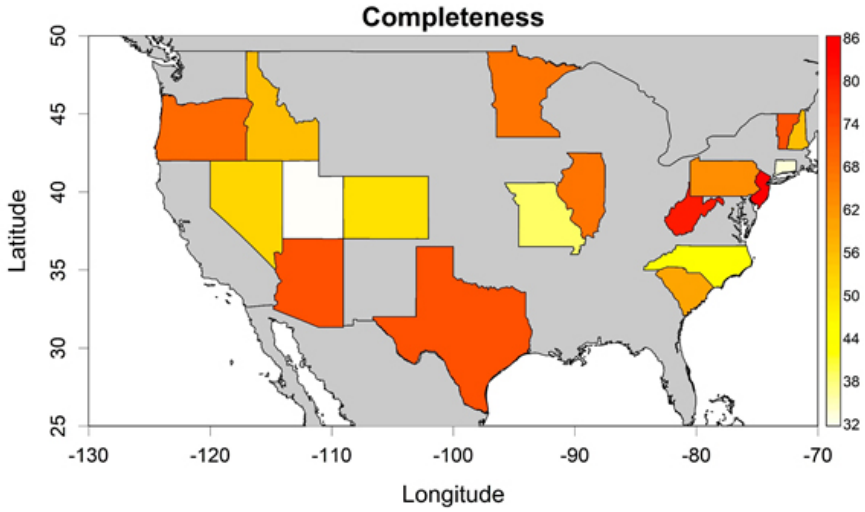
The polygons are depicted with the function spplot of the package sp (Edzer *et al*. 2005; Bivand *et al*., 2013; Pebesma *et al*. 2017).

EXAMPLE

Example 1. The database of the example includes 15,142 records for the 54 Iberian species of the Scarabaeidae (Coleoptera) previously compiled in the so called BANDASCA database (Lobo & MartÃn-Piera, 2002). The following maps show the records obtained in provinces of Spain using the estimator exact and the Rational function to adjust the data.



Example 2. An example using an external shape uploaded by the user (in this case the states of USA). The dataset are the records downloaded from GBIF of the flowering plants of the family Polygonaceae.The states with a grey background had no records, species and/or it was not possible to estimate the slope and/or completeness.



Value

The folowing files are obtained:

RData or CSV files: 1) Species per site, 2) Estimators and 3) Standard error and R2 of the functions.

JPG files with maps: 1)Records, 2) Observed richness, 3) Completeness and 4) Slope.

Source

Spatial database of the location of the world's administrative areas (or administrative boundaries) was obtained from the Web Site http://www.openstreet.org/.

References

Bivand, R.S., Pebesma, E. & Gomez-Rubio, V. 2013. *Applied spatial data analysis with R*. Springer, NY.

Edzer J. Pebesma, E.J. & Bivand, R.S. 2005. Classes and methods for spatial data in R. *R News*, 5: 9-13.

Lobo, J.M. & MartÃn-Piera, F. 2002. Searching for a predictive model for Iberian dung beetle species richness based on spatial and environmental variables. *Conservation Biology*, 16: 158-173.

Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, 70: 16-22.

Monod, J. (1950) La technique de culture continue, theorie et applications. *Annales de l'Institut Pasteur (Paris)*, 79: 390-410.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H. & Wagner, H. 2014. Community Ecology Package. R package version 2.0-10. Available at: https://CRAN.R-project.org/package=vegan.

Pebesma, E., Bivand, R.S., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J. & O'Brien, J. 2017. Classes and Methods for Spatial Data. R package version 1.2-5. Available at: http://CRAN.R-project.org/package=sp.

SoberÃ³n, J. & Llorente, B.J. 1993. The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, 7: 480-488.

Examples

```
## Not run:

#Download records from GBIF of the flowering plants of the family Polygonaceae

library(rgbif)
records<-occ_search(scientificName = "Polygonaceae", limit=5000, return='data',
hasCoordinate=TRUE)

#Data frame with the format A required by the function KnowBPolygon

records<-data.frame(records$species,records$decimalLongitude, records$decimalLatitude)
names(records)<-c("Species","Longitude","Latitude")

#A column is added to the records with the number of counts
#(format A), assuming 1 count per record

dim<-dim(records)
Counts<-rep(1,dim[1])
records<-cbind(records,Counts)

#Running the function

data(States) #State Boundaries of the United States
data(adworld)
KnowBPolygon(data=records, shape=States, admAreas=TRUE, shapenames="NAME", minLon=-130,
maxLon=-70, minLat=25, maxLat=50, colscale=rev(heat.colors(100)), jpg=FALSE)

## End(Not run)
```

`MapCell` *Cell maps*

Description

It allows to depict on a map any of the variables (records richness, observed richess, predicted richness, completeness and the slope) exported by the function KnowB using a CSV, RData or raster file, and with the spatial resolution (cell size) specified in the file.

Usage

```
MapCell(data, Area="World", minLon, maxLon, minLat, maxLat, colbg="#FFFFFF",
colcon="#C8C8C8", colf="black", pro=TRUE, inc=0.005, exclude=NULL,
colexc=NULL, colfexc="black", colscale=c("#C8FFFFFF","#64FFFFFF","#00FFFFFF","#64FF64FF",
"#C8FF00FF","#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"),
legend.pos="y", breaks=9, xl=0, xr=0, yb=0, yt=0, asp, lab=NULL, xlab="Longitude",
```

```
ylab="Latitude", main=NULL,cex.main=1.2, cex.lab=1, cex.axis=0.9, cex.legend=0.9,
family="sans", font.main=2, font.lab=1, font.axis=1, lwdP=0.6, lwdC=0.1, trans=c(1,1),
log=c(0,0), ndigits=0, ini=NULL, end=NULL, jpg=FALSE, filejpg="Map.jpg")
```

Arguments

data	A CSV or RData file exported by the function KnowB (see details) or an ESRI ASCII raster file with the variable (richness, records, etc.).
Area	Only if using RWizard. A character with the name of the administrative area or a vector with several administrative areas (countries, regions, etc.) or river basins. If it is "World" (default) the entire world is plotted. For using administrative areas or river basins, in addition to use RWizard, it is also necessary to replace data(adworld) by @_Build_AdWorld_ (see example 2).
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude.
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude.
colbg	Background color of the map (in some cases this is the sea).
colcon	Background color of the administrative areas.
colf	Color of administrative areas border.
pro	If it is TRUE an automatic calculation is made in order to correct the aspect ratio y/x along latitude.
inc	Adds some room along the map margins with the limits x and y thus not exactly the limits of the selected areas.
exclude	A character with the name of the administrative area or a vector with several administrative areas that may be plotted with a different color on the map (only if using RWizard).
colexc	Background color of areas selected in the argument exclude.
colfexc	Color of borders of the areas selected in the argument exclude.
colscale	Palette color.
legend.pos	Whether to have a horizontal "x" or vertical "y" color scale.
breaks	Number of breakpoints of the color legend.
xl, xr, yb, yt	The lower left and upper right coordinates of the color legend in user coordinates.
asp	The y/x aspect ratio.
lab	A numerical vector of the form c(x, y) which modifies the default way that axes are annotated. The values of x and y give the (approximate) number of tickmarks on the x and y axes.
xlab	A title for the X axis.
ylab	A title for the Y axis.
main	An overall title for the plot.
cex.main	The magnification to be used for main titles relative to the current setting of cex.
cex.lab	The magnification to be used for X and Y labels relative to the current setting of cex.

cex.axis	The magnification to be used for axis annotation relative to the current setting of cex.
cex.legend	The magnification to be used for the color scale relative to the current setting of cex.
family	The name of a font family for drawing text.
font.main	The font to be used for plot main titles.
font.lab	The font to be used for x and y labels.
font.axis	The font to be used for axis annotation.
lwdP	Line width of the plot.
lwdC	Line width of the borders.
trans	It is possible to multiply or divide the dataset by a value. For a vector with two values, the first may be 0 (divide) or 1 (multiply), and the second number is the value of the division or multiplication.
log	It is possible to apply a logarithmic transformation to the dataset. For a vector with two values, the first may be 0 (do not log transform) or 1 (log transformation), and the second number is the value to be added in case of log transformation.
ndigits	Number of decimals in legend of the color scale.
ini	Minimum to be considered in the color scale.
end	Maximum to be considered in the color scale.
jpg	If TRUE the plots are exported to jpg files instead of using the windows device.
filejpg	Name of the jpg file.

Details

It allows to depict on a map any of the files (CSV or RData) exported by the function KnowB: records, observed richness, completeness and slope.

FUNCTIONS

The function color.legend of the package plotrix (Lemon et al., 2014) is used for building the map.

Value

A map is obtained.

References

Lemon, J. (2006) Plotrix: a package in the red light district of R. *R-News*, 6(4):8-12.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duursma, R., Biancotto, E., Levy, O., Dutang, C., Solymos, P., Engelmann, R., Hecker, M., Steinbeck, F., Borchers, H., Singmann, H., Toal, T. & Ogle, D. (2015). Various plotting functions. R package version 3.6-1. Available at: https://CRAN.R-project.org/package=plotrix.

Examples

```
## Not run:

#Example 1. Observed pecies richness of freshwater fishes around the world.

data(RFishes)
data(adworld)
MapCell(data=RFishes, main= "Species richness of freshwater fishes")



#Example 2. Only to be used with RWizard.

data(RFishes)
@_Build_AdWorld_
MapCell(data = RFishes , Area = c("Argentina", "Bolivia", "Brazil", "Chile",
"Colombia", "Ecuador", "French Guiana", "Guyana", "Paraguay", "Peru", "Suriname",
"Uruguay", "Venezuela","Panama","Nicaragua","Costa Rica"),
main = "Species richness of freshwater fishes in South America")

## End(Not run)
```

MapPolygon *Choropleth maps*

Description

It allows to shade the polygons in proportion to any of the variables (records richness, observed richess, predicted richness, completeness and the slope) exported by the function KnowBPolygon in the file "Estimators".

Usage

```
MapPolygon(data, polygonname, var, shape=NULL, shapenames=NULL, admAreas=TRUE,
Area="World", minLon, maxLon, minLat, maxLat, int=30, colbg="#FFFFFF",
colcon="#C8C8C8", colf="black", pro=TRUE, inc=0.005, exclude=NULL, colexc=NULL,
colfexc="black", colscale=c("#C8FFFFFF","#64FFFFFF","#00FFFFFF","#64FF64FF",
"#C8FF00FF","#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"), colm="black",
legend.pos="y", breaks=9, xl=0, xr=0, yb=0, yt=0, asp, lab=NULL, xlab="Longitude",
ylab="Latitude", main=NULL, cex.main=1.6, cex.lab=1.4, cex.axis=1.2, cex.legend=0.9,
family="sans", font.main=2, font.lab=1, font.axis=1, lwdP=0.6, lwdC=0.1,
trans=c(1,1), log=c(0,0), ndigits=0, ini=NULL, end=NULL, jpg=FALSE, filejpg="Map.jpg")
```

Arguments

data	Data file exported by the function KnowBPolygon named "Estimators" with the values of records, observed richness, predicted richness, completeness and slope for each area polygon.

polygonname	A variable available in the data file with the names of the polygons.
var	A variable available in the data file with the values to be used for shading the polygons.
shape	If the estimators in the function link[KnowBR]KnowBPolygon were calculated using an external shape file, it is necessary to indicate the file in this argument. It is not necessary to select any polygon within the file, just to load the whole shape file.
shapenames	Variable in the shapefile with the names of the polygons.
admAreas	If it is TRUE the border lines of the countries are depicted in the map.
Area	Only if using RWizard. A character with the name of the administrative area or a vector with several administrative areas (countries, regions, etc.) or river basins. If it is "World" (default) the entire world is plotted. For using administrative areas or river basins, in addition to use RWizard, it is also necessary to replace data(world) by @_Build_AdWorld_ (see examples).
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude.
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude.
int	Number of intervals into which the variable is splited.
colbg	Background color of the map (in some cases this is the sea).
colcon	Background color of the administrative areas.
colf	Color of administrative areas border.
pro	If it is TRUE an automatic calculation is made in order to correct the aspect ratio y/x along latitude.
inc	Adds some room along the map margins with the limits x and y thus not exactly the limits of the selected areas.
exclude	A character with the name of the administrative area or a vector with several administrative areas that may be plotted with a different color on the map (only if using RWizard).
colexc	Background color of areas selected in the argument exclude.
colfexc	Color of borders of the areas selected in the argument exclude.
colscale	Palette color.
colm	Color of the polygons without information when using when using an external shape file.
legend.pos	Whether to have a horizontal "x" or vertical "y" color scale.
breaks	Number of breakpoints of the color legend.
xl, xr, yb, yt	The lower left and upper right coordinates of the color legend in user coordinates.
asp	The y/x aspect ratio.
lab	A numerical vector of the form c(x, y) which modifies the default way that axes are annotated. The values of x and y give the (approximate) number of tickmarks on the x and y axes.
xlab	A title for the X axis.

ylab	A title for the Y axis.
main	An overall title for the plot.
cex.main	The magnification to be used for main titles relative to the current setting of cex.
cex.lab	The magnification to be used for X and Y labels relative to the current setting of cex.
cex.axis	The magnification to be used for axis annotation relative to the current setting of cex.
cex.legend	The magnification to be used for the color scale relative to the current setting of cex.
family	The name of a font family for drawing text.
font.main	The font to be used for plot main titles.
font.lab	The font to be used for x and y labels.
font.axis	The font to be used for axis annotation.
lwdP	Line width of the plot.
lwdC	Line width of the borders.
trans	It is possible to multiply or divide the dataset by a value. For a vector with two values, the first may be 0 (divide) or 1 (multiply), and the second number is the value of the division or multiplication.
log	It is possible to apply a logarithmic transformation to the dataset. For a vector with two values, the first may be 0 (do not log transform) or 1 (log transformation), and the second number is the value to be added in case of log transformation.
ndigits	Number of decimals in legend of the color scale.
ini	Minimum to be considered in the color scale.
end	Maximum to be considered in the color scale.
jpg	If TRUE the plots are exported to jpg files instead of using the windows device.
filejpg	Name of the jpg file.

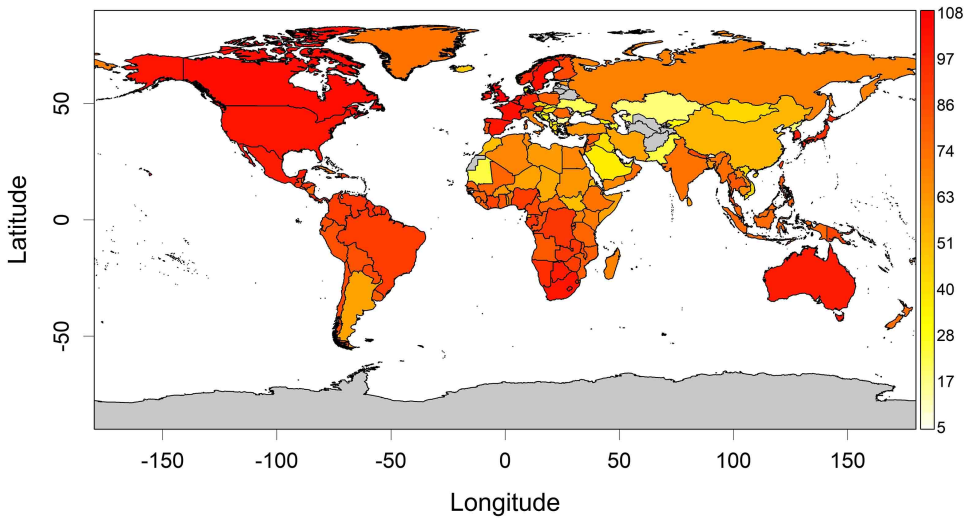
Details

FUNCTIONS

The function color.legend of the package plotrix (Lemon et al., 2014) is used for building the map.

EXAMPLE

Completeness of the records of freshwater fish species in all countries of the world.



Value

A map is obtained.

References

Lemon, J. (2006) Plotrix: a package in the red light district of R. *R-News*, 6(4):8-12.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duursma, R., Biancotto, E., Levy, O., Dutang, C., Solymos, P., Engelmann, R., Hecker, M., Steinbeck, F., Borchers, H., Singmann, H., Toal, T. & Ogle, D. (2015). Various plotting functions. R package version 3.6-1. Available at: https://CRAN.R-project.org/package=plotrix.

Examples

```
data(Estimators)
data(adworld)
MapPolygon(data=Estimators, polygonname="Area", var="Completeness",
colscale=rev(heat.colors(100)))
```

`RFishes` *Species richness of freshwater fishes*

Description

Species richness of freshwater fish species in cells of 1 degree around the world (Guisande et al., 2010).

Usage

```
data(RFishes)
```

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, L.F., Duque, S. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

States *States of USA*

Description

A file with information about the state boundaries of the United States.

Usage

```
data(States)
```

Source

https://www.census.gov/

SurveyQ *Survey quality*

Description

Discriminations among good, fair and poor quality of surveys in cells and polygons.

Usage

```
SurveyQ(data, Longitude=NULL, Latitude=NULL, cell=60, Areas=NULL,
variables=c("Slope","Completeness","Ratio"), completeness=c(50,90),
slope=c(0.02,0.3), ratio=c(3,15), shape=NULL, shapenames=NULL, admAreas=TRUE,
Area="World", minLon, maxLon, minLat, maxLat, main=NULL, PLOTP=NULL,
PLOTB=NULL, POINTS=NULL, XLAB=NULL, YLAB=NULL, XLIM=NULL, YLIM=NULL,
palette=c("blue","green","red"), COLOR=c("red","green","blue"), colm="black",
labels=TRUE, sizelabels=1, LEGENDP=NULL, LEGENDM=NULL, file="Polar coordinates.csv",
na="NA", dec=",", row.names=FALSE, jpg=FALSE, filejpg="Map.jpg")
```

Arguments

data	Data file exported by the functions KnowBPolygon or KnowB named "Estimators" with the values of records, observed richness, predicted richness, completeness and slope for each area polygon.
Longitude	Variable with the longitude of the cells if the file "Estimators" was obtained with the function KnowB, so it is the case when working with cells.
Latitude	Variable with the latitude of the cells if the file "Estimators" was obtained with the function KnowB, so it is the case when working with cells.
cell	Resolution of the cells (spatial units) in minutes on which calculations were carried out. In the present version the user can select any resolution between 1 and 60 minutes.
Areas	Variable with the names of the polygons, if the file "Estimators" was obtained with the function KnowBPolygon, so it is the case when working with polygons.
variables	The slope, completeness and ratio obtained in the file "Estimators", in that order.
completeness	Values of the completeness to define the thresholds for poor, fair and good quality surveys of the cells or polygons.
slope	Values of the slope to define the thresholds for poor, fair and good quality surveys of the cells or polygons.
ratio	Values of the ratio to define the thresholds for poor, fair and good quality surveys of the cells or polygons.
shape	If the estimators in the function link[KnowBR]KnowBPolygon were calculated using an external shape file, it is necessary to indicate the file in this argument. It is not necessary to select any polygon within the file, just to load the whole shape file.
shapenames	Variable in the shapefile with the names of the polygons.
admAreas	If it is TRUE the border lines of the countries are depicted in the map.
Area	Only if using RWizard. A character with the name of the administrative area or a vector with several administrative areas (countries, regions, etc.) or river basins. If it is "World" (default) the entire world is plotted. For using administrative areas or river basins, in addition to use RWizard, it is also necessary to replace data(world) by @_Build_AdWorld_ (see examples of function KnowBPolygon).
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude.
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude.
main	Main title of the map.
PLOTP	It allows to specify the characteristics of the function plot.default of the polar coordinates plot.
PLOTB	It allows to specify the characteristics of the function plot.default of the bubble chart.
POINTS	It allows to modify the points of the bubble chart with the function points.
XLAB	Legend of the X axis.
YLAB	Legend of the Y axis.

XLIM	Vector with the limits of the X axis.
YLIM	Vector with the limits of the Y axis.
palette	The color gradient of the bubble chart may be one of these palettes: "heat.colors", "terrain.colors", "gray.colors", "topo.colors" or "cm.colors".
COLOR	It allows to modify the colors of the map and polar coordinates plot. It must be three colors.
colm	Color of the polygons without information when using when using an external shape file.
labels	If it is FALSE, points are depicted instead of the labels of the polygons in the polar coordinates plot.
sizelabels	Text size of the labels of the polygons in the polar coordinates plot.
LEGENDP	It allows to modify the legend of the polar coordinates plot.
LEGENDM	It allows to modify the legend of the map.
file	CSV FILES. Filename with the polar coordinates.
na	CSV FILE. Text that is used in the cells without data.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
row.names	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.
jpg	If TRUE the map is exported to jpg files instead of using the windows device.
filejpg	Name of the jpg file.

Details

This function has been designed to identify and plot the cells or polygons with good, fair and poor quality surveys. This function uses the file called "Estimators" obtained from the functions KnowBPolygon or KnowB to estimate the polar coordinates of all cells or polygons and to discriminate among cells or polygons according to the quality of the survey.

The variables used by this function are slope, completeness and ratio (number of records/species observed). The default values to identify the cells or polygons with good, fair and poor quality surveys are: slope lower than 0.02, completeness higher than 90% and ratio higher than 15 for good quality surveys, and slope higher than 0.3, completeness lower than 50% and ratio lower than 3 for poor quality surveys.

The order of the variables is important for the estimation of the polar coordinates because a different angle is assigned to each variable. Therefore, the variables must be introduced in this order: slope, completeness and ratio.

All variables are transformed to a scale ranged between -1 and 1. For each value the X and Y polar coordinates are estimated using the following equations:

$$X = \sum_{i=1}^{3} |z_j| cos(\alpha) \quad Y = \sum_{i=1}^{3} |z_j| sin(\alpha)$$

where z is the value of the variable j.

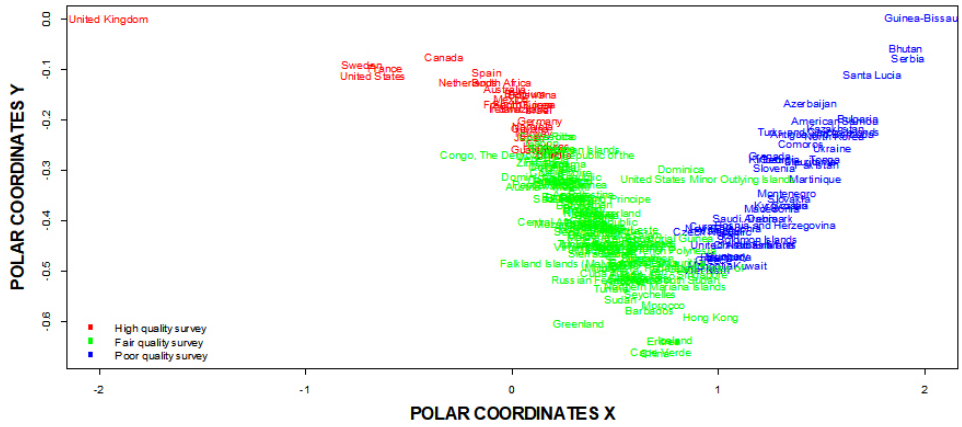
Each variable is assigned an angle (α). The increment value of the angle is always 60. Therefore, the first variable (slope) if the transformed value is ≥ 0 the α value is 60 and if the transformed value is < 0 the value is 240.

For the second variable (completeness) if the transformed value is ≥ 0 the α value is 120 and if the value is < 0 the value is 300.

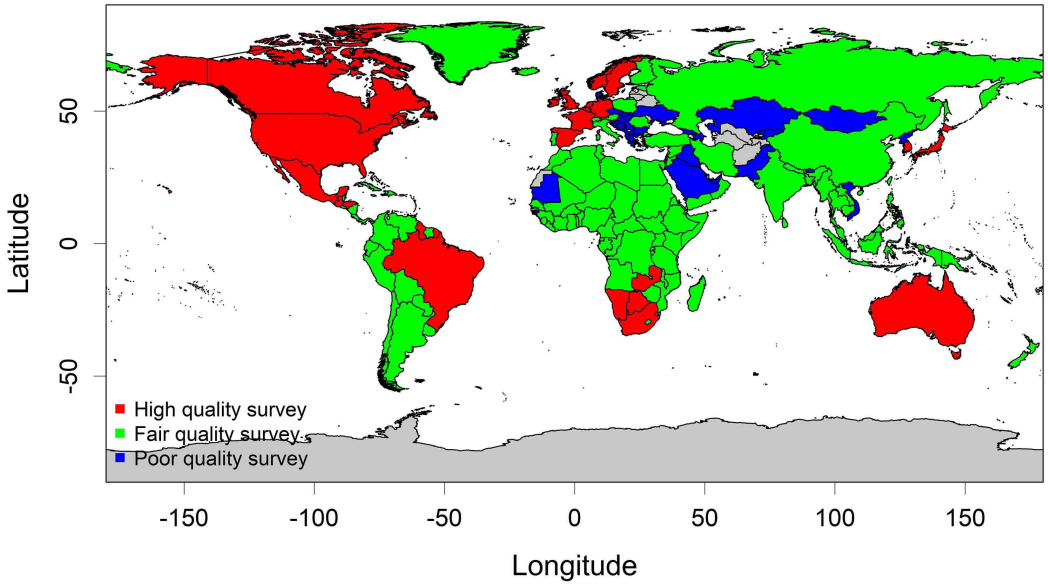
For the third variable (ratio) if the value is ≥ 0 the α value is 180 and if the transformed value is < 0 the value is 360.

Degrees to radians angle conversion is carried out assuming that 1 degree = pi/180 radians.

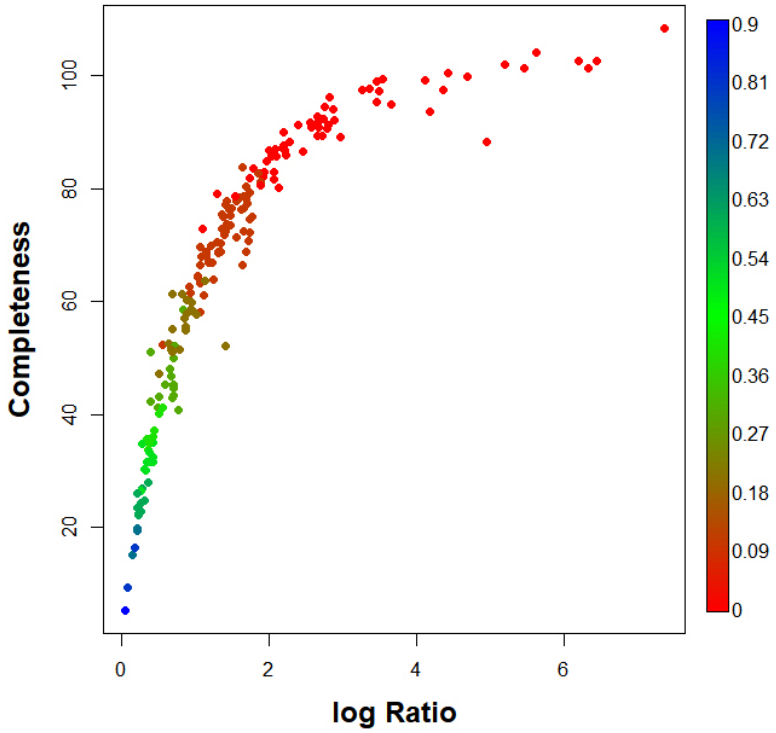
EXAMPLES Polar coordinates of the records of freshwater fish species in all countries of the world.



Quality survey of the records of freshwater fish species in all countries of the world.



Bubble chart of the relationship between log(Ratio) and completeness, being the color gradient the slope value.



Value

It is depicted a plot with the polar coordinates of each polygon or cell, a map with the quality survey of the cells or polygons and a file with the polar coordinates of the cells or polygons.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Examples

```
## Not run:

data(adworld)
data(Estimators)
SurveyQ(data=Estimators, Areas="Area")

## End(Not run)
```

SurveyQCZ *Survey quality of climate zones*

Description

Estimation of the survey quality of different climate zones.

Usage

```
SurveyQCZ(data, Longitude="Longitude", Latitude="Latitude", cell=NULL, hull=TRUE,
Area="World", shape=NULL, shapenames=NULL, aprox=TRUE, VIDTAXA=NULL, por=80, k=NULL,
VIF=FALSE, VARSEDIG=FALSE, BUBBLE=FALSE, variables=c("Slope","Completeness","Ratio"),
completeness=c(50,90), slope=c(0.02,0.3), ratio=c(3,15), minLon=NULL,
maxLon=NULL, minLat=NULL, maxLat=NULL, xlab="Longitude", ylab="Latitude",
colscale=c("#C8FFFFFF","#64FFFFFF","#00FFFFFF","#64FF64FF","#C8FF00FF",
"#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"), colcon="transparent",
breaks=10, ndigits=0, xl=0, xr=0, mfrowBOXPLOT=NULL, mfrowMAP=NULL,
main="Percentage of ignorance/poor\n cells in each cluster", cexCM=0.5,
legpos="bottomleft", jpg=FALSE, filejpg="Survey Quality CZ.jpg", dec=",")
```

Arguments

data	Data file exported by the function KnowB named "Estimators" with the values of records, observed richness, predicted richness, completeness and slope for each area polygon. This file may include the values of the environmental variables for each cell.
Longitude	Variable with the longitude.
Latitude	Variable with the latitude.
cell	Resolution of the cells (spatial units) in minutes. In the present version the user can select any resolution between 1 and 60 minutes. If it is NULL (default), it is used the resolution of the ASC files with the environmental variables. It is not NULL, the ASC files are rescaled to the resolution selected by the user.
hull	If it is TRUE, the extent is estimated as the convex null of the data. It is FALSE, it is used as extent the polygons defined in the arguments *shape* of *Area*.
Area	Only if using RWizard. It allows to plot countries, regions, river basins, etc., as extentd. A character with the name of the administrative area or a vector with several administrative areas (countries, regions, etc.) or river basins. If it is "World" (default) the entire world is plotted. For using administrative areas or river basins, in addition to use RWizard, it is also necessary to replace data(world) by @_Build_AdWorld_.
shape	Optionally it may be used a shape file with the information of the polygons.
shapenames	Variable in the shapefile with the names of the polygons.
aprox	If it is TRUE and there is not environmental data available in the cell, it is used the nearest cell available in the environmental data set.

VIDTAXA	It accesses the VIDTAXA function of the VARSEDIG package.
por	Cut-off threshold specifying the cumulative variance percentage, to determine how many axes are selected from the Principal Components or Correspondence analyses. By default it is 80%, which means that the axes are selected until reaching an accumulated variance percentage of 80%.
k	Number of clusters in which the Dendrogram is divided. If it is NULL, the algorithm select automatically the maximum number of clusters in which the Dendrogram can be divided, which are those groups that are statistically different in at least one variable according to the U Mann-Whitney test. If the are is large, there may be many different climate zones, so with NULL option running time may be long.
VIF	If it is TRUE, the inflation factor of the variance (VIF) is used to select the highly correlated variables and, therefore, not correlated variables are excluded from the Principal Components analysis.
VARSEDIG	If it is TRUE, the VARSEDIG algorithm is performed.
variables	The slope, completeness and ratio obtained in the file "Estimators", in that order.
BUBBLE	If it is TRUE, the BUBBLE plot the VARSEDIG function is depicted.
completeness	Values of the completeness to define the thresholds for poor, fair and good quality surveys of the cells.
slope	Values of the slope to define the thresholds for poor, fair and good quality surveys of the cells.
ratio	Values of the ratio to define the thresholds for poor, fair and good quality surveys of the cells.
minLon, maxLon	Optionally it is possible to define the minimum and maximum longitude.
minLat, maxLat	Optionally it is possible to define the minimum and maximum latitude.
xlab	Legend of the X axis in the map.
ylab	Legend of the Y axis in the map.
colscale	Color of the bar scale.
colcon	Background color of the administrative areas.
breaks	Number of breakpoints of the color legend.
ndigits	Number of decimals in legend of the color scale.
xl, xr	The lower left and right coordinates of the color legend in user coordinates.
mfrowBOXPLOT	It allows to specify the boxplot panel. It is a vector with two numbers, for example c(2,5) which means that the boxplots are put in 2 rows and 5 columns.
mfrowMAP	It allows to specify the map panel. It is a vector with two numbers, for example c(3,2) which means that the map panel are put in 3 rows and 2 columns.
main	Main title of the map with the percentage of ignorance/poor cells.
cexCM	Size of the points in the maps of the climate clusters.
legpos	Legend position with the number of the cluster in the maps of the climate clusters.
jpg	If TRUE the plots are exported to jpg files instead of using the windows device.
filejpg	Name of the jpg file.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".

Details

The aim of this algorithm is to identify different climate zones and to estimate survey quality of these zones. The climate zones are classified from those with higher percentage of ignorance/poor survey quality cells (poor surveyed climate zones), to those with the lower percentage of ignorance/poor survey quality cells (better surveyed climate zones). This function uses the algorithm of the VARSEDIG function (Guisande et al., 2016; 2019; Guisande, 2019, which is briefly summarized as follows.

This function only works if there are ASC files with the environmental variables in the working directory.

In the first step of the algorithm, a Principal Components analysis is performed, being the cases the different cells and the variables the environmental variables. The aim is to determine the environmental variables responsible for the variability observed among the cells.

To detect the potential groups being formed in the Principal Components analysis, a Dendrogram is applied to the scores obtained from the axes that absorb a greater variance. By default, the axes that absorb 80% of the variability are chosen, but this value can be modified by the user.

Subsequently, a Discriminant Analysis is carried out to determine if the clusters that have been generated are well discriminated, that is, to determine the number of correctly identified cases in each cluster.

Next, a U Mann-Whitney test is performed to determine if there are significant differences in the variables between the clusters.

The idea of this function is to find the largest possible number of clusters with the highest discrimination percentage. To do this, the user should perform tests modifying the cut-off threshold by specifying the cumulative variance percentage to determine how many axes are selected from the Main Components (by default *por=80*) and the variables to be included, eliminating those that are not correlated and are not useful in the Principal Components analyses, as well as those that have little discrimination power in the Discriminant Analysis.

Once the different climate zones have been identified, with the file Estimators.CSV obtained from the function KnowB (Lobo et al., 2018; Guisande & Lobo, 2019), with the values of the slope, completeness and the ratio between the number of records and the observed species (R/S) for different cells, it is estimated the percentage of ignorance/poor cells in each climate zone. Ignorance cell are those for which was not possible to estimate the slope, completeness and/or R/S, because there are not records or the number of records in small. Poor quality surveys cells are those that slope > 0.3, completeness < 50

FUNCTIONS

This function uses the algorithm of the VIDTAX function of the VARSEDIG package (Guisande et al., 2016; 2019). The Principal Components Analysis was performed with the prcomp function of the stats package. The vif function of the usdm package was used for the calculation of VIF (Naimi et al., 2014; Naimi, 2017). To perform the *biplot* graph the scatterplot function of the car package was used (Fox et al., 2018). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices. KMO test was performed with the function KMO of the package psych (Revelle, 2018). The U Mann-Whitney test is performed with the *wilcox.test* function of the base stats package. The comparison between clusters with the VARSEDIG algorithm is done with the VARSEDIG function (Guisande et al., 2016). The Linear Discriminant Analysis was performed with the functions candisc of the candisc package (Friendly, 2007; Friendly & Fox, 2017) and lda

of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018). The Quadratic Discriminant Analysis was performed with the function qda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018). The graph with one dimension in the Discriminant analysis was performed with the function plot.cancor of the candisc package (Friendly, 2007; Friendly & Fox, 2017).

EXAMPLE

We used the file Estimators obtained with the function KnowB using a database that includes 121,709 records of freshwater fishes obtained from Pelayo-Villamil et al. (2015), with the Clench estimator and a cell resolution of 5'.

In the working directory, there are the environmental variables BIO1, BO2, BIO4, BIO8, BIO12, BIO14, BIO15, BIO18 and BIO19 of the WorldClim data set (Hijmans et al., 2005), as ASC raster files.

As the argument *VIF=FALSE*, there is not information about VIF, and the first statistic obtained is the KMO test, which tells us if the variables are adequate for the Principal Components. The value must be greater than 0.5. Therefore, all variables that do not have a value greater than 0.5, could be eliminated from the analysis. In the case that the value is exactly 0.5, it means that it is not possible to estimate the KMO. The minimum value was 0.73 for BIO12, so all variables are adequate for the Principal Components.

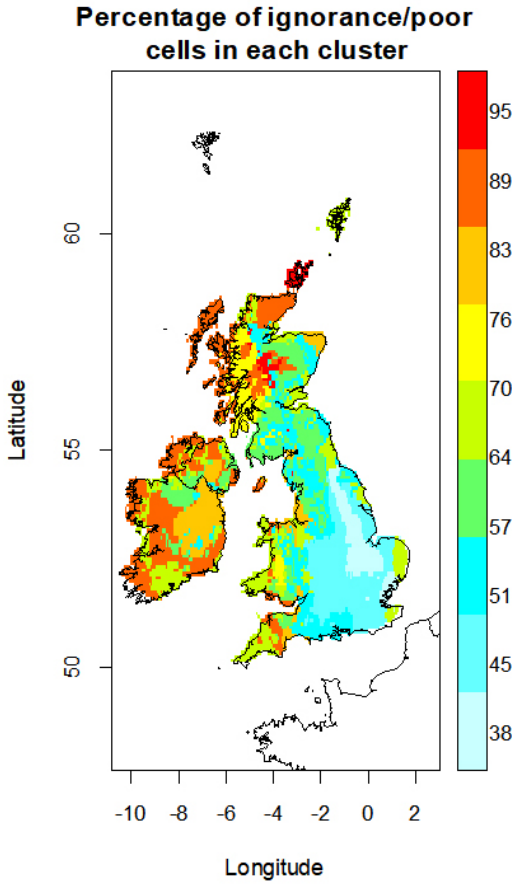
The next statistic that appears is Bartlett's test of sphericity, which tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate. A value p of the contrast smaller than the level of significance allows rejecting the hypothesis and concluding that there is correlation. Therefore, for the Principal Components analysis to be valid, the probability must be less than 0.05, as it is in this case, with a $p<0.001$.

The first figure is the Principal Components. The first axis accounts for 62.9%, the second for 16.5% and the third for 7% of the variance observed. The first three axes explain 86.5% of the variance. Since the default value of *por=80* was selected, these three Principal Component axes are selected.

In the Dendrogram there were 22 clusters statistically different, because the argument *k=NULL* in the script (default option), which means that the algorithm finds the maximum number of statistically different climate zones. The results obtained in the files U Mann-Whitney test.CSV and Descriptive statistics of clusters.CSV show that there are statistical differences among all 22 clusters in at least one variable (U Mann-Whitney test, $p<0.001$).

Other plots and statistics are obtained, which are fully explained in the manuscript that describes the algorithm VIDTAXA (Guisande et al., 2019).

The final plot is the map with the percentage of ignorance/poor cells (see map below). It is clear that the area with a better survey quality is in South-East of United Kingdom.



Value

It is obtained:

1. A TXT file with the VIF (if the argument *VIF=TRUE*), the correlations between variables, the Kaiser-Meyer-Olkin (KMO) test, the Bartlett sphericity test and the results of the Principal Components or Correspondence analyses. The file is called by default "Output.TXT".

2. A CSV FILE with the coordinates for each case of the Principal Components or Correspondence analyses. The file is called by default "Cat loadings.CSV".

3. A CSV FILE with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram. The file is called by default "Descriptive statistics of clusters.CSV".

4. A CSV FILE with the original data of the variables and the cluster to which each case belongs. The file is called by default "Original data and cluster number.CSV".

5. A CSV FILE with the coordinates of the variables in the Linear Discriminant Analysis plot. The file is called by default "Var loadings-Linear.csv"

6. A CSV FILE with the coordinates of the categories in the Linear Discriminant Analysis plot. The file is called by default "Cat loadings-Linear.csv".

7. A CSV FILE with the predictions table using the cross-validation of Linear Discriminant Analysis. The file is called by default "Table cross-validation-Linear.csv".

8. A CSV FILE with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Linear Discriminant Analysis. The file is called by default "Cases cross-validation-Linear.csv".

9. A CSV file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Table cross-validation-Quadratic.csv".

10. A CSV file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Cases cross-validation-Quadratic.csv".

11. A CSV file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test. The file is called by default "U Mann-Whitney test.csv".

12. A CSV file called by default "Priorization.csv", with the clusters of the different climate zones arranged from the cluster with the higher percentage of cells catalogued as ignorance/poor (lowest quality survey) to the cluster with the lower percentage of cells catalogued as ignorance/poor (highest quality survey).

13. A scatterplot of the Principal Components or Correspondence analyses.

14. A Dendrogram grouping by clusters according to the scores of the Principal Components or Correspondence analyses.

15. A graphic panel with a boxplot for each variable comparing the values of these variables between each of the clusters obtained in the Dendrogram.

16. A Graph of the Discriminant Analysis showing the influence of the variables on the discriminant axis I, differentiating the different clusters.

17. A graph of the Discriminant Analysis showing the scores of the discriminant axes I and II, differentiating the different clusters.

18. If the argument *BUBBLE=TRUE*, a bubble chart with the number of variables that are statistically different between clusters.

19. The maps with the climate zones.

20. The map with the percentage of ignorance/poor cells.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018) Companion to Applied Regression. R package version 3.0-0. Available at: https://CRAN.R-project.org/package=car.

Friendly, M. & Fox, J. (2017) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. Available at: https://CRAN.R-project.org/package=candisc.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Guisande, C., Vari, R.P., Heine, J., GarcÃa-RosellÃ³, E., GonzÃ¡lez-Dacosta, J., PÃ©rez-Schofield, B.J., GonzÃ¡lez-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162: 571-580.

Guisande, C., Rueda-Quecho, A.J., Rangel-Silva, F.A., Heine, J., GarcÃa-RosellÃ³, E., GonzÃ¡lez-Dacosta, J. & Pelayo-Villamil, P. (2019) VIDTAXA: an algorithm for the identification of statistically different groups based on variability obtained in factorial analyses. *Ecological Informatics*, 46: 62-68.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25: 1965-1978.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: https://CRAN.R-project.org/package=IDPmisc.

Naimi, B. (2017) Uncertainty analysis for species distribution models. R package version 1.1-18. Available at: https://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Pelayo-Villamil, P., Guisande, C., Vari, R.P., ManjarrÃ©s-HernÃ¡ndez, A., GarcÃa-RosellÃ³, E., GonzÃ¡lez-Dacosta, J., Heine, J., GonzÃ¡lez-Vilas, L., Patti, B., Quinci, E.M., JimÃ©nez, L.F., Granado-Lorencio, C., Tedesco, P.A., Lobo, J.M. (2015) Global diversity patterns of freshwater fishes-Potential victims of their own success. *Diversity and Distributions*, 21: 345-356.

Revelle,W. (2018) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.8.4. Available at: https://CRAN.R-project.org/package=psych.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2018) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-50. Available at: https://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. https://www.stats.ox.ac.uk/pub/MASS4/.

Examples

```
## Not run:
####This script only works if there are ASC files, with
####environmental variables, in the working directory

data(FishIrelandUK)

data(adworld)

SurveyQCZ(data=FishIrelandUK, maxLon=3, mfrowBOXPLOT=c(3,3), cexCM=0.2)

## End(Not run)
```

Index