Package 'VARSEDIG'

December 7, 2018

Version 1.9

Title An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy

Author Cástor Guisande González

Maintainer Cástor Guisande González <castor@uvigo.es>

Description An algorithm which identifies the morphometric features that significantly discriminate two taxa and validates the morphological distinctness between them via a Monte-Carlo test, polar coordinates and overlap of the area under the density curve.

License GPL (>= 2)

Encoding latin1

Depends R (>= 3.2)

Suggests adehabitatHS, kulife, MASS, car, ade4, IDPmisc, REdaS, ca, ltm, psych, usdm

Repository CRAN

NeedsCompilation no

Date/Publication 2018-12-07 11:40:08 UTC

R topics documented:

characiformes *MORPHOMETRIC VARIABLES OF CHARACIFORMS*

Description

Morphometric data of several freshwater fish species of the order Characiforms, as the length of the dorsal fin base (M12), body height (M11), etc. For details see Guisande et al. (2010).

Usage

```
data(characiformes)
```

Format

An array (matrix) with 31 columns: taxonomic data (order, family, genus and species) and 27 morphometric variables.

Source

http://www.ipez.es.

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.

scorpaeniformes *MORPHOMETRIC VARIABLES OF SCORPAENIFORMES*

Description

Morphometric data of several marine fish species of the order Scorpaeniformes, as the length of the dorsal fin base (M12), body height (M11), etc. For details see Guisande et al. (2010).

Usage

```
data(scorpaeniformes)
```

Format

An array (matrix) with 31 columns: taxonomic data (order, family, genus and species) and 27 morphometric variables.

Source

http://www.ipez.es.

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.

VARSEDIG — *Variable selection to discriminate among taxonomic groups*

Description

This function performs an algorithm for morphometric characters selection and statistical validation in morphological taxonomy.

Usage

```
VARSEDIG(data, variables, group, group1, group2, method="overlap", stepwise=TRUE,
VARSEDIG=TRUE, minimum=TRUE, kernel="gaussian", cor=TRUE, ellipse=FALSE, convex=TRUE,
DPLOT=NULL, SCATTERPLOT=NULL, BIVTEST12=NULL, BIVTEST21=NULL, Pcol="red",
colbiv="lightblue", br=20, sub="", lty=1, lwd=2.5, ResetPAR=TRUE, PAR=NULL, XLABd=NULL,
YLABd=NULL, XLIMd=NULL, YLIMd=NULL, COLORd=NULL, COLORB=NULL, LEGENDd=NULL, AXISd=NULL,
MTEXTd= NULL, TEXTd=NULL, XLABs=NULL, YLABs=NULL, XLIMs=NULL, YLIMs=NULL,
PCHs=NULL, COLORs=NULL, LEGENDs=NULL, MTEXTs=NULL, TEXTs=NULL, LEGENDr=NULL,
MTEXTr= NULL, TEXTr=NULL, arrows=TRUE, larrow=1, ARROWS=NULL, TEXTa=NULL, devnew=TRUE,
model="Model.rda", file1="Overlap.csv", file2="Coefficients.csv",
file3="Predictions.csv", file4="Polar coordinates.csv", file="Output.txt",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variables to be selected.

`group` Variable with the groups to be discriminated.

`group1` First group.

`group2` Second group.

`method` Three different methods for prioritizing the variables according to their capacity for discrimination can be used. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

`stepwise` If TRUE, the logistic regression is applied by steps, in order to eliminate those variables that are not significant. The Akaike information criterion (*AIC*) is used to define what are the variables that are excluded.

`VARSEDIG` If it is TRUE, the variables are added for the estimation of polar coordinates in the priority order according to the method "overlap", "Monte-Carlo", or "logistic regression" and the variable is selected if it significantly contributes to discriminate between both groups. See details section for further information.

`minimum` If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. If it is FALSE, the algorithm selects all significant variables, and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`kernel` A character string giving the smoothing kernel to be used for estimating the overlap of the area under the curve between groups. This must be one of "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine". For further details about the estimation of the density curve see the details section of the function density of base stats package.

`cor` If it is TRUE the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one that has a greater positive correlation.

`ellipse` If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *group* is depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

`DPLOT` It allows to specify the characteristics of the function plot.default of the density plot.

`SCATTERPLOT` It accesses the function scatterplot of the car package, with the graph *biplot* that performs the X an Y polar coordinates.

`BIVTEST12` It accesses the function biv.test of the package adehabitatHS, which performs the bivariate plot that displays the results of a bivariate randomisation test. From all values of group 2, it shows the value with higher probability to belong to group 1.

BIVTEST21 As in the argument *BIVTEST12*, but from all values of group 1, it shows the value with higher probability to belong to group 2.

Pcol Color or name for the observation of group 2 in the BIVTEST12 plot and for the value of group 1 in the BIVTEST21 plot.

colbiv Color or name of all values of group 1 in the BIVTEST12 plot and all values of group 2 in the BIVTEST21 plot.

br Numbers of breaks of the histograms in the BIVTEST plots.

sub Title in the BIVTEST plots.

lty Type of line of the density curve for each group. If it is a vector, it must be as many as different categories of the variable *group*.

lwd Line width relative to the default (default=1), so 2 is twice as wide of the density curve.

ResetPAR If it is FALSE, the default condition of the functio PAR is not placed and maintained those defined by the user in previous graphics.

PAR It accesses the function PAR that allows to modify many different aspects of the graph.

XLABd Legend of the X axis in the density plot.

YLABd Legend of the Y axis in the density plot.

XLIMd Vector with the limits of the X axis in the density plot.

YLIMd Vector with the limits of the Y axis in the density plot.

COLORd Color of the density curves in the density plot. It must be as many as different categories of the variable *group*. As the color has transparency, the plot must be copy as bitmap and not metafile.

COLORB Color of the lines in the density plot. It must be as many as different categories of the variable *group*.

LEGENDd It allows to modify the legend of the density plot. If it is FALSE the legend is not shown.

AXISd It allows to add axes to the density plot.

MTEXTd It allows to add text on the margins of the density plot.

TEXTd It allows to add text in any area of the inner part of the density plot.

XLABs Legend of the X axis in the scatterplot.

YLABs Legend of the Y axis in the scatterplot.

XLIMs Vector with the limits of the X axis in the scatterplot.

YLIMs Vector with the limits of the Y axis in the scatterplot.

PCHs Vector with the symbols of the scatterplot, that should be as many as different groups the variable *group* has. If NULL, they are automatically calculated starting with the symbol 15.

COLORs It allows to modify the colors of the scatterplot. It must be as many as different categories of the variable *group*.

LEGENDs It allows to modify the legend of the scatterplot.

`MTEXTs` It allows to add text on the margins of the scatterplot.

`TEXTs` It allows to add text in any area of the inner part of the scatterplot.

`LEGENDr` It allows to modify the legend of the BIVTEST plot. If it is FALSE the legend is not shown.

`MTEXTr` It allows to add text on the margins of the BIVTEST plot.

`TEXTr` It allows to add text in any area of the inner part of the BIVTEST plot.

`arrows` If it is TRUE the arrows are shown in the scatterplot with the polar coordinates. These arrows show the vector of the variables selected when calculating the polar coordinates.

`larrow` It modifies the length of the arrows.

`ARROWS` It accesses the function Arrows of the package IDPmisc, which performs the arrows.

`TEXTa` It allows to modify the labels at the end of the arrows.

`devnew` If it is TRUE, each plot is depicted in a different window.

`model` Filename with the model of the binomial logistic regression.

`file1` CSV FILE. Filename with the overlap of the area under the curve between both categories for all variables.

`file2` CSV FILES. Filename with regression coefficients of the binomial logistic regression.

`file3` CSV FILES. Filename with the predictions of the binomial logistic regression.

`file4` CSV FILES. Filename with the polar coordinates for both categories of the variable *group*.

`file` TXT FILE. Name of the output file with the results of the binomial logistic regression, the variables that significantly discriminate between the two groups and Euclidean distance between the two groups considering 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.

Details

Classification methods such as logistic regression and discriminant analysis are probably the best available methods for the identification of the variables optimally able to predict group membership (Guisande et al. 2011; Guisande & Vaamonde 2012). Classification and Regression Trees (CARTs) are useful for identifying the variables that best discriminate groups, it is impossible using those methods to test the significance of the variables or to predict group membership (Guisande & Vaamonde 2012).

There are three advantages of logistic regression over discriminant analysis (Guisande et al., 2011): 1) the logistic regression is much more relaxed and flexible in its assumptions than the discriminant analysis because, unlike the discriminant analysis, the logistic regression does not have the requirements of the independent variables to be normally distributed, linearly related, nor equal variance within each group; 2) logistic regression may be more powerful and efficient analytic strategy if there are qualitative variables among predictors; 3) it is possible to use a stepwise logistic regression and, therefore, to select only those variables that significantly discriminate between groups. Discriminant analysis, however, does not have a statistical test of the coefficients of individual independent variables comparable to logistic regression, so it is not possible to test significance of variables and, therefore, to select only the variables that significantly predict group membership. Actually, to include variables with low discrimination capacity leads to reduce the identification success of the discriminant analysis.

The disadvantages of logistic regression are mainly also three: 1) the lack of a graphical representation of the results; 2) to evaluate the predictability of the final model chosen from the analysis it is not enough with the information about the percentage of cases correctly identified; 3) when the assumptions mentioned above regarding the distribution of predictors are met, discriminant function analysis may be more powerful and efficient analytic strategy than logistic regression (Tabachnick & Fidell, 1996)

This function performs an algorithm for: 1) prioritizing the variables by their discrimination capacity using three different methods, 2) selecting only those variables that significantly discriminate between two groups, 3) evaluating the predictability of the final model chosen with a Monte-Carlo test and 4) the results are graphically depicted in four different plots.

1. Prioritizing the variables by their discrimination capacity

Three different methods for prioritizing the variables according to their capacity for discrimination can be used.

1. If the argument *method="overlap"*, a density curve is obtained for each variable and the overlap of the area under the curve between the two groups is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. This information is saved in *file1="Overlap.csv"*.

2. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity).

3. If the argument *method="logistic regression"*, then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC). The model of the regression is saved in *model="Model.rda"*, the coefficients in *file2="Coefficients.csv"* and the predictions of the regression in *file3="Predictions.csv"*.

2. Polar coordinates

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:

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

where *z* is the value of the variable *j* and *n* the number of variables.

Each variable is assigned an angle (α). The increment value of the angle is always $\frac{360}{n*2}$. If for instance the number of variables is 5, the increment angle is 36. Therefore, for the first variable if the value is ≥ 0 the α value is 36 and if the value is < 0 the value is 36+180, for the second variable if the value is ≥ 0 the α value is 36+36 and if the value is < 0 the value is 36+36+180, etc. Conversion of degrees to radians angle is carried out assuming that 1 degree = 0.0174532925 radians.

The order of the variables is consequently important because a different alpha value is assigned. If the argument *cor=TRUE*, this order is established calculating the correlation matrix of the variables and by ordering them such that each variable is followed by the variable to which it is highly correlated. The goal is to favor a larger dispersion of the data in the resulting polar coordinates system.

3. Algorithm for variables selection

The variables are added for the estimation of polar coordinates in the priority order according to *method="overlap"*, *method="Monte-Carlo"* or *method="logistic regression"*.

Mean X and Y polar coordinates are estimated for both groups and via these means the Euclidean distance is calculated between both groups.

In the case of the X and Y polar coordinates, a Monte-Carlo test is used for testing the statistical hypothesis if a value of one group is significantly higher or lower that the values of the other group. The test is performed for both X and Y polar coordinates and compares all values of one group with those of the other group. For instance, when all values of group 1 are compared with group 2, and the mean X polar coordinate of group 1 is higher than the one of group 2, the alternative hypothesis of the Monte-Carlo test is *greater*, and the p-value is estimated as (number of random values equal to or greater than the observed one + 1)/(number of permutations + 1). The null hypothesis is rejected if the p-value is less than the significance level. If the mean X polar coordinate of group 1 is lower than the one of group 2, the alternative hypothesis is *smaller*, a p-value is estimated as (number of random values equal to or less than the observed one + 1)/(number of permutations + 1). Again, the null hypothesis is rejected if the p-value is less than the significance level. The same process is applied when comparing all values of group 2 with those of group 1.

A variable is selected if it both: 1) contributes to increase Euclidean distance between both groups compared with the Euclidean distance obtained with the set of previously selected variables; and 2) the p-values of the Monte-Carlo test for X and Y coordinates when comparing both group 1 with group 2 and group 2 with group 1 are smaller than the p-values obtained with the set of previous selected variables. Therefore, from the pool of all independent variables, only those variables with the highest significant contribution to discriminating between both groups are selected.

The variables selected are saved in the *file="Output.txt"* and the polar coordinates of all values of both groups estimated with the variables selected are depicted in a scatterplot and saved in *file4="Polar coordinates.csv"*.

At the end of the process, it is selected the value with the highest p-value. Therefore, if this p-value is close or lower than the significance level of 0.05, it may be concluded that any of the values of one group may be identified as belonging to the other group.

Two plots are obtained with the value of the group 1 with the highest p-value of belonging to group 2 and the value of the group 2 with the highest p-value of belonging to group 1, respectively. In both plots, the x-axis corresponds to the X polar coordinates and the y-axis corresponds to Y polar coordinates.

If p-value is close or lower than 0.05 for X or Y polar coordinates, but in both cases when comparing group 1 with group 2 and group 2 with 1, it may be concluded that the variables selected are significantly contributing to discriminate between both groups, so with these variables is possible to achieve a 100% of identification success when predicting group membership.

FUNCTIONS

The density plot is performed with the function plot.default of base graphics package. The density curve is estimated with the function density of base stats package. The area under the curve is estimated with the function auc of the package kulife (Ekstrom et al., 2015). The random test was performed with the function as.randtest of the package ade4 (Chessel et al., 2004; Dray et al., 2007; 2015). The bivariate plot that displays the results of a bivariate randomisation test, for which the p-values are computed with the function as.randtest (one-sided tests), was performed with the function biv.test of the package adehabitatHS (Calenge, 2006; 2015). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The scatterplot is performed with the function scatterplot of the car package (Fox & Weisberg, 2011; Fox et al., 2014). The convex hull is estimated with the function chull of the package grDevices.

EXAMPLES

For the example, morphometric data of three families of freshwater fishes, as the distance from the origin of the dorsal fin to the origin of the anal fin (M13), the length of the dorsal fin base (M12), body height (M11), etc., are used. For details see Guisande et al. (2010).

Figure shows the plots obtained with VARSEDIG (Guisande et al., 2016), in an example comparing the species *Moenkhausia dichroura* and *Moenkhausia oligolepis*.

The variables that better discriminate between both species are the M26 (interorbital width) and M11 (distance from the dorsal-fin origin to the dorsal limit of the pelvic-fin base). Between these two variables, a density plot is depicted for the quantitative variable with lower overlap between both groups and, thus, the highest discrimination capacity: in this example M26 (Figure 1A).

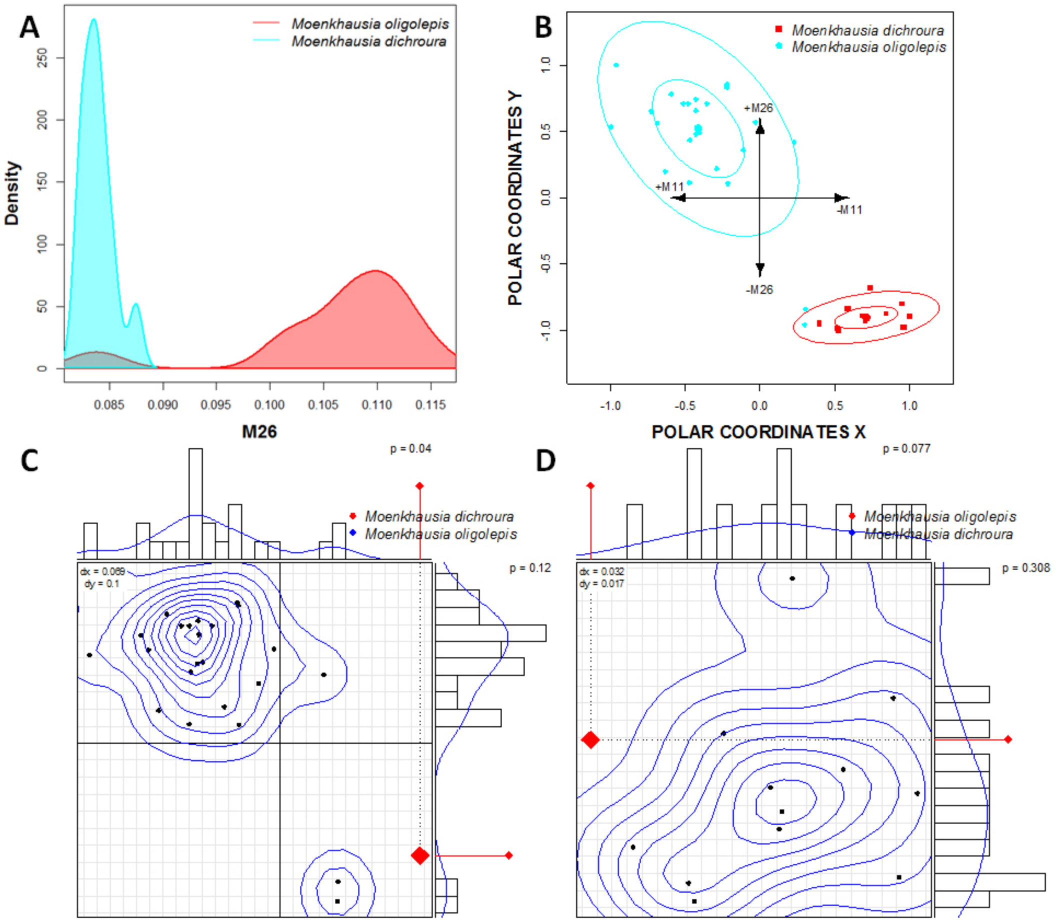


Figure 1B shows the scatterplot of the polar coordinates obtained for both species using variables M26 and M11. The arrows show the vector of the variables with both of these variables higher in *M. oligolepis*.

This example illustrates that the VARSEDIG algorithm is not only useful for identifying the variables that better discriminate between two taxa, but also may be informative when it comes to finding misidentified individuals. In the example, it appears that two individuals identified as *M. oligolepis* are *M. dichroura* (Figure 1B).

Figure 1C displays the results of a bivariate randomisation test. From all individuals of the species *M. dichroura*, the figure shows the individual of *M. dichroura* (red point) with higher probability to be identified as belonging to the M. oligolepis. Kernel density is estimated to indicate the contours of the distribution of randomised values. The two marginal histograms correspond to the univariate tests on each axis, for which the p-values (one-sided tests) are computed. As p-value is lower than 0.05 for X axis (p = 0.04), the null hypothesis is rejected. Consequently the X polar coordinates of all individuals of the of the species *M. dichroura* are significantly different than those of the species *M. oligolepis* and, therefore, none of the individuals designated as *M. dichroura* may be identified as belonging to the species *M. oligolepis*.

Figure 1D also displays the results of a bivariate randomisation test but, in this case, from all individuals of the species *M. oligolepis*, the figure shows the individual (red point) with higher probability to belong to the species *M. dichroura*. Both p-values are higher than 0.05, so null hypothesis is accepted for both X and Y polar coordinates. This that some individuals of the species *M. oligolepis* may be identified as belonging to the species *M. dichroura*.

It is not necessary a p-value lower than 0.05 for both X and Y, but it is just necessary and p-value lower than 0.05 for X or Y when comparing both group 1 with 2 and group 2 with 1. Therefore, if p-value is close or lower than the significance level of 0.05 for X or Y polar coordinates in both cases comparing group 1 with 2 and group 2 with 1, it would mean a 100% of identification success between both groups. In this example, however, with the variables M16 and M11 is not possible to predict group membership with a 100% of accuracy because, although none of the individuals of the species *M. dichroura* may be identified as belonging to the species *M. oligolepis*, some individuals of the species *M. oligolepis* may be identified as belonging to the species *M. dichroura*. The failure to reach 100% may be due to the possible misidentification of two individuals of *M. dichroura* as *M. oligolepis*.

Value

It is depicted 4 plots: 1) a density plot with the overlap of the area under de curve between the two groups for the variable that better discriminates between both groups, 2) a scatter plot with the polar coordinates for both groups, 3) a bivariate plot that shows from all values of group 2 the value with higher probability to belong to group 1, and 4) a bivariate plot that shows from all values of group 1 the value with higher probability to belong to group 2. Moreover, 5 files are saved: 1) overlap of the area under the curve between both categories for all variables, 2) regression coefficients of the binomial logistic regression, 3) predictions of the binomial logistic regression, 4) polar coordinates for both categories of the variable *group*, and 5) a TXT file with the results of the binomial logistic regression, the variables that better discriminate between the two groups and the Euclidean distance between groups considering the polar coordinates.

Author(s)

Cástor Guisande González, Universidad de Vigo, Spain.

References

Calenge, C. (2006) The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197, 516-519.

Calenge, C. (2016) Analysis of Habitat Selection by Animals. R package version 0.3.12. Available at: https://CRAN.R-project.org/package=adehabitatHS.

Chessel, D., Dufour, A.B. and Thioulouse, J. (2004) The ade4 package-I- One-table methods. *R News*, 4, 5-10.

Dray, S. & Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1-20.

Dray, S. & Dufour, A.B. and Chessel, D. (2007) The ade4 package-II: Two-table and K-table methods. *R News*, 7(2), 47-52.

Dray, S., Dufour, A-B. & Thioulouse, J. (2015) Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences. R package version 1.7-2. Available at: https://CRAN.R-project.org/package=ade4.

Ekstrom, C., Skovgaard, Ib M. & Martinussen, T.(2015) Datasets and functions from the (now non-existing). R package version 0.1-14. Available at: https://CRAN.R-project.org/package=kulife.

Fox, J. & Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

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. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: https://CRAN.R-project.org/package=car.

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.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162 (3), 571-580

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.

Tabachnick, B.G. & Fidell, L.S. (1996) *Using Multivariate Statistics*. NY, HarperCollins.

Examples

```
data(characiformes)

VARSEDIG(data = characiformes , variables = c("M2","M3","M4","M5","M6","M7","M8","M9","M10",
"M11","M12","M13","M14","M15","M16","M17","M18","M19","M20","M21","M22","M23",
"M24","M25","M26","M27","M28"), group="Species" , group1= "Moenkhausia oligolepis",
group2="Moenkhausia dichroura", LEGENDd=c("x='topright'", "legend = dati",
"col = COLORB", "lty=lty", "bty='n'", "cex=1.2", "text.font= 3"),
LEGENDs=c("x='topright'", "legend=unique(datosF[,'Group'])", "col = color1",
"pch = pcht", "bty='n'", "cex=1.2", "text.font=3"), LEGENDr=c("x='topright'",
"legend = dati", "col=col", "pch= c(16,16)", "bty='n'", "cex=1.2", "text.font=3"),
XLIMs=c(-1.2,1.2), YLIMs=c(-1.3,1.3), BIVTEST12=c("br=br", "cex=1.1",
"col=colbiv", "sub=sub", "Pcol=Pcol"), BIVTEST21=c("br=br", "cex=1.1",
"col=colbiv", "sub=sub", "Pcol=Pcol"), colbiv="blue", ellipse=TRUE, convex=FALSE)
```

VARSEDIM *Variable selection to discriminate many taxonomic groups*

Description

This function performs an algorithm for morphometric characters selection and statistical validation in morphological taxonomy among many taxonomic groups.

Usage

```
VARSEDIM(data, variables, group, method="overlap", stepwise=TRUE,
VARSEDIG=TRUE, minimum=TRUE, kernel="gaussian", cor=TRUE, ellipse=FALSE,
convex=TRUE, file="Plots VARSEDIG.pdf", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variables to be selected.

`group` Variable with the groups to be discriminated.

`method` Three different methods for prioritizing the variables according to their capacity for discrimination can be used. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

`stepwise` If TRUE, the logistic regression is applied by steps, in order to eliminate those variables that are not significant. The Akaike information criterion (*AIC*) is used to define what are the variables that are excluded.

`VARSEDIG` If it is TRUE, the variables are added for the estimation of polar coordinates in the priority order according to the method "overlap", "Monte-Carlo", or "logistic regression" and the variable is selected if it significantly contributes to discriminate between both groups. See details section for further information.

`minimum` If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. If it is FALSE, the algorithm selects all significant variables,and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`kernel` A character string giving the smoothing kernel to be used for estimating the overlap of the area under the curve between groups. This must be one of "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine". For further details about the estimation of the density curve see the details section of the function density of base stats package.

`cor` If it is TRUE the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one that has a greater positive correlation.

`ellipse` If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *group* is depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

`file` PDF FILE. Filename with the plots of the function VARSEDIG.

`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.

Details

The difference with the function VARSEDIG is that all the different taxa of the variable group are compared with each other, instead of just comparing two taxa. It uses the same algorithm described in the function VARSEDIG.

Value

It is obtained a PDF file with the plots of the function VARSEDIG.

Author(s)

Cástor Guisande González, Universidad de Vigo, Spain.

Examples

```
## Not run:
data(characiformes)
VARSEDIM(data=characiformes, variables= c("M2", "M3", "M4",  "M5", "M6",
"M7", "M8", "M9", "M10", "M11", "M12", "M13", "M14", "M15", "M16", "M17",
"M18", "M19", "M20", "M21", "M22", "M23", "M24", "M25", "M26", "M27", "M28"),
group="Genus")

## End(Not run)
```

VIDTAXA *IDENTIFICATION OF TAXA BASED ON MORPHOLOGICAL VARIABILITY*

Description

Identification of the different taxa based on the morphological variability observed in a Principal Components Analysis or a Correspondence Analysis.

Usage

```
VIDTAXA(data, var, labels, cat=NULL, analysis="PCA", por=80, k=NULL,
pthreshold=0.05, ellipse=FALSE, convex=FALSE, dim=c(1,2), size=c(1,5),
showCluster=TRUE, VIF=FALSE, threshold=10, method="overlap", minimum=TRUE,
ResetPAR=TRUE, PAR=NULL, PCA=NULL, SCATTERPLOT=NULL, HCLUST=NULL,
CLUSTER=NULL, BOXPLOT=NULL, mfrowBOXPLOT=NULL, LabelCat=NULL, COLOR=NULL,
COLORC=NULL, COLORB=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL,
ylabBOXPLOT=NULL, LEGEND=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL,
arrows=TRUE, larrow=0.7, colArrows="black", quadratic=FALSE, file1="Output.txt",
file2="Cat loadings.csv", file3="Descriptive statistics of clusters.csv",
file4="Original data and cluster number.csv", file5="Var loadings-Linear.csv",
file6="Cat loadings-Linear.csv", file7="Table cross-validation-Linear.csv",
file8="Cases cross-validation-Linear.csv", file9="Table cross-validation-Quadratic.csv",
file10="Cases cross-validation-Quadratic.csv", file11="Plots VARSEDIG.pdf",
file12="U Mann-Whitney test.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`var` Variables that are included in the analysis.

`labels` Variable that allows to display a label for each case.

`cat` Optionally, it is possible to specify a variable to show a grouping in the plot of the Principal Components or Correspondence analyses.

`analysis` If it is "PCA" a Principal Components analysis is carried out, whereas a Correspondence analysis is performed if the selection is "CA".

`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.

`pthreshold` Threshold probability of the U Mann-Whitney test.

`ellipse` If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*. If it is TRUE, the ellipses of the clusters in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm are also calculated.

`convex` If it is TRUE, the convex hull is calculated for each category in the plot of the Principal Components or Correspondence analyses, but only if some variable has been selected in the argument *cat*. If TRUE, the convex hull of the clusters is also calculated in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm.

`dim` Vector with two values indicating the axes that are shown in the plot of the Principal Components or Correspondence analyses.

`size` Size range of bubbles. Two values: minimum and maximum size.

`showCluster` If it is TRUE, the number of each cluster is shown in the Dendrogram.

`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.

`threshold` Cut-off value for the VIF.

`method` Three different methods for prioritizing the variables according to their capacity for discrimination can be used in the VARSEDIG algorithm. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words, from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

`minimum` If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. If it is FALSE, the algorithm selects all significant variables, and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses the PAR function that allows to modify many different aspects of the graphs.

`PCA` It accesses the prcomp function of the stats package.

`SCATTERPLOT` It accesses the function scatterplot of the car package.

`HCLUST` You may access the function hclust of the stats package.

`CLUSTER` Access to the function that allows to modify the graphic representation of the Dendrogram.

`BOXPLOT` Allows to specify the characteristics of the boxplot.

`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.

`LabelCat` It allows to specify a vector with the names of the clusters in the boxplots. They must be as many as clusters.

`COLOR` It allows to modify the colours of the graphic in the in the plot of the Principal Components or Correspondence analyses, but they must be as many as different groups have the variable *cat*.

`COLORC` It allows to modify the colours of the clusters in the Dendrogram, but they must be as many as clusters.

`COLORB` It allows to modify the colours of the clusters in the boxplots, but they must be as many as clusters.

`PCH` Vector with the symbols in the plot of the Principal Components or Correspondence analyses, which must be as many as different groups have the variable *cat*. If it is NULL they are calculated automatically starting with the symbol 15.

`XLIM, YLIM` Vectors with the axes limits *X* and *Y* in the plot of the Principal Components or Correspondence analyses.

`XLAB, YLAB` Legends of the axes *X* and *Y* in the plot of the Principal Components or Correspondence analyses.

`ylabBOXPLOT` You can specify a vector with the legends of the axes *Y* of the boxplots. They should be as many as the number of variables.

`LEGEND` It allows to include or to modify a legend in the plot of the Principal Components or Correspondence analyses.

`MTEXT` It allows to add text in the margins in the plot of the Principal Components or Correspondence analyses.

`TEXTvar` It allows to modify the labels of the variables in the plot of the Principal Components or Correspondence analyses.

`TEXTlabels` It allows to modify the labels of the cases in the plot of the Principal Components or Correspondence analyses plot.

`arrows` If it is TRUE the arrows are shown in the scatterplot in the plot of the Principal Components or Correspondence analyses.

`larrow` It modifies the length of the arrows in the plot of the Principal Components or Correspondence analyses.

`colArrows` Colours of the arrows in the plot of the Principal Components or Correspondence analyses.

`quadratic` If TRUE, a Quadratic Discriminant Analysis is performed, in addition to the Linear Discriminant Analysis.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the output file with the coordinates of the cases in the plot of the Principal Components or Correspondence analyses.

`file3` CSV FILE. Name of the output file with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram.

`file4` CSV FILE. Name of the output file with the original data of the variables and the cluster to which each case belongs.

`file5` CSV FILE. Name of the output file with the coordinates of the variables in the Linear Discriminant Analysis plot.

`file6` CSV FILE. Name of the output file with the coordinates of the categories in the Linear Discriminant Analysis plot.

`file7` CSV FILE. Name of the output file with the prediction table using the cross-validation of the Linear Discriminant Analysis.

`file8` CSV FILE. Name of the output 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.

`file9` CSV FILE. Name of the output file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis.

`file10` CSV FILE. Name of the output 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.

`file11` PDF File. Name of the output file with the graphics obtained from the VARSEDIG algorithm.

`file12` CSV FILE. Name of the output file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if a comma "," or a dot "." is used as decimal separator.

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

The aim of this analysis is to determine what statistically different groups are formed by applying a Principal Components or Correspondence analyses.

The first axis in a Principal Components analysis or Correspondence analysis is the linear combination of the original variables that has maximum variance. The second component is the linear combination of the original variables with maximum variance with the added condition that it is independent of the first (orthogonal), and so on, all the main components can be obtained, which, being independent of each other, contain different information. The independence or absence of correlation means that the new variables or components do not share common information. Each main component, therefore, explains the maximum possible residual variability (which has not already been explained above). Therefore, in a Principal Components or Correspondence analyses the cases are differentiated according to the variables that have greater variability. The idea of the analysis is to determine if statistically different groups are formed associated to the variability observed in the variables.

This analysis can be useful to find different groups when you really do not know what they are. For example, find different species using morphometric variables, without really knowing how many potential species there are and to what species each individual belongs. However, it is important to note that only different groups will be detected if the variables that have more variability give rise to different groups. It is possible that a variable does not present a great variability, but it is important for discriminating groups. This type of differentiation based on variables that do not have high variance, would not be detected in this analysis.

To detect the potential groups being formed, 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.

Finally, the algorithm of the VARSEDIG function is applied (see for more details (Guisande, 2018). With this algorithm it is possible to determine if all the cases of each cluster are statistically different from the other 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 *by=80*) and the variables to be included, eliminating those that are not correlated and are not useful in the Principal Components or Correspondence analyses, as well as those that have little discrimination power in the Discriminant Analysis.

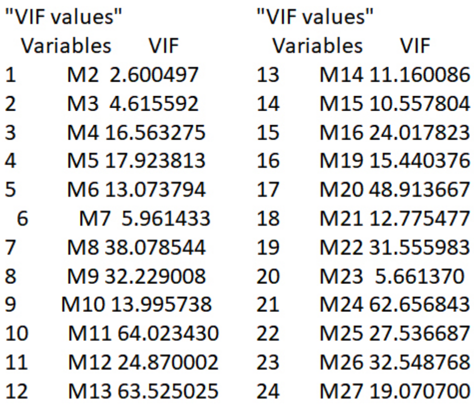
FUNCTIONS

The Correspondence analysis was performed with the ca function of the package ca (Greenacre & Pardo, 2006; Greenacre, 2007; Nenadic & Greenacre, 2007; Greenacre, 2013). 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). Bartlett's test sphericity was performed with the function bart_spher of the package REdaS (Maier, 2015). 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 VARSEDIM function of the VARSEDIG package (Guisande et al., 2016: Guisande, 2018). 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

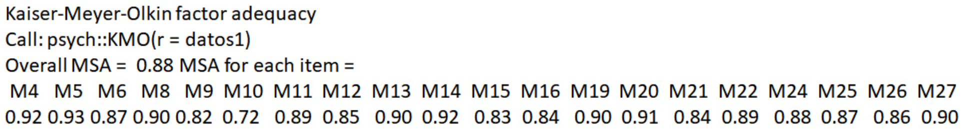
The example consisted of analysing the morphometric variability of several species of scorpaeniformes. The aim is to find how many groups are statistically different based on the morphometric variability observed in the Principal Components analysis. For purposes only of graphic presentation in the Principal Components, the genus is used as a category *cat="Genus"*. It is important to highlight that the category is not used for any statistical analysis and it is simply used to group the cases with ellipses or with the convex hull in the Principal Components graphic.

The analysis is performed by eliminating the variables that are not correlated, for which it is specified *VIF=TRUE*. Therefore, the first result obtained is the VIF values of the variables. Those variables with a VIF lower than the threshold are no included in the Principal Components analysis.



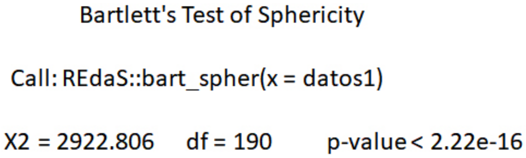
```
"VIF values"                  "VIF values"
  Variables      VIF            Variables      VIF
1        M2  2.600497         13      M14 11.160086
2        M3  4.615592         14      M15 10.557804
3        M4 16.563275         15      M16 24.017823
4        M5 17.923813         16      M19 15.440376
5        M6 13.073794         17      M20 48.913667
 6        M7  5.961433        18      M21 12.775477
7        M8 38.078544          19      M22 31.555983
8        M9 32.229008          20      M23  5.661370
9       M10 13.995738          21      M24 62.656843
10       M11 64.023430         22      M25 27.536687
11       M12 24.870002         23      M26 32.548768
12       M13 63.525025         24      M27 19.070700
```

The second 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.



```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.88 MSA for each item =
 M4   M5   M6   M8   M9  M10  M11  M12  M13  M14  M15  M16  M19  M20  M21  M22  M24  M25  M26  M27
0.92 0.93 0.87 0.90 0.82 0.72 0.89 0.85 0.90 0.92 0.83 0.84 0.90 0.91 0.84 0.89 0.88 0.87 0.86 0.90
```

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.



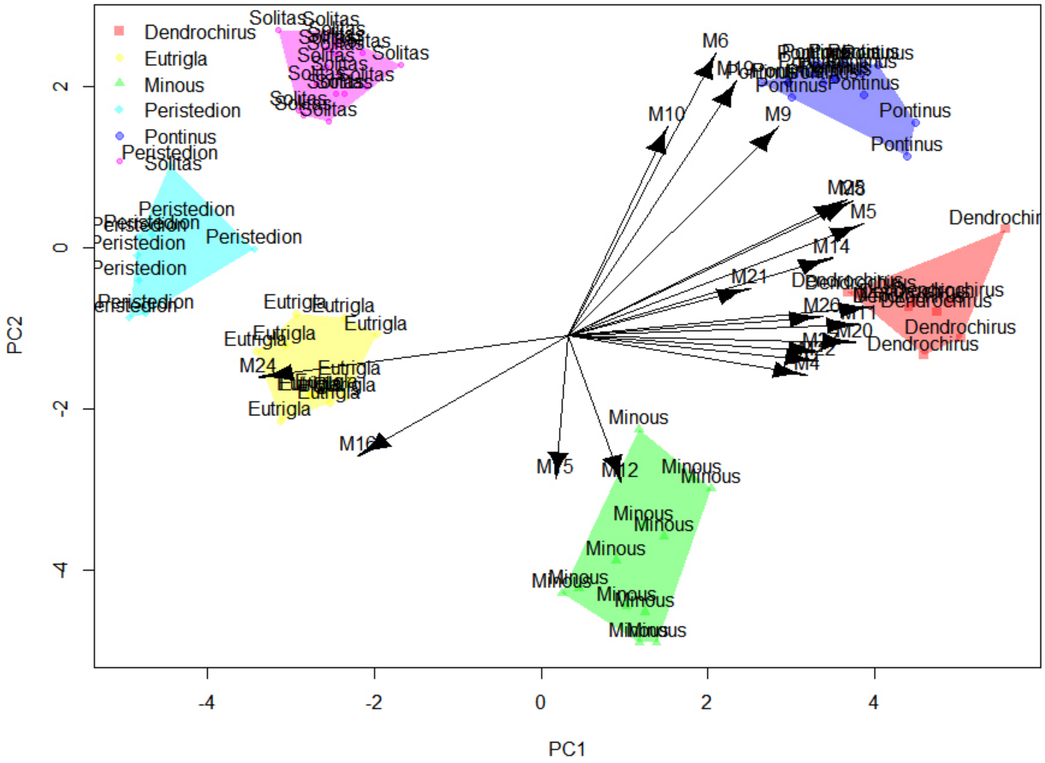
```
          Bartlett's Test of Sphericity

Call: REdaS::bart_spher(x = datos1)

X2 = 2922.806     df = 190        p-value < 2.22e-16
```

Figure VIDTAXA.1 shows that the variability observed in the Principal Components analysis allows to clearly differentiate among the genera.

Figure VIDTAXA.1. Principal Components analysis showing the variability observed in the genera.



The first axis accounts for 54%, the second for 25.3% and the third for 8.5% of the variance observed. The first three axes explain 87.8% of the variance. Since the default value of *by=80* was selected, these three Principal Component axes are selected.

Figure VIDTAXA.2 shows the Dendrogram where 6 clusters are grouped, which are the six genera used in this example.

Figure VIDTAXA.2. Dendrogram with the scores of the axes selected from the Principal Components analysis.

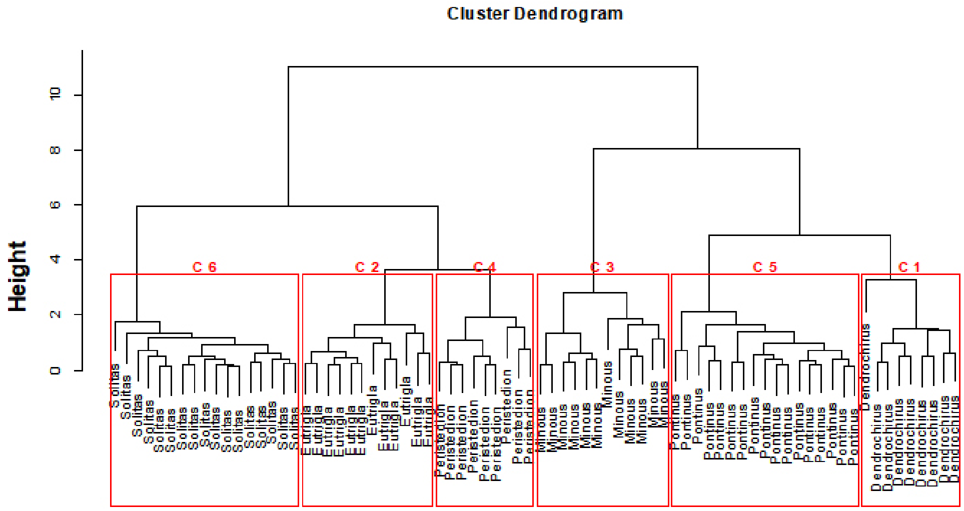
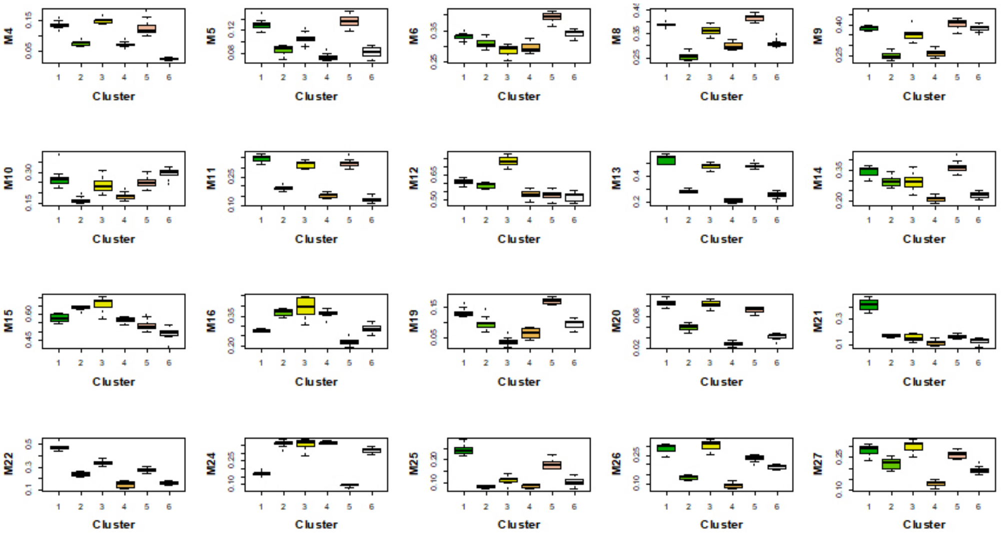


Figure VIDTAXA.3 shows the differences between clusters for each of the variables. It is clear, for instance, the difference in M21 for cluster 1, in M6 for cluster 5, etc.

Figure VIDTAXA.3. Boxplot obtained for each of the variables with the averaged values for each cluster.



The Discriminant Analysis shows that it is possible to correctly discriminate 100% of cases by cross-validation with the Linear method. The first discriminant axis explains most of the variability and discriminates well between the 6 clusters (Figure VIDTAXA.4). Many variables seem to be important for the discrimination since the arrows are not small. Figure VIDTAXA.5 shows the first two discriminant axes and shows the differences between the 6 clusters.

Figure VIDTAXA.4. Axis I of the Discriminant analysis

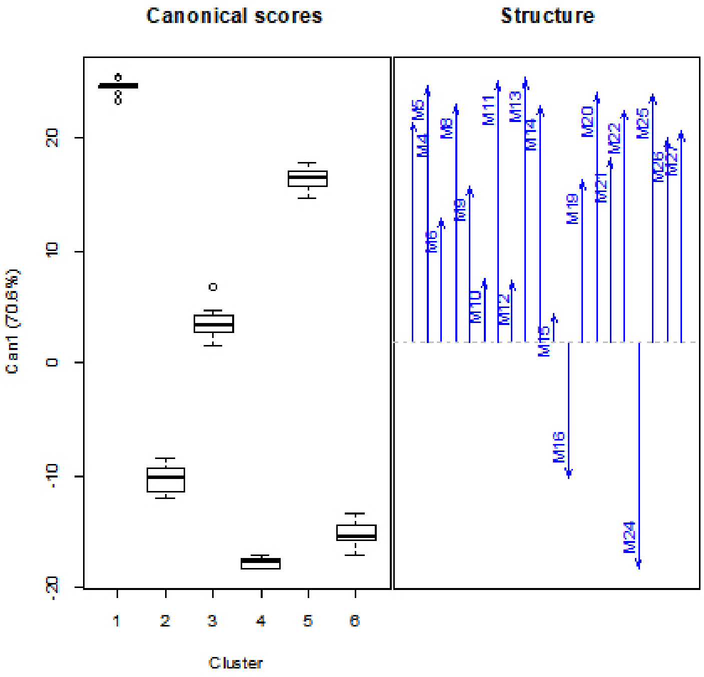
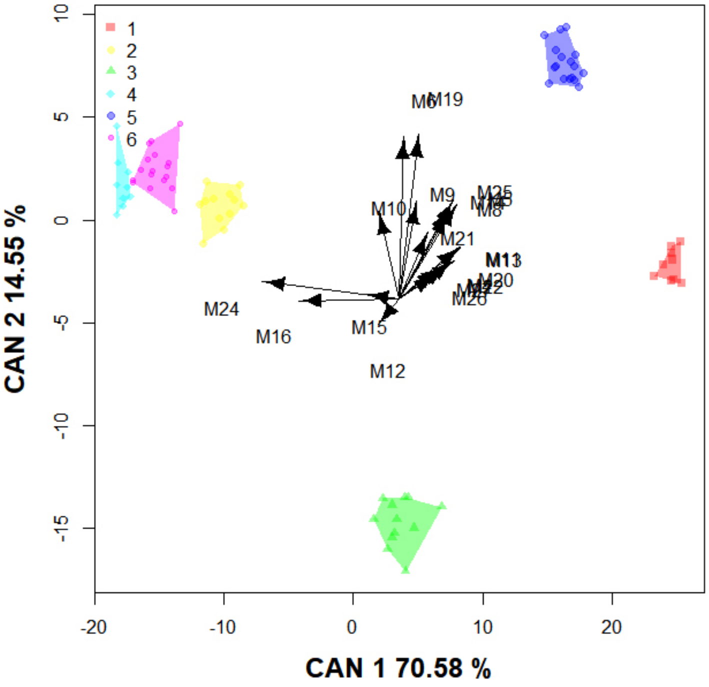
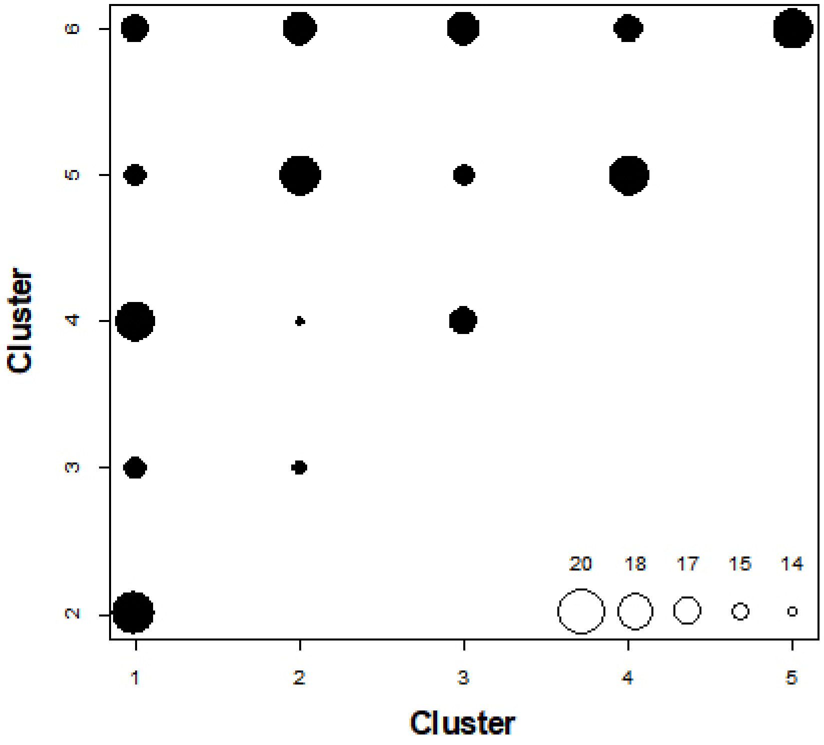


Figure VIDTAXA.5. Axes I and II of the Discriminant analysis



The next test to determine if the clusters are statistically different was the comparison of the variables between the clusters. The results of the U Mann-Whitney test are shown in Figure VIDTAXA.6. For clusters to be different, there must be at least one statistically different variable when comparing each cluster with all the others. In the graph it is noted that in the comparison between all the clusters there is always a point, that is, there is always at least one variable that is different. In fact, between cluster 2 and cluster 4, the smaller number of statistically different variables was observed, a total of 14 variables. Therefore, from the comparison of the variables between clusters with the U Mann-Whitney test, it is concluded that the clusters are statistically different from each other.

Figure VIDTAXA.6. Plot where the bubbles show the number of variables, that are statistically different (p <= 0.05) between clusters.



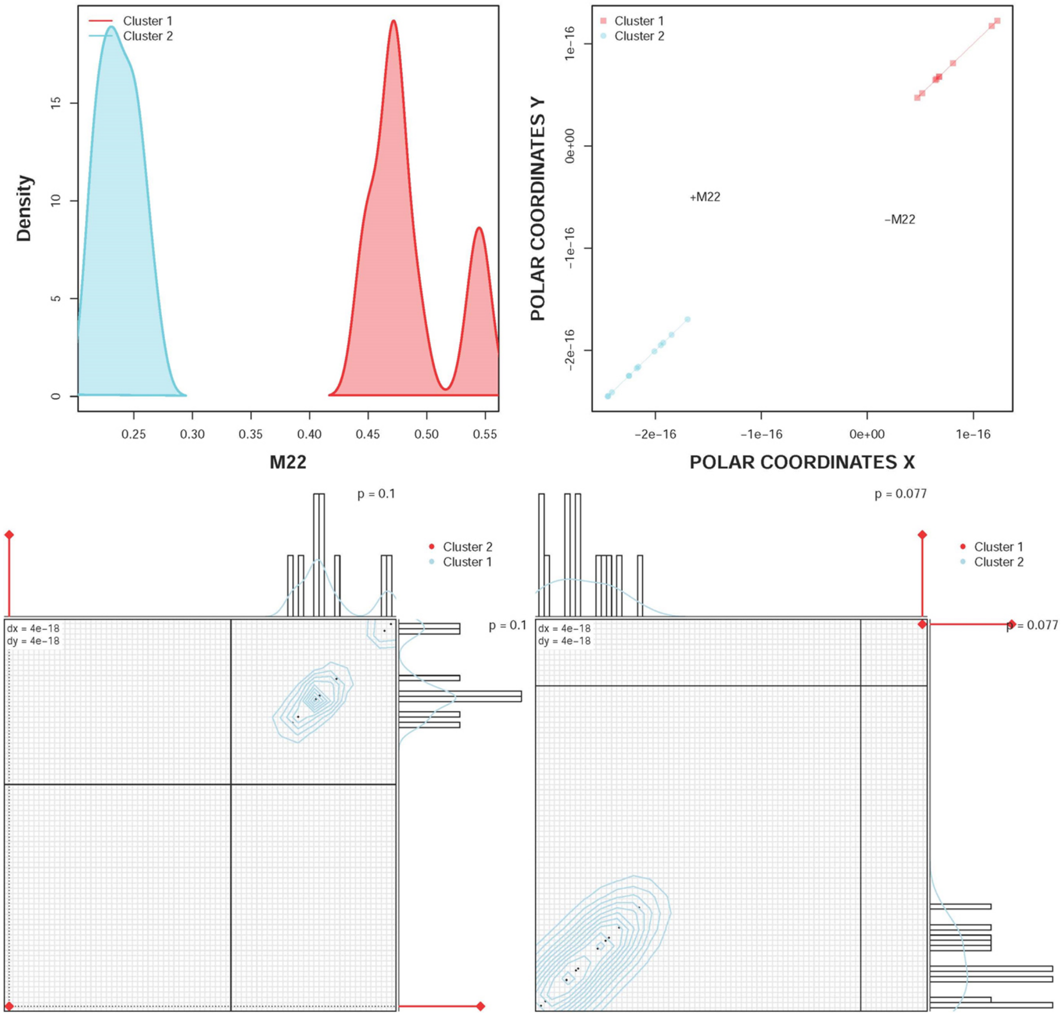
Finally, in a pdf, the plots obtained from applying the VARSEDIG algorithm are saved, whose objective is to compare all the clusters with each other.

Figure VIDTAXA.7 shows the example of the comparison of cluster 1 with 2. It is observed that the variable that discriminate significantly between both clusters is M22 (upper right panel). The Monte-Carlo test showed that the individuals that most resembles cluster 2 in cluster 1 (lower left panel) does not have significant differences in the polar coordinate axes X and Y (p = 0.1).

The individual that most resembles cluster 1 to cluster 2 (bottom right panel), it is very close to the significance threshold on both the polar coordinate axes X and Y (p = 0.077). Therefore, it cannot be concluded that cluster 1 and 2 are different. The same process would be done to compare the rest of the clusters.

Figure VIDTAXA.8. Plots obtained from the algorithm VARSEDIG.

It is shown the comparison between the cluster 1 and 2.



Therefore, according to the Discriminant Analysis and the tests performed with the U Mann-Whitney test, the clusters are statistically different from each other, but the VARSEDIG algorithm showed that not all clusters are statistically different. However, it is very important to emphasize that the VARSEDIG algorithm considers two statistically different groups if the case that most resembles each group is statistically different using the Monte-Carlo test. The Monte-Carlo test needs a large number of cases in each group for detecting significant differences. That is, it is possible that, as it was shown in the comparison of cluster 1 with cluster 2, the cases of both groups that resemble each other are not within the point cloud of the other group, but due to the low number of cases in each group, it is not possible to determine that the difference is not due to chance.

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.

12. A PDF file with the graphics obtained from the VARSEDIG algorithm.

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. A bubble chart with the number of variables that are statistically different between clusters.

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.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC.

Greenacre, M. (2013). Simple, Multiple and Joint Correspondence Analysis. R package version 0.53. Available at: https://CRAN.R-project.org/package=ca.

Greenacre, M.J. & Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, 35: 193-218.

Guisande, C. (2018) An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy. R package version 1.8. Available at: https://CRAN.R-project.org/package=VARSEDIG.

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.

Maier, M.J. (2015) Companion Package to the Book 'R: Einfuehrung durch angewandte Statistik. R package version 0.9.3. Available at: https://CRAN.R-project.org/package=REdaS.

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.

Nenadic, O. & Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20: 1-13.

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.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17: 1-25.

Rizopoulos, D. (2018) Latent Trait Models under IRT. R package version 1.1-1. Available at: https://CRAN.R-project.org/package=ltm.

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:

data(scorpaeniformes)

VIDTAXA(data=scorpaeniformes, var=c("M2","M3","M4","M5","M6","M7",
"M8","M9","M10","M11","M12","M13","M14","M15","M16","M19","M20",
"M21","M22","M23","M24","M25","M26","M27"), labels="Genus",
cat="Genus", VIF=TRUE, convex=TRUE)

## End(Not run)
```

Index