Package 'FactorsR'

February 4, 2019

Version 1.4

Title Identification of the Factors Affecting Species Richness

Author Cástor Guisande González

Maintainer Cástor Guisande González <castor@uvigo.es>

Description It identifies the factors significantly related to species richness, and their relative contribution, using multiple regressions and support vector machine models. It uses an output file of 'ModestR' (<http://www.ipez.es/ModestR>) with data of richness of the species and environmental variables in a cell size defined by the user. The residuals of the support vector machine model are shown on a map. Negative residuals may be potential areas with undiscovered and/or unregistered species, or areas with decreased species richness due to the negative effect of anthropogenic factors.

License GPL (>= 2)

Encoding latin1

Depends R (>= 3.0.0)

Suggests plotrix, relaimpo, nortest, lmtest, car, kernlab, usdm, hier.part

NeedsCompilation no

Repository CRAN

Date/Publication 2019-02-04 22:13:29 UTC

R topics documented:

1

---

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.

Factors   *Identification of the factors affecting species richness*

Description

This function provides tools for the identification of most likely causal factors significantly corre-
lated with species richness by following these steps: 1) quantifying the severity of multicollinearity
among the factors, 2) estimating the contribution of factors to species richness by using hierar-
chical partitioning, 3) selecting the factors with higher contribution among auto-correlated factors
and finally, 4) among all factors selected in the previous steps, selecting those factors significantly
correlated with species richness using a stepwise multiple regression. Predictions and residuals of
a support vector machine (SVM) model performed with the factors selected are shown on a map.
Negative residuals may be potential areas with undiscovered and/or unregistered species, or areas
with decreased species richness due to the negative effect of anthropogenic factors. This function
uses an output file of ModestR (www.ipez.es/ModestR) with data of richness of the species and
environmental variables in a cell size defined by the user.

Usage

```
Factors(data, varY, varX, Area="World", outliers=FALSE, quant1=0.05, quant2 = 0.95,
stepwise=FALSE, direction="both", deplog=FALSE, indlog=FALSE, zero=TRUE,
vifmethod="hier.part", threshold=10, PAIRS=TRUE, CoVa=FALSE, SVM=TRUE, Map=TRUE, TXT=TRUE,
jpg=TRUE, jpg1="Correlation matrix.jpg", jpg2="Contribution in the regression model.jpg",
jpg3="Residuals.jpg", jpg4="Actual richness.jpg", jpg5="Predicted richness.jpg",
jpg6="Contribution by Hierarchical Partitioning-All variables.jpg",
jpg7="Contribution by Hierarchical Partitioning-Selected variables.jpg", pairs=NULL,
```

```
colhist="#00FFFFFF", ksvm=NULL, typeV="lmg", rela=TRUE, b=1000, names.abbrev=15,
ylimV=NULL, mainV=NULL, cex.title=1.5, minLon, maxLon, minLat, maxLat, colbg="#FFFFFF",
colcon="#C8C8C8",colf="black", pro=TRUE, inc=0.005, exclude=NULL, colexc=NULL,
colfexc="black", colscale1=c("#FFFFFFFF","#FF0080FF","#FF00FFFF","#00FFFFFF",
"#64FF64FF","#C8FF00FF","#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"),
colscale2=c("#FFFFFFFF","#C8FFFFFF","#64FFFFFF","#00FFFFFF","#64FF64FF",
"#C8FF00FF", "#FFFF00FF","#FFC800FF","#FF6400FF","#FF0000FF"), legend.pos="y",
breaks=10, xl=0, xr=0, yb=0, yt=0, asp, lab=NULL, xlab="Longitude",
ylab="Latitude", main1="Residuals of SVM model", main2="Actual richness",
main3="Richness predicted by SVM model",  cex.main=1.6, cex.lab=1.4,
cex.axis=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, ini, end=NULL, file1="Output.txt",
CSV1="Residuals of SVM model.csv", CSV2="Actual richness.csv",
CSV3="Richness predicted by SVM model.csv",
CSV4="Residuals of the multiple regression model.csv", na="NA", dec=",",
row.names=FALSE)
```

Arguments

|	
data	A matrix or data frame with the species richness and, environmental variables and/or biogeographical indexes (see details section).
varY	Dependent variable. In the CSV obtained using the software ModestR, this variable is Species.Richness.
varX	Independent variables.
Area	A character with the name of the administrative area or a vector with several administrative areas (see details).
outliers	If TRUE the outliers are removed.
quant1	Lower quantile for removing outliers.
quant2	Higher quantile for removing outliers.
stepwise	If TRUE applies a stepwise regression to eliminate those variables that are not significant in the multiple regression. It uses the Akaike Information Criterion (*AIC)*) to define what are the variables that are excluded. The variables selected are also used in the SVM model.
direction	The mode of stepwise search, can be one of: "both" (default), "backward" or "forward".
deplog	If TRUE the logarithm is applied to the dependent variable using the following equation: $ln(variable - (min(variable) - 1))$.
indlog	If TRUE the logarithm of all independent variables is calculated and these transformed variables are added to the other variables in order to perform the multiple regression and SVM models with the transformed and untransformed independent variables. It uses the equation described in the argument *deplog*.
zero	If TRUE (default) the zero values of the dependent variable (cells without species) are included in both multiple regression and SVM models.
vifmethod	There are four methods for selecting variables uncorrelated: "hier.part" (default), "vif", "vifcor" and "vifstep" (see details). If NULL correlated variables are not

	eliminated and, therefore, included in both the multiple regression and SVM models.
threshold	A number specifying the VIF threshold(see details section).
PAIRS	If TRUE (default) a matrix of scatterplots is produced.
CoVa	If TRUE the plot showing the relative contribution of the independent variables in the multiple regression model is shown (only if the number of independent variables is greater than two). It is a slow process and, therefore, it is advisable to use a small number of variables (maximum 15).
SVM	If TRUE (default) the Support Vector Machine (SVM) analysis is performed. It is a time-consuming process.
Map	If TRUE (default) the maps with the residuals of the SVM model, actual richness and richness predicted by SVM model are shown.
TXT	If TRUE (default) the TXT file with the statistical analyses is generated.
jpg	If TRUE (default) the plots are exported to jpg files instead of using the windows device.
jpg1	Name of the jpg file with the matrix of scatterplots.
jpg2	Name of the jpg file with the plot of the contribution of the independent variables in the regression model.
jpg3	Name of the jpg file with the map of the residuals of the SVM model.
jpg4	Name of the jpg file with the map of the actual richness.
jpg5	Name of the jpg file with the map of the richness predicted by SVM model.
jpg6	Name of the jpg file with the plot of the contribution of all independent variables estimated by hierarchical partitioning.
jpg7	Name of the jpg file with the plot of the contribution of selected independent variables estimated by hierarchical partitioning after removing auto-correlated variables.
pairs	It allows access to the function pairs to modify the arguments of the matrix of scatterplots.
colhist	The filling color for the histograms in the matrix of scatterplots.
ksvm	It allows access to the function ksvm to modify the arguments of the SVM model.
typeV	Plot with the relative contribution of each independent variable of the argument *CoVa*. It can be a character string, character vector or list of character strings. It is the collection of metrics that are to be calculated. Available metrics: lmg, pmvd (non-US version only), last, first, betasq, pratt, genizi and car. For brief sketches of their meaning see details section of the function calc.relimp.
rela	Plot with the relative contribution of each independent variable of the argument *CoVa*. It is a logical argument requesting relative importances summing to 100% (rela=TRUE, default). If rela is FALSE, some of the metrics sum to R^2 (lmg, pmvd, pratt), others do not have a meaningful sum (last, first, betasq).
b	Plot with the relative contribution of each independent variable of the argument *CoVa*. It is the number of bootstrap runs requested on boot.relimp (default: b=1000). Make sure to set this to a lower number, if you are simply testing code.

|	
names.abbrev	Plot with the relative contribution of each independent variable of the argument *CoVa*. It is an integer that provides the number of characters to which the bar labels are shortened.
ylimV	Plot with the relative contribution of each independent variable of the argument *CoVa*. The plot routines try to use appropriate scaling. If adjustments are needed, ylim can be used as usually on a plot.
mainV	Plot with the relative contribution of each independent variable of the argument *CoVa*. The plot routine uses a default title based on the reponse name. If adjustments are desired, main can be used for specifying a different title. Note that only the first title is affected.
cex.title	Plot with the relative contribution of each independent variable of the argument *CoVa*. It specifies the text size for the overall title. The par option cex.main can be used for specifying the size of individual plot titles.
minLon, maxLon	Maps. Optionally it is possible to define the minimum and maximum longitude (see details).
minLat, maxLat	Maps. Optionally it is possible to define the minimum and maximum latitude (see details).
colbg	Maps. Background color of the map (in some cases this is the sea).
colcon	Maps. Background color of the administrative areas.
colf	Maps. Color of administrative areas border.
pro	Maps. If it is TRUE an automatic calculation is made in order to correct the aspect ratio y/x along latitude.
inc	Maps. Adds some room along the map margins with the limits x and y thus not exactly the limits of the selected areas.
exclude	Maps. 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.
colexc	Maps. Background color of areas selected in the argument exclude.
colfexc	Maps. Color of borders of the areas selected in the argument exclude.
colscale1	Maps. Palette color of the map with the residuals of the SVM model.
colscale2	Maps. Palette color of the maps with the actual and richness predicted by SVM model.
legend.pos	Maps. Whether to have a horizontal "x" or vertical "y" color gradient.
breaks	Maps. Number of breakpoints of the color legend.
xl,xr,yb,yt	Maps. The lower left and upper right coordinates of the color legend in user coordinates.
asp	Maps. The y/x aspect ratio.
lab	Maps. 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	Maps. A title for the x axis.
ylab	Maps. A title for the y axis.

main1	Maps. An overall title for the map of the residuals.
main2	Maps. An overall title for the map of the actual richness.
main3	Maps. An overall title for the map of the richness predicted by SVM model.
cex.main	Maps. The magnification to be used for main titles relative to the current setting of cex.
cex.lab	Maps. The magnification to be used for x and y labels relative to the current setting of cex.
cex.axis	Maps. The magnification to be used for axis annotation relative to the current setting of cex.
family	Maps. The name of a font family for drawing text.
font.main	Maps. The font to be used for plot main titles.
font.lab	Maps. The font to be used for x and y labels.
font.axis	Maps. The font to be used for axis annotation.
lwdP	Maps. Line width of the plot.
lwdC	Maps. Line width of the borders.
trans	Maps. 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	Maps. Number of decimals in legend of the color scale.
ini	Maps. Minimum to be considered in the maps of the actual and predicted richness.
end	Maps. Maximum to be considered in the maps of the actual and predicted richness.
file1	TXT file. Name of the txt file with the results of the statistical analysis.
CSV1	CSV file. Name of the CSV file with the residuals of SVM model.
CSV2	CSV file. Name of the CSV file with the values of the actual richness.
CSV3	CSV file. Name of the CSV file with the values of richness predicted by SVM model.
CSV4	CSV file. Name of the CSV file with the residuals of the multiple regression model.
na	CSV file. The string to use for missing values in the data.
dec	CSV file. The string to use for decimal points in numeric or complex columns: must be a single character.
row.names	CSV file. Either a logical value indicating whether the row names of x are to be written along with x, or a character vector of row names to be written.

Details

The file required in the argument *data* may be obtained using ModestR (available at the web site www.ipez.es/ModestR), as it is shown in the following screenshot (Export/Export maps of the select branch/To RWizard Applications/To FactorsR).
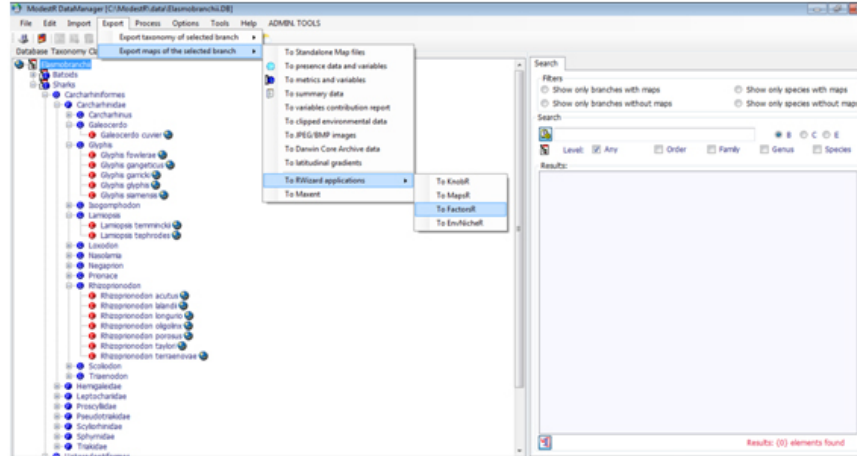
The menu shown in the following screenshot is obtained, where it is possible to select one or several biogeographical indexes (Rarity index, AOO index, etc.) and one or several environmental variables, in addition to species richness.
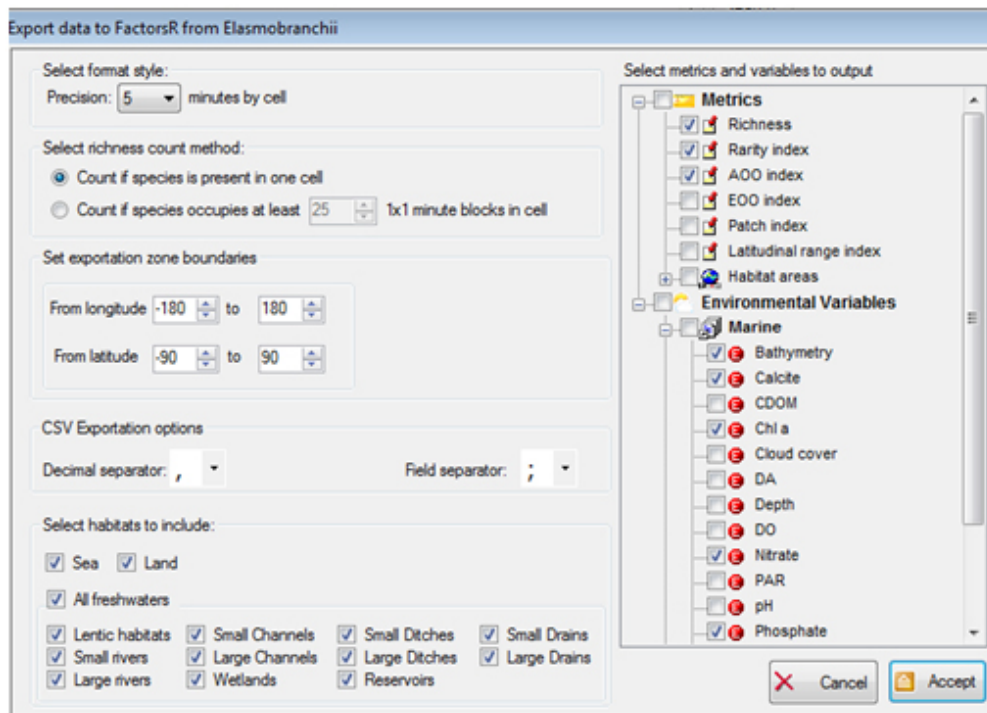

The format of the CSV file obtained is shown in the following screenshot, and this CSV file is the one required in this argument.

Latitude	Longitude	Species.Richness	Bathymetry	Calcite	Chl.a	Nitrate	Phosphate
80.6667	-179.9167	0	-1688.189	5.00E-05	0.207	2.502	0.6373
80.5833	-179.9167	0	-1618.8977	5.00E-05	0.207	2.495	0.6375
80.5	-179.9167	0	-1417.3229	5.00E-05	0.207	2.489	0.6376
80.4167	-179.9167	0	-1480.315	5.00E-05	0.207	2.483	0.6378
80.3333	-179.9167	0	-1707.0866	5.00E-05	0.207	2.476	0.638
80.25	-179.9167	0	-1782.67716	5.00E-05	0.207	2.47	0.6382
80.1667	-179.9167	0	-1782.67716	5.00E-05	0.207	2.463	0.6384
80.0833	-179.9167	0	-1694.48818	5.00E-05	0.207	2.457	0.6385
80	-179.9167	0	-1417.3229	5.00E-05	0.207	2.45	0.6387
79.9167	-179.9167	0	-1511.811	5.00E-05	0.207	2.443	0.6389
79.8333	-179.9167	0	-1587.40156	5.00E-05	0.207	2.437	0.639
79.75	-179.9167	0	-1700.78738	5.00E-05	0.193	2.43	0.6392

If the argument *Area = "World"* (default) the entire world is plotted in the maps of the residuals, actual richness and predicted richness.

Clicking on the icon world, as it is shown in the following screenshot, would display a menu with all countries and their regions. It is possible to select one or several countries and/or regions and the selected administrative areas would be only shown on the map.

If the arguments *minLon*, *maxLon*, *minLat* and *maxLat* are not specified, they are calculated automatically based on the countries and/or regions selected. The latitude and longitude of the map may be delimited, by just specifying the arguments *minLon*, *maxLon*, *minLat* and *maxLat*.
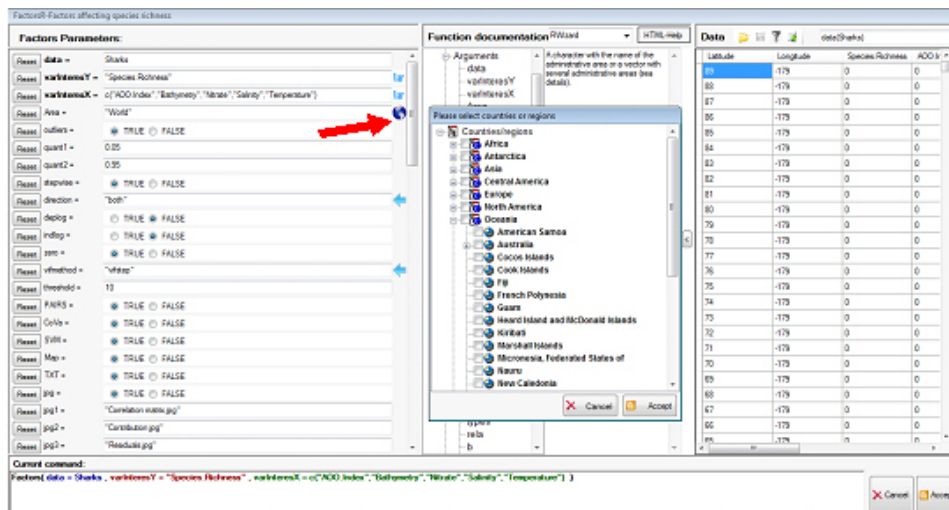
In the function four different methods for selecting uncorrelated variables can be chosen. If "vif" is selected, a pop-up window comes out where variables with VIF above threshold indicated by the user in the argument *threshold* are reported, and the user decides which variables are removed. The advantage of the manual method is that the user decides which variables to delete from the correlated pairs.

In the "vifcor" and "vifstep" options, an automatic selection of the variables to exclude is made, which is obviously much faster and comfortable. If "vifcor" is selected, you must first find the pair of variables that has the maximum linear correlation, and eliminate those having the greater VIF,

and the process should be repeated until there is no variable with a high correlation coefficient with another variable. If "vifstep" is selected, the VIF must be calculated for all variables, and exclude which has the greater VIF, always considering the threshold defined by the user in the argument *threshold*, the process should be repeated until no variable is correlated with another. The problem of "vifcor" and "vifstep" methods is that these are not taken into account when eliminating variables, which is the one that has a greater contribution to explain the observed variability in the dependent variable and, therefore, are not the best methods for obtaining the best regression model.

The default is "hier.part" in which the contribution of all independent variables to the regression model is calculated by the hierarchical partitioning method (Chevan & Sutherland, 1991) and subsequently between the variables that are above VIF threshold, eliminating those that contribute less to the regression model. Therefore, among the auto-correlated variables, those that best explain the observed variability in the dependent variable are selected. The problem is that one can only use this method if the number of dependent variables, including those log-transformed and untransformed, is not greater than 12. If the number is greater, then the "vifstep" method is automatically applied.

The default option of the argument *outliers=TRUE*, means that outliers are eliminated.

If the argument *stepwise=TRUE* is selected, it means that a regression is performed by steps using Akaike Information Criterion (*AIC*) to define which are the variables that are excluded from the regression model, since they do not contribute to explain the dependent variable. The criterion of Akaike (*AIC*) is used to decide how many explanatory variables are chosen, and what they should be. The method of Akaike values the goodness of the model by setting a penalty because of their complexity, so a simpler model is preferable to another with more independent variables that only explains a small additional portion of the variability of the dependent variable. In general $AIC = 2k - 2ln(L)$, where k is the number of explanatory variables or parameters of the model and *L* the likelihood or probability associated with the sample used in accordance with model, so that *AIC* is smaller than the lower the number of variables and the greater the likelihood; between alternative models the one with the lowest value of *AIC* must be chosen. At each step the variable that, when it is incorporated into the model, makes smaller the value *AIC* is chosen, until any unused variables allows to reduce it.

If the argument *indlog = TRUE*, the logarithm of all independent variables were calculated using the following equation $LnVariable = ln(variable - (min(variable) - 1))$, and all these variables are added to the non-transformed ones. The aim is to perform a regression model in which the transformed and non-transformed variables are included, to find the best model, because sometimes the transformed variable shows a better relationship with the dependent variable that the not transformed logarithmically.

FUNCTIONS

The goodness of fit measures for the entire hierarchy of models using all combinations of N independent variables is calculated using the function hier.part of the package hier.part (Walsh & Mac Nally, 2014).

The function lm is used for performing the multiple regression and the function step for applying the stepwise regression, both of the package stats.

The function lillie.test of the package nortest (Gross, 2013) is used for performing Kolmogorov-Smirnov Normality test with Lilliefors correction.

The function dwtest of the package lmtest (Hothorn et al., 2013) that tests autocorrelation with Durbin-Watson test and the function bptest of the package lmtest (Hothorn et al., 2013) that tests homoscedasticity with Breusch-Pagan test are also used.

The functions vif, vifcor and vifstep of the package usdm (Naimi, 2014; Naimi et al., 2014) and vif of the package car (Fox & Weisberg, 2011; Fox et al., 2014) are used to estimate the variance inflation factor (VIF).

The function calc.relimp is used for the estimation of the relative contribution of the independent variables in the multiple regression (Grömping, 2006; 2013).

The functions panel.hist and panel.reg of the package SciViews (Grosjean, 2014) and pairs are used for the matrix of scatterplots.

The function ksvm is used for the estimation of the SVM model of the package kernlab (Karatzoglou et al., 2004; Karatzoglou, 2014)

Finally, for the maps the functions image and color.legend of the package plotrix (Lemon et al., 2014) is necessary.

EXAMPLE

The range maps of shark species were elaborated following the methodology described by Guisande et al. (2013) and these maps were inputted into ModestR (García-Roselló et al., 2013).

The Bio-ORACLE global dataset 90_N 90_S real values were used as environmental factors (Tyberghein et al., 2012). The variables used were sea surface temperature, nitrate, salinity and depth.

In addition to the variables mentioned above, the mean area of occupancy was also included (AOO index, see Guisande et al. 2013 for further details). In this example, the number of variables used was lower than it in Guisande et al. (2013) to shorten the execution of the script.

In the script the argument *indlog=TRUE*, so the logarithm of all independent variables is calculated and these transformed variables are added to the other variables in order to select, among transformed and untransformed independent variables, the variables with higher contribution to the model. The argument *threshold=12*, so higher than the defult value.

The first table of the following screenshot shows VIF values higher than 12 and, therefore, there is multicollinearity among the independent variables, which is logical for this to happen because there is the same variable transformed and untransformed, for instance «Temperature» and «LNTemperature». The second table shows the contribution of the variables to the model estimated by using hierarchical partitioning. As the script uses the default option *vifmethod="hier.part"*, the auto-correlated variables (variables with VIF values higher than the threshold) are removed following the criteria of selecting the variable with higher contribution to the model between two auto-correlated variables. The third table shows the variables selected, which are not auto-correlated.

```
[1] "VIF FOR ALL VARIABLES"

[[2]]
        Variables       VIF
2       AOO.Index  8.452278
3      Bathymetry 25.762898
4         Nitrate 15.672724
5        Salinity 11.745162
6     Temperature 24.405538
7     LNAOO.Index  7.737827
8    LNBathymetry 21.452271
9        LNNitrate 15.597387
10      LNSalinity  9.811083
11   LNTemperature 34.179316

[[3]]
[1] "CONTRIBUTION OF VARIABLES ESTIMATED"

[[4]]
[1] "BY USING HIERARCHICHAL PARTITIONING (I)"

[[5]]
                        I
AOO.Index        7.582696
Bathymetry       3.536790
Nitrate          8.662267
Salinity         3.247465
Temperature     27.428611
LNAOO.Index      9.179692
LNBathymetry     2.773240
LNNitrate       12.450973
LNSalinity       1.857250
LNTemperature   23.281016
```

```
[1] "VIF FOR SELECTED VARIABLES"

[[7]]
      Variables       VIF
5   Temperature 11.276643
7     LNNitrate  3.437600
6   LNAOO.Index  4.697777
4      Salinity  7.613697
3    Bathymetry  2.175291
```

The multiple regression shows that sea surface temperature, nitrate, salinity, depth and AOO index explained 66.8% of the variance observed in species richness of sharks. The following figures show the contribution of all variables to the model and the contribution of the variables selected once those auto-correlated variables have been removed. The contribution is estimated by using hierarchical partitioning. Temperature was the variable with a higher contribution to the variance observed in the species richness of sharks.

All variables

Selected variables
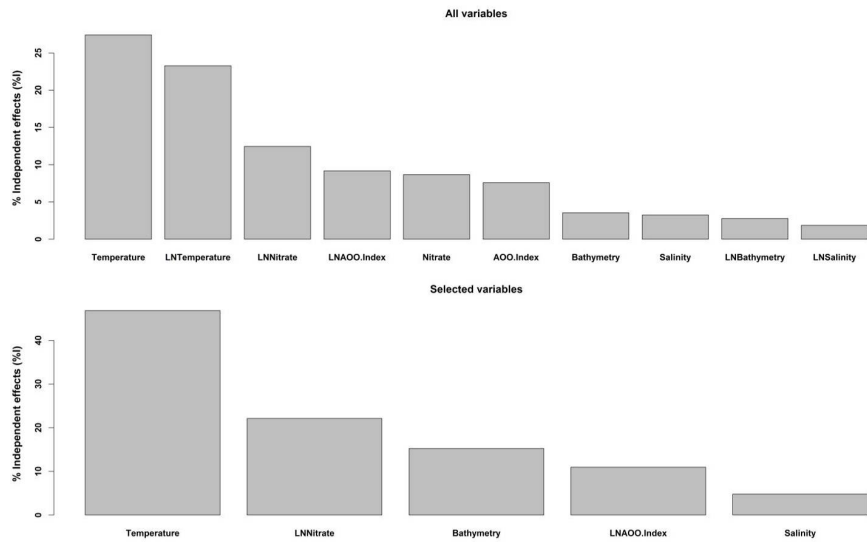
The following figure shows the matrix of scatterplots among the dependent variable (Species richness) and the dependent variables selected in the stepwise regression (in this case all the independent variables used for the analysis).
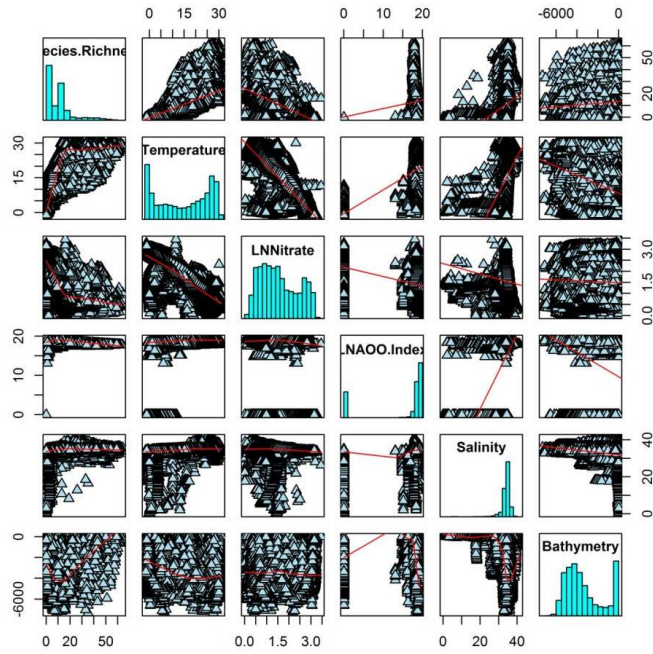

The following map shows the species richness of sharks around the world with a raster of 60´ x 60´.

Actual richness
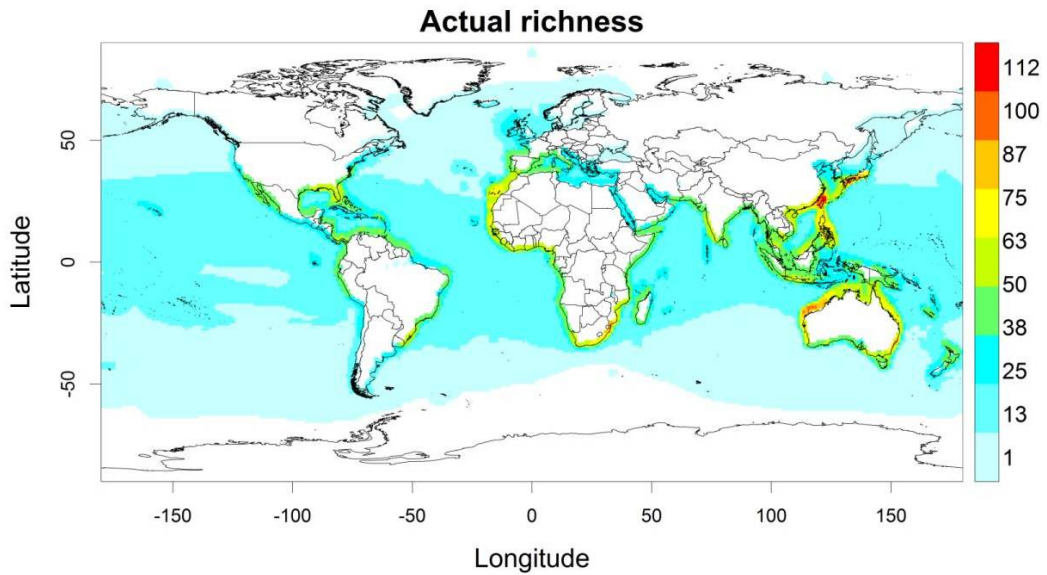
The following map shows the species richness of sharks around the world with a raster of 60′ x 60′ predicted by SVM model. The coefficient of determination of the model performed with SVM showed that the variables used in the model explained the 96.7% of the variance observed in the species richness of sharks.
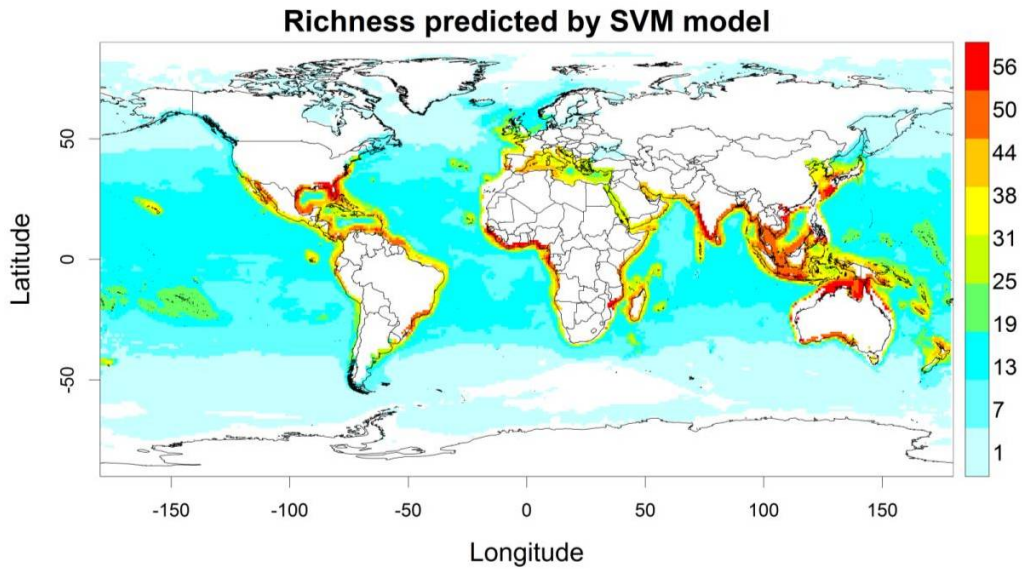
Residuals of the SVM model around the world in cells of 60′x 60′ are shown in the following map. Negative residuals of the SVM model may be a potential indicator of areas with undiscovered and/or unregistered species, or areas with lower species richness due to the negative effect of anthropogenic factors. See Guisande et al. (2013) for further details about this example.
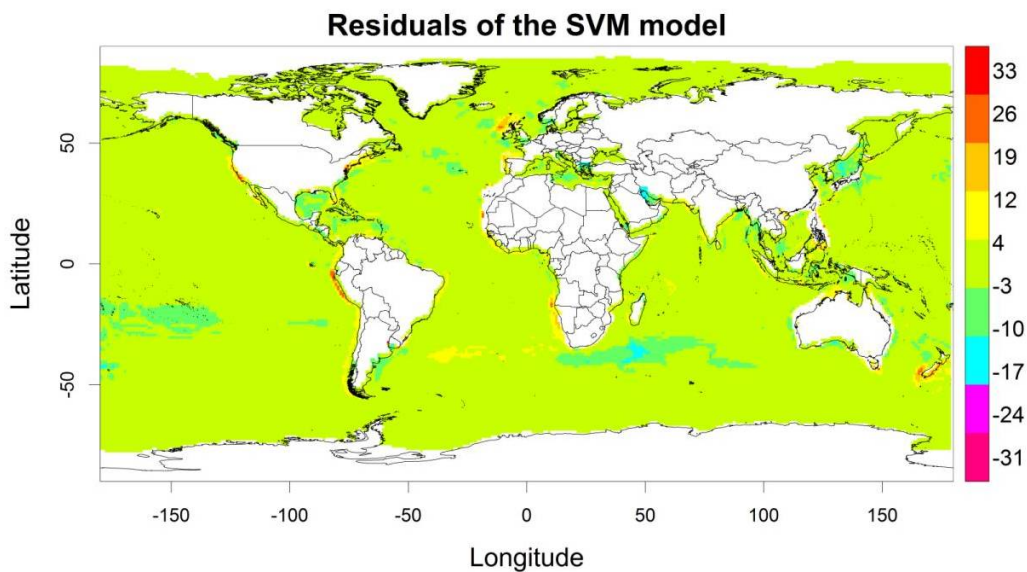
Value

A TXT file is obtained with the VIF values of all variables, the VIF of the selected variables after removing those with colinearity, the VIF of the variables that are finally in the regression model, the normality test, homogeneity of variances and homoscedasticity and the regression model. A matrix of scatterplots, a plot with the relative contribution of the independent variables estimated by using hierarchical partitioning and a plot with the relative contribution of the independent variables in the multiple regression if the argument *CoVa=TRUE*. Also maps of the residuals obtained from the SVM and, actual richness and richness predicted by the SVM model.

Author(s)

Cástor Guisande González, Universidad de Vigo, Spain.

References

Durbin, J. & Watson G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, 38: 159-171.

Fox, J. & Weisberg, S. (2011) *An R Companion to Applied Regression*. Second Edition, Sage.

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. (2013) Companion to Applied Regression. R package version 2.0-21. Available at: http://CRAN.R-project.org/package=car.

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., Vaamonde, A. & Granado-Lorencio, C. (2013) ModestR: a software tool for managing and analyzing species distribution map databases. *Ecography*, 36: 1202-1207.

Groemping, U. (2006) Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*, 17: 1-27.

Grömping, U. (2014)Relative importance of regressors in linear models. R package version 2.2-2. Available at: http://CRAN.R-project.org/package=relaimpo.

Grosjean, P. (2014) SciViews GUI API - Main package. R package version 0.9-5. Available at: http://CRAN.R-project.org/package=SciViews.

Gross, J. (2014) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Guisande, C., Patti, B., Vaamonde, A., Manjarrés-Hernández, A., Pelayo-Villamil, P., García-Roselló, E., González-Dacosta, J., Heine, J. & Granado-Lorencio, C. (2013) Factors affecting species richness of marine elasmobranchs. *Biodiversity and Conservation*, 22: 1703-1714.

Hothorn, T. et al., (2014) Testing Linear Regression Models R package version 0.9-33. Available at: http://CRAN.R-project.org/package=lmtest.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11: 1-20.

Karatzoglou, A. (2014) Kernel-based Machine Learning Lab. R package version 0.9-19. Available at: http://CRAN.R-project.org/package=kernlab.

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. (2014). Various plotting functions. R package version 3.5-7. Available at: http://CRAN.R-project.org/package=plotrix.

Naimi, B. (2014) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://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.

Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, de Clerck O (2012) Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21: 272-281.

Walsh, C. & Mac Nally, R. (2014) Hierarchical Partitioning. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=hier.part.

Examples

```
## Not run:

data(Sharks)
data(adworld)

Factors(data = Sharks, varY = "Species.Richness", varX = c("AOO.Index",
"Bathymetry","Nitrate","Salinity","Temperature"), outliers=TRUE, stepwise=TRUE,
indlog = TRUE, threshold=12)

#Remove the data set
rm(Sharks)
rm(adworld)

## End(Not run)
```

Sharks   *Species richness of marine sharks and environmental variables*

Description

Species richness of sharks and environmental variables in cells of 60 minutes x 60 minutes around the world.

Usage

```
data(Sharks)
```

Format

A matrix of shark species richness in cells of 60 minutes x 60 minutes and the preveiling mean values of sea surface temperature, nitrate, salinity, depth and mean area of occupancy (AOO index) in each cell around the world.

Source

The range maps of the species were elaborated following the methodology described by Guisande et al. (2013) and these maps were inputted into ModestR (www.ipez.es/ModestR). The Bio-ORACLE global dataset 90_N 90_S real values were used as environmental factors (Tyberghein et al., 2012). The variables used were sea surface temperature, nitrate, salinity and depth. In addition to the variables mentioned above, the AOO index was also included. The output file from ModestR with the species richness and the environmental variables is a CSV file which was converted to a RData file.

References

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., Vaamonde, A. & Granado-Lorencio, C. (2013) ModestR: a software tool for managing and analyzing species distribution map databases. *Ecography*, 36: 1202-1207.

Guisande, C., Patti, B., Vaamonde, A., Manjarrés-Hernández, A., Pelayo-Villamil, P., García-Roselló, E., González-Dacosta, J., Heine, J. & Granado-Lorencio, C. (2013) Factors affecting species richness of marine elasmobranchs. *Biodiversity and Conservation*, 22: 1703-1714.

Pelayo-Villamil, P., Guisande, C., González-Vilas. L., Carvajal-Quintero. J.D., Jiménez-Segura, L.F., García-Roselló, E., Heine, J., González-Dacosta, J., Manjarrés-Hernández, A., Vaamonde, A., Granado-Lorencio, C. (2012) ModestR: Una herramienta infromática para el estudio de los ecosistemas acuáticos de Colombia. *Actualidades Biológicas*, 34: 225-239.

Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, de Clerck O (2012) Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21: 272-281.

Index