'MaxentVariableSelection' vignette

Alexander Jueterbock

2015-09-14

Contents

Citation

To cite the package 'MaxentVariableSelection' in publications, use:

Jueterbock A, Smolina I, Coyer JA and Hoarau, G (2015) The fate of the Arctic seaweed *Fucus distichus* under climate change: an ecological niche modelling approach. Submitted manuscript

Introduction

Complex niche models show low performance in identifying the most important range-limiting environmental variables and in transferring habitat suitability to novel environmental conditions (Biology 2011; Dan L. Warren et al. 2014). This vignette demonstrates how to constrain complexity and increase performance of Maxent niche models by identifying the most important set of uncorrelated environmental variables and by fine-tuning Maxent's regularization multiplier.

Requirements and input data

Users of this package should be familiar with Maxent Niche Modelling. If you are not, you can find a great tutorial here.

ASCII Grids of environmental variables

To test for the importance of environmental variables in discriminating suitable from non-suitable habitat, these variables must be available as ASCII grids (in ESRI's .asc format). All variables must have the same extent and resolution.

For example, biologically meaningful variables are freely available from the WorldClim dataset (R J Hijmans et al.) for terrestrial systems and from the Bio-ORACLE dataset (Tyberghein et al. 2012) for marine systems.

Maxent jar file

This package runs Maxent to evaluate the performance of models built with different subsets of environmental variables. The Maxent program, however, doesn't come with this R package. It has to be downloaded separately from here. The program is free but you will have to enter your name, institution, and email prior to downloading. This package has been tested with Maxent version 3.3.3k.

Files of occurrence and background locations

Geo-coded background and occurrence records have to be provided in a format that is referred to in the Maxent tutorial as SWD format. The data sets `Backgrounddata.csv` and `Occurrencedata.csv`, which were used in (Jueterbock et al. 2015) and are included in this package, exemplify the required content.

For each location, this file lists the longitude and latitude values, as well as the values for each environmental variable.

Let's have a look at the content of `Occurrencedata.csv`:

```
occurrencelocations <- system.file("extdata", "Occurrencedata.csv",
                                   package="MaxentVariableSelection")
occurrencelocations <- read.csv(occurrencelocations,header=TRUE)
head(occurrencelocations)
```

```
##          species  longitude latitude  calcite parmean salinity sstmax
## 1 Fucusdistichus  -20.24000 74.31639 0.002835  32.733   31.574  2.305
## 2 Fucusdistichus  144.18307 44.04696 0.004014  31.303   32.714 19.643
## 3 Fucusdistichus -176.56049 51.93829 0.000151  25.391   32.780  8.637
## 4 Fucusdistichus  -58.83333 55.01667 0.000362  32.034   30.774  9.107
## 5 Fucusdistichus  -75.89486 78.86478 0.002603  35.241   32.193  1.101
## 6 Fucusdistichus  -70.61184 42.99216 0.000936  31.433   31.107 18.801
```

For each of 98 occurrence records of the macroalga *Fucus distichus* (`Fd`), this dataset contains `longitude` and `latitude` values, as well as values for four environmental variables (`calcite`, `parmean`, `salinity`, and `sstmax`).

Such files are easy to create with the *extract* function of the R package 'raster' (Robert J Hijmans 2015) once we have longitude and latitude values of the occurrence records.

You can install the 'raster' package with the following command

```
install.packages('raster')
```

Let's extract variables from the Bio-ORACLE database (Tyberghein et al. 2012) for each of the 98 occurrence records of *Fucus distichus*. Downloading the variables (from here) may take a while as the rasters sum to about 160 MB in size. the variables are compressed as RAR files. Extract them and save them in a folder called `BioORACLEVariables`. The following R code shows how to obtain the value of these variables for each of the occurrence records

```
# load the raster package
library(raster)

# Load the occurrence records
occurrencelocations <- system.file("extdata", "Occurrencedata.csv",
                                   package="MaxentVariableSelection")
occurrencelocations <- read.csv(occurrencelocations,header=TRUE)
LonLatData <- occurrencelocations[,c(2,3)]

# Then load the environmental variables into R with the help of the
# stack function of the 'raster' package. You can not just copy the
# following line but have to adjust the filepath to your own.

files <- list.files("/home/alj/Downloads/BioORACLEVariables",pattern='asc',
full.names=TRUE)
Grids <- raster::stack(files)

# Extracting the variables for all occurrencelocations
VariablesAtOccurrencelocations <- raster::extract(Grids,LonLatData)

# Combining the extracted values with the longitude and latitude values
Outfile <- as.data.frame(cbind("Fucusdistichus", LonLatData,
                               VariablesAtOccurrencelocations))
colnames(Outfile) <- c("species","longitude","latitude",
                       colnames(VariablesAtOccurrencelocations))


#writing this table to a csv file:

write.csv(Outfile, file =
"VariablesAtOccurrencelocations.csv", append = FALSE,sep = ",", eol =
"\n", na = "NA", dec = ".",col.names = TRUE,row.names=FALSE)
```

Workflow tutorial

The central function of this package is called `VariableSelection`. The basic usage of this function is:

```
VariableSelection(maxent, outdir, gridfolder,
occurrencelocations, backgroundlocations, additionalargs,
contributionthreshold, correlationthreshold, betamultiplier)
```

The following sections guides you through an example with the possible settings of each argument.

maxent

Specify the file path to the `maxent.jar` file, which has to be downloaded from here. Not that you can not just copy the following line but instead have to specify the file path to the folder containing `maxent.jar` on your own computer.

```
maxent <- ("/home/alj/Downloads/maxent.jar")
```

If you are working on a Windows platform, your filepath might rather look similar to the following, where ... should be replaced by the hierarchical file structure that leads to `maxent.jar` on your computer.:

```
maxent <- ("C:/.../maxent.jar")
```

outdir

Filepath to the output directory, including the name of the directory, like:

```
outdir <- ("/home/alj/Downloads/OutputDirectory")
```

Again, you need to adjust the filepath to the correct path and folder on your own computer. All result files will be written into this folder. Please don't put important files in this folder as all files but the output files of the `VariableSelection` function will be deleted from this folder.

gridfolder

Here, you specify the filepath to the folder containing all your ASCII grids of environmental variables that you consider to be potentially relevant in setting distribution limits of your target species. All variables must have the same extent and resolution.

If you are puzzled what grids I am referring to, have a look at the section **ASCII Grids of environmental variables** above.

Like for the `maxent` and `outdir` arguments, you also can not just copy the following line as the location of ASCII grids is likely different on your computer. Adjust the filepath to the folder where you stored your ASCII grids. In this example I provide the filepath to the `Bio-ORACLE` rasters that I downloaded from here.

```
gridfolder <- ("/home/alj/Downloads/BioORACLEVariables")
```

occurrencelocations

Here, you need to specify the filepath to the csv file of occurrence locations (see the section **Files of occurrence and background locations** above for the required file format). An example file of occurrence locations for the macroalga *Fucus distichus* is included in this package. You load it with:

```
occurrencelocations <- system.file("extdata", "Occurrencedata.csv",
                                   package="MaxentVariableSelection")
```

Instead, if the file with your occurrence locations is stored on your own computer, set here the filepath to it. For example:

```
occurrencelocations <- "/home/alj/Downloads/Occurrencedata.csv"
```

backgroundlocations

The same applies for the `backgroundlocations` argument as for the `occurrencelocations` argument.

If you want to use the background locations that were used in (Jueterbock et al. 2015), type:

```
backgroundlocations <- system.file("extdata", "Backgrounddata.csv",
                                   package="MaxentVariableSelection")
```

Instead, if you have your own csv file with backgroundlocations specify the filepath here, similar to:

```
backgroundlocations <- "/home/alj/Downloads/Backgrounddata.csv"
```

additionalargs

Maxent arguments can be specified with the `additionalargs` argument. You find an overview of possible arguments in the end of the Maxent help file.

The following settings are preset in this package to calculate information criteria and AUC values. Don't change these values in the `additionalargs` string:

- `autorun=true`: starts running with the startup of the program.
- `plots=false`: does not create any plots.
- `writeplotdata=false`: does not write output data that are required to plot response curves.
- `visible=false`: does not show the Maxent user interface but runs the program from the command line.
- `randomseed=true`: the background data will be split into different test/train partitions with each new run.
- `writebackgroundpredictions=false`: does not write .csv files with predictions at background points.
- `replicates=10`: number of replicate runs; this is only set for the calculation of AUC values. The calculation of AICc values doesn't require to partition the occurrence sites in test and training data, thus, also does not require to replicate runs.
- `replicatetype=subsample`: uses random test/training data for replicated sample sets; this is only set for the calculation of AUC values.
- `randomtestpoints=50`: 50 percent of presence points are randomly set aside as test data; this is only set for the calculation of AUC values.
- `redoifexists`: redoes the analysis if output files exist already.
- `writemess=false`: does not write multidimensional similarity surfaces (MESS), which inform on novel climate conditions in the projection layers.
- `writeclampgrid=false`: does not write a grid on the spatial distribution of clamping.
- `askoverwrite=false`: does not ask if the data shall be overwritten if they exist already.
- `pictures=false`: does not create .png images of the output grids.

- `outputgrids=false`: does not write output grids for all replicate runs but only the summary grids; this is only set for the calculation of AUC values.
- `outputformat=raw`: write output grids with raw probabilities of presence; this is only set for calculation of information criteria, like AICc.

Instead, what additional Maxent argument you might want to set are the features to be used, by selecting true or false for the arguments `linear`, `quadratic`, `product`, `threshold`, and `hinge`.

For this example, we are using only `hinge` features, with the following setting:

```
additionalargs="nolinear noquadratic noproduct nothreshold noautofeature"
```

`contributionthreshold`

This sets the threshold of model contribution below which environmental variables are excluded from the Maxent model. Model contributions range from 0% to 100% and reflect the importance of environmental variables in limiting the distribution of the target species.

In this example, we set the threshold to 5%, which means that all variables will be excluded when they contribute less than 5% to the model:

```
contributionthreshold <- 5
```

`correlationthreshold`

This sets the threshold of Pearson's correlation coefficient (ranging from 0 to 1) above which environmental variables are regarded to be correlated (based on values at all background locations). Of the correlated variables, only the variable with the highest contribution score will be kept, all other correlated variables will be excluded from the Maxent model. Correlated variables should be removed because they may reflect the same environmental conditions, and can lead to overly complex or overpredicted models. Also, models compiled with correlated variables might give wrong predictions in scenarios where the correlations between the variables differ.

Here, we are setting the threshold to 0.9:

```
correlationthreshold <- 0.9
```

`betamultiplier`

This argument sets the values of beta multipliers (regularization multipliers) for which variable selection shall be performed. The smaller this value, the more closely will the projected distribution fit to the training data set. Models that are overfitted to the occurrence data are poorly transferable to novel environments and, thus, not appropriate to project distribution changes under environmental change. Performance will be compared between models created with the beta values given in this `betamultiplier` vector. Thus, providing a range of beta values from 1 (the default in Maxent) to 15 or so, will help you to spot the optimal beta multiplier for your specific model.

Here, we are testing betamultipliers from 2 to 6 with steps of 0.5 with the following setting:

```
 betamultiplier=seq(2,6,0.5)
```

Variable selection procedure

With the argument seetings that we specified above, we can start the variable selection. The following function starts the model-selection procedure

```
library("MaxentVariableSelection")
VariableSelection(maxent,
                  outdir,
                  gridfolder,
                  occurrencelocations,
                  backgroundlocations,
                  additionalargs,
                  contributionthreshold,
                  correlationthreshold,
                  betamultiplier
                  )
```

The run will take a while (ca. one hour). The resulting output files will be saved in the folder that you specified in the argument `outdir`.

The following paragraphs explain what steps are actually taken in this procedure to identify the set of most relevant environmental variables. First, an initial Maxent model is compiled with all provided environmental variables. Then, all variables are excluded which have a relative contribution score below the value set with `contributionthreshold`. Then, those variables are removed that corrleate with the variable of highest contribution (at a correlation coefficient $>$ `correlationthreshold` or $<$ `-correlationthreshold`).

The remaining set of variables is then used to compile a new Maxent model. Again, variables with low contribution scores are removed and remaining variables that are correlated to the variable of second-highest contribution are discarded. This process is repeated until left with a set of uncorrelated variables that all had a model contribution above the value set with contributionthreshold. These steps are performed for the range of betamultipliers specified with `betamultiplier`.

The performance of each Maxent model is assessed with the sample-size-adjusted Akaike information criterion (AICc) (Akaike 1974), the area under the receiver operating characteristic (AUC) estimated from test data (Fielding and Bell), and the difference between AUC values from test and training data (AUC.Diff), which estimates model overfitting. AICc values are estimated from single models that include all occurrence sites, based on code in ENMTools (D L Warren, Glor, and Turelli). AUC values, instead, are averaged over ten replicate runs which differ in the set of 50% test data that are randomly sub-sampled from the occurrence sites and withheld from model construction.

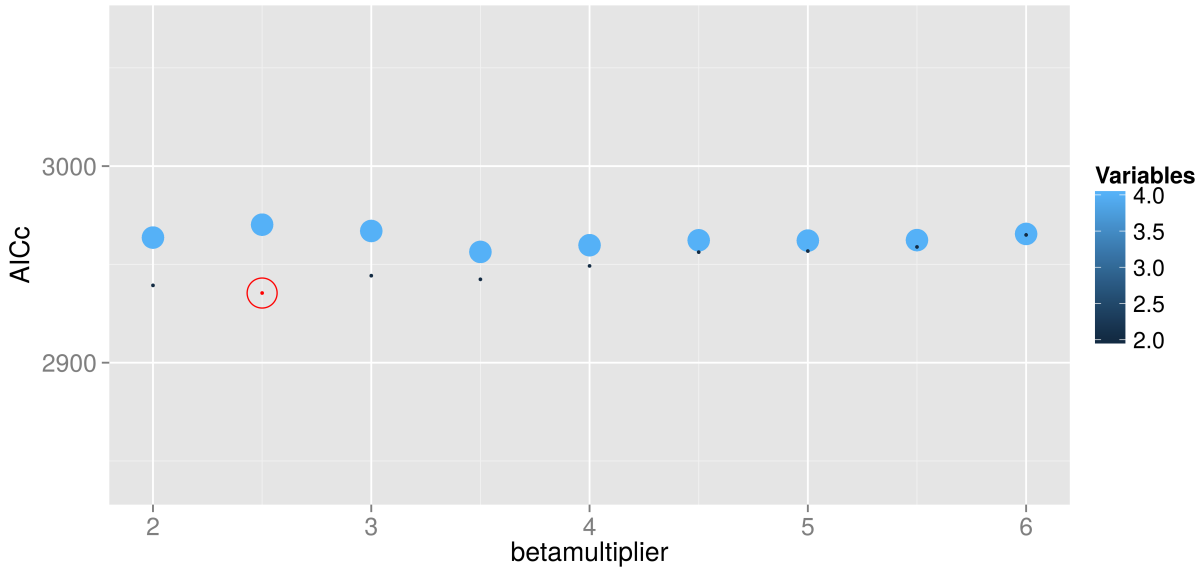
Maximization of test AUC values generally favors models that can well discriminate between presence and absence sites (Fielding and Bell; Biology 2011). This, however, bears the risk to overpredict the realized niche of a species (Jiménez-Valverde). Instead, minimization of AICc values generally favors models that recognize the fundamental niche of a species and that are better transferable to future climate scenarios (Biology 2011). The optimal set of variables along with the best beta multiplier is then identified as the model of highest AUC or lowest AICc values (the results of the `VariableSelection` functions gives you both options).

Result files

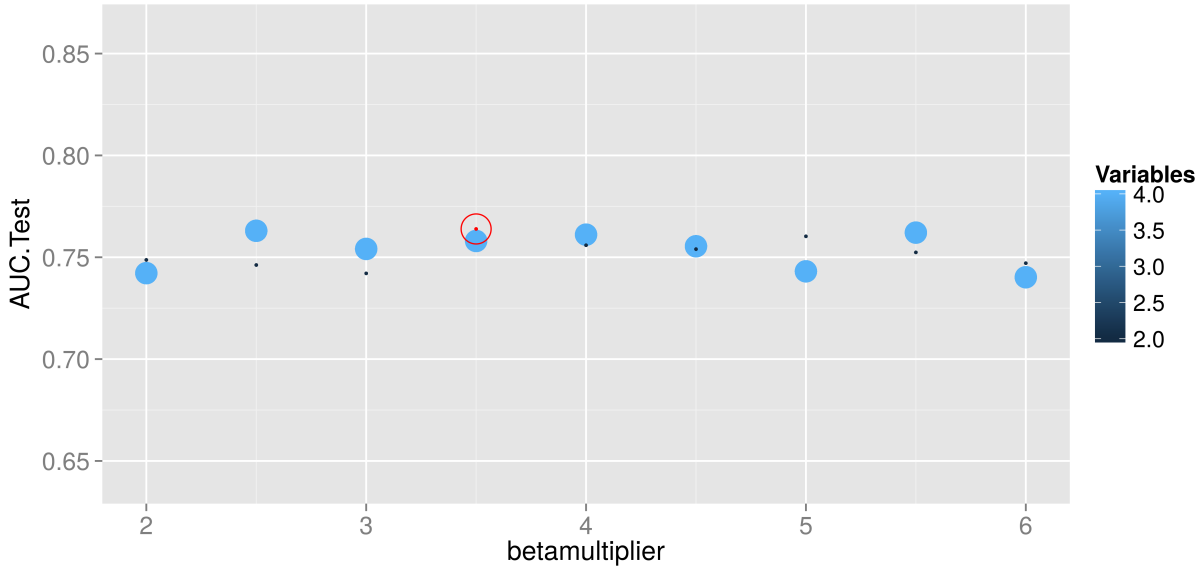
The results are saved in the directory that was specified in the argument `outdir`. If you were patient enough to wait for the example variable selection to finish (see above), you will find seven tables and four figures in the output directory.

The output gives you the opportunity to decide whether you choose AUC.Test values or AICc values as performance criterion to identify the model of highest performance. If we choose AICc values as performance criterion, the following result files are most relevant:

The figure `ModelSelectionAICc_MarkedMinAICc.png` shows that the model with beta-multiplier 2.5 and with only 2 environmental variables shows highest performance (lowest AICc).



Instead, if you would decide to use AUC.Test values as performance criterion, the most relevant figure is `ModelSelectionAUCTest_MarkedMaxAUCTest.png`. If you compare the two figures, you find that the model with maximum AUC.Test value differs from the model with minimum AICc value. Although both models chose 2 variables, the model with maximum AUC.Test value had a betamultiplier of 3.5, while the model with minimum AICc value had a betamultiplier of 2.5.



The figures also shows that the maximum number of variables in any of the models is four , although the `gridfolder` contained many more variables. Only those four variables that were extracted for the `occurrencelocations` and for the `backgroundlocations` were included in the initial Maxent models. If you would like to identify if other than these four variables (`calcite`, `parmean`, `sstmax`, and `salinity`) are important in setting distribution limits, they first have to be extracted for the `occurrencelocations` and the `backgroundlocations` as described above in the section 'Files of occurrence and background locations' above.

The table `ModelWithMinAICc.txt` lists performance indicators of the model with lowest AICc value. This table is a subset of the table `ModelPerformance.txt`, which lists performance indicators of all created Maxent models. The first five columns show that the best performing model is model number 4, with a betamultiplier of 2.5 and only 2 variables:

Model	betamultiplier	variables	samples	parameters	loglikelihood
4	2.5	2	97	10	-1456.44

The following six columns provide the performance indicators of this model:

AIC	AICc	BIC	AUC.Test	AUC.Train	AUC.Diff
2932.87	2935.43	2958.62	0.75	0.78	0.035

The table `ModelWithMaxAUCTest.txt`, instead, lists performance indicators of the model with maximum AUC.Test value.

The table `VariableSelectionMinAICc.txt` shows contributions of and correlations between environmental variables for those models that lead directly to the model of lowest AICc value. This table and the table `VariableSelectionMaxAUCTest.txt` are subsets of `VariableSelectionProcess.txt`, which lists variable contributions and correlations for all created Maxent models.

The content of `VariableSelectionMinAICc.txt` is:

Test	Contributions	Correlation	Contributions	Correlation
Model	3	3	4	4
betamultiplier	2.5	2.5	2.5	2.5
calcite	55.6044000	1.0000000	58.9015000	0.1314525
parmean	0.0332	NA	NA	NA
salinity	1.4447	NA	NA	NA
sstmax	42.9176000	0.1314525	41.0985000	1.0000000

Here, the variable selection process that leaded to the model with minimum AICc value started with model 3. Two other models with betamultiplier 2 were created before. Since the `contribtionthreshold` was set to 5, all variables with a lower contribution (`parmean` and `salinity`) were excluded. That's why they have `NA` values from the first correlation test on (`NA` stands for 'not available'). The remaining variables (`calcite` and `sstmax`) were kept in the end because the correlation between them was below the `correlationthreshold` of 0.9.

In conclusion, the optimal model settings (based on the performance-indicator AICc) are: 1) a beta-multiplier of 2.5; and 2) a subset of two environmental variables: `calcite` and `sstmax`.

References

Akaike, H. 1974. "'A New Look at the Statistical Model Identification'." *IEEE Transactionson Automatic Control* 19 (6): 716–723. doi:10.1109/TAC.1974.1100705.

Biology, Integrative. 2011. "Ecological Niche Modeling in Maxent : the Importance of Model Complexity and the Performance of Model Selection Criteria." *Ecological Applications : a Publication of the Ecological Society of America* 21 (2): 335–342.

Fielding, A H, and J F Bell. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation* 24 (01): 38–49.

Hijmans, R J, S E Cameron, J L Parra, P G Jones, and A Jarvis. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology* 25 (15): 1965–1978.

Hijmans, Robert J. 2015. "raster: Geographic Data Analysis and Modeling." http://cran.r-project.org/package=raster.

Jiménez-Valverde, A. "Insights into the Area Under the Receiver Operating Characteristic Curve (AUC) as a Discrimination Measure in Species." *Ecology* 5: 498–507. doi:10. 1111/j. 1466-8238. 2011. 00683. x.

Jueterbock, Alexander, Irina Smolina, James A Coyer, and Galice Hoarau. 2015. "The Fate of the Arctic Seaweed Fucus Distichus Under Climate Change: an Ecological Niche Modelling Approach." *Submitted Manuscript*.

Tyberghein, Lennert, Heroen Verbruggen, Klaas Pauly, Charles Troupin, Frederic Mineur, and Olivier De Clerck. 2012. "Bio-ORACLE: a Global Environmental Dataset for Marine Species Distribution Modelling." *Global Ecology and Biogeography* 21 (2) (February): 272–281. doi:10. 1111/j. 1466-8238. 2011. 00656. x. http://doi.wiley.com/10.1111/j.1466-8238.2011.00656.xhttp://dx.doi.org/10.1111/j.1466-8238.2011.00656.x.

Warren, D L, R E Glor, and M Turelli. "ENMTools: a Toolbox for Comparative Studies of Environmental Niche Models." *Ecography* 33 (3): 607–611. doi:10. 1111/j. 1600-0587. 2009. 06142. x. http://dx.doi.org/10.1111/j.1600-0587.2009.06142.x.

Warren, Dan L., Amber N. Wright, Stephanie N. Seifert, and H. Bradley Shaffer. 2014. "Incorporating Model Complexity and Spatial Sampling Bias into Ecological Niche Models of Climate Change Risks Faced by 90 California Vertebrate Species of Concern." *Diversity and Distributions* 20 (3): 334–343. doi:10.1111/ddi.12160.